Landmark Recognition: State-of-the-Art Methods in a Large-Scale Scenario*

Magdalena Rischka and Stefan Conrad

Institute of Computer Science
Heinrich-Heine-University Duesseldorf
D-40225 Duesseldorf, Germany
rischka@cs.uni-duesseldorf.de
conrad@cs.uni-duesseldorf.de

Abstract. The recognition of landmarks in images can help to manage large image collections and thus is desirable for many image retrieval applications. A practical system has to be scalable with an increasing number of landmarks. For the domain of landmark recognition we investigate state-of-the-art CBIR methods on an image dataset of 900 landmarks. Our experiments show that the kNN classifier outperforms the SVM in a large-scale scenario. The examined visual phrase concept has shown not to be as effective as the classical Bag-of-Words approach although the most landmarks are objects with a relatively fixed composition of their (nearby) parts.

Keywords: Image Retrieval, Large-Scale, Landmark Recognition, Bag-of-Words, Bag-of-Phrases

1 Introduction

The ongoing development of personal electronic devices like digital cameras, mobile phones or tablets with integrated camera and high-capacity memory cards, as well as their decreasing prices enable taking photos everywhere and at any time. Collecting and storing photos as well as sharing photos with others on online social network platforms leads to huge photo collections in personal households and to a much greater extent on the world wide web. To manage and reuse these images in an useful way (e.g. for search purposes) it is necessary to capture the images' content, i.e. to annotate the images with meaningful textual keys. A large amount of the collections' images are photos shot in the photographer's vacations and trips showing (prominent) places and landmarks the photographer visited. The detection and recognition of landmarks in images offers several advantages regarding applications: the above-mentioned annotation constitutes

 In: T. Seidl, M. Hassani, C. Beecks (Eds.): Proceedings of the LWA 2014 Workshops: KDML, IR, FGWM, Aachen, Germany, 8-10 September 2014, published at http://ceur-ws.org

a foundation for a search or can be used as a suggestion for a photo description to the user. Another usage is the identification of locations the photographers visited, for example to summarize personal image collections by offering an overview of places. The application of mobile landmark recognition enables tourists to look up sights in real-time to obtain informations on them. Capturing images' content by manual annotation of images with landmarks however is very time-consuming, in the scale of these collections even inconvertible, therefore an automatic solution is needed. Several systems for automatic landmark recognition have been proposed [2–6] differing in the focus of application scenario, the initial situation referring metadata, problem definition and implemented techniques. For example the authors of [2] create a database from geo-tagged Flickr photos and Wikipedia. Object-level recognition is performed with the aid of an index and candidate images are ranked using a TF-IDF scheme. [3] also creates a dataset from Flickr images and then derives scene maps of landmarks which are retrieved with an inverted index. [4] creates the database by crawling travel guide websites and then builds a matching graph out of the feature matches of the images. For retrieval a kd-tree is used. We concentrate on images without any metadata, thus on content-based methods only. Several state-of-the-art methods in CBIR have been examined and applied successfully on small or average size datasets. Our focus is on the large-scale aspect of a landmark recognition system and the usability in real world scenarios, thus our contribution is the comparison of these methods with reference to scalability.

The remainder of this paper is organized as follows: in the next section we outline and formalize the problem of landmark recognition by defining the landmark term, describing the characteristics of landmark images, specifying the landmark recognition task and presenting the components of the landmark recognition system evaluated in section 3. In section 4 we summarize our results and discuss future work.

2 Landmark Recognition Problem and System

A *landmark* is a physical object, created by man or by nature, with a high recognition value. Usually a landmark is of remarkable size and is located on a fixed position of the earth. Examples of landmarks are buildings, monuments, statues, parks, mountains and other structures and places. Due to their recognition value, landmarks often serve as geographical points for navigation and localisation. The largest amount of photos of landmarks contain only one landmark, which in the most cases takes in 80% of the photo area, in very few cases it takes only a small part of the photo (when it is taken from apart). A marginal part of photos show two or more landmarks. A landmark recognition system has to conduct the following task automatically:

Definition 1 (Landmark Recognition Task). *Given a set of L landmarks $\mathcal{L} = \{l_1, ..., l_L\}$ and an image i whose semantic content is unknown. The task is*

to assign a set of landmarks to the image:

$$i \rightarrow \begin{cases} \emptyset & \textit{if image } i \textit{ does not contain any landmark} \\ \{l_{j_1}, ..., l_{j_n}\} & \textit{if image } i \textit{ contains landmarks } \{l_{j_1}, ..., l_{j_n}\} \end{cases} \quad (1)$$

This definition implies a multi-label classification problem with a decision refusal. We simplify the multi-label classification problem defined in (1) by building our system on a single-label classification approach, thus we accept a possible misclassification of images containing more than one landmark. We focus on the classification step, the decision refusal which is usually performed with a post-processing verification algorithm (like RANSAC) is beyond this work. The main components of our landmark recognition system, which are the image representation and the classifier are discussed in the following paragraphs.

