Multimedia Multimodal Geocoding

Lin Tzy Li[1,2], Daniel Carlos Guimarães Pedronette[1], Jurandy Almeida[1], Otávio A. B. Penatti[1], Rodrigo Tripodi Calumby[1,3], and Ricardo da S. Torres[1]
[1]RECOD Lab, Institute of Computing, University of Campinas (UNICAMP), Campinas, SP – Brazil, 13083-852
[2]Telecommunications Res. & Dev. Center, CPqD Foundation, Campinas, SP – Brazil, 13086-902
[3]Department of Exact Sciences, University of Feira de Santana (UEFS), Feira de Santana, BA – Brazil, 44036-900
{lintzyli, dcarlos, jurandy.almeida, penatti, tripodi, rtorres}@ic.unicamp.br

ABSTRACT

This work is developed in the context of the placing task of the MediaEval 2011 initiative. The objective is to geocode (or geotag) a set of videos, i.e., automatically assign geographical coordinates to them. This paper presents an architecture for multimodal geocoding that exploits both visual and textual descriptions associated with videos. This work also describes our efforts regarding the implementation of this architecture to demonstrate its applicability. Conducted experiments show how our multimodal approach enhances the results compared to relying on a single modality.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; H.2.8 [**Database Management**]: Database Applications—*Spatial databases and GIS*; I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding—*Video analysis*

General Terms

Algorithms, Experimentation

Keywords

multimodal geotagging, rank aggregation, video retrieval

1. INTRODUCTION

The great amount of geographical entities available on the Web has created a great interest in locating them on maps. Geographical information is often enclosed into digital objects (e.g., documents, images, and videos) and using it to support geographical queries is of great interest. Associating latitude and longitude with a digital object is often called *geocoding*, *geotagging*, or *georeferencing* [13].

Textual information is commonly used for multimedia geocoding [13], however, approaches based on textual information usually lack of objectivity and completeness, since understanding the visual content of a multimedia object may change according to the experience and perception of each subject. Other problems include lexical and geographical ambiguities in recognizing place names [8]. An interesting alternative to address those shortcomings is to use the image/video content in the geocoding process. The objective is to explore image/video visual properties (such as texture, color, and movement) as alternative cues for geotagging. Furthermore, having multiple sources of information for multimedia geocoding also opens the opportunity to combine them by using data fusion approaches.

In this paper, we present an architecture that allows us to explore the combination of visual-based and text-based approaches in order to improve video geocoding. The validation of this proposal is performed in the context of the Placing Task of the MediaEval 2011 initiative, whose aim is to automatically assign geographical coordinates (latitude and longitude) to a set of annotated videos.

2. RELATED WORK

Geocoding approaches based on visual clues are proposed in the context of landmark recognition [13], as well as for non-landmark images [9]. Usually, those approaches are modeled as image classification problem in content-based image retrieval (CBIR), assisted by a huge database of geotagged image as knowledge base [7].

Researches on video geocoding have been done for the Placing Task. There were three main approaches as summarized in [11]: (a) geoparsing and geocoding video metadata assisted by a gazetteer of geographic name; (b) propagation of the lat/long of a similar video of the development database to the test video; (c) division of the training set in geographical regions by clustering or exploiting fixed-size grids, using a model to assign items to each group (based on textual metadata and visual clues). In 2010 and 2011, the best result of the Placing Task at MediaEval [11] was from VanLaere et al. [19], who used only metadata of images and videos, combining approaches (b) and (c).

As mentioned before, this work exploits data fusion (visual and textual) to enhance video geocoding rather than focus on visual modality as we did in our previous work [12, 17]. Related work on multimodal on video geocoding [3, 10] explore hierarchical approach by first finding geographical candidate locations using textual data, and then applying visual features to match the query video to videos of the development set (their knowledge database) that are found inside the geographic locations defined in the first step.

In our work, we propose a late fusion approach, which merges the monomedia similarity information by means of

ACM SIGSPATIAL GIS '12, November 6-9, 2012. Redondo Beach, CA, USA

aggregation functions.

Rank aggregation methods combine scores/rankings generated by different features (from different modalities) to obtain a more accurate one. The features are homogeneously and seamlessly combined, representing an important advantage of our approach, since other features can be easily added to the fusion step. The best combinations occur when all systems being combined have good performance and the inputs are independent and non-correlated [4].

Score-based rank aggregation approaches has been used in the multimodal image retrieval context [16] and also used for associating photos to georeferenced textual documents [2]. In this paper, we aim at using a score-based rank aggregation approach for combining features of different modalities.

