SocialSensor: Finding Diverse Images at MediaEval 2013

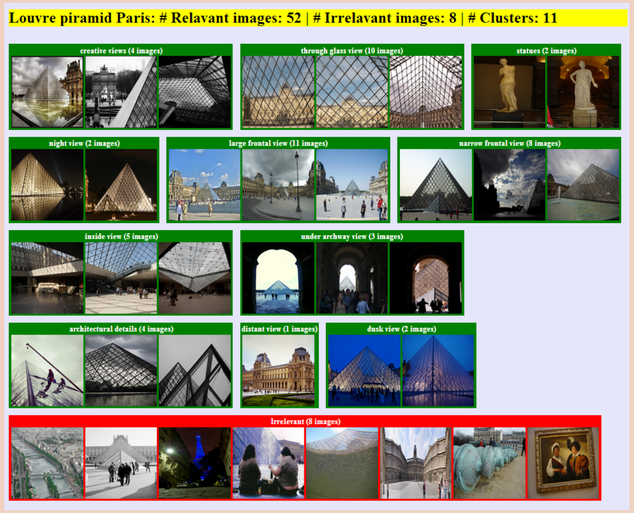
David Corney, Carlos Martin, Ayse Göker
IDEAS Research Institute
Robert Gordon University, Aberdeen
[d.p.a.corney|c.j.martin-dancausa|a.s.goker]@rgu.ac.uk

Eleftherios Spyromitros-Xioufis, Symeon Papadopoulos, Yiannis Kompatsiaris
Information Technologies Institute,
CERTH, Thessaloniki, Greece
[espyromi|papadop|ikom]@iti.gr

Luca Aiello, Bart Thomee
Yahoo! Research Barcelona,
08018 Barcelona, Spain
[alucca|bthomee]@yahoo-inc.com

Summary

- Participation in all 5 runs
- A different algorithm for each feature type
- A common criterion for model selection: best CR@10 calculated using leave-one(-location)-out cross-validation on the devset locations
- A simple visualization tool for getting more familiar with the problem at hand!



Run 1: Visual-only features

A single feature: **VLAD+SURF** vectors [1] with **multiple vocabulary aggregation** (k=4x128) and joint dimensionality reduction (to 1024d) with **PCA and whitening** [2]. Implementation publicly available at: `https://github.com/socialsensor/multimedia-indexing`

Relevance & Diversity (RD) method

A greedy optimization algorithm that selects a fixed-size subset S of the set of images $I = \{im_1, ..., im_N\}$ that is (approximately) optimal with respect to the following criterion [3] that accounts for both **relevance to the query** location and **diversity within** S.

$$U(S|l) = \sum_{im_{si} \in S} RD_{si} = \sum_{im_{si} \in S} w * R(im_{si}|l) + (1 - w) * D(im_{si}|S)$$

Relevance: The definition of [3] ($R(im_{si}|l) = 1 - d(im_{si}, im_q)$) would not work, especially when using only visual information.

Wikipedia image of Louvre Pyramid in Paris (left) and a **relevant** image of a statue inside Louvre (right).

Wikipedia image of Basilica of St. Mary of Health in Venice (left) and an **irrelevant** image with a human in focus (right).

Instead we use as $R(im_{si}|l)$ the output of a supervised classifier trained on the devset images. It tries to capture the notion of relevance as defined in this task, e.g.: **out-of-focus or human-in-focus = irrelevant / drawings = relevant**

Diversity in [3] is defined as: $D(im_{si}|S, l) = \frac{1}{|S|} \sum_{im_{sj} \in S, j \neq i} d(im_{si}, im_{sj})$. This definition is not ideal because a single image im_{sj} in S that has a high similarity with im_{si} suffices to reduce the diversity of the set.
Thus, we define diversity as: $D(im_{si}|S, l) = \min_{j, j \neq i} d(im_{si}, im_{sj})$, i.e the dissimilarity of im_{si} to the most similar image in S.

Optimization Algorithm: First adds the image with the highest relevance score in S and then sequentially adds the image which has the highest RD score among the remaining images.