Image Representation To describe images we extract the popular SIFT [7] features. The SIFT algorithm extracts local features by detecting stable points and then describing the (small) surrounding area around each point by an histogram of gradients. An image i is represented by a set of local SIFT points: $SIFT(i) = \{p_1, ..., p_P \mid p = (x, y, s, d)\}$ with x, y are the coordinates of the point p in the image, s is the scale and d the 128-dimensional descriptor. We analyse two types of image representation based on the local SIFT features: Bag-of-Words (BoW) and the Bag-of-Phrases (BoP) model based on the visual phrase concept. Although visual phrases have been used in general object recognition applications, they raised less attention in the domain of landmark recognition. We like to analyse if visual phrases improve the BoW classification results.

The *Bag-of-Words* model is a classical approach to create a compact image representation based on local features. The idea is to aggregate local features to one global descriptor and thus to avoid the expensive comparison of images by matching local descriptors against each other. The BoW descriptor bases on a dictionary of visual words which is obtained by partitioning the descriptor-space. Then each partition is represented by an instance of this partition, usually the center of the partition, which is called the *visual word*. Several methods for partitioning the descriptor-space have been proposed, a simple and most used one is the k-Means clustering algorithm which requires the input parameter k (which onwards is denoted as D to differentiate between the kNN parameter k) for the number of clusters (visual words) to obtain.

Definition 2 (Bag-of-Words Model).
Given a dictionary $\mathcal{D} = \{(w_1, c_1), ..., (w_D, c_D)\}$ of D visual words (w_j, c_j) (w_j is label, c_j the center of the partition) and an image in SIFT representation. Each SIFT point p of the image is assigned to its visual word w_p by:

$$w_p := w_j = \underset{(w_j, c_j) \in \mathcal{D}}{argmin} \; (EuclideanDistance(d, c_j)) \quad (2)$$

The Bag-of-Words image representation is given by:

$$BoW(i) = \{f_1, ..., f_D\} \text{ with } f_j = \frac{1}{P} \sum_{p=1}^{P} \begin{cases} 1, & w_p = j \\ 0, & else \end{cases} \tag{3}$$

Visual phrases catch spatial relations in local neighborhood by considering pairs of nearby local features or visual words to support more semantic, analogously to phrases in text retrieval. We follow [8] and define the visual phrase and the Bag-of-Phrases model as follows:

Definition 3 (Bag-of-Phrases Model).
Given the visual dictionary $\mathcal{D} = \{(w_1, c_1), ..., (w_D, c_D)\}$. A visual phrase $ph_{j,k}$ is a pair of visual words: $ph_{j,k} = (w_j, w_k)$ with $j \leq k$. An image in SIFT representation with its visual words $SIFT'(i) = \{p_1, ..., p_P \mid p = (x, y, s, d, w)\}$ contains the phrase $ph_{j,k}$ if there exist two SIFT points p_m and p_n with their visual words w_j and w_k and it holds

$$EuclideanDistance((x_m, y_m), (x_n, y_n)) \leq max(\lambda \cdot s_m, \lambda \cdot s_n) \tag{4}$$

for a fixed scale factor λ. The Bag-of-Phrases image representation is given by:

$$BoP(i) = \{f_1, ..., f_{D_2}\} \quad with \quad D_2 = \frac{D \cdot (D+1)}{2} \tag{5}$$

with f_j is the relative frequency of the visual phrase ph_j in image i.

Classifier For the choice on the classifier, we evaluate three well-known classifiers, the Support-Vector-Machine (SVM), the k-Nearest Neighbor (kNN) and the Nearest Center classifier (NC). The SVM is a popular classifier as it provides better classification results than other standard classifiers in the most (computer vision) classification tasks. However the drawback of the SVM classifier is the long classifier learning time, especially with an increasing training data size. In addition to that [1] has shown that the superiority of the SVM over the kNN classifier (with regard to classification quality) can swap with an increasing number of classes. As our focus is on the large-scale landmark recognition with reference to an increasing number of landmarks, we investigate the landmark recognition on both classifiers. The kNN classifier has no training part, instead the classification time is linear in the number of training examples, which in a scenario of over 100.000 images and a system implementation without the use of appropriate and efficient access structures can put a strain on the user. However the classifier NC can be seen as a lightweight classifier, as both the learning and the classification time is linear in the number of classes. For the SVM we use the RBF kernel and the one-vs-one mode, for the kNN we set $k = 5$ (as a result of preliminary experiments on different k), for the kNN and the NC classifier we choose the histogram intersection as the similarity function.