3. PROPOSED FRAMEWORK FOR MULTIMODAL GEOCODING

The proposed architecture for dealing with multimodal geocoding is composed by three modules (Figure 1): a) **text-based geocoding** is responsible for all text processing, that is, it uses GIR geocoding techniques to predict a location, given available metadata of a video; b) **content-based geocoding module** is in charge of predicting a location based on visual similarity of the test images/videos with regard to a knowledge database (development dataset); and c) **data fusion** combines the geocoding results generated by the previous modules and assigns a location to a target video. The idea is to rely on text and image/video data whenever possible. The final result is a combination of the results from each modality. A data fusion module is responsible for geocoding results of textual (metadata) and visual (frames) parts of a video.

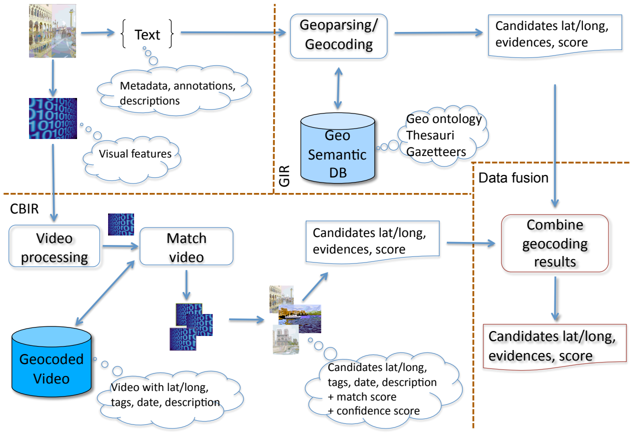
Text
Metadata, annotations, descriptions
Visual features
GIR
Geoparsing/ Geocoding
Candidates lat/long, evidences, score
Geo Semantic DB
Geo ontology
Thesauri
Gazetteers
Data fusion
CBIR
Video processing
Match video
Candidates lat/long, evidences, score
Combine geocoding results
Candidates lat/long, evidences, score
Geocoded Video
Video with lat/long, tags, date, description
Candidates lat/long, tags, date, description + match score + confidence score

Figure 1: Multimodal geocoding proposal.

In the following sections, we present how each module is implemented.

3.1 Text Retrieval & GIR

For processing textual information associated with videos, we propose 1) Geographic Information Retrieval (GIR) techniques to recognize and associate location with digital objects, based on their textual content and 2) Information Retrieval (IR) classical matching functions to retrieve similar digital objects.

In the context of the Placing Task, since there were metadata – such as title, description, and keywords – associated with videos, they are processed by the text module of our framework. The current implementation of this module relies on classical IR text matching methods: vector space model and text similarity functions [14], such as cosine, bag-of-words (normalized documents terms intersection), dice, okapi, and tf-idf sum. Along this document, they are named *ACos*, *ABow*, *ADice*, *AOkapi*, and *Atdfidf*, respectively.

3.2 Visual Information Retrieval

To encode video visual properties we have used two approaches. One is based on video frames and do not consider transitions between them, which is called *bag-of-scenes* (BoS) [17]. The other approach encodes motion information by using *histogram of motion patterns* (HMP) [1].

The *bag-of-scenes* (BoS) approach [17] lies in the idea that video frames are like pictures from places. Therefore, if we have a dictionary of pictures from places of interest and we assign the video frames to one (or more) of the pictures in the dictionary, we create a simple representation that gives good insights of the video location. The video feature vector, called bag-of-scenes, works like a place activation vector. To create such a representation, we can use the same strategies that are solid in the classic bag-of-visual-words model. In this work, we tested the BoS based on CEDD descriptor with 50 (BoS50CEDD or BoS^{50}_{CEDD}), 500 (BoS500CEDD or $\mathrm{BoS}^{500}_{CEDD}$), 5000 scenes (BoS5000CEDD or $\mathrm{BoS}^{5000}_{CEDD}$).

Our second approach for encoding video visual properties is based on a *histogram of motion patterns* (HMP) [1]. Different from the bag-of-scenes model, this approach considers the 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, which has proven to be a powerful tool for describing the video content [1].

3.3 Information Fusion

Given a query video whose location will be determined, it is compared with all those in dataset, considering the different features defined for each modality, which produces a different score. The goal of the data fusion module is to combine the scores produced by features of different modalities in order to produce a more effective score. We use a rank aggregation method based on a multiplication approach, initially proposed for multimodal image retrieval [16].

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 considered. The new aggregated score sim_a is computed by multiplying individual feature scores as follows:

$$sim_a(v_q, v_i) = \frac{\sqrt[m]{(sim_1(v_q, v_i) + 1) \times \cdots \times (sim_m(v_q, v_i) + 1)}}{2}. \tag{1}$$

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

4. EXPERIMENTS & RESULTS

The aim of conducted experiments is to evaluate the proposed framework for multimodal geocoding. Our evaluation concerns the video geocoding problem, in the context of the Placing Task at MediaEval 2011 [18].

