RECOD Working Notes for Placing Task MediaEval 2011*

Lin Tzy Li, Jurandy Almeida, and Ricardo da S. Torres
Institute of Computing, University of Campinas – UNICAMP
13083-852, Campinas, SP – Brazil
{lintzyli, jurandy.almeida, rtorres}@ic.unicamp.br

ABSTRACT

This work is developed in the context of placing task at MediaEval 2011. It consists in automatically assigning geographical coordinates to a set of videos. Our group proposed an architecture design for the multimodal geocoding. In this paper, we focused on implementing a simple content-based approach, which is part of the proposed framework. The reported results show our strategy compared to those from previous year participant using only visual content to accomplish this task.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]

1. INTRODUCTION

The geographic information is present in people's daily life, thus it is not surprising that there is a huge amount of data on the Web about geographical entities and a great interest in localizing them on maps. That information is often enclosed in digital objects (e.g., documents, image, and videos). Once they are geocoded (i.e., associated to a latitude or longitude), one can perform geographical queries.

Current solutions for geocoding multimedia material are usually based on textual information [2, 6]. Such a strategy depends on the human intervention to tag textual descriptions of the data. However, there is a lack of objectivity and completeness of those descriptions, since the understanding of the visual content of multimedia data may change according to the experience, and perception of each subject, not to mention lexical and geographical problems in recognizing place names [5]. This opens new venues for the investigation of methods that use image/video content in the geocoding process. Furthermore, data fusion/rank aggregation approaches could be also used for combining evidences found in both textual and visual content.

In this paper, we present an approach for visual content-based geocoding, although we aim to explore the combination of textual and visual content of digital objects in order to improve their geocoding. The idea here is to test how well video similarity in term of its motion sequence would fit our purposes of predicting their location.

This work is developed in the context of Placing Task at MediaEval 2011. The goal of such a task is to automatically assign geographical coordinates (latitude and longitude) to a set of annotated videos. More details regarding data, task, and evaluation are described in [7].

*We thank FAPESP, CNPq, and CAPES for financial support.

MediaEval 2011 Workshop, September 1-2, 2011, Pisa, Italy

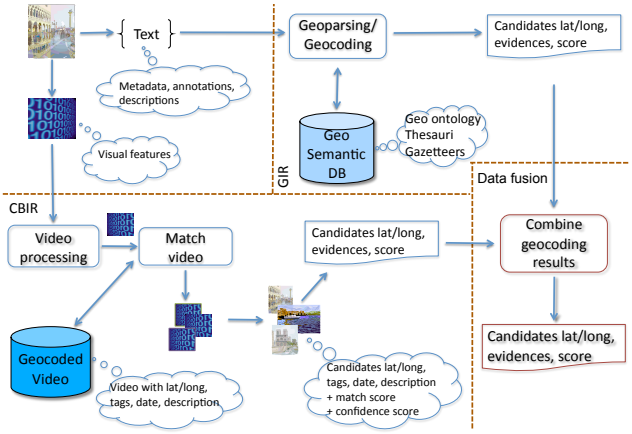

Figure 1: Multimodal geocoding proposal

2. THE PROPOSED FRAMEWORK

The proposed architecture for dealing with multimodal geocoding is composed by three modules (Figure 1): (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 textual part of the digital object. Content-based geocoding module is responsible for dealing with and geocoding based on its visual content. Finally, the rank aggregation-based module combines the results generated by the previous modules and gives the final result of the geocoding. The idea is to rely on text and image whenever possible.

In this paper, we focused on the second module, exploring a method to identify similar videos whose visual content indicates where those videos were filmed. Although it is allowed to use all the metadata associated to the given video, such as descriptions and tags provided by users, we focused on geocoding based on visual features of the videos.

2.1 Extracting & Comparing Visual Features

Instead of using any keyframe visual features provided by the organizers, we adopted a simple and fast algorithm to compare video sequences described in [1]. It consists of three main steps: (1) partial decoding; (2) feature extraction; and (3) signature generation.

For each frame of an input video, motion features are extracted from the video stream. For that, 2×2 ordinal matrices are obtained by ranking the intensity values of the four luminance (Y) blocks of each macroblock. This strategy is employed for computing both the spatial feature of the 4-blocks of a macroblock and the temporal feature of corresponding blocks in three frames (previous, current, and next). Each possible combination of the ordinal measures

Table 1: Results using only videos visual content (distance between ground truth and estimated)

Radius (km)	1	10	20	50	100	200	500	1000	2000	5000	10000
Dev set: % in range	14.42	16.02	16.44	16.93	17.51	18.36	21.20	26.04	34.76	46.76	84.29
Test set: % in range	0.21	1.12	1.59	1.93	2.71	3.33	6.08	12.16	22.11	37.78	79.45

is treated as an individual pattern of 16-bits (i.e., 2-bits for each element of the ordinal matrices). Finally, the spatio-temporal pattern of all the macroblocks of the video sequence are accumulated to form a normalized histogram. For a detailed discussion of this procedure, refer to [1].

