Learning Crowdsourced User Preferences for Visual Summarization of Image Collections

Stevan Rudinac, Martha Larson, *Member, IEEE*, and Alan Hanjalic, *Senior Member, IEEE*

***Abstract*—In this paper we propose a novel approach to selecting images suitable for inclusion in the visual summaries. The approach is grounded in insights about how people summarize image collections. We utilize the Amazon Mechanical Turk crowdsourcing platform to obtain a large number of manually created visual summaries as well as information about criteria for image inclusion in the summary. Based on these large-scale user tests, we propose an automatic image selection approach, which jointly utilizes the analysis of image content, context, popularity, visual aesthetic appeal as well as the sentiment derived from the comments posted on the images. In our approach we do not describe images based on their properties only, but also in the context of semantically related images, which improves robustness and effectively enables propagation of sentiment, aesthetic appeal as well as various inherent attributes associated with a particular group of images. We discuss the phenomenon of a low inter-user agreement, which makes an automatic evaluation of visual summaries a challenging task and propose a solution inspired by the text summarization and machine translation communities. The experiments performed on a collection of geo-referenced Flickr images demonstrate the effectiveness of our image selection approach.**

***Index Terms*—Crowdsourcing, image aesthetic appeal, image content and context, image set evaluation, learning to rank, sentiment analysis, social media, user-informed visual summarization.**

I. Introduction

RAPID growth of the amount of digital multimedia data available in personal and professional collections as well as the content sharing and social networking websites, has created the need for powerful tools enabling analysis, representation, abstraction and summarization of data for more efficient and effective browsing and retrieval. Summarization techniques, in particular, aim at providing a compact representation of a single multimedia data document or data collection. Depending on the type of data and the application domain, summaries may consist of text, images, video segments or a combination of these.

In this paper we focus on visual summaries. Visual summaries serve to abstract a video [1], [2], set of videos [3] or an image collection [4]–[6] and usually consist of video segments or images (e.g., photos or video keyframes). Although humans in general intuitively understand the concept of a (visual) summary, giving a single and universal definition of the summary appears to be difficult [7]. While intuitively the structure and content of a summary should depend on the purpose it should fulfill [8], the final assessment of its quality can only be made based on its compatibility with the expectations of the human users. Therefore, given a particular application and use case, the specific criteria reflecting the user's perception of the summarization quality should be identified and used to steer the summarization algorithm. In other words, a summarization algorithm should be *user informed* in order to be successful.

Existing methods for visual summarization have typically been guided by studies (e.g., [9]) of users' preferences in terms of a tradeoff between the relevance and representativeness of the information included in the summary and the ability of the summarization algorithm to diversify the included visual content [4]–[6], [10]. The notions of relevance, representativeness and diversity, as well as the interplay among the three are, however, too general to be modeled successfully in a given summarization scenario, and especially across scenarios. Furthermore, although the quality of visual summaries generated using the existing approaches is sometimes judged by human evaluators (e.g., [4]), explicit information on how humans create visual summaries has hardly been inferred or taken into account while developing summarization algorithms. Therefore, the insights obtained so far can be considered insufficient to serve as guidelines for developing a solid visual summarization approach.

In this paper we demonstrate how user-informed visual summarization algorithms can be facilitated by relying on *crowdsourcing*. We first run a large-scale crowdsourcing experiment to obtain insight into how users perform visual summarization. Then we use this insight to decide on the appropriate features, based on which images in the collection can be ranked. The ranking reflects the suitability of an image as a candidate for inclusion in the summary, that is, how likely an image would be selected for the summary by the users.

We take the problem of visual summarization of geographic areas as the sample use case in this paper to demonstrate the benefits of the proposed user-informed image selection concept. We foresee, however, that the material presented here will be of use in a wide range of summarization problems. The paper makes the following main contributions, whose implications transcend our specific choice of use case:

- We show how to deploy crowdsourcing to acquire implicit and explicit criteria humans find important when performing visual summarization.

Manuscript received September 07, 2012; revised January 31, 2013 and April 18, 2013; accepted April 24, 2013. Date of publication May 03, 2013; date of current version September 13, 2013. This work was supported by the European Commission's 7th Framework Programme (FP7) under grant agreement no. 216444 (NoE PetaMedia). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Daniel Gatica-Perez.

The authors are with the Multimedia Information Retrieval Lab, Delft University of Technology, Delft, The Netherlands (e-mail: s.rudinac@tudelft.nl; m.a.larson@tudelft.nl; a.hanjalic@tudelft.nl).

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

Digital Object Identifier 10.1109/TMM.2013.2261481

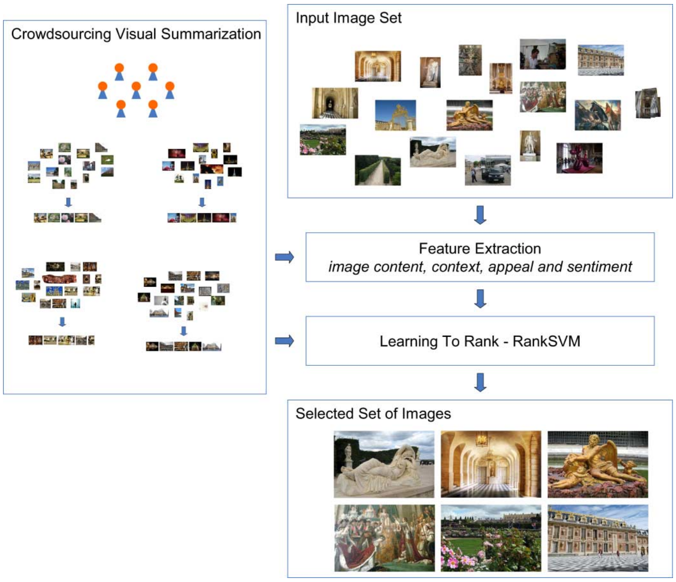

Crowdsourcing Visual Summarization
Input Image Set
Feature Extraction
image content, context, appeal and sentiment
Learning To Rank - RankSVM
Selected Set of Images

Fig. 1. Illustration of the proposed user-informed approach to image selection for creating visual summaries. All images are downloaded from Flickr under CC license.

- We propose a novel approach for embedding the derived criteria into descriptive features and learning to distinguish between images based on the likelihood of their appearance in the human-created visual summaries.
- In order to match the criteria inferred from human-created summaries, we expand the scope of features used to represent the image collection beyond those that are typically deployed for visual summarization. This expansion encompasses, in particular, features related to the context, aesthetic appeal, sentiment and popularity of an image.
- We provide new insights regarding the applicability of some standard image aesthetic appeal features in a general summarization scenario.
- We demonstrate that the existence of multiple "optimal" visual summaries leads to a low inter-user agreement that makes image set evaluation difficult. We therefore propose an automatic evaluation protocol based on the pyramid approach and motivated by the experience from the text domain that has been documented in the literature by the text summarization and machine translation communities.

In Section II we provide an overview of the proposed image selection approach and explain in more detail the rationale behind it. In Section III we report on related work. In Section IV our crowdsourcing experiment is described and then in Section V we present the features used to represent images. Our approach to user-informed image selection is introduced in Section VI. Section VII details the pyramid approach to summary/image set evaluation, while in Sections VIII and IX we present the experimental results. Finally, Section X concludes the paper.

II. Approach Overview and Rationale

Our approach to user-informed image selection for the purpose of summarizing an image collection is illustrated in Fig. 1. To allow us to develop a deeper understanding of how people create summaries of image collections, we first run a crowdsourcing experiment on the Amazon Mechanical Turk[1] platform and collect a large number of manually created visual summaries. The participants of the study were also asked to indicate the reasons for selecting a particular image for the summary, which helped us acquire insight into the general criteria that should be satisfied by an automatic summarization algorithm.

In the next step, we map these criteria on a number of features used to represent the images in the collection, both in terms of their individual properties and in the context of other images in the collection. Feature selection is steered by two main observations derived from the crowdsourcing experiment. First, we observed that the number of semantically related images in the original collection plays an important role when selecting an image for the summary (e.g., related to the paradigms of diversity and representativeness as introduced in the previous work [4], [6]). We consider images to be semantically related if they are captured at nearby locations (e.g., having the same or similar geo-coordinates) and are also visually similar to each other (e.g., depict the same scenes, objects or events). Images captured at the same geo-location, but with different depicted content are considered semantically different. Based on this understanding of semantic similarity, we consider geo-coordinates and standard images features, which reflect the saliency of the depicted visual content (object, scene) as the input for geo-visual clustering that reveals semantic links among the images in the collection.

We observed, however, that some other more subtle criteria also played an important role when the human summarizers were deciding on which images to select for the summaries. While typically a low inter-user agreement is expected regarding the inclusion of a specific image in a summary (probability is inversely proportional to the number of equally qualifying candidate images), it was striking to see that some images were selected by many users, far more often than other images. Based on the comments the users provided with their summaries, we concluded that an explanation of the criteria for image selection in these cases could be linked to the notions of *image aesthetic appeal* [11]–[13], *affect* and *sentiment* [14], [15] that have been investigated in various research contexts, such as e.g., image processing and computer vision, affective computing, natural language processing and social network analytics.

