

Domain-specific Image Geocoding: a Case Study on Virginia Tech Building Photos

Lin Tzy Li^{1,2}, Otávio A. B. Penatti¹, Edward A. Fox³ and Ricardo da S. Torres¹

¹RECOD Lab, Institute of Computing, University of Campinas (UNICAMP), Campinas, SP – Brazil, 13083-852

²Telecommunications Res. & Dev. Center, CPqD Foundation, Campinas, SP – Brazil, 13086-902

³Digital Library Research Laboratory, Department of Computer Science, Virginia Tech, Blacksburg, VA 24061
{lintzyli, penatti, rtorres}@ic.unicamp.br, fox@vt.edu

ABSTRACT

The use of map-based browser services is of great relevance in numerous digital libraries. The implementation of such services, however, demands the use of geocoded data collections. This paper investigates the use of image content local representations in geocoding tasks. Performed experiments demonstrate that some of the evaluated descriptors yield effective results in the task of geocoding VT building photos. This study is the first step to geocode multimedia material related to the VT April 16, 2007 school shooting tragedy.

Categories and Subject Descriptors

H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing; I.2.10 [Artificial Intelligence]: Vision and Scene Understanding

Keywords

geocoding; map-based browsing; content-based video retrieval

1. INTRODUCTION

Since geographic information is involved in people's daily lives, there is a great amount of data about geographical entities on the Web. This information is often found in digital objects (e.g., documents, images, and videos) of several digital libraries. The process of associating a geographic location with photos, videos, and documents is called *geocoding* by the Geographic Information Retrieval (GIR) community, but it is also known as *geotagging* or *georeferencing* [9]. When a digital object is geocoded, it is related to some place on Earth, and therefore it can be browsed on a map. That opens new opportunities for establishing new relations based on geographic location.

In this work, we are interested in geocoding documents about the Virginia Tech (VT) April 16, 2007 school shooting tragedy, where 32 people were murdered before the shooter killed himself. A collection of documents about this event

was archived by CTRnet (Crisis, Tragedy and Recovery Network)¹, a digital library network for providing a range of services relating to different kinds of tragic events [3]. The aim of our present work is to support the creation of map-based browsing services based on photo content. The first step towards that is to be able to geocode images. This work describes preliminary experiments related to the evaluation of image descriptors in image geocoding tasks.

2. RELATED WORK

From 1995 to 2004, a project led by the University of California (UC Santa Barbara) called ADEPT - Alexandria Digital Earth ProtoType [4] focused on geocoding digital library objects by taking into account textual metadata.

Geocoding approaches based on visual clues are proposed in the context of landmark recognition, as well as for non-landmark images [9]. Usually, those approaches are modeled as image classification or content-based image retrieval (CBIR) problems. Those approaches often take advantage of a huge collection of geotagged images that is used as a knowledge base [6].

In our previous work [10], we explored a strategy based on global descriptors to tackle the challenge of geocoding videos based only on visual features. In more recent studies [7, 8], we combined visual and textual features for video geocoding.

Previous initiatives, such as [12], perform matching of local descriptors to find similar regions within images of a dataset of buildings. Therefore, although they are not explicitly geocoding images, their approaches could be used for that purpose. The approach presented in [12] works with buildings from the University of Oxford². After describing images with a scheme based on a visual vocabulary (quantized local features), a matching strategy is performed between a given query image and images from the dataset. They compare the performance of different vocabulary sizes, as well as vocabularies generated by different methods. In this work, we employ a similar strategy. Our work aims at evaluating the performance of local image description approaches in the task of geocoding building photos.

3. EXPERIMENT

This section presents experiments aiming at evaluating the effectiveness of visual-dictionary-based descriptors in the task of assigning locations to building photos.

¹<http://www.ctrnet.net/> (as of May 13, 2013).

²<http://www.robots.ox.ac.uk/~vgg/data/oxbuildings/> (as of May 13, 2013).

