Ghent University-iMinds at MediaEval 2013 Diverse Images: Relevance-Based Hierarchical Clustering

Baptist Vandersmissen¹ baptist.vandersmissen@ugent.be

Abhineshwar Tomar¹ abhineshwar.tomar@ugent.be Fréderic Godin¹ frederic.godin@ugent.be

Wesley De Neve^{1,2} wesley.deneve@ugent.be Rik Van de Walle¹ rik.vandewalle@ugent.be

¹ Multimedia Lab, ELIS, Ghent University – iMinds, Ghent, Belgium ² Image and Video Systems Lab, KAIST, Daejeon, South Korea

ABSTRACT

In this paper, we attempt to tackle the MediaEval 2013 Retrieving Diverse Social Images challenge, which is a filter and refinement problem on a Flickr-based ranked set of social images. We developed three different approaches, using visual data, textual data and a combination thereof, respectively. Hierarchical clustering on highly relevant images, combined with a greedy approach to complement the ranking, forms the basis of our approach.

1. INTRODUCTION

In this paper, we describe our approach for tackling the MediaEval 2013 Retrieving Diverse Social Images Task [2]. This task focuses on result diversification in the context of social image retrieval. We refer to [2] for a complete task overview.

We suggest a cluster-based approach for the visual run and a semantic similarity-based approach for the textual run. The third run focuses on hierarchical clustering of relevant images and represents a combination of the purely visual and textual techniques.

2. VISUAL RUN

We propose a hierarchical clustering-based approach for the ranking of images in accordance with their relevance and diversity for a specific location. This method builds on the approach provided in [1]. It introduces an inter-cluster ranking machanism and differs on use of feature vectors, distance measure and Synthetic Representative Image (SRI) calculation method. We want to refine a set of m images retrieved from Flickr to a ranking of size n.

• The set of m images is hierarchically clustered to produce k clusters. Similarity between two images x_i and x_j (represented by a CN3x3 and LBP3x3 feature vector) is measured using a Gaussian kernel:

$$s(x_i, x_j) = exp\left(-\frac{||x_i - x_j||_2^2}{\sigma^2}\right).$$
 (1)

• For each of the k clusters produced, we calculate a Synthetic Representative Image (SRI). The SRI acts as a representative image for all the images within a particular

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

cluster. The SRI for a set of images is calculated by taking the mean of all the corresponding feature vectors of these images. Intra-cluster ranking of all of the m images is produced by calculating their Euclidean distance to the SRI. The image with rank 1 (topmost rank) is closest to the SRI of the cluster.

• Subsequently, we rank the k different clusters, again by calculating the distance of the cluster SRIs to a general SRI value, which is the mean over all cluster SRIs. Finally, n images are selected by iterating over the ranked clusters and taking the topmost ranked image within each cluster.

3. TEXTUAL RUN

The textual run makes use of information derived from tags and other textual metadata. This approach aims at diversifying the results by reranking the images retrieved from Flickr using textual relevance and semantic similarity. Our solution is based on [3] and makes use of an adapted performance metric to improve the ranking characteristics. Images for a query can then be ordered by directly optimizing the performance metric. This metric is named Average Diverse Precision (ADP) and is derived from the conventional Average Precision metric by adding a diversity component. We refer to [3] for a comprehensive overview.

We implemented a greedy approach that optimizes an estimation of the ADP measurement. Let $\boldsymbol{\tau}$ denote an ordering of the images, and let $\tau(i)$ be the image at the position of rank *i* (a lower number indicates an image with a higher rank). With the top i - 1 documents established, we can derive that the *i*th image should be decided as follows:

$$\tau(i) = \arg \max_{x \in \mathscr{D} - \mathscr{S}} \left\{ \frac{Rel(x)}{i} Div(x)(C + Div(x)) \right\}, \quad (2)$$

where

$$\mathscr{S} = \{\tau(1), \tau(2), \dots, \tau(i-1)\},$$
(3)

$$C = \sum_{k=1}^{i-1} Rel(\tau(k))Div(\tau(k)), \qquad (4)$$

with Rel(x) and Div(x) denoting the estimated relevance and diversity of the image, respectively.

3.1 Relevance Estimation

We estimate the relevance of an image by making use of textual metadata (number of views, number of comments)

Copyright is held by the author/owner(s).