Experiments:

- With different classifiers using cross-validation and AUC for model selection; best results obtained with linear SVMs
- We used the w that gave the best results for CR@10 on the devset ($\simeq$ 0.56), for producing the test set predictions

Run 2: Text-only features

- **Image relevance**: We built a forest of 100 random decision trees [4] using most of the textual descriptors available in the datasets. We used both direct image features, such as number of comments and views, and also derived features from the description, tag and title image fields separately, such as the number of words in the field and the normalised sum of tf-idf, social tf-idf and probabilistic values of each word. All continuous variables were discretized.
- **Diverse images**: We used hierarchical clustering to find 15 clusters for each location.

Within each cluster, images are ranked by the predicted relevance using the random forest. We then stepped through the clusters iteratively selecting the most relevant remaining image until (up to) 50 had been selected.

Run 3: Visual-text fusion

A simple late fusion scheme: The union of the images returned for each location by Run 1 & 2, ordered in ascending average rank.

Run 4: Human-machine hybrid approach

Task

To improve computer-generated short-lists of 15 images by filtering out 5 images as being either poor-quality or near-duplicates with any of the remaining images, leaving 10 images per location. Short-lists were generated using the text-only method.

Human participants

Not expected to be familiar with any of the locations, nor allowed to consult other sources. Two participants carried out the annotation on a total of 46 locations, around 12% of the total test set.

Run 5: Device and local weather data

The following data sources are combined to get pictures that are diverse in terms of distance from the landmark, angle of the shot, weather conditions and time of the day:

1. date and time the photo was taken, generally reliable at the granularity of one day
2. *f-stop* (aperture size of the shutter) and the exposure time (shutter speed), that can be combined as EV=f-stop^2·exposure, used previously to differentiate indoor from outdoor pictures [5]
3. geo-location of the device when the photo was taken, from which we compute the angle and distance to the photographed landmark
4. We also query a public database of historical weather data (`www.ncdc.noaa.gov`) to get the weather of the day the picture was taken, which indicates the main weather conditions (e.g. sun, fog, rain, snow, haze, thunderstorm, tornado)

We input the features to the k-means algorithm (k=10). Inside each cluster, when multiple candidates photos are available, we select the photo with the highest number of Flickr favourites. We verified that including the number of favourites as an additional feature to the k-means is beneficial for the selection of diverse images.

Results

	Combined test set			Keyword test set			GPS test set		
Method	*P@10*	*CR@10*	*F1@10*	*P@10*	*CR@10*	*F1@10*	*P@10*	*CR@10*	*F1@10*
Run 1	0.733	**0.429**	**0.521**	0.621	**0.415**	0.477	0.803	**0.438**	**0.549**
Run 2	0.732	0.390	0.491	0.639	0.393	0.467	0.791	0.388	0.506
Run 3	**0.785**	0.405	0.510	0.681	0.406	0.485	**0.850**	0.405	0.526
Run 4	0.750	0.408	0.508	**0.683**	0.414	**0.487**	0.792	0.405	0.522
Run 5	0.733	0.406	0.504	0.649	**0.415**	0.485	0.787	0.401	0.516

- Best performance in terms of CR@10 and F1@10 for our visual run (run 1)
- Human-machine hybrid (run 4) run improves the textual run (run 2)

References

[1] E. Spyromitros-Xioufis, S. Papadopoulos, I. Kompatsiaris, G. Tsoumakas, and I. Vlahavas, "An empirical study on the combination of SURF features with VLAD vectors for image search," in *WIAMIS*, 2012.

[2] H. Jégou and O. Chum, "Negative evidences and co-occurences in image retrieval: The benefit of PCA and whitening," in *ECCV*, 2012.

[3] T. Deselaers, T. Gass, P. Dreuw, and H. Ney, "Jointly optimising relevance and diversity in image retrieval," in *ACM CIVR '09*, (New York, USA), ACM, 2009.

[4] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[5] B. N. Lee, W.-Y. Chen, and E. Y. Chang, "A scalable service for photo annotation, sharing, and search," in *ACM MULTIMEDIA '06*, (Santa Barbara, CA, USA), pp. 699–702, ACM, 2006.

Acknowledgements

This work is supported by the SocialSensor FP7 project, partially funded by the EC under contract number 287975.