Therefore, similar to e.g., [12] we extract several image aesthetic appeal features (e.g., image colorfulness, aspect ratio) and consider image popularity indicators as well (i.e., view count and number of comments). For consistency reasons, we adopt notation from related work, where image aesthetic appeal features are considered to be those that influence aesthetic rating of an image [11]–[13]. Regarding the sentiment, similar to [15] we conjecture that the useful information might be derived from the comments posted on images, which often have an affective dimension. For the reasons of consistency with the related work, we refer to this particular step as the image sentiment analysis. Our sentiment analysis approach is based on publicly available Whissell's Dictionary of Affect in Language [16], attempting to quantify emotions in natural language. Finally, we investigate

[1] https://www.mturk.com/mturk/welcome

whether the targeted levels of appeal and sentiment can also be detected indirectly using various popularity indicators that can be derived from popular online image sharing sites.

The selected features serve as input into our proposed image selection approach. This approach aims at learning inherent properties that make images more or less likely to be selected for the summary by humans. We start from the reference summaries obtained through crowdsourcing and train a RankSVM [17] for each collection subset, providing frequently selected images as the positive and least frequently selected images as the negative examples. The final image ranking, which could be used as input when producing a visual summary, is generated by rank aggregation as explained in detail in Section VI.

III. Related Work

In this section we discuss previous work related to the problems and technologies addressed in the paper.

A. Visual Summarization

Generally, visual summarization aims at building a compact representation of a single video, set of videos or an image collection. Informedia [18] was probably one of the earliest projects addressing video summarization. More recently, TRECVID benchmark series run the BBC rushes summarization evaluation pilot (e.g., [1]), where the benchmark participants were provided 40 BBC rushes video files for each of which they were expected to generate visual summaries with up to 2% of the duration of the original file.

With the growing popularity of social media, a number of approaches for generating summaries of collections of community-contributed images have been proposed. Kennedy and Naaman [4] propose a multimodal approach to providing representative and diverse views of landmarks using Flickr images. In [10] travelogues and Flickr images are used for creating the summaries of touristic cities. Popescu *et al.* [19] make use of Flickr images and associated metadata for discovery and recommendation of tourist trips. Cao *et al.* [5] first cluster Flickr images using associated geo-coordinates and then represent each geo-cluster by the most representative images and the most frequent tags. In our previous work [6] we presented a multimodal approach to visual summarization of geographic areas using community contributed images. The approach makes use of visual content of the images, associated annotations (i.e., title, description and tags) as well as the information about users and their social network to select representative, but diverse images of a geographic area within a predefined radius from a selected location.

Visual summarization of data recorded by the wearable capturing devices is another example of application domain rapidly gaining popularity in the research community. For example, given a video recording of a wearable camera, Lee *et al.* [20] propose an approach, which jointly utilizes saliency detection and temporal event analysis for automatically generating visual summaries depicting the most important people and objects appearing in the video.

B. Summary Evaluation

Automatic summary evaluation has been a topic of intensive research in the (text) information retrieval community [7], [8] and although many different metrics have been proposed over the years, the evaluation problem still poses significant challenges. Since 2001, the Document Understanding Conference (DUC) series [21] and the successor series, Text Analysis Conference (TAC) have been the epicenter of research in the field of automatic summarization and summary evaluation [22]. The majority of the proposed metrics for summary evaluation have relied on the assumption that a good summary should be as similar as possible to one, or preferably more, human-created reference summaries. In [23], BLEU, an algorithm for automatic evaluation of machine translation was proposed. The main idea behind BLEU is to compare candidate translation with several reference translations (e.g., translations made by humans) using n-gram co-occurrence statistics. ROUGE [24] is another well-known example of the metric for evaluation of machine translation and automatic summarization, based on a similar idea.

A common problem with the automatic summary evaluation metrics such as e.g., ROUGE is a low agreement between human-created reference summaries. Therefore, based on the assumption that some summarization content units (SCUs) are more important and therefore should be given a higher weight when scoring summaries, a pyramid evaluation approach was proposed [25] and later adopted by TAC as the official summary evaluation metric [22]. Although it shows a high correlation with the human judgment about the quality of an automatically generated summary, the pyramid approach has the drawback that the SCUs need to be manually annotated.

Compared to the field of document summarization, the multimedia community has made relatively few attempts to systematically evaluate visual summaries. The participating video summarization systems in TRECVID BBC rushes benchmark [1] were evaluated using common metrics. However, the automatic evaluation was not the focus of the initiative and the summaries were judged on several parameters by the human evaluators.

Inspired by the well known BLEU [23] and ROUGE [24] metrics, Li and Merialdo [3] proposed VERT, an algorithm targeting automatic evaluation of video summaries. While BLEU and ROUGE compare candidate summary with several human-created reference summaries in terms of e.g., n-gram co-occurrence statistics, as a unit for comparison VERT analogously uses the “group of n keyframes” as an alternative. However, as will be illustrated in Section VII, a very low overlap between human-created reference summaries deems the evaluation metrics such as BLEU, ROUGE and VERT inapplicable to the task addressed in this paper.

C. Image Aesthetic Appeal and Sentiment Analysis

Estimating image aesthetic appeal as well as the sentiment that images evoke is a complex problem that has been a subject of intensive research. Approaches to image aesthetic appeal estimation aim at measuring the image properties that make it appealing to the user. In [26] a user study was conducted to identify those properties, which led to several categories of features related to e.g., people, composition/subject, quality (blur, contrast, etc.) and redundancy. Example image properties found by the similar studies to be correlated with the aesthetic appeal include image colorfulness, sharpness, rule of thirds, size, aspect ratio and face appeal features amongst many other [11]–[13], [27], [28].

As a result of the increased popularity of social media in the recent years, the analysis of sentiment evoked by multimedia content is becoming increasingly more sophisticated and easier to carry out. For example, from the comments posted in relation to a YouTube video or a Flickr image, it is often possible to understand whether users perceive the multimedia item as e.g., pleasing, happy or sad. Recently, the publicly available lexical resources such as e.g., Whissell's Dictionary of Affect in Language (DAL) [16] and SentiWordNet [29] have been proven effective in sentiment analysis of digital content. In the process of the creation of the DAL, a large number of words were annotated with regard to their *pleasantness (valence)* , *activation* and *imagery*. Similarly, in SentiWordNet, each synset of WordNet lexical database [30] is accompanied by *positivity*, *negativity* and *objectivity* sentiment scores. For example, in [31] DAL was successfully utilized for detection of narrative peaks in documentary videos, while in [32] SentiWordNet was deployed for predicting the rating of YouTube comments. In another recent study, Siersdorfer *et al.* [15] make use of SentiWordNet to analyze user-generated comments associated with the Flickr images and quantify their sentiment.

IV. Crowdsourcing for Visual Summarization

Our automatic image selection approach is informed by the large-scale user tests, which are carried out to investigate the criteria that guide user's selection of images for the visual summary. Below we first describe the image dataset used in the study and then elaborate on the setup and the lessons learned from the crowdsourcing experiment.

A. Image Collection

For the user tests we make use of Flickr image collection described in detail in our recent work [6]. We initially selected 500 geo-locations in Paris, France output by a location recommender system [33] and downloaded at most 100 creative commons (CC) licensed images captured within 1 km of each location together with the associated metadata such as e.g., title, keywords, description, comments, geotags (latitude and longitude), information on uploader and commenters. Finally, we kept only 207 locations for which 100 images were available. Downloaded images were selected based on a high Flickr popularity score, which ensures reasonable quality and relevance. The images were not pre-filtered according to the type or topic and thus reflect a wide spectrum of users' interests, such as e.g., landmarks, various types of events in both indoor and outdoor setting as well as the people in their everyday activities. Underlying variations in semantic density and visual homogeneity of 207 selected locations have a similar effect as varying the area size or sampling a varying number of images.

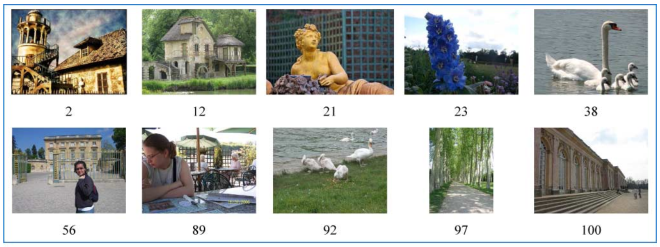
B. Crowdsourcing Experiment

Recently, crowdsourcing platforms such as e.g., Amazon Mechanical Turk (MTurk) and CrowdFlower[2] have emerged as the powerful tools for efficient and relatively inexpensive completion of tasks that require human intelligence. On MTurk, such tasks are called Human Intelligence Tasks (HITs) and can take various forms, such as e.g., translating text from one language to another, rating or tagging images, videos and music. While in the beginning, the majority of MTurk workers were US-based, the recent studies suggest a rapid internationalization of the MTurk labor force [34]. A number of studies have shown that with appropriate design of the HIT, a crowdsourcing platform will yield the same annotations or answers as conventional approaches for collecting judgments from users, e.g., in a laboratory setting [35], [36]. Since the crowdsourcing is a relatively young discipline, to assure a high quality of results and avoid spamming, the HIT design should be approached carefully. Namely, as suggested by [37] the quality of results depends on the factors such as e.g., payment amount per HIT, task complexity and worker qualification/reputation. However, the same study suggests that there is no universal recipe on how to choose those parameters. For example, increasing payment per HIT generally results in a higher quality of results, but it also attracts workers with a more sophisticated spamming methods. Similarly, while increasing the task complexity (effort) might lead to a higher amount of spam, it also yields a higher quality of results after the spam is removed. The study presented in [38] investigates techniques that help detect malicious workers and consequently reduce amount of spam. For example, the study suggests that the malicious users are less inclined to accept tasks involving free text inputs than e.g., those with check boxes.

