A Multimodal Approach for Video Geocoding*

Lin Tzy Li, Jurandy Almeida, Daniel Carlos Guimarães Pedronette,
Otávio A. B. Penatti, and Ricardo da S. Torres
Institute of Computing, University of Campinas – UNICAMP
13083-852, Campinas, SP – Brazil
{lintzyli, jurandy.almeida, dcarlos, penatti, rtorres}@ic.unicamp.br

ABSTRACT

Developed in the context of placing task at MediaEval 2012, this work addresses the problem of automatically assigning geographical coordinates to videos. This year our group extended the implementation of our framework for multimodal geocoding for combining textual and visual descriptors. In this paper, we describe our approach and report the results for 2012 datasets.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

1. INTRODUCTION

Geographic information is often associated with digital objects (e.g., documents, images, and videos) that once are geocoded (i.e., associated with a latitude and longitude), they can be found by means of geographical queries or visualized on maps.

Current solutions for geocoding multimedia material are usually based on textual information [3]. Those solutions depend on people to tag textual descriptions (a laborious and time consuming task). That scenario opens new venues for investigating methods that also use image/video content in the geocoding process.

In this paper, we present data fusion/rank aggregation approaches by combining evidences found in both textual and visual content. This work is developed in the context of the Placing Task at MediaEval 2012 whose goal is to automatically assign geographical coordinates (lat, long) to a set of annotated videos. Details about data, task, and evaluation protocols can be found in [7].

2. THE PROPOSED FRAMEWORK

The proposed architecture for dealing with multimodal geocoding is composed of three modules [2]: (1) text-based geocoding; (2) content-based geocoding; and (3) data fusion/rank aggregation-based geocoding. The first module is in charge of geocoding based solely on the textual part of the digital object. The content-based geocoding module is responsible for dealing with geocoding based on visual content. Finally, the rank aggregation-based module combines the results generated by the previous modules and assigns a location to the target video.

In this paper, we focus on modules (2) and (3), exploring a method that combines textual (title, description, and keywords of video) and visual features.

2.1 Textual features

For textual features, we used title, description, and keywords to calculate textual similarities between two videos. The text similarity functions used were Okapi and Dice [4]. In one of conducted experiment, we used as training set, the tags of $\sim$3 million Flickr photos from the development set.

2.2 Visual Information Retrieval

To encode video visual properties, we have used two approaches. The *bag-of-scenes* (BoS) [6] is based on static video frames and the *histogram of motion patterns* (HMP) [1] encodes motion information. The HMP was our team's approach last year [2].

The *bag-of-scenes* (BoS) approach [6] is based on the idea that video frames are like pictures from places. Therefore, by having a dictionary of pictures from places of interest, we can assign the video frames to one (or more) of the pictures in the dictionary. The video feature vector, called bag-of-scenes, works like a place activation vector. To create such a representation, we can use the same well-established strategies in the bag-of-visual-words model. In this work, we tested it based on the CEDD descriptor with dictionary made of 5000 scenes (BoS^{5000}_{CEDD}) and 500 scenes (BoS^{500}_{CEDD}).

The *histogram of motion patterns* (HMP) [1], unlike the bag-of-scenes model, considers the video movement by the transitions between frames. For each frame of an input sequence, motion features are extracted from the video stream. After that, each feature is encoded as a unique pattern, representing its spatio-temporal configuration. Finally, those pattens are accumulated to form a normalized histogram.

2.3 Data Fusion/Rank aggregation

We use a rank aggregation method based on a multiplication approach, initially proposed for multimodal image retrieval [5]. Let v_q be a query video that is being compared to another video v_i from the dataset. Let $sim_0(v_q, v_i)$ be a function defined in the interval $[0, 1]$ that computes a similarity score between the videos v_q and v_i, where 1 denotes a perfect similarity. Let $\mathcal{S} = \{sim_1, sim_2, \ldots, sim_m\}$ be a set of m similarity functions defined by different features. The new aggregated score sim_a is computed by multiplying

*We thank FAPESP, CNPq, and CAPES for financial support and the notes and comments from the organizers.

Copyright is held by the author/owner(s).
MediaEval 2012 Workshop, October 4-5, 2012, Pisa, Italy

individual feature scores as follows:

$$sim_a(v_q, v_i) = \frac{\sqrt[m]{\prod_{k=1}^{m}(sim_k(v_q, v_i) + 1)}}{2} \quad (1)$$

By multiplying the different similarity scores, high scores obtained by one feature are propagated to the others, leading to high aggregated values.

2.4 Geocoding approach

Our method to predict an unseen query video is divided into three steps: text processing, visual processing, and data/information fusion. We use the videos of the development set (15,563 videos) as geo-profiles, in the sense that they are compared to each test video (4,182 videos).