3.1 Image collections

We use two datasets in our experiments. One is used as our visual knowledge base and will be referred to as training data. The other is the test data images whose locations will be predicted by the proposed geocoding system.

Training dataset

The training data is a subset of 4,852 photos from VT’s University Relations (UniRel) Photo Library. Each photo has some metadata associated, such as keywords, caption describing the scene, date, camera model, and photographer’s name. For our purpose, we filtered the photos by the content of keywords and caption fields. As we were interested in the university buildings, we searched for photos whose metadata (keywords or caption) contains building and place names (e.g., Duck Pond). The building/place names list was built up from both the VT site³ and the campus building database maintained by GIS staff for campus facilities. The resulting training set contains photos of buildings or places with their location. Figure 1a shows the spatial distribution of buildings whose photos are in the training set.

Test dataset

The test dataset⁴ contains 565 photos of VT buildings. Most of them were obtained from personal collections while some others were downloaded from the VT Web site³. The photos were obtained under different angle and light conditions. The locations of these photos are shown in Figure 1b. Note that the test set covers a smaller area (near to the Drillfield in the campus center) when compared to the training set.



(a) Training set.

(b) Test set.

Figure 1: Spatial distribution of photos used as (a) training and (b) test set. Generated using <http://www.sethoscope.net/heatmap/> (as of May 2013). Map tiles by Stamen Design, under CC BY 3.0. Data by OpenStreetMap, under CC BY SA.

Ground truth

The ground truth for the images in training and test data sets, that is, the “correct” location for each of them, was

³<http://www.vt.edu/about/buildings/> (as of May 13, 2013).

⁴<http://www.recod.ic.unicamp.br/VTBuildings/> (as of May 13, 2013)

inferred from the corresponding building/place name associated with the photo. For the training photos, we used the place/building name that appears in their metadata. For the test photos, we use the name that we manually labeled each photo. The ground truth for these photos is based on the lat/long from the VT site³, as well as on building names geocoded by Google’s geocoding service (via the geopy toolbox⁵, a Python wrapper). However, if no matches were found by the geocoding service or if disambiguations were needed, the place/building was manually located and confirmed in the Google Map using its name. Additionally, some photos and some of the resulting geocoding locations were visually and manually inspected to decide their final location and/or coordinates.

3.2 Geocoding Process

The geocoding scheme adopted is based on performing K-nearest neighbor (KNN) searches. In this exploratory study, the location of a test photo is defined based on the geographic coordinates of the most similar image in the training set, i.e., is defined in terms of the location of the 1-nearest neighbor (i.e., $K = 1$) of the test photo. The visual distances between an input test image and all training images are computed. Training images are then ranked in ascending order of their visual distance to the input test image, and the latitude and longitude coordinates of the top-ranked training photo is assigned to the test image.

3.3 Evaluation criteria

The evaluation criterion used here is inspired by the evaluation procedure adopted in the Placing Task at MediaEval [13]. The effectiveness of a method is based on the geographic distance (great-circle distance) of the estimated geo-coordinates of a digital object to its corresponding ground truth location, in a series of widening circles of radius. An estimated point is counted as correct if it is within a particular circle size, that is, a radius value or precision level.

In our case, we are interested in determining as accurately as possible the location for a photo image. Furthermore, our area of interest is restricted to the Virginia Tech campus, so our precision level should be in meters. Taking into account that the two farthest points of the town of Blacksburg (where VT is located) are about 10 km apart, we can accept that two points on the VT campus should not be further apart than 5 km. The precision levels adopted are 1, 50, 100, 150, 200, 250, 300, 350, 400, 450, 500, 550, 600, 650, 700, 750, 800, 850, 900, 950, 1000, 1100, 1200, 1300, 1400, 1500, 1600, 1700, 1800, 1900, 2000, 3000, 4000, and 5000 meters.

3.4 Setup

First, the visual content properties of each provided image are encoded into feature vectors, considering all evaluated descriptors. Then, the visual distances between the photos in the test set and all photos in the training set are computed. Finally, for each test photo, a ranked list of training photos is produced.