Considering these and related recommendations for ensuring a high quality of results, we designed our crowdsourcing task as follows. We recruited 20 different MTurk workers per location for manual creation of reference summaries. As some of them repeated the HIT for the other locations as well, the total number of workers used for the task was 697. The images of a given location were displayed to the worker in 10 rows with 10 images each. To get a better overview of the entire location and fit images to the width of a computer screen, the height of each displayed image was set to 60 pixels. The workers were able to scroll vertically and horizontally and click on the image to see it in full resolution. In the task description, we avoided steering the workers towards any specific criteria for summary creation or to bias them by revealing information about the location. The precise wording of the task was: " *In this task we will show you a set of 100 images and ask you to select 10 of them for a "visual summary". The summary should capture the essence of the larger 100-image set. In other words, by looking at the 10-image visual summary, you should gain the same overall impression as given by the larger 100-image set.* "

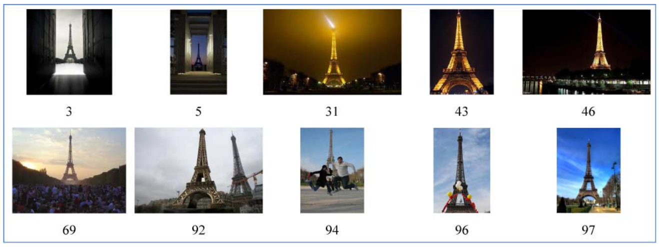
After the 10-image summary was created, the worker was asked to sort the selected images in the order of importance and briefly explain reasons for selecting each image using a free text input form. Beside helping us to understand the criteria for summary creation, sorting images in the order of importance and providing reasons for image inclusion in the summary using the free text form served also as another spam control mechanism. Further, the worker was expected to answer several questions about the properties of the original 100-image set, such as e.g., whether it was difficult to create summary of a given image set, whether the presented images in worker's personal opinion

[2] http://crowdflower.com/



38	There are several pictures of the swans, and this shows an adult with many children in the water
2	This shows a great work of public architecture
21	This shows an important sculpture
23	This shows the native flora of the land
12	This shows a great architectural work which appears to be a residential house
56	This is one of the tourists next to a landmark
97	This is a great picture, photographically speaking. Shows nature from a great angle
100	Another great photographic picture, which also includes an architectural landmark
89	Shows one of the tourists in a public restaurant or cafe
92	Shows the swans again, only this time some are on land

Fig. 2. An example visual summary manually generated by an MTurk worker. The images are further sorted in order of importance and the reasons for their inclusion in the summary are indicated.



97	Classic image of the tower, day
92	Closer, different angle, day
3	Framed, day shot of the full tower
5	Framed, evening shot of the full tower
94	Enjoying the sights, day shot
69	Nice shot with the horizon
31	Evening shot fully lit
43	Closer evening shot
46	Nice full shot with the lights in the evening
96	Picture of a celebration, decorations

Fig. 3. An example of behavior exhibited by a smaller number of workers to represent a particular collection by the images of its most dominant/representative landmark or event.

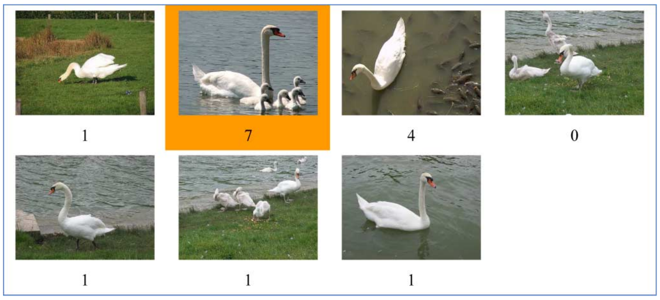


Fig. 4. An example showing several semantically related images captured in the area around a particular location. The numbers below each image indicate how many out of 20 workers selected that particular image for the visual summary.

show significant or important things and whether they are diverse. The answers were provided via a 4-point Likert scale. Finally, the free form text input was left for the feedback on task complexity, user friendliness, question ambiguity etc. An example visual summary made by a worker is shown in Fig. 2.

C. How Do Users Approach Visual Summarization

We first perform qualitative analysis of the manually generated visual summaries as well as the criteria for image selection reported by the MTurk workers. The analysis reveals that most of them select images that are semantically similar to many other images in the collection, making sure at the same time that as many semantically different images as possible are included in the summary. In this respect, this observation is in line with the previous user studies such as e.g., [4] and suggests that a trade-off between representativeness and diversity was targeted by the workers. However, we avoid making such explicit hypotheses in this paper as the analysis also revealed that humans often have distinct and individual perspectives on representativeness and diversity. Imposing the general expectations on a summary and using them to steer the design of a summarization algorithm would therefore be rather artificial and distract the summarization approach from reaching its goal. As an example we compare the summaries in Figs. 2 and 3 that have both been generated from the sets of highly diverse images showing various objects and events. Since the worker in Fig. 3 decided to include images of the Eiffel Tower only, considering exclusively representativeness and diversity as defined in the previous work would lead to intuitive conclusion that the worker does not consider diversity as an important criterion and that this summary is qualitatively worse than the one in Fig. 2. However, the worker in Fig. 3 does consider diversity, but at another semantic level (e.g., different views are selected, the images are captured during daytime and nighttime etc.). Such behavior is more frequently observed in the case of image collections including images of well-known objects or events (cf. Fig. 3).

Furthermore, we observed that the semantically similar images (e.g., showing the same object or event) were not necessarily considered by the workers as equally suitable for inclusion in the summary. For example, in a particular location for which a summary is shown in Fig. 2, 7 out of 100 images are depicting swans. As shown in Fig. 4, one of those images was included in the visual summary by 7 (out of 20) workers, which indicates an unusually high consensus (inter-annotator agreement). As already indicated in Section II, we relate this to the notions of image aesthetic appeal and sentiment, which we assume to have influenced the workers during the summary creation.

V. Feature Extraction

Based on the insights derived from the study in Sections III and IV, we propose an approach for automatic user-informed selection of images serving as input for visual summary. Here we first describe the categories of features we extract from the images in the collection. Then, in Section VI, we elaborate on the algorithm that deploys these features to learn to rank the images based on their suitability for the visual summary.

As mentioned in the introduction, one of the particular novelties of our approach is that we do not describe each image based on its properties only, but also in the context of the other semantically related images from e.g., the same geo-visual cluster. More particularly, we represent each image i with a feature vector $\mathbf{x}_i$ based on its “importance”, popularity, aesthetic appeal and sentiment evoked in the users, but also with the mean and variance of those features computed for the images within the same geo-visual cluster. The mean is expected to improve robustness of representation by propagating properties within a group of semantically similar images, while the variance indicates to what degree such propagation is justifiable (e.g., how much a particular feature varies among semantically similar images).

A. Geo-Visual Clustering

For each of 207 geographic areas (cf. Section IV-A), similar to [5], we first cluster images using their geo-coordinates. To cluster images into a certain number of geo-clusters, we make use of affinity propagation clustering [39], which was proven effective for the similar tasks in our previous work [6] as well as in [5] and [10]. Another property that makes the affinity propagation clustering preferable to some alternatives is its effectiveness in automatically determining the number of clusters.

The inputs into affinity propagation clustering are the similarities between images computed as

$$\mathbf{S}_g(i,j) = \mathrm{sim}(\mathbf{g}_i, \mathbf{g}_j) = e^{-\delta(lat_i, lon_i, lat_j, lon_j)} \tag{1}$$

where $\delta(lat_i, lon_i, lat_j, lon_j)$ is the great circle distance between geo-locations $\mathbf{g}_i = (lat_i, lon_i)$ and $\mathbf{g}_j = (lat_j, lon_j)$ associated with the images i and j.

After the geo-clusters are created, we produce the final geo-visual clusters by clustering images belonging to the same geo-cluster based on their visual features. The images are represented using a popular bag of visual words model (BoW) based on scale-invariant feature transform (SIFT) descriptors [40]. First, a certain number of keypoints are detected and described using the SIFT detector and descriptor. Further, k-means clustering is used to cluster the descriptors extracted from all images of a certain geographic area into 500 clusters (visual words). Finally, an image is represented with a 500-bin histogram, where each bin corresponds to a visual word in the codebook. In the following step, we cluster images from a particular geo-cluster into a certain number of visual clusters. For that, we again utilize the affinity propagation clustering using as the input image visual similarities

$$\mathbf{S}_v(i,j) = \mathrm{sim}(\mathbf{f}_i, \mathbf{f}_j) = e^{-\|\mathbf{f}_i - \mathbf{f}_j\|^2} \tag{2}$$

where $\mathbf{f}_i$ and $\mathbf{f}_j$ are the BoW feature vectors (histograms) of images i and j.