MediaEval 2011 Data set & Evaluation criteria

The dataset is composed by a development set with 10,216 videos and a test set with 5,347 videos. Both sets are provided with their extracted keyframes and corresponding pre-extracted low-level visual features, and metadata [18]. Latitude and longitude are provided only for the development set that can be used for training.

Participants in the Placing Task 2011 were allowed to use image/video metadata, audio and visual features, as well as external resources. The result evaluation was based on the distance to the ground truth geographic coordinate point, in a series of widening circles of radius (in km): 1, 10, 20, 50, 100, 200, 500, 1000, 2000, 5000, and 10000. Thus, an estimated location is considered correct at a particular circle size, if it lies within a given circle radius. That can be seen as quality or precision measurement.

Experimental Setup

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 as geo-profiles, in the sense that they are compared to each test video.

We first evaluate our framework in the context of using only one modality (textual or visual). In this phase, different (textual and visual) descriptors are explored, and the descriptors with the best results are then used in the data fusion module.

The visual processing module describes the visual content of each provided video. All videos in the set are compared against each other and, for each 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).

In our approach, 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. In the development set, considering that a query video is always the best match to itself – thus it will be the first in this list – the second video is regarded as being the top video.

For the test result, the same procedure done in development set is applied, but now each video in test set is compared to those in the development set, and the most similar video of the development set transfers its lat/long to the query test video.

Results

Experiments were performed both for the development and test sets using each modality (text and visual) individually. Due to space constraints, we report the results using only the development set.

For textual data, we employ the similarity functions described in Section 3.1, considering the different metadata fields: title, description, and keywords. It is worth noting that, in this experiment, we used the best parameters of the bag-of-scenes (BoS) and the histogram of motion patterns (HMP) approaches presented in [1, 17].

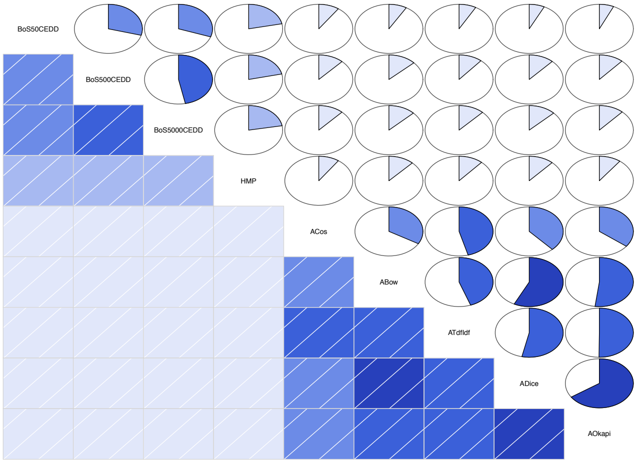

Figure 2: Correlation graph for pairs of methods evaluated (development set). The darker the colors, the higher the correlation. The diagonal holds the name of each methods for corresponding row & column.

The individual results show that text-based approaches outperforms visual-based approaches, being okapi the best method, followed by dice. Considering only the visual-based approaches, HMP is slightly better than BoS$^{5000}_{CEDD}$ (see Figure 3).

The correlation among different features and modalities is used as additional cues for the fusion module. Figure 2 shows the correlation graph *(corrgrams [6])* for the development set.

The correlograms indicate higher correlation among the different textual descriptors. The same behavior can be observed for the correlation scores among the different visual descriptors. However, the correlation between textual and visual descriptors is very low. As we stated before, the best combinations occur when the inputs are independent and non-correlated [4]. Therefore, the combination of textual and visual-based methods can be considered.

Based on individual modality results and the correlation analysis, the results of the top two best methods for each modality were chosen for combination.

Figure 3 shows the stacked histograms comparing the methods in each widening circle used in the Placing Task evaluation. Each stack (rectangle) in the histograms represents the amount of correctly geocoded video for a given radius in the set of widening circles (from bottom to top) used to measure the performances in the Placing Task.

