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

ABSTRACT

We describe the participation of the SocialSensor team in the Retrieving Diverse Social Images Task of MediaEval 2013. We submitted entries for all five runs after developing independent algorithms for visual features, text features and internet features (including local weather data). Our best CR@10 results came in the visual-only run, while the vision-text fusion run produced a slightly higher precision.

1. INTRODUCTION

The goal here is to produce a ranked list of images that are both relevant and diverse in response to a location-based query [3]. Throughout our work, we aimed to maximise the CR@10 score based on leave-one(-location)-out cross-validation results from the 50 devset locations. Below, we describe our methods for the five runs in turn before briefly summarising and discussing the results.

2. APPROACHES

2.1 Run 1: Visual-only features

For the visual-only run, each image is represented using optimized VLAD+SURF vectors. Compared to standard VLAD+SURF vectors [6], these vectors include multiple vocabulary aggregation (four visual vocabularies with $k = 128$ centroids each) and joint dimensionality reduction (to 1024 dimensions) with PCA and whitening [4].

Relevance & Diversity Method: Given a set of images $I = \{im_1, ..., im_N\}$, we developed an algorithm that selects a fixed-size set $S \subset I$ that is (approximately) optimal with respect to both relevance to the query location and diversity within S. We define the utility U of a set of images S with respect to a query location l as: $U(S|l) = \sum_{im_{si} \in S} w * R(im_{si}|l) + (1-w) * D(im_{si}|S)$ where $R(im|l)$ is the relevance score for im given the location and $D(im|S)$ is the diversity score within S. The same joint criterion, which we call Relevance & Diversity (RD), was used in [2]. However, we use different definitions for $R(im|l)$ and $D(im|S)$ that are more suitable for this task. While *relevance* in [2] is defined using a similarity measure between each image and a given query image, we use the ground truth data to train a classifier whose prediction for an image is used as the relevance score. We use all relevant images as positive and all irrelevant images as negative examples. *Diversity* in [2] is defined as: $D(im_{si}|S, l) = \frac{1}{|S|} \sum_{im_{sj} \in S, j \neq i} d(im_{si}, im_{sj})$ where $d(im_{si}, im_{sj})$ is a dissimilarity measure between im_{si} and im_{sj}. We found that this definition is not ideal because a single image im_{sj} in S that has a high similarity with im_{si} reduces the diversity of the set. Instead, we define diversity as: $D(im_{si}|S, l) = \min_{j, j \neq i} d(im_{si}, im_{sj})$ which defines it as the dissimilarity of im_{si} to the most similar image in S. As a dissimilarity measure we use the Euclidean distance between the VLAD vectors representing each image.

Optimization & Experiments: To find a set S that approximately optimizes U, we use the greedy optimization algorithm of [2]. This algorithm first adds to S the image with the highest relevance score and then sequentially adds the remaining image which has the highest RD score. We experimented with several types of relevance classifiers used in the RD method. Area Under ROC (AUC) was used for model selection by applying cross-validation. We applied the greedy optimization algorithm with the best performing classifier for several values of the weight w and chose the parameters that gave the best results for CR@10 ($\simeq 0.56$) on the devset, for producing the test set predictions.

2.2 Run 2: Text-only features

To predict the relevance of an image, we built a forest of 100 random decision trees [1] using most of the textual descriptors available in the datasets. The textual descriptors used for classification were: number of comments and views; Flickr ranking; author name. We also derived features from the description, tags and title fields separately: the number of words in the field; the normalised sum of tf-idf, social tf-idf and probabilistic values of each word (as provided by the organisers); the normalised sum of tf-idf values of each keyword where each value is the tf-idf value of each word from the Wikipedia page of the corresponding location, and using the remaining locations as the full corpus; and the average of the previous four values. We also discretized the continuous variables; the Flickr ranking and author were already discrete.

Independently, we used hierarchical clustering to find 15

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

clusters for each location. Within each cluster, we then ranked the images 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.

We found some cases where groups of images have identical text features but had different ground truth labels. These include casual holiday pictures where the Flickr user provided the same tags, descriptions etc. for a whole set of images, despite their diversity. Any deterministic text-only approach will fail to label these images correctly.

2.3 Run 3: Visual-text fusion

In order to leverage both visual and textual information we developed a simple late fusion scheme that combines the outputs of the visual and textual approaches described in the previous subsections. This is done by taking the union of the images returned for each location by the two approaches and ordering them in ascending average rank, i.e. the average of the ranks that they receive by each approach. Preliminary experiments indicated that early fusion (i.e. taking the individual features derived from each aspect of the data and combining them before making any decisions about relevance or diversity) was less effective.

2.4 Run 4: Human-machine hybrid approach

We developed a very simple approach to combine human and computer responses in an attempt to make use of people's natural visual processing abilities and their abilities to make rapid judgements from incomplete data. The test set comprised a total of 38,300 images from 346 locations. To obtain any form of human response requires either a large number of people (e.g. through crowd-sourcing) or a substantial reduction in the number of images. We chose the latter and presented the participants with computer-generated short-lists of images and asked them to improve it. Specifically, we used the text-only methods (Section 2.2) to list the top 15 relevant and diverse images. The human participant then had to select five of these 15 as being either poor-quality images or images that (nearly) duplicate any of the remaining set. Participants were not expected to be familiar with any of the locations, nor did they consult other sources. The final submission for each location consisted of the 10 remaining images, followed by the 5 "rejected" images. Two participants carried out the annotation on a total of 46 locations, around 12% of the total test set.

2.5 Run 5: Device and local weather data

Multimedia objects captured with modern cameras and smartphones are labeled with *Exif* metadata generated directly from the mobile device at the time the photo or video is taken. For this task, among all the data available we consider i) date and time the photo was taken, generally reliable at the granularity of one day; ii) *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]; iii) geo-location of the device when the photo was taken, from which we compute the angle and distance to the photographed landmark. 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).

	Expert			**Crowd-sourced**		
Method	*P@10*	*CR@10*	*F1@10*	*P@10*	*CR@10*	*F1@10*
Run 1	0.733	**0.429**	**0.521**	0.729	**0.764**	0.723
Run 2	0.732	0.390	0.491	0.702	0.760	0.691
Run 3	**0.785**	0.405	0.510	**0.800**	0.763	**0.753**
Run 4	0.750	0.408	0.508	0.725	0.738	0.698
Run 5	0.733	0.406	0.504	0.702	0.696	0.672

Table 1: Results for test set for top 10 results, using expert and crowd-sourced ground truth sets.

We combine all these data sources to get pictures that are diverse in terms of distance from the landmark, angle of the shot, weather conditions and time of the day. We input the feature 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.

3. RESULTS AND DISCUSSION

Table 1 summarises the results when returning the top 10 images per location compared to the expert and crowd-sourced ground truth. Our strongest results came from the visual features (run 1); a slight improvement in precision came when these were combined with text features (run 3). Our results are close for all five runs, despite the variety of features and algorithms used. This could indicate that the inherent signal/noise ratio of the data is a limiting factor, although further algorithmic development and optimisation could also improve matters. Future work includes the use of concept detection algorithms to improve diversity by explicitly including images matching different concepts (e.g. exterior; detail; night-time etc.).

4. ACKNOWLEDGEMENTS

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

5. REFERENCES

[1] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

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

[3] B. Ionescu, M. Menéndez, H. Müller, and A. Popescu. Retrieving diverse social images at MediaEval 2013: Objectives, dataset and evaluation. In *MediaEval 2013 Workshop*, Barcelona, Spain, October 18-19 2013.

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

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

[6] 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.