We conjecture that a frequency of appearance of an object or event in the images throughout the collection indicates its importance for the visual summary. Therefore, given the detected geo-visual clusters $C_l, l = 1 \ldots k$, for an image i from the cluster C_l, we define the first component of the feature vector $\mathbf{x}_i$ of image i as $x_{i1} = |C_l|/N$, where N is the total number of images per location (here set to 100, as explained in Section IV-A).

B. Image Popularity

In photo sharing websites such as e.g., Flickr, image view count and number of comments are generally believed to be correlated, at least weakly, with the user-perceived image aesthetic appeal. As such information is usually relatively easy to obtain and does not imply additional computational costs, without going into a deeper analysis of the factors that influence popularity of social media, we decided to include it in our image representation.

View Count: An image is represented by its view count (x_{i2}) as well as the mean and variance of the view counts of images in the same geo-visual cluster (x_{i3} and x_{i4}).

Number of Comments: Number of comments posted on an image together with the mean and variance of the number of comments associated with the images belonging to the same cluster are added as x_{i5}, x_{i6} and x_{i7}.

C. Image Aesthetic Appeal

To model image aesthetic appeal we make use of proven and computationally inexpensive aesthetic appeal indicators, i.e., image aspect ratio, colorfulness, luminance and sharpness.

Aspect Ratio: Our user study indicates that the users have a strong preference towards "landscape" image orientation or in other words the images having larger width than height. The exceptions are e.g., images of a particularly tall building such as Eiffel Tower (cf. Fig. 3). We compute the aspect ratio as $x_{i8} = w/h$, where w and h are the image width and height. Additionally, we represent an image with the mean and variance of the aspect ratio of all images from the same geo-visual cluster (x_{i9} and x_{i10}).

Colorfulness: Image colorfulness is evaluated using a metric proposed in [28], which shows a high correlation with human perception. Then, an image i is represented with its estimated colorfulness (x_{i11}) as well as the mean and variance of the colorfulness of the images belonging to the same geo-visual cluster (x_{i12} and x_{i13}).

Luminance: To calculate the global luminance of an image, we first convert it from the RGB to YCbCr color space and then compute the mean value of the Y-channel in all pixels. The image i is represented with its luminance (x_{i14}) as well as the mean and variance of the luminance of all images belonging to the same geo-visual cluster (x_{i15} and x_{i16}).

Sharpness: Image sharpness is evaluated using the publicly available software [41], which computes the cumulative probability of blur detection (CPBD) at the edges in the image [42]. Similar to colorfulness and luminance, we represent each image with its estimated sharpness (x_{i17}) as well as the mean and variance of sharpness of semantically related images from the same geo-visual cluster (x_{i18} and x_{i19}).

D. Sentiment Analysis

Compared to some other content sharing websites, such as e.g., YouTube, Flickr images are associated with a smaller average number of comments, which are often not very polarized. While in YouTube a controversial semantic theme of a video might cause an intensive discussion amongst visitors, such behavior is less frequently observed in Flickr. Still, as recently suggested in [15], Flickr comments might carry a valuable information for estimating sentiment of an image.

Since Flickr comments are often written in different languages, we first translate them all into English using Google Translate service. Further, for the terms appearing in the Whissell's Dictionary of Affect in Language (DAL) we obtain the valence, activation and imagery scores. Valence value indicates the level of pleasantness or unpleasantness that a particular word expresses, activation indicates the associated arousal level and the imagery designates whether a particular word is easy or hard to imagine. For example, the word *beautiful* is associated with a maximum valence value 3, while the word *terrible* has the lowest valence of 1. Contrary to the word *love*, associated with a relatively high activation of 2.6, the word *scenery* has an activation of only 1.2. Finally, an example of the word with the lowest imagery of 1 is *like*, while the words designating objects, such as e.g., *camera* or *house* are associated with a high imagery value of 3. Although in e.g., narrative peak detection scenarios [31] usually only valence and activation are utilized, we conjecture that even imagery could provide a potentially valuable information for determining sentiment of a comment. For example, a high imagery of the words in the comments on a Flickr image might indicate an absence of feedback containing strong sentiments or rather descriptive nature of the comments.

Although, in general, natural language processing (NLP) may prove beneficial for the sentiment analysis, here we choose not to perform it for several reasons. Namely, as already mentioned earlier in this section, the Flickr comments are relatively short, seldom polarized and frequently express appreciation of the image, which simplifies the sentiment analysis and reduces the need for NLP. Additionally, since the sentiment analysis is not the main focus of this paper, we choose to perform it in a simple and computationally inexpensive manner that was proven effective in related work such as e.g., [43].

We compute the mean valence, activation and imagery values for the words in a comment and then average it over all comments posted on that image (feature vector components x_{i20}, x_{i21} and x_{i22}). Finally, we also represent an image with the mean and variance of valence (x_{i23} and x_{i24}), activation (x_{i25} and x_{i26}) and imagery (x_{i27} and x_{i28}) features across images belonging to the same geo-visual cluster.

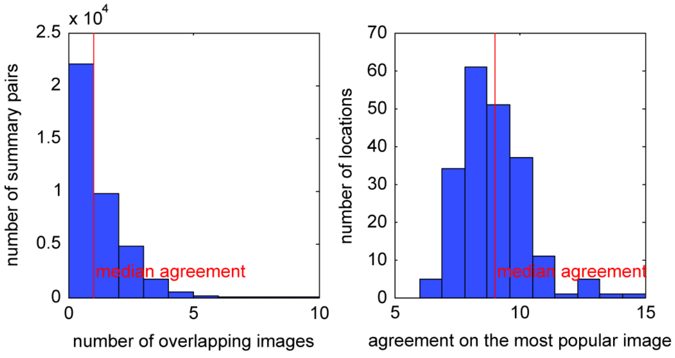
VI. User-Informed Image Selection

To facilitate the selection of images for the summary we set as our target to produce a ranked list of images per location, where the rank position of an image serves as an indicator of its suitability for the visual summary. We approach learning to generate the ranked list in a user-informed fashion, first by selecting the training images from the human-created reference summaries and then by learning the ranking function taking the features from the previous section as the input.

We start the training data selection by sorting the images per location (collection subset consisting of 100 images) according to the number of MTurk workers that selected them for their summaries. Further, we choose a set of image preference pairs, $(i, j) \in \mathcal{P}$, each consisting of a top ranked and a bottom ranked image. In the selected set of preference pairs $\mathcal{P}$, a top or bottom-ranked image i can appear in only one preference pair and for each preference pair $(i, j) \in \mathcal{P}$, image i is preferred over image j. Then, to learn the ranking model, a well-known RankSVM method [44] could be used. In the method originally proposed by Joachims in [44], the RankSVM model is based on minimizing the following objective function

$$\frac{1}{2}\|\mathbf{w}\|^2 + C \sum_{(i,j)\in\mathcal{P}} \ell(\mathbf{w}^\top \mathbf{x}_i - \mathbf{w}^\top \mathbf{x}_j) \tag{3}$$

where $\mathbf{x}_i$ and $\mathbf{x}_j$ are the feature vectors representing images i and j, respectively, C is a regularization parameter and ℓ is a loss function, such as e.g., $\ell(z) = \max(0, 1 - z)$ in case of SVMLight implementation [44]. However, due to the relatively high computational costs associated with training of SVMLight, here we make use of a fast RankSVM method described in [17], whose clear notation we adopt in (3). The method is based on Newton optimization and avoids explicit computing of all possible difference vectors $\mathbf{x}_i - \mathbf{x}_j$ to significantly reduce the RankSVM training time.



(a) (b)

Fig. 5. Illustration of the agreement between human-made reference summaries; (a) Histogram of the size of overlap between reference summaries produced by different workers, which shows the number of summary pairs having a particular number of images in common. (b) Histogram of the level of agreement on the most popular image, which shows the number of locations for which a particular number of workers selected the most popular image for their summary. As the most popular image in a given location (collection subset), we consider an image that appears in the largest number of summaries.

As described in Section IV-A, the locations in Paris at which the images were captured are often rather different in terms of both semantic density and visual homogeneity. We conjecture that the images selected for the visual summary by an MTurk worker must be considered in the context of images of that particular geographic area. For example, their diversity and representativeness strongly depends on e.g., the diversity of the starting image set, whether the objects and events depicted in the images are perceived as significant or important etc. Also, an image might be selected not because it is particularly appealing, but simply because most of the other images are perceived as unappealing. Therefore, we train RankSVM separately for each of t locations (collection subsets) in the training set. Given a location from the test image set, we apply the trained models and produce t lists of images ranked according to their suitability for the visual summary. Finally, a rank aggregation algorithm is applied to produce the final image ranking.

