ARTEMIS @ MediaEval 2013: A Content-Based Image Clustering Method for Public Image Repositories

Andrei Bursuc Titus Zaharia
Institut Mines-Télécom; Télécom SudParis; ARTEMIS Department
UMR CNRS 8145 MAP5
9, rue Charles Fourier, 91011 Evry Cedex
{Andrei.Bursuc, Titus.Zaharia}@telecom-sudparis.eu

ABSTRACT
In this paper, we describe our approach and its results as part of the MediaEval 2013 Retrieving Diverse Social Images Task evaluation. We illustrate a content-based technique relying on a single type of visual descriptors that makes it possible to identify groups of similar instances of a given landmark and select the most representative images from each such group out of a set of relatively noisy or redundant images. This method builds for each landmark a matching graph through quantized interest points similarities and identifies groups of similar instances of the landmark as connected components. In this approach we do not make use of the textual metadata of the images or any other external source of information.

1. INTRODUCTION
The goal of the Retrieving Diverse Social Images task [1] is to identify a subset of meaningful representative images for a landmark, given a corpus of images that are retrieved through textual queries and GPS coordinates from popular image repositories such as Flickr. While most of the images crawled for each landmark (up to 150) are relevant, many instances are redundant and some other ones are noisy. This set should be then reduced to a compact collection of representative images, while taking into account both the relevance and the diversity of the selected images for the given landmark. For this purpose a collection of pictures from 396 locations from all around the world is available together with raw metadata, visual features and textual models. We chose to tackle this task from the visual description perspective in order to find visually meaningful instances of a given landmark.

2. METHODOLOGY
Our approach leverages on the visual descriptors extracted from interest points for identifying precise correspondences between images depicting the same landmark. In order to enhance the speed of the computation of the image correspondences, we employ the Bag of Visual Words model (BoVW) [2] which clusters the interest point descriptors into a reduced set of descriptors. The images are then transposed into a landmark matching graph, where images are nodes and edges connect similar images. Multiple instances of the given landmark are identified as connected components in the landmark graph, and from each such component the dominant images are chosen as being representative. The advantage of this method lies in the fact that image similarities can be computed quickly through the BoVW vectors intersection. In addition, during the similarity computation we identify common descriptors between pairs of images; descriptors which can be further used to build a rich visual description of the identified representative images from each cluster. Thus, new collections of images of the same landmark can be assigned to a cluster quickly as they are matched only with the visual descriptor built for each cluster.

MediaEval 2013 Workshop, October 18-19, 2013, Barcelona, Spain

2.1 Visual Description
In order to identify reliable correspondences between images, we employ the Hessian-Affine co-variant regions [3] along with the RootSIFT descriptors [4], which have proven their effectiveness in various visual retrieval tasks.

The matching of the interest point descriptors between all pairs of images for each landmark is usually very lengthy. In order to reduce the computational time while keeping the accuracy of interest point matches, we employ the BoVW model together with a large vocabulary. Thus, we cluster the descriptors extracted from all images (approx. 40M RootSIFT descriptors) into a vocabulary of 500K visual words and then quantize all descriptors into BoVW feature vectors. Such a large vocabulary reduces the quantization errors that can typically occur for the BoVW model (*i.e.*, assigning multiple different descriptors to the same visual word). This allows us to transfer the interest point matching into the quantization of the BoVW vectors. Two descriptors that have been quantized to the same visual word, are likely to be similar and to be also identified as a corresponding pair when performing one-to-one matching at the images level. This operation is highly useful when new images are added to the existing corpus and similar images are thus quickly identified, since the matching is done only once for the quantization. The spatial consistency of the matches if verified through fast geometric consistency verification based on LO-RANSAC [5] which rejects most of the outliers.

2.2 Matching graph
In order to discard the false positive, noisy or redundant images and to identify the most representative images for a given landmark, we identify the correspondences between all pairs of images in the visual descriptor space. We consider that two images contain the same instance of the landmark if they have at least M_{min} geometrically verified and back-projected correspondences.

The role of this exhaustive matching procedure is twofold. Firstly, it makes it possible to reject the false positive images that have been retrieved. Usually such false positives are quite different from the rest of the true positive images and they will be cleared out when matched with the rest of true positives. Secondly, the different instances of the landmark (*e.g.*, different viewpoints, different weather conditions) that are found in the raw image sets can be grouped together in clusters through the matching.

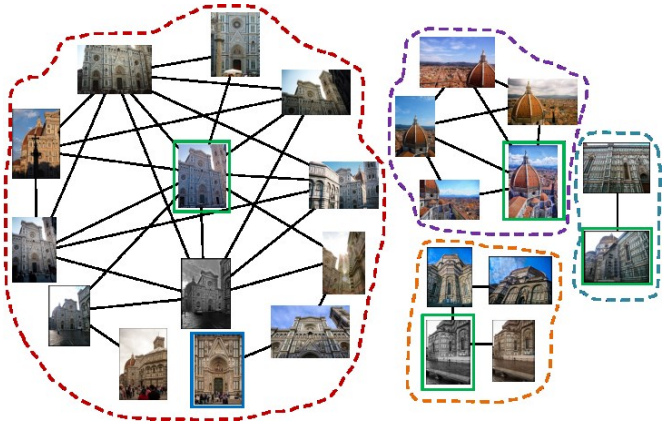

Figure 1. Landmark graph for the topic *214 - Saint Mary of the Flower*. 4 clusters are identified with 4 corresponding representative images (green bounding box). An extra diverse image (blue bounding box) is selected from the largest cluster (red dashed line).

