STYLE-CENTRIC IMAGE SUMMARIZATION FROM PHOTOGRAPHIC VIEWS OF A CITY

Wei-Yi Chang and Yu-Chiang Frank Wang

Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan

ABSTRACT

Visual summarization addresses the task of selecting images from an image collection, so that the sampled images would contain representative information which sufficiently highlights the collected visual data. In this paper, we solve the problem of style-centric visual summarization using photographic landmark images of a city. Different from existing works which typically retrieve landmark images based on salient visual appearances, our proposed method is able to produce different sets of summarized images, while each set corresponds to a particular image style. This is achieved by performing unsupervised clustering on images within and across landmark categories, which discovers the common photographic styles from the input image collection. Our experiments will confirm that, compared to standard clustering algorithms, our approach is able to achieve satisfactory summarization outputs with style consistency.

***Index Terms*—** Visual Summary, Image Understanding, Clustering

1. INTRODUCTION

With the rapid growth of the Internet, a large number of photographic images can be found from online albums (e.g., *Flickr* [1]), which allows one to collect and summarize the images of any topic of interest. For example, one can collect the images of a celebrity or a series of photos taken at a ceremony. However, when the number of such images is remarkably large, it will be very time consuming for the user to decide which images to collect. Moreover, one would also expect high variations for the summarized outputs due to the diversity of online images.

In this paper, we focus on summarizing the photographic images of a city with particular style preferences. Among image styles, we particularly consider the styles of *atmosphere* and *color* as defined in [1], not those associated with optical techniques (e.g., HDR) or composition. Using the photos of different landmarks taken at a particular city, we aim to develop an algorithm for clustering the collected images, with the ability to identify the representative landmark images with style consistency. While our work is very different from existing location-based visual applications [2] (e.g., landmark retrieval [3, 4] or trip planning [5]), our developed approach can be further integrated into the above tasks for improving the user's quality of experience.

Fig. 1. Random vs. style-centric photographic summarization for the city of Rome.

In order to visually summarize a city, we consider the photographic images of its landmarks. With the help of social networks, previous works have been proposed to discover the landmarks by mining from blogs [6] or geo-tagged images and check-in data [7]. With the attractive landmarks determined, Chen *et al.* [8] proposed a framework for automatically generating tourist maps with landmark icons, while these icons were generated from the representative images. Since a landmark could exhibit different visual appearances due to time and weather changes, Min *et al.* [9] applied topic models to generate visual summarization of landmarks by considering both viewpoints and scenes, which resulted in more comprehensive summarization results. In addition, Papadopoulos *et al.* [10] utilized photo clusters to identify the corresponding landmarks and events in a city. And, Rundinac *et al.* [11] organized the photos of a city with representative and diverse (e.g., hotel, store, and landmark) images. However, to the best of our knowledge, existing visual summarization works does not consider the style preferences during summarization.

As illustrated in Figure 1, instead of performing visual summarization of a city without any style consistency, we aim at producing the summarized outputs with particular pref-

[1] http://www.flickr.com/

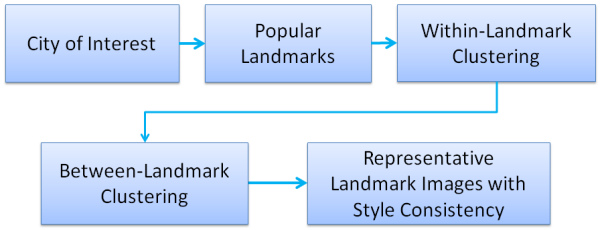


Fig. 2. The proposed framework for style-centric visual summarization.

erences in styles. To achieve this, we need to identify the foreground regions of each landmark image, followed by the discovery of image styles within and across landmark categories. With the common styles across landmark categories determined, the representative landmark image of each style will be selected for final visual summarization.

2. OUR PROPOSED METHOD

We now detail our proposed method for style-centric visual summarization, which performs within and between-landmark clustering to identify common image styles. The flowchart of our proposed method is illustrated in Figure 2.

2.1. Within-Landmark Clustering

When summarizing the photographic city images using their landmark photos, one can expect the collection of multiple images of the same landmark. Thus, before determining the common styles for the city images across landmarks, we need to first identify the images of the same landmark into subgroups, each corresponds to images taken under similar setting (e.g., time, lighting, etc.). Since it is not practical to assume the number of such subgroups to be known in advance, we apply an agglomerative hierarchical clustering to group the similar images. Once this bottom-up clustering process is complete, the dominant groups would be applied for cross-landmark style discovery (as discussed in Section 2.2).

Figure 3 shows an example of within-landmark clustering for the images of *Castel Sant Angelo*. It can be seen that, different clusters in Figure 3 are associated with images of distinct photographic styles. As noted above, we only consider the dominant clusters (i.e., the clusters with image numbers above a predetermined threshold) for performing cross-landmark style discovery. This also alleviates the undesirable effects of outliers (i.e., images with rare styles) for visual summarization.