VII. The Pyramid Approach to Set Evaluation

As discussed in Section III-B, a common problem in evaluation of e.g., document summaries and machine translations is a low inter-user agreement (e.g., [25]). Fig. 5(a) shows a histogram of the level of agreement between summaries manually produced by the MTurk workers. The histogram indicates that the agreement is in general very low, with the mean of 1.5 and median of 1. In other words, two reference summaries usually have only one image in common, which makes the evaluation algorithms such as e.g., BLEU [23], ROUGE [24] and VERT [3] practically inapplicable. However, we also observe a high inter-user agreement in case of some images. We conjecture that

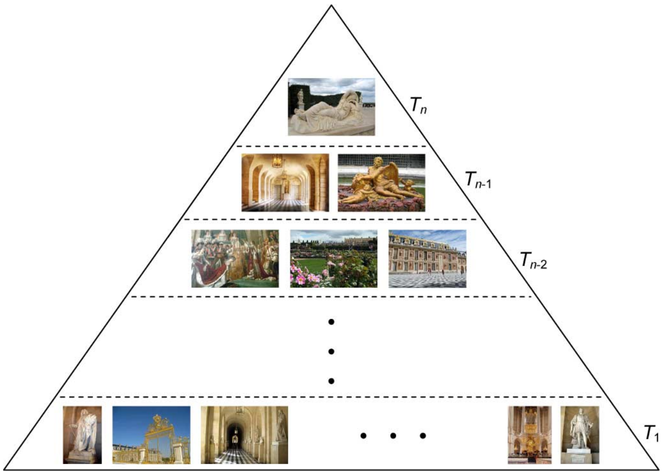


Fig. 6. Illustration of the pyramid structure, where each tier consists of the images appearing in the same number of reference summaries. Each image in the bottom tier T_1 appears in only one reference summary, while the images in the top tier T_n are those most frequently selected for the visual summary.

those images, frequently appearing in the reference summaries are indeed the most important for the visual summary. The histogram in Fig. 5(b) shows in how many locations the workers agree on the most popular image, where the image is considered as the most popular if it appears in the largest number of reference summaries.

We observe that in each collection subset, there is at least one image that has been selected for the visual summary by at least 6 workers and that the median agreement on the most popular image per location is 9. To optimally exploit the inter-user agreement, we follow the idea of [25] and propose a pyramid approach for evaluating the suitability of images for the visual summary. As illustrated in Fig. 6, each pyramid tier consists of the images appearing in the same number of visual summaries. The most frequently selected images are placed in the top tier, while the bottom tier is composed of images that were selected by a single MTurk worker only. Images that do not appear in any of 20 reference summaries generated for a given location are considered unimportant and therefore discarded. For example, in the particular case of location for which an illustration is shown in Fig. 6, the pyramid has 9 tiers and the image in the top tier appears in 11 out of 20 reference summaries.

We conjecture that an *optimal* set should include all images from the upper tiers and draw the remaining images from the last tier needed to reach a specified set size. In case of pyramid depicted in Fig. 6, an optimal 5-image set should include all images from the tiers T_n and T_{n-1} as well as 2 images from the tier T_{n-2}. Obviously, several optimal sets can be created as described above and in this particular example the number of such optimal sets is 3. According to the pyramid approach an optimal set $\tilde{R}$ with N_R images would receive the maximum score $d_{\max}$ computed as follows

$$d_{\max} = \sum_{i=\theta+1}^{n} i \times |T_i| + \theta \times \left(N_R - \sum_{i=\theta+1}^{n} |T_i| \right),$$

$$\theta = \max_{i} \left(\sum_{j=i}^{n} |T_j| \geq N_R \right). \qquad (4)$$

Then, an arbitrary set R with N_R images receives the score d

$$d = \frac{1}{d_{\max}} \times \sum_{i=1}^{n} i \times |T_i \cap R|. \qquad (5)$$

For example, as the pyramid depicted in Fig. 6 has 9 tiers, the optimal 5-image set would receive the maximum score $d_{\max} = 1 \times 9 + 2 \times 8 + 2 \times 7 = 39$.

In Section IX-A we will demonstrate that the pyramid score is indeed effective in evaluating the quality of an image set.

VIII. Experimental Setup

A. Baselines Used for Evaluation of the Pyramid Score

In Section IX-A the effectiveness of the pyramid score is evaluated through comparison of the values obtained for reference summaries and the summaries composed of either the least popular images or the summaries output by the approaches that do not take into account image popularity, aesthetic appeal or sentiment. More particularly, for comparison we use the following baselines.

Low View Count: The images with the lowest view count are selected.

RWR-RD [6]: The approach utilizes random walk with restarts over a multi-layer graph modeling text associated with the images, visual features extracted from them as well as the information about users and their social network to select a set of representative and diverse images of a particular geographic area. The approach is designed such to show various aspects of the area, but it is unaware of image popularity, aesthetic appeal or sentiment.

MA Clustering: The approach is based on the same multi-layer graph [6] as the RWR-RD approach described above and utilizes random walk with restarts algorithm to compute multi-modal image similarities. The images are further clustered using the affinity propagation clustering [39] based on the computed similarities and the cluster centroids are selected for the result image set. Like RWR-RD, the approach does not focus on aesthetic properties of the images and their popularity.

Ensemble Clustering: The images are first clustered independently using the low-level visual features and the text associated with them [6] and then the ensemble clustering approach [45] is applied to produce a single, reinforced clustering. Finally, the clusters' visual centroids are selected for the visual summary of a collection. The approach does not make use of information about image popularity, aesthetics or sentiment.

B. Baselines Used for Image Selection Evaluation

In Section IX-A we evaluate our proposed image selection approach by means of the pyramid score and compare it with two intuitive control baselines (Random and High VC) as well as the proven visual summarization approaches (MAC-VC and EC-VC).

Random: Images are randomly sampled from the collection. We find it important to report the performance of a random baseline in scenarios such as the one described in this paper, to investigate whether the performance of the tested approaches differs significantly from random.

High VC: Images are selected based on a high view count. Although view count in general might be considered as an unreliable popularity indicator due to e.g., ease of manipulation and bias towards highly popular content causing the long tail problem [46], it is usually considered to be (weakly) correlated with the aesthetic appeal and sentiment.

MAC-VC: A modification of MA Clustering approach described in the previous section. Instead of choosing cluster centroids for the final results list, an image with the highest view count is selected to represent each cluster.

EC-VC: A variant of Ensemble Clustering approach described in the previous section, which, instead of choosing visual centroids, samples an image with the highest view count from each cluster for the final results list.

C. Training RankSVM and Rank Aggregation

As explained in Section VI, we train RankSVM model separately for all t locations in the training set and produce t ranked lists of images for a test location. We experimentally set the number of preference pairs $|\mathcal{P}| = 20$ (cf. Section VI) as a tradeoff between three factors—the number of training samples (preferably larger), the quality of samples (preferably only a small fraction of top and bottom ranked samples should be used) and the total number of images per collection subset (in this particular case—100). Once the individual ranked lists are produced, the final ranking is generated through rank aggregation. In the past decade a number of approaches to rank aggregation have been proposed [47], [48]. In our exploratory experiments the approach proposed by Pihur *et al.* [48] yielded a good performance, but due to a high computational complexity and the fact that the main focus of this paper are not the approaches for rank aggregation, we opted for a lightweight alternative. Here we perform the rank aggregation by simply computing the average rank of an image across all t lists. In our exploratory experiments such approach was proven to yield insignificantly lower performance than computationally intensive alternatives such as e.g., [48].

IX. Experimental Results

Through the experiments presented in this section we aim to answer the following research questions:

A. Is the pyramid score introduced in Section VII effective in estimating the quality of an image set?
B. Does our proposed approach succeed in selecting a set of images suitable for visual summarization?
C. Is the performance well distributed across locations/collection subsets?
D. Which features are the most important for isolating images with desired properties?
E. What is the relationship between different features?
F. Is our proposed approach applicable in case of image collections missing information richness of social media?

A. Evaluation of the Pyramid Score

We conjecture that a good evaluation metrics should yield a significantly higher scores for the reference summaries manually generated by the MTurk workers than for apparently lower-quality image sets or image sets automatically generated without taking into account sophisticated features, such as those related to e.g., image aesthetic appeal and sentiment. Our goal is also to investigate how the scores change with the varying number of reference summaries used to create a pyramid. Therefore, we vary the number of reference summaries used for pyramid building from 2 to 18 and compute the scores for the remaining reference summaries and three summarization approaches described in Section VIII-A: LVC, RWR-RD, MAC and EC. The scores obtained for the reference summaries are simply averaged for easier comparison. All scores obtained for a particular approach under the same setting are averaged across all locations.

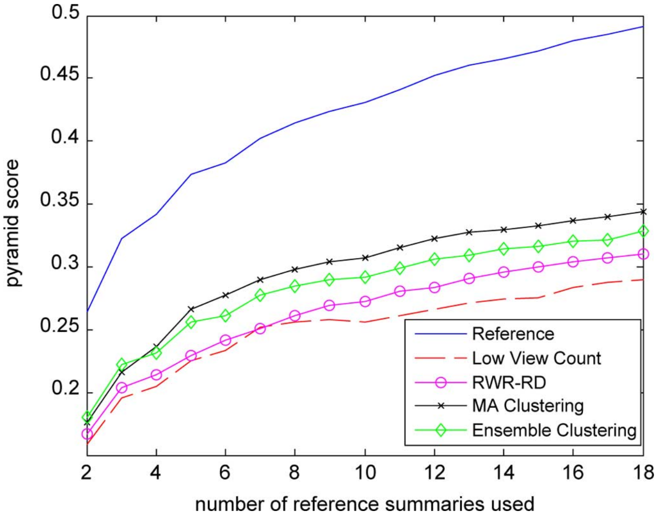


Fig. 7. Variation of pyramid score depending on the number of reference summaries used for pyramid construction. The scores are computed for the remaining reference summaries and the four visual summarization approaches.