To represent each image, we used the bag-of-visual-word model [14]. In that model, after extracting low-level features with local descriptors, we quantize the feature space in order to obtain a visual dictionary (codebook) and then we represent each local description according to the dictionary. For low-level feature extraction, we used: dense SIFT

⁵<https://github.com/geopy> (as of May 13, 2013).

(6 pixels) [15], sparse SIFT (Harris-Laplace detector) [15], and sparse SURF (Fast-Hessian detector) [1]. We randomly quantized the feature space [17], generating two dictionary sizes: 1,000 and 10,000 visual words.

To compute the bag-of-words representation, we used soft assignment ($\sigma=60$ for SIFT and $\sigma=0.08$ for SURF) [16] and two pooling methods: max pooling [2] and Word Spatial Arrangement (WSA) [11]. WSA was used only over the sparse SIFT, while max pooling was used for all low-level features. Table 1 lists the evaluated methods.

Table 1: Image representations evaluated.

Acronym	Method
D.SIFT.1k	dense SIFT, 1,000 words, soft assignment ($\sigma=60$), max pooling
D.SIFT.10k	dense SIFT, 10,000 words, soft assignment ($\sigma=60$), max pooling
S.SIFT.1k	sparse SIFT, 1,000 words, soft assignment ($\sigma=60$), max pooling
S.SIFT.10k	sparse SIFT, 10,000 words, soft assignment ($\sigma=60$), max pooling
S.SURF.1k	sparse SURF, 1,000 words, soft assignment ($\sigma=0.08$), max pooling
S.SURF.10k	sparse SURF, 10,000 words, soft assignment ($\sigma=0.08$), max pooling
WSA.1k	sparse SIFT, 1,000 words, soft assignment ($\sigma=60$), WSA
WSA.10k	sparse SIFT, 10,000 words, soft assignment ($\sigma=60$), WSA

4. RESULTS

Figure 2 presents the geocoding results for evaluated methods, considering different precision levels. Observe that the S.SURF.10k descriptor yields better results starting from the 150 m precision level on, followed by the D.SIFT.1k descriptor. Given that there are neighboring buildings in VT that are apart from each other (measured from their centroid) about 100 to 200 meters, it is reasonable to tolerate a maximum estimation error around 200 m. The S.SURF.10k method geocoded correctly 20% of the photos within 150 m precision radius and almost 80% within 600 m. For the precision level of 1 m, WSA.10k geocoded more images.

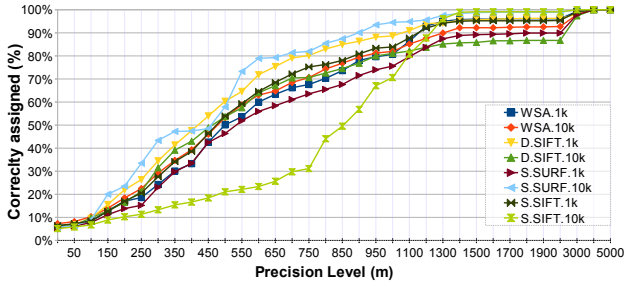


Figure 2: Correctly predicted test photos.

Correlation Analysis

In our previous study [7], we have shown that combining individual non-correlated descriptors may improve geocoding results. A correlation analysis helps evaluating the most promising descriptors to be combined. In order to do that, we analyze the results for each descriptor evaluated on the training set. We perform experiments considering each image of the training set as a query photo. In this case, given that the query photo always is the best match to itself (thus it will be the first in this list), we use the second photo of available ranked lists to define the final location.

Figure 3 shows the correlation graph (*corrgrams* R package) for the results of the training set. In this case, for each method and query image, its geocoding result is the geographic distances between a predicted point and its ground truth. Thus, we studied the correlation of these results for the evaluated methods.