2.2. Between-Landmark Clustering

2.2.1. Common style selection

Given images with distinctively dominant photographic styles, our goal is to discover the common styles across the

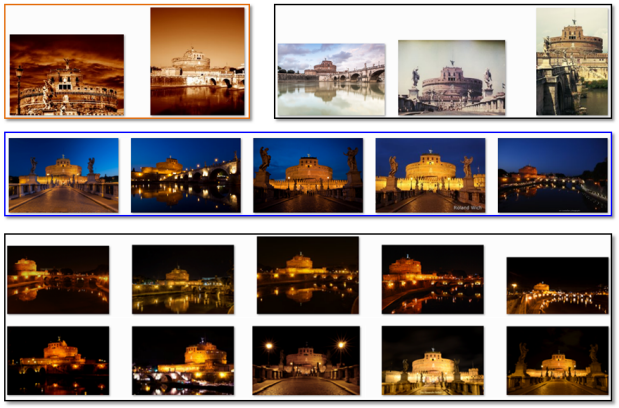


Fig. 3. Example results of within-landmark clustering for *Castel Sant Angelo*.

images of different landmarks. However, performing clustering simply over landmark images will not be able to solve this task. This is due to the fact that, if doing so, there is no guarantee that the output cluster would contain all landmark.

To address the above problem, we decompose all the landmark images into foreground (i.e., landmark) and background regions by the saliency detection approach of [12]. Then, we consider the background image regions across all landmark categories for common style selection, as detailed below.

Given N images $(x_1, x_2, \ldots, x_N)$ across L different landmarks, we determine the K-nearest neighbors ($\mathcal{N}_i^K$) for each image x_i. We then calculate a L-dimensional histogram for x_i, in which entry denotes the number of its neighbors belonging to the associated landmark category. Thus, the histogram matrix $\boldsymbol{H} \in \mathbb{R}^{N \times L}$ can be constructed, in which each element is defined as:

$$h_{ij} = \sum_{x_k \in \mathcal{N}_i^K} I(x_k, j), \text{ where } I(x_k, j) = \begin{cases} 1, & \text{if } x_k \in j \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

Here, $I(x_k, j)$ indicates whether photo x_k belongs to landmark j. With $\boldsymbol{H}$, we denote each background image with neighbors from the associated landmark category. To further determine whether such images belong to a common or unique style, we calculate the *landmark diversity* as follows:

$$div(x_i) = \sum_{j=1}^{L} I(h_{ij}), \text{ where } I(h_{ij}) = \begin{cases} 1, & \text{if } h_{ij} > T_L \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

In (2), the diversity value of image x_i returns the number of landmarks (between 1 and L), which observes more than T_L images as the K-nearest neighbors of x_i. According to the landmark diversity, we mark an image with a common style label if its landmark diversity value is above a predetermined threshold T_C. On the other hand, if its landmark diversity is below T_C, this image will be marked as a unique style label (and not considered for later clustering/summarization). In our work, we fix $T_L = 3$ and $T_C = 2$ for all our experiments.

(a) Common photo styles

(b) Unique photo styles

Fig. 4. Example clusters of (a) common and (b) unique photo styles for images of Rome. Note that four landmark categories are available.

2.2.2. *Common style discovery with diversity information*

Once the images of common photographic styles are marked, the problem of common style discovery turns into a clustering task. That is, to discover the styles (clusters) that contain all selected landmarks, we apply the images with common style label for clustering.

It is worth noting that, in addition to visual features, we also incorporate the landmark diversity information into our clustering process by normalizing each row in $\boldsymbol{H}$ as an additional feature vector. Then, inspired by [13], we advance a multi-view k-means clustering (MVKMC) algorithm on both visual feature and landmark diversity features to discover the common styles from landmark images.

Let $\boldsymbol{X}_{(v)} \in \mathbb{R}^{d_v \times N}$ denote the d_v-dimensional feature in v-th view and M types of heterogenous features, MVKMC is performed by solving:

$$\min_{D_{(v)}, A, \alpha_{(v)}} \sum_{v=1}^{M} \alpha_{(v)} \|\boldsymbol{X}_{(v)} - \boldsymbol{D}_{(v)}\boldsymbol{A}\|_{2,1}$$
$$s.t.\ A_{kj} \in \{0,1\}, \sum_{k=1}^{K} A_{kj} = 1, \sum_{v=1}^{M} \alpha_{(v)} = 1, \qquad (3)$$

where $\alpha_{(v)}$ is the weight factor for the v-th view, $\boldsymbol{D}_{(v)} \in \mathbb{R}^{d_v \times K}$ is the centroid matrix for v-th view, and $\boldsymbol{A} \in \mathbb{R}^{K \times N}$ is the consensus common cluster indicator matrix. Figure 4 shows example results of our clustering outputs.