The graphs in Fig. 7 show that the computed scores generally grow with the increasing number of reference summaries used to construct the pyramid. Further, the scores averaged over remaining reference summaries are significantly higher than those computed for a set of images selected based on a low view count and the baselines that do not take into account image aesthetic appeal and sentiment.

In Fig. 8 we show for which percentage of locations image set produced in a particular way yields the highest score. This percentage increases with the increasing number of summaries used to construct the pyramid. Again, the pyramid score appears to be effective in discriminating between the high quality image sets manually created by the MTurk workers and those created automatically.

B. Evaluation of the Proposed Image Selection Approach

Here we compare the performance of our proposed approach for user-informed image selection with the performance of several competitive baselines described in Section VIII-B: Random, High VC, MAC-VC and EC-VC. As the Figs. 7 and 8 indicate that the margin between scores computed for different approaches increases with the increasing number of reference summaries, for pyramid construction we make use of all 20 manually created summaries. We opt for a "leave-one-out" strategy simultaneously training RankSVM on $t = 206$ collection subsets and apply the trained model on the remaining

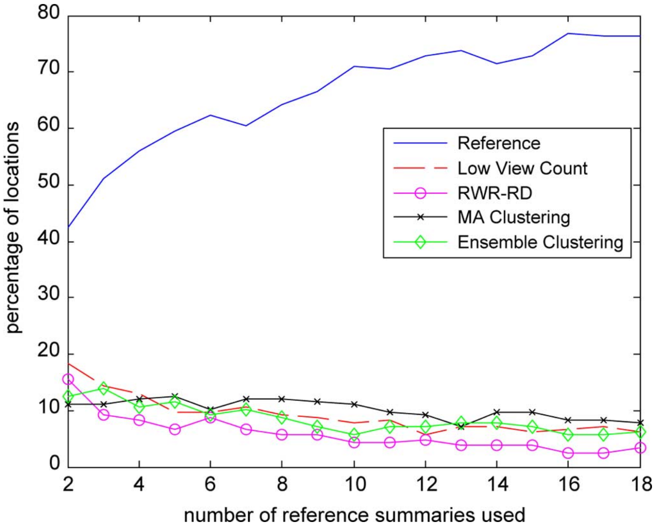


Fig. 8. Comparison of the pyramid scores assigned to the reference summaries and the four visual summarization approaches. The percentage of locations for which a particular approach yields the highest score is reported.

TABLE I
PERFORMANCE OF OUR RSVM-CAS SELECTION APPROACH AND THE FOUR BASELINES REPORTED IN TERMS OF PYRAMID SCORE AVERAGED OVER ALL 207 LOCATIONS

Selection Method	$N_R = 5$	$N_R = 10$	$N_R = 15$	$N_R = 20$
Random	0.2530	0.2559	0.3008	0.3500
High VC	0.4135	0.4497	0.4778	0.5015
MAC-VC	0.3450	0.3881	0.4320	0.4565
EC-VC	0.3622	0.4104	0.4417	0.4643
RSVM-CAS	**0.5660**	**0.5736**	**0.5961**	**0.6216**

subset (location). Finally, for easier comparison we report the scores averaged over all 207 locations.

The performance comparison of our RSVM-CAS selection approach and the four baselines in terms of pyramid score averaged over all 207 locations is presented in Table I. Our proposed approach clearly selects higher-quality image sets of various sizes N_R. Further, although Random image selection yields a reasonable collection sampling in terms of e.g., representativeness and diversity [6], this approach does not take into account criteria found important by the users when creating visual summaries, such as e.g., image aesthetic appeal and sentiment. Finally, view count might be considered as a solid selection strategy in cases when a low computational complexity is required. However, view count alone is often seen as an unreliable popularity indicator as it can be unavailable and manipulated, but it can also lead to a bias towards the mainstream content. Although our proposed RSVM-CAS approach makes use of view count and number of comments, we conjecture that the other features modeling image aesthetic appeal, sentiment and context make it more robust to those and similar negative factors.

C. Performance Distribution Across Image Collection

TABLE II
PERFORMANCE OF OUR RSVM-CAS SELECTION APPROACH AND THE FOUR BASELINES REPORTED IN TERMS OF PERCENTAGE OF LOCATIONS FOR WHICH A PARTICULAR APPROACH IS THE BEST PERFORMER

Selection Method	$N_R = 5$	$N_R = 10$	$N_R = 15$	$N_R = 20$
Random	2.4	0.5	0.5	0.5
High VC	16.9	18.8	15.5	15.9
MAC-VC	10.6	4.3	7.7	5.4
EC-VC	8.3	5.8	9.6	7.7
RSVM-CAS	**61.8**	**70.6**	**66.7**	**70.5**

TABLE III
RANKED LIST OF FEATURES SORTED BY THEIR EFFECTIVENESS IN DISCRIMINATING BETWEEN IMAGES APPEARING MOST FREQUENTLY IN THE REFERENCE SUMMARIES AND THOSE SELECTED LEAST FREQUENTLY

Rank	Feature	Rank	Feature
1	aspect ratio (x_{i8})	15	mean sharpness (x_{i18})
2	mean aspect ratio (x_{i9})	16	luminance (x_{i14})
3	colorfulness (x_{i11})	17	var view count (x_{i4})
4	view count (x_{i2})	18	sharpness (x_{i17})
5	nr comments (x_{i5})	19	cluster size (x_{i1})
6	valence (x_{i20})	20	mean luminance (x_{i15})
7	activation (x_{i21})	21	var nr comments (x_{i7})
8	mean view count (x_{i3})	22	var imagery (x_{i28})
9	imagery (x_{i22})	23	var valence (x_{i24})
10	mean colorfulness (x_{i12})	24	var activation (x_{i26})
11	mean activation (x_{i25})	25	var sharpness (x_{i19})
12	mean valence (x_{i23})	26	var colorfulness (x_{i13})
13	mean nr comments (x_{i6})	27	var aspect ratio (x_{i10})
14	mean imagery (x_{i27})	28	var luminance (x_{i16})

To investigate whether the performance of our proposed RSVM-CAS approach is well distributed across the collection, we compute the percentage of locations for which a particular approach performs better then the alternatives. As shown in Table II, our proposed RSVM-CAS approach is the best performer in the largest number of locations for various sizes N_R of the output image set.

D. Analysis of Feature Discriminativeness

Here we compare the effectiveness of each feature used in discriminating between images that appear frequently in the reference summaries and the least popular ones. For each location we select 20 images appearing most frequently in the reference summaries and treat them as the positive class. Similarly, for the negative class we select 20 images that appear least frequently in the reference summaries. Further, we perform the forward feature selection for classification using the 1-Nearest Neighbor error criterion, which first selects a single most discriminative feature and which further iteratively selects the feature that improves most the discriminativeness of the feature set. Once the list of features sorted according to their discriminativeness is produced for each location, we perform the rank aggregation by averaging the rank of each feature across all 207 ranked lists. The ranked list of features is shown in Table III.

Surprisingly, image aspect ratio and colorfulness features emerge as the most discriminative, which further confirms our assumption that the users are to a large extent driven by image aesthetic appeal when selecting images for the visual summary. For example, the most frequently occurring aspect ratios (w/h) in the entire image collection are 1.2723, 1.4164, 0.7300, 1.4085, 0.6521 and 0.9540, while the most frequent aspect ratios amongst the images selected by the MTurk workers are

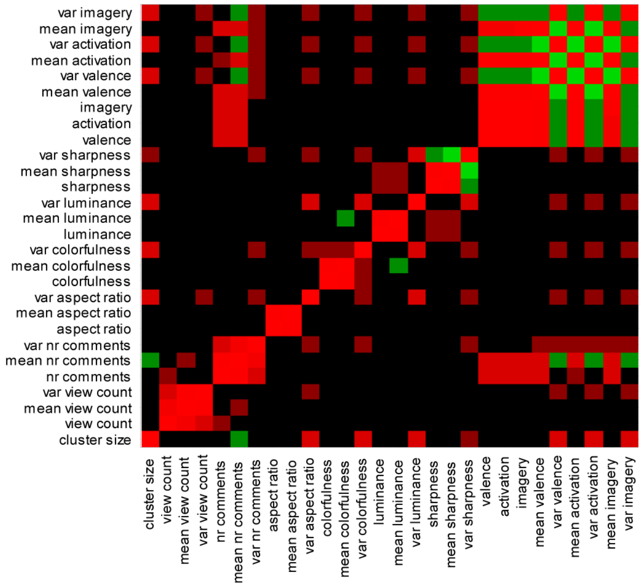


Fig. 9. Relationship between features expressed in terms of median correlation coefficient computed over all locations. Red and green colors indicate positive and negative correlation.