	Develop	Test Set						
	+ GPS		- GPS		+ GPS		- GPS	
	CR@10	P@10	CR@10	P@10	CR@10	P@10	CR@10	P@10
Visual	43.6(2.4)	79.2 (-6.8)	48.4(2.0)	71.6(2.8)	37.5	76.1	34.7	56.8
Textual	44.2(2.9)	81.6 (-4.4)	51.6(5.2)	67.2 (-1.6)	39.7	74.9	37.5	58.6
Combined	49.8 (8.5)	85.6 (-0.4)	$51.7 \ (5.3)$	74.8 (6.0)	41.3	80.5	42.8	66.7

Table 1: Results (in %). Numbers between brackets denote the relative difference to the Flickr ranking and CR denotes the cluster recall of the ranking.

and social tags. This data was provided together with three textual features: TF–IDF, Social TF–IDF and a probabilistic feature type. We suggest a linear combination of this normalized data.

$$Rel(x) = \alpha \times tags_x + \beta \times views_x + \gamma \times comments_x, \quad (5)$$

where $views_x$ and $comments_x$ denote normalized number of views and number of comments, respectively and

$$tags_x = \frac{\sum_{t \in \mathscr{T}_x} a \operatorname{prob}_{t,x} + b \operatorname{tfidf}_{t,x} + c \operatorname{stfidf}_{t,x}}{|\mathscr{T}_x|}, \quad (6)$$

with \mathscr{T}_x denoting the set of tags linked to image x. We found that the precision of image orderings, which are purely based on relevance information, can be maximized by setting parameters a and γ to zero and increasing the impact of the parameters linked to 'tag' and 'views' information.

3.2 Diversity Estimation

Diversity of an image to a ranking is defined as the minimal difference with the other images in that ranking:

$$Div(x) = \min_{i \in \{1, \dots, n\}} \{1 - s(x, \tau(i))\}$$
(7)

We use a semantic similarity measure based on Google distance to assess the difference between two images. The average of the summation of the similarities between the different tags of both images gives us a value that describes the semantic similarity between both images.

4. COMBINED RUN

To estimate the relevance of an image, we use the textualbased method (cf. Section 3.1). Similarity between images is based on visual image features and a Gaussian kernel method to measure the difference between two image vectors (cf. Equation 1). In order to provide both a relevant and diverse ordering, we make use of hierarchical clustering techniques. We, again, try to refine a set of m images retrieved from Flickr to a ranking of size n.

- First, *l* images are selected with the highest estimated relevance. *l* is an arbitrary number that depicts a subset of the final ranking. The larger *l* becomes, the more the focus will shift from relevance to diversity.
- Next, these *l* images are hierarchically clustered based on a distance matrix calculated with the above described visual similarity. Per cluster, the most relevant item is selected and added to the final ranking.
- Depending on the number of remaining places of the first *l* spots in the final ranking, images are greedily added

based on a **gain** score. This score is higher for images that maximize the diversity and relevance related to the current ranking.

• When the first *l* spots in the ranking are filled, the algorithm runs from the start until all *n* spots are taken.

Table 2: Results on subset of 50 locations from test set (in %) with crowd-sourcing ground truth.

	GT_1		GT_2		GT_3	
	CR@10	P@10	CR@10	P@10	CR@10	P@10
Visual	83.0	72.5	78.7	72.5	69.9	72.5
Textual	77.7	65.7	72.1	65.7	63.5	65.7
Combined	83.6	69.4	77.4	69.4	66.1	69.4

5. EXPERIMENTS

In Table 1, we can see the results of our algorithms (three runs), compared to the regular Flickr ranking and the results on the test set. We notice that the combined method creates the most precise and diverse rankings based on both precision and cluster recall measures (averages 42.1% cluster recall on test set). Analysis of the results on the development set indicate the importance of finding the right balance between relevance and diversity. We also observe the similarity of the visual run and the textual run in terms of effectiveness. Table 2 lists the results on 50 locations of the test set evaluated with crowd-sourcing ground truths.

6. CONCLUSIONS

We observe that a method clustering images focused on high relevance outperforms all others. This method uses textual data for the relevance estimation and visual features for the diversity assessment.

7. REFERENCES

- R. Anca-Livia, J. Stöttinger, B. Ionescu, M. Menéndez, and F. Giunchiglia. Representativeness and diversity in photos via crowd-sourced media analysis. 2012.
- B. Ionescu, M. Menéndez, H. Müller, and A. Popescu. Retrieving diverse social images at mediaeval 2013: Objectives, dataset and evaluation. In *MediaEval 2013* Workshop, CEUR-WS.org, Barcelona, Spain, 2013.
- [3] M. Wang, K. Yang, X.-S. Hua, and H.-J. Zhang. Towards a relevant and diverse search of social images. *Multimedia*, *IEEE Transactions on*, 2010.