The visual processing module describes the visual content of each provided video. All videos in the test set are compared to those in the development set and, for each test video, a list of videos – ranked by similarity in descending order – is produced. The textual processing module works similarly, except for the kind of information considered (textual content instead of visual).

The fusion module takes as input different lists produced and generates a brand new list using our rank aggregation. The resulting new list is the one to be used to estimate the lat/long of a query video.

In the current implementation, we consider that the video on the top of each list is the one that should transfer its known lat/long to the query video.

3. EXPERIMENTAL RESULTS

Our approach is focused on data fusion, so most of our submissions considered combined results, although the use of a single modality was also evaluated for comparison purposes.

run 1: combines textual descriptors Okapi and Dice, considering three implementations: Okapi applied to three textual metadata (title, description, keywords) associated with a video (Okapi_all); Okapi applied to the keywords field (Okapi_keywords), as well as Dice applied to the keywords (Dice_keywords).

run 2: combines two textual and two visual features: Okapi_all, Okapi_keywords, HMP, and BoS^{5000}_{CEDD}.

run 3: only considers HMP. We used this method last year, and it is used here as baseline.

run 4: combines results from 3 visual descriptors: HMP, BoS^{5000}_{CEDD}, and BoS^{500}_{CEDD}.

run 5: uses 3,185,258 Flickr keywords as geo-profile. The keywords of test video were compared to this geo-profile using the Okapi similarity function.

The evaluation results are shown in Table 1. Note that, by relying just on video similarity based on visual content, the HMP algorithm alone (run 3) reaches 81.73% only when accepting an error of 10,000 km, 24.77% when the threshold is 1,000 km, 16.62% for 100 km. The combination of HMP and the two BoS configurations (BoS^{500}_{CEDD} and BoS^{5000}_{CEDD}) is slightly better (81.73%, 25.47%, and 17.07% for 10,000 km, 1,000 km, and 100 km, respectively).

We noticed that the result of classical text vector space combined (run 1) is twice as good as the use of visual cues alone (run 4). On the other hand, the combination of different textual and visual descriptors (run 2) leads to statistically significant improvements (confidence ≥ 0.99) over results of the method that relies only on textual clues (run 1).

Table 1: Results for the test set

Precision	Results (%)				
(km)	run1	run2	run3	run4	run5
1	21.40	22.29	15.81	15.93	9.28
10	30.68	31.25	16.07	16.09	19.44
100	35.39	36.42	16.62	17.07	24.13
500	41.77	43.35	19.68	19.97	29.29
1000	45.38	47.68	24.77	25.47	33.91
5000	62.29	66.91	45.34	45.34	65.73
10000	85.27	87.95	81.95	81.73	87.69

Interestingly, run 3 – using only HMP (our last year approach) – performs much better with this year's data set. For example, for 1 km, the obtained score is 15.81% in 2012 and only 0.21% in 2011. One possible reason is the use of a larger development set (∼5,000 videos were added to the development set), which might have made the geo-profile richer this year.

Results of run 5, where photos metadata worked as geo-profile for estimating lat/long for test videos, are worse than using visual information only at 1 km precision. However, for other radii, run 5 is better than runs 3 and 4.

4. CONCLUSIONS

In our approach, we combined textual information found in video metadata (e.g., descriptions) and visual features. We used the video similarity between videos in the development set and those in the test set to estimate the location of the latter. Obtained results demonstrate that this approach is promising as it yields better results than those observed for a single modality. Future works include the investigation of other strategies for combining different modalities and considering other information sources, such as Geonames and Wikipedia, to filter out noisy data from ranked lists.

5. REFERENCES

[1] J. Almeida, N. J. Leite, and R. da S. Torres. Comparison of video sequences with histograms of motion patterns. In *ICIP*, pages 3673–3676, 2011.

[2] L. T. Li, J. Almeida, and R. da S. Torres. RECOD working notes for placing task MediaEval 2011. In *MediaEval 2011 Workshop*, volume 807, 2011.

[3] J. Luo, D. Joshi, J. Yu, and A. Gallagher. Geotagging in multimedia and computer vision–a survey. *MTAP*, 51:187–211, Jan. 2011.

[4] C. D. Manning, P. Raghavan, and H. Schtze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.

[5] D. C. G. Pedronette, R. da S. Torres, and R. T. Calumby. Using contextual spaces for image re-ranking and rank aggregation. *MTAP*, pages 1–28, 2012.

[6] O. A. B. Penatti, L. T. Li, J. Almeida, and R. da S. Torres. A Visual Approach for Video Geocoding using Bag-of-Scenes. In *ICMR*, 2012.

[7] A. Rae and P. Kelm. Working Notes for the Placing Task at MediaEval 2012. In *MediaEval 2012 Workshop*, 2012.