1.3333, 1.5015, 0.7500, 1.4970, 0.6660 and 1.0000. Further, a high discriminativeness of mean aspect ratio feature confirms our assumption about importance of image context. We conjecture that in the case of e.g., "panoramic" spots many images will have a similar aspect ratio that best captures the content of the scene. In that sense, a similar aspect ratio of the images in a particular geo-visual cluster might be (implicitly) indicative of e.g., interestingness or visual appeal of a view from that location. Lightweight popularity indicators such as e.g., view count and number of comments are also positioned high in the ranked list. Finally, sentiment features extracted from image comments, namely valence, activation and imagery fall into the group of the most discriminative features as well. Valence and activation appear to be more important than imagery, which is not surprising, since those features provide more explicit information about sentiment of a word.

On the other hand, sharpness and luminance appear to be less important than the other aesthetic appeal and sentiment features. Also, a relatively low rank of cluster size feature might indicate that aesthetic attributes of the image as well as the sentiment it evokes play a more important role than the representativeness and diversity. Finally, when considering contextual features (i.e., mean and variance of a particular feature computed for semantically similar images, e.g., those in the same geo-visual cluster), mean is to be preferred to variance. This observation may suggest that the feature variability within a set of images belonging to the same geo-visual cluster is relatively small.

E. Relationship Between Different Features

To complement the experiment from the previous section, here we investigate the correlation between different features. The heat map in Fig. 9 visualizes the relationship between features expressed in terms of median correlation coefficient computed over all 207 locations.

As shown in Fig. 9, there is no apparent correlation between view count and the image sentiment features—valence, activation and imagery. Also, image aesthetic appeal features including image aspect ratio and colorfulness, which emerged as the most discriminative features in the previous section, seem to be uncorrelated with the view count and number of comments. However, the number of comments shows a certain degree of correlation with the valence, activation and imagery, which is somewhat expected considering the fact that those features were extracted from the image comments. Finally, we observe a high correlation between valence, activation and imagery features.

TABLE IV
PERFORMANCE OF OUR RSVM-CA AND RSVM-CAS SELECTION APPROACHES REPORTED IN TERMS OF PYRAMID SCORE AVERAGED OVER ALL 207 LOCATIONS

Selection Method	$N_R = 5$	$N_R = 10$	$N_R = 15$	$N_R = 20$
RSVM-CA	0.5294	0.5342	0.5602	0.5781
RSVM-CAS	**0.5660**	**0.5736**	**0.5961**	**0.6216**

F. Extension to Non-Annotated Image Collections

Compared to rich social media, offline collections are often poorly (if at all) annotated and images are lacking the useful information such as e.g., title, description, tags, comments and view count. Here we investigate the effectiveness of our approach in such cases when only information automatically captured by the camera is available, i.e., image content and automatically captured geo-coordinates. Although the geo-tags available in Flickr are sometimes manually inserted by the users, for the purpose of this experiment we consider them all to be automatically generated by the capturing device. We conjecture that the increasing availability of capturing devices (e.g., cameras and smart phones) with a high positioning accuracy, make the scenario realistic. In the cases when the geo-coordinates are not available at all, a clustering described in Section V-A could be performed based on e.g., visual features only.

Following the scenario described above, we retain only the following features: cluster size (x_{i1}), aspect ratio (x_{i8}), mean aspect ratio (x_{i9}), var aspect ratio (x_{i10}), colorfulness (x_{i11}), mean colorfulness (x_{i12}), var colorfulness (x_{i13}), luminance (x_{i14}), mean luminance (x_{i15}), var luminance (x_{i16}), sharpness (x_{i17}), mean sharpness (x_{i18}) and var sharpness (x_{i19}). Performance of our proposed approach utilizing contextual and image aesthetic appeal features only (RSVM-CA) is shown in Table IV.

Comparing the results in Table IV with those in Table I we observe that the RSVM-CA manages to outperform the approaches utilizing rich information available in social media, while being agnostic to image aesthetic appeal and image sentiment. However, the performance drop compared to RSVM-CAS confirms the importance of popularity indicators and image sentiment features.

X. DISCUSSION AND FUTURE WORK

We have used information about how humans select images for visual summaries, which was collected with a large-scale crowdsourcing study, as the basis for a novel method for automatically selecting images for visual summarization. The crowdsourcing study revealed inherent properties of images that are important for humans and also provided us with training data. Our approach uses features based on these properties and RankSVM method to generate a list of images ranked by their suitability for inclusion in a visual summary. As such, the selected image set can be used as a "general purpose" visual summary or as a starting point in building a summary with particular properties.

We discuss a phenomenon of a low inter-user agreement and prove effectiveness of the metric based on the pyramid score in evaluating the quality of a selected set of images. Both the evaluation metric and our image selection approach are tested on a collection of geo-referenced Flickr images. Under various conditions our approach has proven effective in generating image sets composed of images that are frequently selected for the visual summaries by humans. The approach shows a potential for use in both information-rich social media environments as well as in the case of non-annotated image collections.

Both our large-scale user study and the analysis of feature discriminativeness indicate the effectiveness of the computationally inexpensive image aesthetic appeal features. Our analysis places image popularity indicators and sentiment features in the group of most discriminative features and their use brings an additional improvement in the system. Surprisingly, no apparent correlation has been found between image aesthetic appeal features and the popularity indicators, which might indicate that some other properties have a larger impact on the popularity of social media. We leave a deeper study of the relations between different features for the future work.

Although we prove the effectiveness of the pyramid score in evaluating the quality of selected image sets from the aspect of selection of images found by users as suitable for visual summarization, it does not explicitly evaluate attributes such as e.g., image set diversity. We believe that the largest potential for better incorporating diversity into the evaluation metric is a more sophisticated means of determining the semantic similarity between images. Further, we demonstrate that the amount of data and a low inter-user agreement deem the traditional evaluation metrics such as e.g., ROUGE and BLEU practically inapplicable, creating a need for the metrics taking into account specificities of multimedia content. Also, as we show in this paper, the way the users perceive the image selection criteria and their interplay are often more complex than the related work in the field often suggests. In our future work we will further investigate those criteria and the means to evaluate them.

Currently we are estimating sentiment of the comments posted in response to the images only, but we plan to investigate whether useful affective information might be extracted from image title, tags and description generated by the uploader. Finally, our future work will also include a deeper analysis of the factors that influence effectiveness of our approach.

REFERENCES

[1] P. Over, A. F. Smeaton, and G. Awad, "The TRECVID 2008 BBC rushes summarization evaluation," in *Proc. 2nd ACM TRECVID Video Summarization Workshop, ser. TVS '08*. New York, NY, USA: ACM, 2008, pp. 1–20.

[2] D. Simakov, Y. Caspi, E. Shechtman, and M. Irani, "Summarizing visual data using bidirectional similarity," in *Proc. 2012 IEEE Conf. Computer Vision and Pattern Recognition*, 2008, vol. 0, pp. 1–8.

[3] Y. Li and B. Merialdo, "VERT: automatic evaluation of video summaries," in *Proc. Int. Conf. Multimedia, ser. MM '10*. New York, NY, USA: ACM, 2010, pp. 851–854.

[4] L. S. Kennedy and M. Naaman, "Generating diverse and representative image search results for landmarks," in *Proc. 17th Int. Conf. World Wide Web, ser. WWW '08*. New York, NY, USA: ACM, 2008, pp. 297–306.

[5] L. Cao, J. Luo, A. Gallagher, X. Jin, J. Han, and T. Huang, "A worldwide tourism recommendation system based on geotagged web photos," in *Proc. 2010 IEEE Int. Conf. Acoust. Speech and Signal Processing (ICASSP)*, Mar. 2010, pp. 2274–2277.

[6] S. Rudinac, A. Hanjalic, and M. Larson, "Generating visual summaries of geographic areas using community contributed images," *IEEE Trans. Multimedia*, IEEE Early Access Article.

[7] D. R. Radev, E. Hovy, and K. McKeown, "Introduction to the special issue on summarization," *Comput. Linguist.*, vol. 28, no. 4, pp. 399–408, Dec. 2002.

[8] D. Harman and P. Over, "The DUC summarization evaluations," in *Proc. 2nd Int. Conf. Human Language Technol. Res., ser. HLT '02*, 2002, pp. 44–51, Morgan Kaufmann Publishers.

[9] J. Carbonell and J. Goldstein, "The use of mmr, diversity-based reranking for reordering documents and producing summaries," in *Proc. 21st Annu. Int. ACM SIGIR Conf. Res. and Develop. in Inform. Retrieval, ser. SIGIR '98*. New York, NY, USA: ACM, 1998, pp. 335–336.

[10] 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 Understand.*, vol. 115, no. 3, pp. 352–363, Mar. 2011.

[11] A. K. Moorthy, P. Obrador, and N. Oliver, "Towards computational models of the visual aesthetic appeal of consumer videos," in *Proc. 11th Eur. Conf. Comput. vision: Part V, ser. ECCV'10*. Berlin, Germany: Springer-Verlag, 2010, pp. 1–14.

[12] P. Obrador and N. Moroney, "Low level features for image appeal measurement," *Proc. SPIE, Image Quality and System Performance VI, ser. IS&T/SPIE*, vol. 7242, pp. 72420T–1, 2009.

[13] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Studying aesthetics in photographic images using a computational approach," in *Proc. 9th Eur. Conf. Computer Vision—Volume Part III, ser. ECCV'06*. Berlin, Germany: Springer-Verlag, 2006, pp. 288–301.