This kind of plot is presented in [5] and shows the correlation values for each pair of methods as a square matrix. A darker color indicates higher correlation value. In the lower triangle, the correlation values are encoded by the intensity of the cell color. In the upper triangle, the same values are represented by the painted area in the circles and its color intensity. The diagonal of this matrix holds the name of the descriptor corresponding to each row and column. In this case, the lowest correlations are between S.SIFT.1k and the others, as shown by the lightest colors in the first column and the smallest painted area in circles in the first line.

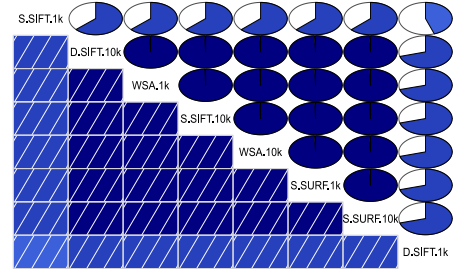


Figure 3: Correlation among evaluated descriptors.

Figure 4 shows the geocoding results of evaluated descriptors on the training set. We can observe that WSA.10k and S.SIFT.1k yield the best performance, followed by WSA.1k and D.SIFT.10k. On the other hand, S.SIFT.1k is less correlated with other descriptors (see Figure 3), which suggests that its combination with other descriptors may improve the geocoding results.

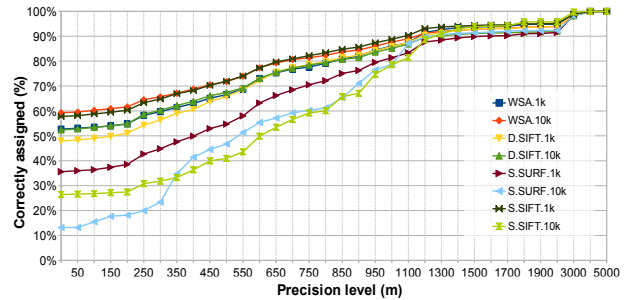


Figure 4: Correctly predicted training photos.







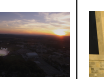











Comparing the geocoding results on both test and training sets, we found some surprising results. S.SURF.10k, for example, yields the worst results in the training set, but performs very well on the test set. One possible explanation relies on the differences between test set and training set images. In the test set, there are more close up photos, whereas the training set includes pictures in a wider frame, i.e., the images depict more distant objects (buildings).

Examples of geocoding results

Table 2 shows two examples of query photos of the test set and its corresponding top-similar images in the training set for each visual descriptor. Each table cell also presents the geographic distance between the top-ranked training image and the query photo's ground truth.

Consider, for example, the top-ranked image in the case of query P1080710 (picture of Torgersen Bridge). The S.SURF

Table 2: The best visual match for each query image and its geocoding result. Values below the photo thumbnail refer to the geographic distance (in meters) to the ground-truth location of the query image.

Query	Building Name	D.SIFT.1k	D.SIFT.10k	S.SIFT.1k	S.SIFT.10k	S.SURF.1k	S.SURF.10k	WSA.1k	WSA.10k
	Lane Hall								
P1080012		424.08	86.87	349.62	161.13	217.67	1231.06	238.44	238.44
	Torgersen Bridge								
P1080710		906.41	2801.90	74.33	1070.43	0.00	0.00	275.15	74.33

(1k and 10k) descriptor is able to match it to a photo that only pictures a detail of that building, whereas WSA.10k and S.SIFT.1k match that to a photo from the same building but under a different light (darker) condition. However, as this photo was labeled as Torgersen Hall instead of Torgersen Bridge, its geographic distance was not zero. The query P1080012 (Lane Hall) shows an example where S.SURF.10k performed very badly. D.SIFT.10k, in turn, matched it to a photo of a building (Shanks Hall) that is close (86.87 m) to the Lane Hall, while S.SIFT.10k found the query similar to a picture of Torgersen Bridge (161.13 m from there) at night. In summary, each descriptor provides different, but potentially complementary information that could be combined to improve geocoding results.