2.3. Summarization of Landmark Images

After discovering the common styles (clusters) from the background regions of images across landmarks, the final task is to select a representative landmark image for each style for completing the visual summarization process. For each cluster C_a associated with a common style, we select the representative landmark image based on the following strategy:

$$x_r = \arg\min_{x_r} \sum_{x_i \in j} dist(x_r, x_i), \text{ where } x_r, x_i \in C_a. \qquad (4)$$

We note that, for this final stage of visual summarization, we consider each image with foreground regions (i.e., landmark) presented. From (4), we see that image x_r will be selected for representing landmark j, if the distance $dist(x_r, x_i)$ between it and other images x_i of the same landmark in the same style cluster is minimum.

Once a particular common style is of interest to the user, our proposed method is able to output representative landmark images of that style, which completes the process of style-centric visual summarization.

Table 1. Selected cities and their landmarks (based on the information available on *Foursquare*).

City	landmarks
Rome	Altare della Patria, Basilica di San Pietro, Castel Sant Angelo and Colosseum
Paris	Arc de Triomphe, Eiffel Tower, Notre Dame and Louvre Museum

3. EXPERIMENT

3.1. Dataset and Settings

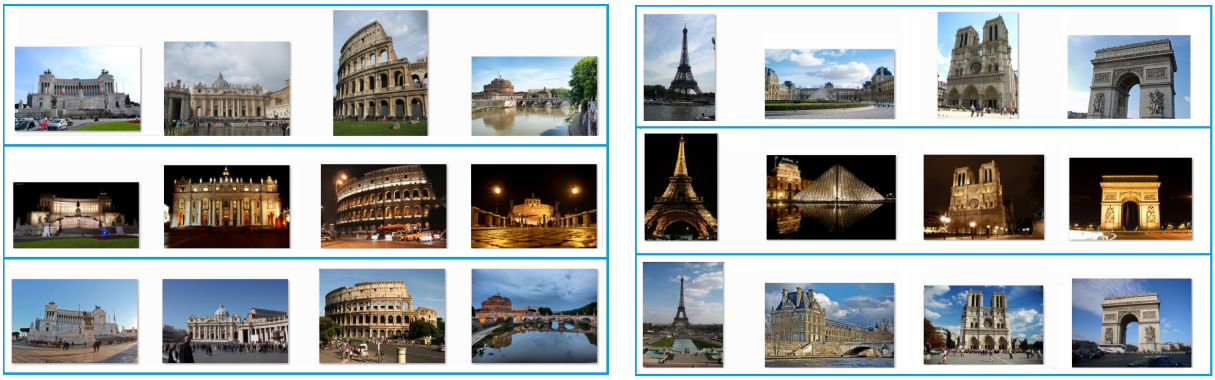
We now evaluate the performance of our proposed method. Since we focus on summarizing city photographic images, we consider the popular location-based social network of *Foursquare* [2] to collect popular landmarks for the city of interest (about 220 images for each landmark). In our experiments, we consider the cities of Paris and Rome. Once the landmarks are determined, we search and collect the corresponding images from both *Flickr* and *Flickr15K* [14] as our image data. The selected cities and their landmarks are listed in Table 1. To describe each landmark image, a 1024-dimensional Lab color histogram is considered.

3.2. Evaluation

To compare the summarization performance of different approaches, we consider different clustering algorithms, including k-means (KM), spectral clustering (SC), and affinity propagation (AP), for discovering common image styles. For all these approaches, once the clustering results are obtained, we apply (4) to visualize the clustering/summarization results.

Since we expect that the summarized landmark images are style-consistent, we define and apply *similarity sum* (*SimSum*)

[2] http://foursquare.com/



(a) our style-centric approach for Rome

(b) our style-centric approach for Paris

Fig. 5. Example visual summarization results for (a) Rome and (b) Paris. Note that in each figure, each row represents a common photographic style, and each column shows the selected landmark images.

Table 2. Results of similarity sum (*SimSum*) for selected summarization from two cities.

SimSum	KM	SC	AP	Ours
Rome-C1	4.7191	4.9597	4.8158	4.9597
Rome-C2	4.5331	4.5331	4.2824	4.5862
Rome-C3	4.2675	3.7990	4.5637	4.3036
Paris-C1	4.3949	4.1844	4.3949	4.7152
Paris-C2	5.9172	5.9141	5.7746	5.9090

to evaluate the performance of visual summarization. Given L landmark images as summarized output, *SimSum* is calculated by summing up the similarity between images:

$$SimSum = \sum_{i=1}^{L} \sum_{j=i+1}^{L} sim(i,j), \tag{5}$$

where $sim(i,j)$ is the cosine similarity of image feature between images i and j.