[14] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," in *Proc. Conf. Human Language Technol. and Empirical Methods in Natural Language Processing, ser. HLT '05*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2005, pp. 347–354.

[15] S. Siersdorfer, E. Minack, F. Deng, and J. Hare, "Analyzing and predicting sentiment of images on the social web," in *Proc. Int. Conf. Multimedia, ser. MM '10*. New York, NY, USA: ACM, 2010, pp. 715–718.

[16] C. Whissell, "Using the revised dictionary of affect in language to quantify the emotional undertones of samples of natural language," *Psychol. Rep.*, vol. 105, pp. 509–521, Oct. 2009.

[17] O. Chapelle and S. S. Keerthi, "Efficient algorithms for ranking with SVMS," *Inf. Retriev.*, vol. 13, no. 3, pp. 201–215, Jun. 2010.

[18] H. D. Wactlar, T. Kanade, M. A. Smith, and S. M. Stevens, "Intelligent access to digital video: Informedia project," *IEEE Computer*, vol. 29, no. 5, pp. 46–52, 1996.

[19] A. Popescu, G. Grefenstette, and P.-A. Moëllic, "Mining tourist information from user-supplied collections," in *Proc. 18th ACM Conf. Inform. and Knowledge Manage., ser. CIKM '09*. New York, NY, USA: ACM, 2009, pp. 1713–1716.

[20] Y. J. Lee, J. Ghosh, and K. Grauman, "Discovering important people and objects for egocentric video summarization," in *Proc. 2012 IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, Jun. 2012, pp. 1346–1353.

[21] H. T. Dang, "Overview of DUC 2006," in *Proc. Document Understanding Conf., ser. DUC '06*, 2006.

[22] K. Owczarzak and H. T. Dang, "Overview of the TAC 2011 summarization track: Guided task and AESOP task," in *Proc. Text Anal. Conf.*, 2011.

[23] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting on Association for Computational Linguistics, ser. ACL '02*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002, pp. 311–318.

[24] C. Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Proc. ACL-04 Workshop Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, 2004, pp. 74–81.

[25] A. Nenkova and R. J. Passonneau, "Evaluating content selection in summarization: The pyramid method," *HLT-NAACL*, pp. 145–152, 2004.

[26] A. E. Savakis, S. P. Etz, and A. C. Loui, "Evaluation of image appeal in consumer photography," *Proc. SPIE Human Vision and Electron. Imaging V*, vol. 3959, pp. 111–120, Jun. 2000.

[27] S. Winkler, "Visual fidelity and perceived quality: Towards comprehensive metrics," in *Proc. SPIE Human Vision and Electron. Imaging Conf., ser. Lecture Notes in Comput. Sci.. SPIE*, 2001, vol. 4299, pp. 114–125.

[28] D. Hasler and S. E. Süsstrunk, "Measuring colorfulness in natural images," in *Society of Photo-Optical Instrumentation Engineers (SPIE) Conf. Series*, Jun. 2003, vol. 5007, pp. 87–95.

[29] A. Esuli and F. Sebastiani, "SentiWordNet: A publicly available lexical resource for opinion mining," in *Proc. 5th Conf. Language Resources and Evaluation, ser. LREC '06*, 2006, pp. 417–422.

[30] G. A. Miller, "WordNet: a lexical database for English," *Commun. ACM*, vol. 38, no. 11, pp. 39–41, Nov. 1995.

[31] B. Jochems, M. Larson, R. Ordelman, R. Poppe, and K. P. Truong, "Towards affective state modeling in narrative and conversational settings," in *Proc. Interspeech 2010*, Sep. 2010, pp. 490–493.

[32] S. Siersdorfer, S. Chelaru, W. Nejdl, and J. S. Pedro, "How useful are your comments?: analyzing and predicting YouTube comments and comment ratings," in *Proc. 19th Int. Conf. World Wide Web, ser. WWW '10*. New York, NY, USA: ACM, 2010, pp. 891–900.

[33] M. Clements, "Personalised access to social media," Ph.D. dissertation, TU Delft, Delft, The Netherlands, 2010.

[34] J. Ross, L. Irani, M. S. Silberman, A. Zaldivar, and B. Tomlinson, "Who are the crowdworkers?: shifting demographics in mechanical turk," in *CHI '10 Extended Abstracts on Human Factors in Computing Systems, ser. CHI EA '10*. New York, NY, USA: ACM, 2010, pp. 2863–2872.

[35] S. Nowak and S. Rüger, "How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation," in *Proc. Int Conf. Multimedia Inform. Retrieval, ser. MIR '10*. New York, NY, USA: ACM, 2010, pp. 557–566.

[36] G. Paolacci, J. Chandler, and P. G. Ipeirotis, "Running experiments on Amazon Mechanical Turk," *Judgment and Decision Making*, vol. 5, no. 5, pp. 411–419, Aug. 2010.

[37] G. Kazai, "In search of quality in crowdsourcing for search engine evaluation," in *Advances in Information Retrieval, ser. Lecture Notes in Computer Science*. Berlin, Germany: Springer, 2011, vol. 6611, pp. 165–176.

[38] C. Eickhoff and A. de Vries, "How crowdsourcable is your task?," in *Proc. CSDM '11*, 2011.

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

[40] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[41] N. D. Narvekar and L. J. Karam, CPBD Sharpness Metric Software. [Online]. Available: http://ivulab.asu.edu/Quality/CPBD.

[42] N. Narvekar and L. Karam, "A no-reference image blur metric based on the cumulative probability of blur detection (CPBD)," *IEEE Trans. Image Process.*, vol. 20, no. 9, pp. 2678–2683, Sep. 2011.

[43] K. Denecke, "Using SentiWordNet for multilingual sentiment analysis," in *Proc. IEEE 24th Int. Conf. Data Eng. Workshop, 2008 (ICDEW 2008)*, Apr. 2008, pp. 507–512.

[44] T. Joachims, "Optimizing search engines using clickthrough data," in *Proc. 8th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, ser. KDD '02*. New York, NY, USA: ACM, 2002, pp. 133–142.

[45] A. Strehl and J. Ghosh, "Cluster ensembles—a knowledge reuse framework for combining multiple partitions," *J. Mach. Learn. Res. (JMLR)*, vol. 3, pp. 583–617, Dec. 2002.

[46] Y.-J. Park and A. Tuzhilin, "The long tail of recommender systems and how to leverage it," in *Proc. 2008 ACM Conf. Recommender Syst., ser. RecSys '08*. New York, NY, USA: ACM, 2008, pp. 11–18.

[47] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar, "Rank aggregation methods for the web," in *Proc. 10th Int. Conf. World Wide Web, ser. WWW '01*. New York, NY, USA: ACM, 2001, pp. 613–622.

[48] V. Pihur, S. Datta, and S. Datta, "Rankaggreg, an R package for weighted rank aggregation," *BMC Bioinformat.*, vol. 10, no. 1, pp. 62–62, 2009.

Stevan Rudinac received the Dipl. Ing. degree in electrical engineering from the University of Belgrade, Serbia in 2006 and the Ph.D. degree in computer science from the Delft University of Technology, The Netherlands in 2013. From 2006 to 2007 he conducted research at the University of Belgrade and from 2007 to 2008 at the Eindhoven University of Technology, The Netherlands. In the period 2008–2012 he was a Ph.D. Researcher in the Multimedia Signal Processing Group at the Delft University of Technology. Since 2012, he has been working as a Forensic Scientist at the Netherlands Forensic Institute. His research interests lie in the field of multimedia information retrieval with focus on video search and visual summarization.

Martha Larson holds an MA and Ph.D. in theoretical linguistics from Cornell University and a B.S. in Mathematics from the University of Wisconsin. Her research interest and expertise lie in the area of speech- and language-based techniques for multimedia information retrieval. She is co-founder of the MediaEval Multimedia Benchmark and has served as organizer of a number of workshops in the areas of spoken content retrieval and crowdsourcing. She has authored or co-authored over 100 publications. Currently, Dr. Larson is assistant professor in the Multimedia Information Retrieval Lab at Delft University of Technology. Before coming to Delft, she researched and lectured in the area of audio-visual retrieval at Fraunhofer IAIS and at the University of Amsterdam.

Alan Hanjalic is a professor, holder of the Antoni van Leeuwenhoek Chair and head of the Multimedia Signal Processing Group at the Delft University of Technology, The Netherlands. His research focus is on multimedia search, recommender systems and social media analytics. Prof. Hanjalic is a member of the IEEE TC on Multimedia Signal Processing and the Steering Committee of the IEEE TRANSACTIONS ON MULTIMEDIA. He was a Keynote Speaker at the Pacific-Rim Conference on Multimedia 2007 and the International Multimedia Modeling Conference 2012. He has been a member of Editorial Boards of several scientific journals in the multimedia field, including the IEEE TRANSACTIONS ON MULTIMEDIA, the IEEE TRANSACTIONS ON AFFECTIVE COMPUTING and the *International Journal of Multimedia Information Retrieval*. He has also held key positions (General or Program (Co-)Chair) in the organizing committees of leading multimedia conferences, among which the ACM Multimedia, ACM CIVR/ICMR and IEEE ICME.