The comparison of histograms can be performed by any vectorial distance function like Manhattan (L_1) or Euclidean (L_2) distances. In this work, we compare video sequences by using the histogram intersection, which is defined as

$$d(\mathcal{H}_{V_1}, \mathcal{H}_{V_2}) = \frac{\sum_i \min(\mathcal{H}^i_{V_1}, \mathcal{H}^i_{V_2})}{\sum_i \mathcal{H}^i_{V_1}},$$

where $\mathcal{H}_{V_1}$ and $\mathcal{H}_{V_2}$ are the histograms extracted from the videos V_1 and V_2, respectively. This function returns a real value ranging from 0 for situations in which those histograms are not similar at all, to 1 when they are identical.

2.2 Geocoding the Visual Content

We used 10,216 videos from the development set released by Placing Task organizer as geo-profiles against which each test video was compared to.

In order to assess how well we did, only relying on visual content during the development phase, we extracted the visual content of each provided video, then we compared all videos of the set against each other, and finally, for each video, we produced a list of videos ordered by similarity in descending order. Considering that a query video always is the best match to itself, thus it will be the first in this list, we took the second video from the top list as the one that will transfer its known lat/long to the query video.

For the test result, applying the visual feature extraction and the similarity computation explained previously, each video in test set (5,347) was compared with those in the development set. Then, for each test video, an ordered list of similar videos from the development set was produced along with its similarity score to that given test video. Finally, we picked the most similar video of this list as the one that will transfer its known lat/long to the query test video, and reported that lat/long as the one to be given to test video.

3. EXPERIMENTAL RESULTS

For this task, we performed one submission for the run that considered just visual content. The evaluation results are shown in Table 1. Note that, by relying just on video similarity based on its visual content, our algorithm will hit 79.45% only when accepting an error of 10,000 km between the ground truth and the assigned point. However, when considering 100 km of error, it predicts lat/long correctly for only 2.71%. These results underperform those from the reference algorithm for this task (winner of the last year), which just analyzes user-contributed tags for predicting the geotag of a video (73.6% of videos are within 100 km) [4].

However, we are interested in comparing to other results using only video content to accomplish the placing task. For instance, Kelm et al. [3], who also reported their results when only visual content of test videos were used to predict their location on Earth, have used visual features of the development set for training a multi-class SVM classifier with RBF kernel. Their best results were achieved by a hierarchical clustering with a diameter threshold of 100 km, which determined 317 classes for the SVM with the descriptors CED, FCTH, and Gabor. They presented their results for video's location correctly predicted within radius of 50 km, 100 km, 200 km, 750 km, and 2,500 km.

In order to compare our results to those presented by Kelm et al. [3], we aggregated the evaluation results presented in Table 1 according to their experimental protocol. This regrouping was possible due to the placing task organizers, who made available to all participants of that task: their tool to calculate the distance (Haversine distance formula) from ground truth to the estimated location for each result; and the test videos ground truth.

Table 2 compares our approach with the results reported by Kelm et al. (adopted from their Table 6) [3]. Notice that our method, although simpler, shows high precision relative to their clustering-and-classification method. The key advantage of our technique is its computational efficiency. Unlike them, we did not use any data to train any classifier.

Table 2: Our regrouped test results vs. Kelm et al.

Radius (km)	50	100	200	750	2500
Our approach %	1.93	2.71	3.33	9.18	24.48
Kelm et al. %	3.38	5.26	6.23	10.65	19.92

4. CONCLUSIONS

Relying just on video content to estimate its location still poses a challenge. It seems that this task requires using textual information found in video metadata such as descriptions, user tags, external knowledges bases as shown by some related works.

Our method used the video similarity between videos in development set and those in test set to estimate location of those. The similarity in this work is given by motion patterns extracted from the video streams. This algorithm is simple and achieved comparable results to those more complex presented by previous work that also was based just on video visual clues.

We believe that we can improve the results by developing new video similarities approaches as well as new combining methods for image and textual evidences in the context of geocoding digital objects.

5. REFERENCES

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

[2] C. B. Jones and R. S. Purves. Geographical information retrieval. *Int. J. Geogr. Inf. Sci.*, 22(3):219–228, 2008.

[3] P. Kelm, S. Schmiedeke, and T. Sikora. Multi-modal, Multi-resource Methods for Placing Flickr Videos on the Map. In *ICMR*, pages 52:1–52:8, 2011.

[4] M. Larson, M. Soleymani, P. Serdyukov, S. Rudinac, et al. Automatic tagging and geotagging in video collections and communities. In *ICMR*, pages 51:1–51:8, 2011.

[5] R. R. Larson. Geographic information retrieval and digital libraries. In *ECDL*, volume 5714/2009, pages 461–464, 2009.

[6] J. Luo, D. Joshi, J. Yu, and A. Gallagher. Geotagging in multimedia and computer vision–a survey. *Multimedia Tools Appl.*, 51(1):187–211, 2011.

[7] A. Rae, V. Murdock, P. Serdyukov, and P. Kelm. Working Notes for the Placing Task at MediaEval 2011. In *MediaEval*, 2011.