Table 2 shows the *SimSum* results of different clustering algorithms, where each row represents the results generated from similar clusters. Due to space limit, we only present selected clusters with higher similarity as those produced by our approach (note that the cluster similarity is calculated by Jaccard coefficient). From Table 2, we see that most results produced by our method were with larger *SimSum* values. It means that, compared to other existing clustering approaches, our proposed method was able to achieve proper visual summarization with style consistency. Figure 5 shows example results of our visual summarization for the cities of Rome and Paris, respectively. From our summarization outputs with different image styles, one can easily pick the set of selected landmark images whose style is most preferable to him/her.

Since we only consider the image styles that contain all the landmark of a city for summarization, we further define the full landmark ratio (FLR) for evaluating the effectiveness of common style discovery:

Table 3. Full landmark ratios (FLR) for different methods.

	KM	SC	AP	Ours
Rome	0.4611	0.4778	0.5000	**0.5625**
Paris	0.3571	0.3524	0.1905	**0.5000**
Average	0.4091	0.4151	0.3453	**0.5313**

$$FLR = \frac{\#\text{cluster contain all landmark}}{\#\text{cluster}}. \tag{6}$$

As shown in Table 3, we see that our method was able to better produce full landmark clusters (with a larger FLR ratio). That means, improved visual summarization results can be achieved by our proposed method. From the above experiments, the use of our approach for style-centric summarization can be successfully verified.

4. CONCLUSIONS

In this paper, we proposed an unsupervised clustering framework for visual summarization. Given different landmark images of a city, our method performed clustering of images within and between landmark categories, which identified the common image styles from the image collection, while the rare and unique ones were disregarded automatically. With the extracted common image styles, representative images for each landmark would be automatically selected, which completes the process of style-centric visual summarization. In our experiments, we considered landmark images of Rome and Paris. For each city, our method was able to produce summarized image outputs with style consistency, which supports the use of our work for practical visual summarization tasks with style preferences.

Acknowledgement This work is supported in part by the Ministry of Science and Technology of Taiwan via MOST103-2221-E-001-021-MY2.

5. REFERENCES

[1] S. Karayev, M. Trentacoste, H. Han, A. Agarwala, T. Darrell, A. Hertzmann, and H. Winnemoeller, "Recognizing image style," *BMVC*, 2014.

[2] R. Ji, Y. Gao, W. Liu, X. Xie, Q. Tian, and X. Li, "When location meets social multimedia: A survey on vision-based recognition and mining for geo-social multimedia analytics," *ACM TIST*, 2015.

[3] Y. Avrithis, Y. Kalantidis, G. Tolias, and E. Spyrou, "Retrieving landmark and non-landmark images from community photo collections," *ACM Multimedia*, 2010.

[4] D.-T. Dang-Nguyen, L. Piras, G. Giacinto, G. Boato, and F. G. B. De Natale, "A hybrid approach for retrieving diverse social images of landmarks," *IEEE ICME*, 2015.

[5] X. Lu, C. Wang, J.-M. Yang, Y. Pang, and L. Zhang, "Photo2trip: generating travel routes from geo-tagged photos for trip planning," *ACM Multimedia*, 2010.

[6] R. Ji, X. Xie, H. Yao, and W.-Y. Ma, "Mining city landmarks from blogs by graph modeling," *ACM Multimedia*, 2009.

[7] J. Liu, Z. Huang, L. Chen, H. T. Shen, and Z. Yan, "Discovering areas of interest with geo-tagged images and check-ins," *ACM Multimedia*, 2012.

[8] W.-C. Chen, A. Battestini, N. Gelfand, and V. Setlur, "Visual summaries of popular landmarks from community photo collections," *ACM Multimedia*, 2009.

[9] W. Min, B.-K. Bao, and C. Xu, "Scene and viewpoint based visual summarization for landmarks," *IEEE ICIP*, 2014.

[10] S. Papadopoulos, C. Zigkolis, S. Kapiris, Y. Kompatsiaris, and A. Vakali, "ClustTour: City exploration by use of hybrid photo clustering," *ACM Multimedia*, 2010.

[11] S. Rudinac, A. Hanjalic, and M. Larson, "Generating visual summaries of geographic areas using community-contributed images," *IEEE TMM*, 2013.

[12] W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency optimization from robust background detection," *IEEE CVPR*, 2014.

[13] X. Cai, F. Nie, and H. Huang, "Multi-view k-means clustering on big data," *IJCAI*, 2013.

[14] R. Hu and J. Collomosse, "A performance evaluation of gradient field hog descriptor for sketch based image retrieval," *Comput. Vis. Image Understad.*, 2013.