For example, the first rectangle in the bottom of the stack refers to the 1km radius. Darker colors was used to code the amount of correctly geocoded videos for more strict precision (smaller radius), therefore, the taller the darker rectangles, the better or more precise the methods. In this figure, we report the results for the best methods of each modality (dice, okapi, BoS$^{5000}_{CEDD}$, HMP), the results for the combination of all the methods, and the combination of only the best ones. Observe that when combining the results of different modalities, the results are better. Moreover, when fusing all the methods presented previously (Fusion$_{all}$), we get an im-

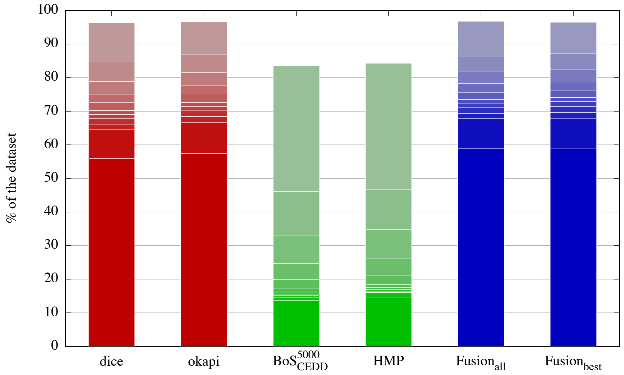

Figure 3: Stacked histograms showing the performances of the best methods in each modality and their fusion in the *development* set.

provement over the best text-based method (okapi), mainly for higher precision radius.

Another interesting result is that we get almost the same improvement when fusing only the best methods of each modality (Fusion$_{best}$). This indicates that our correlation analysis was effective to help select the best methods to be combined.

5. CONCLUSIONS

This paper has presented an architecture for geocoding videos by taking into account multiple modalities. In our architecture, textual and visual features are combined by a rank aggregation approach.

The proposed architecture has been validated in the context of the Placing Task of the MediaEval initiative. Conducted experiments have demonstrated that the use of the proposed multimodal approach improves results when compared with those based on a single modality (either textual or visual feature). The potential of this framework is that each module can be improved separately, leading to better results. For example, our text module can use more sophisticated text processing/analysis, or the fusion module can apply more elaborated approaches (e.g., [5, 15, 16]).

Future work includes the investigation of other strategies for combining different modalities, as well as other visual descriptors for representing videos can also be explored. Finally, we also plan to consider other information sources, such as Geonames and Wikipedia, to filter out noisy data from ranked lists.

Acknowledgements

The authors are grateful to CAPES, FAPESP, and CNPq, as well as CPqD Foundation.

6. 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] R. Candeias and B. Martins. Associating relevant photos to georeferenced textual documents through rank aggregation. In *Int. Semantic Web Conf. - Terra Cognita Workshop*, 2011.

[3] J. Choi, H. Lei, and G. Friedland. The 2011 ICSI video location estimation system. In *Working Notes Proc. MediaEval Workshop*, volume 807, 2011.

[4] W. B. Croft. Combining approaches to information retrieval. In *Adv. in Inf. Retrieval*, volume 7, pages 1–36. Springer US, 2002.

[5] F. A. Faria, A. Veloso, H. M. de Almeida, E. Valle, R. da S. Torres, M. A. Gonçalves, and W. M. Jr. Learning to rank for content-based image retrieval. In *ACM MIR*, pages 285–294, 2010.

[6] M. Friendly. Corrgrams: Exploratory displays for correlation matrices. *The American Statistician*, 56(4):316–324, 2002.

[7] J. Hays and A. A. Efros. im2gps: estimating geographic information from a single image. In *CVPR*, 2008.

[8] C. B. Jones and R. S. Purves. Geographical information retrieval. *Int. J. Geo. Info. Science*, 22(3):219–228, 2008.

[9] Y. Kalantidis, G. Tolias, Y. Avrithis, M. Phinikettos, E. Spyrou, P. Mylonas, and S. Kollias. Viral: Visual image retrieval and localization. *Mult. Tools and App.*, 51:555–592, 2011.

[10] P. Kelm, S. Schmiedeke, and T. Sikora. A hierarchical, multi-modal approach for placing videos on the map using millions of flickr photographs. In *Workshop on Social and behavioural networked media access*, pages 15–20, 2011.

[11] M. Larson, M. Soleymani, P. Serdyukov, S. Rudinac, C. Wartena, V. Murdock, G. Friedland, R. Ordelman, and G. J. F. Jones. Automatic tagging and geotagging in video collections and communities. In *ICMR*, pages 51:1–51:8, 2011.

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

[13] J. Luo, D. Joshi, J. Yu, and A. Gallagher. Geotagging in multimedia and computer vision–a survey. *Mult. Tools and App.*, 51:187–211, 2011.

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

[15] D. C. G. Pedronette and R. da S. Torres. Exploiting clustering approaches for image re-ranking. *J. Vis. Lang. and Comp.*, 22(6):453–466, 2011.

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

[17] 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.

[18] A. Rae, V. Murdock, P. Serdyukov, and P. Kelm. Working notes for the placing task at MediaEval 2011. In *Working Notes Proc. MediaEval Workshop*, volume 807, 2011.

[19] O. Van Laere, S. Schockaert, and B. Dhoedt. Finding locations of flickr resources using language models and similarity search. In *International Conference on Multimedia Retrieval*, pages 48:1–48:8, 2011.