5. CONCLUSIONS

This paper presented an evaluation of several image content local descriptors in the context of geocoding building photos. Our objective is to determine appropriate image content description approaches to be used to geocode images related to the Virginia Tech April 16, 2007 school shooting tragedy. Geocoded photos can be used later in the construction of map-based browsing services in the context of the CTRnet (Crisis, Tragedy and Recovery Network) project. Performed experiments show that some of the evaluated descriptors (e.g., S.SURF.10k, WSA.10k, D.SIFT.1k) yield effective results in geocoding tasks. Future work will focus on the use of data fusion techniques to combine non-correlated descriptors. We also plan to consider available textual descriptions in geocoding tasks and new geocoding schemes.

Acknowledgements

The authors are grateful to CAPES, FAPESP (grant number 2009/10554-8), and CNPq, as well as the CPqD Foundation for their support. We also thank VT University Relations for providing access to some of the photographs used in this work. Additionally, we would like to thank the Center for Geospatial Information Technology (CGIT) and the GIS Manager for VT campus facilities for providing access to the campus building database. Finally, we thank the National Science Foundation (IIS-0916733) for supporting our ‘Crisis, Tragedy and Recovery Network’ research¹.

6. REFERENCES

- [1] H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. In *ECCV*, pages 404–417, 2006.
- [2] Y.-L. Boureau, F. Bach, Y. LeCun, and J. Ponce. Learning mid-level features for recognition. In *CVPR*, pages 2559–2566, 2010.
- [3] E. A. Fox, C. Andrews, W. Fan, J. Jiao, A. Kassahun, S. C. Lu, Y. Ma, C. North, N. Ramakrishnan, A. Scarpa, and Others. A Digital Library for Recovery, Research, and Learning From April 16, 2007, at Virginia Tech. *Traumatology*, 14(1), 2008.
- [4] M. Freeston. The Alexandria Digital Library and the Alexandria Digital Earth Prototype. In *JCDL*, page 410, 2004.
- [5] M. Friendly. Corrgrams: Exploratory displays for correlation matrices. *The American Statistician*, 56(4):316–324, 2002.
- [6] J. Hays and A. A. Efros. IM2GPS: estimating geographic information from a single image. In *CVPR*, 2008.
- [7] L. T. Li, J. Almeida, D. C. G. Pedronette, O. A. B. Penatti, and R. da Silva Torres. A multimodal approach for video geocoding. In *Working Notes Proc. MediaEval Workshop*. CEUR-WS.org, 2012.
- [8] L. T. Li, D. C. G. Pedronette, J. Almeida, O. A. B. Penatti, R. T. Calumby, and R. da Silva Torres. Multimedia Multimodal Geocoding. In *ACM SIGSPATIAL GIS*, pages 474–477, 2012.
- [9] J. Luo, D. Joshi, J. Yu, and A. Gallagher. Geotagging in multimedia and computer vision—a survey. *MTA*, 51:187–211, 2011.
- [10] 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 *ACM ICMR*, 2012.
- [11] O. A. B. Penatti, E. Valle, and R. da S. Torres. Encoding spatial arrangement of visual words. In *CIARP*, volume 7042, pages 240–247, 2011.
- [12] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *CVPR*, pages 1–8, 2007.
- [13] A. Rae and P. Kelm. Working Notes for the Placing Task at MediaEval 2012. In *MediaEval*, 2012.
- [14] J. Sivic and A. Zisserman. Video Google: a text retrieval approach to object matching in videos. In *ICCV*, volume 2, pages 1470–1477, 2003.
- [15] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluating color descriptors for object and scene recognition. *TPAMI*, 32(9):1582–1596, 2010.
- [16] J. C. van Gemert, C. J. Veenman, A. W. M. Smeulders, and J.-M. Geusebroek. Visual word ambiguity. *TPAMI*, 32(7):1271–1283, 2010.
- [17] V. Viitaniemi and J. Laaksonen. Experiments on selection of codebooks for local image feature histograms. In *VISUAL*, pages 126–137, 2008.