In order to identify the different instances of the landmark, we employ the matching results and construct a landmark graph. The nodes of this graph are images and the edges connect similar images which have at least M_{min} verified matches. An example of such a graph is illustrated in Figure 1. A set of images have been discarded, as such images have been identified as isolated nodes with little similarity to other images in the data set. In general, the number of images to be considered is significantly reduced, since we keep only the images from the graph.

For the weighting of the edges that link similar images, or images containing the same object, we consider three similarity measures: the number of verified matches, the number of verified matches relative to the number of interest point descriptors from both matched images and the cosine similarity measure of the corresponding BoVW feature vectors.

2.3 Representative images

In Figure 1 we can notice that images containing similar instances of the landmark of interest are strongly inter-connected. In addition, the clusters of inter-connected regions can be easily identified as connected components in the landmark graph. Here, we can extract 4 connected components, each consisting of images with similar instances of the Saint Mary of Flower (*e.g.*, main entrance, street level side view, roof top view). In general, the less representative instances are either completely rejected in the matching sequence or compose small clusters with poor interconnectivity.

Let us notice that the value of M_{min} that decides whether two images are similar or not, has direct influence on the number of clusters identified in the graph. For lower values, more images will be linked thus reducing the number of clusters and the diversity, while higher values of M_{min} will lead to a reduced number of links and to multiple less meaningful clusters. Since the matching is performed via the visual words, the number of verified matches is lower, as repetitive descriptors are counted only once. In our experiments, we have noticed that a good trade-off between relevance and variety is obtained for $M_{min} = 5$. M_{min} can be varied for specific applications or users who could adjust this parameter themselves according to their preferences.

From each such cluster we then select a dominant/representative image as the one yielding the highest similarity scores cumulated over its matches. Some clusters, might cover a higher number of images with respect to the number of images initially downloaded from Flickr. For such clusters, containing at least τ_{perc} of the images, we take the least representative image (*i.e.*, the image having the lowest cumulated similarity score) to be added to the list of representative images. In our runs $\tau_{perc} = 15\%$.

Table 1. Official results for visual descriptors only run. The crowd sourced evaluation is carried only on a subset of 50 locations from the test set, while the expert evaluation is conducted over all 396 locations.

	Ground Truth	Avg. P@10	Avg. CR@10	Avg. F1@10
Visual only	expertGT dev set	0.542	0.3442	0.41135
	expertGT test set	0.5383	0.2921	0.3653
	crowdGT1 test set	0.6449	0.8098	0.6897
	crowdGT2 test set	0.6449	0.7647	0.6665
	crowdGT2 test set	0.6449	0.6784	0.6282

3. EXPERIMENTAL RESULTS

We have submitted for evaluation a single run, employing only the visual descriptors mentioned above. The results of our run are shown in Table 1. We can notice that our method performs better on the crowd sourced ground truth, while the performance on the expert ground truth is rather modest.

Our method fails on images that depict a unique perspective of the landmark and for which no other similar image has been found. Some other representative instances are lost in large clusters due to the transitivity effect between multiple images. For example, in Figure 1, the tower and close-ups of the main entrance are integrated in the same cluster due to images depicting the entrance from distance that are matched with them. Thus, this shared element includes them in the same cluster. This leads to a reduced number of clusters and a lower cluster recall.

4. CONCLUSION

Our results show that a purely visual approach to such a complex problem can lead to average results and can also integrate new downloaded images quickly. Such a method is effective when no other metadata is available. However, improved results could be obtained by leveraging on other types of information (number of views, tags, author), if available, for further refining of the results.

5. REFERENCES

[1] Ionescu, B., Menéndez, M., Müller, H., and Popescu, A. 2013. Retrieving Diverse Social Images at MediaEval 2013: Objectives, Dataset and Evaluation. In *Proc. MediaEval 2013 Workshop*, Barcelona, Spain, October 18-19.

[2] Sivic, J., and Zisserman, A. 2003. Video Google: A text retrieval approach to object matching in videos. In *Proc. IEEE International Conference on Computer Vision*.

[3] Perdoch, M., Chum, O., and Matas, J. 2009. Efficient Representation of Local Geometry for Large Scale Object Retrieval. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*.

[4] Arandjelovic, R., Zisserman, A. 2012. Three things everyone should know to improve object retrieval. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*.

[5] Philbin, J., Chum, O., Isard, M., Sivic, J., and Zisserman, A. 2007. Object retrieval with large vocabularies and fast spatial matching. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*