3 Evaluation

Evaluation Dataset For the evaluation we use a self-provided dataset of landmarks. We gathered landmark terms from several websites which lists landmarks from all over the world, including the website of [4][1]. Our dataset consists of 900 landmarks from 449 cities and 228 countries. To get images for the training and test sets, we queried the google image search engine with each landmark term (specified by its region - city or country) and then downloaded the results from the original source. For scalability analysis we derived training sets of four different sizes: 45, 300, 600 and 900 landmarks. For each size three training sets (A, B, C) have been created. To create a challenging test set we have chosen images manually from the results of the google image search: The images show the landmarks in their canonical views, under different perspective changes, distortions and lighting conditions (also at night) as well as indoor shootings and parts of the landmarks. The test set consists of 900 landmark images (45 well- and lesser-known landmarks from Europe with 20 test images per landmark). The test images have been proofed to be disjoint from the training set.

Evaluation Measures The outcome of a single-label classifier on a test image is the predicted landmark. To retain a fine-grained evaluation of a test image, we are also interested in a ranking of landmarks, as the ranking reveals how far away is the groundtruth landmark from the top ranking position. The SVM delivers us a ranking based on the probability values of the one-vs-one voting, the NC classifier based on the histogram intersection similarity. The kNN returns only the predicted landmark. To evaluate the results of the classifiers we use two (instance-based) evaluation measures: the (instance-based) recall on the predicted landmark and the MAP measure (which is equal to MRR measure in this case) for the landmarks ranking. We finally report the average value over all test images.

Experiments The first experiment evaluates the Bag-of-Words model in combination with the three classifiers and the four training sets of size 45, 300, 600 and 900. The Bag-of-Words model has one parameter which is the visual dictionary size. We examine the following six different visual dictionary sizes: 500, 1000, 2000, 4000, 6000 and 8000. Figure 1 shows the results of this experiment. The recall and MAP values reported are averages over the training sets A,B,C of the corresponding training set size. Table 1 shows the average recognition time for a test image on the Bag-of-Words model with a visual dictionary size of 8000 depending on the classifier and the training set size. The average recognition time does not include the processing time for image representation computation. Experiments are performed on an usual Intel i7 960 3.2 GHz (64-bit) architecture with 16 GB memory size. The system (kNN) is implemented without the use of any efficient access structure. Considering the recall values of all classifiers for all

[1] `http://mingzhao.name/landmark/landmark_html/demo_files/1000_landmarks.html`

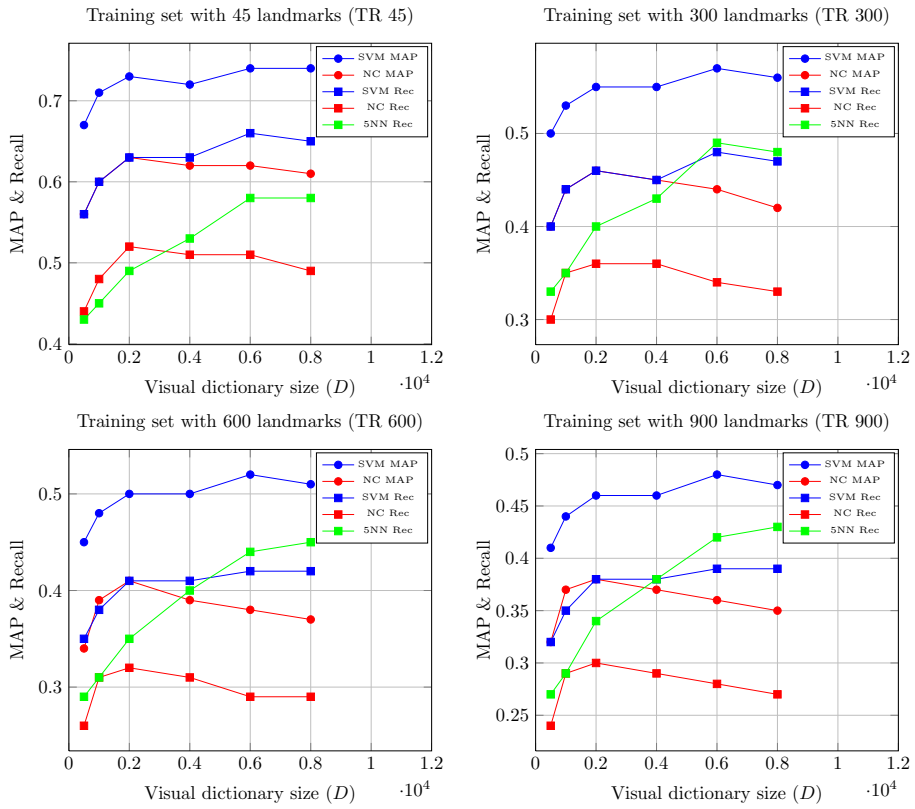

Fig. 1. Classification results of the BoW model depending on the parameters visual dictionary size D (x-axis), classifier with evaluation measure (plots) and training set size (subimages)

training set sizes, we can see that the best values are achieved by the SVM and the 5NN with a visual dictionary size of 6000 and 8000. On the training set TR 45 the SVM gets the best recall value with 0.65 ($D = 6000$). From the training set TR 300 on the 5NN outperforms slightly the SVM resulting in a difference of 4% on TR 900 and $D = 8000$. Furthermore the 5NN shows the tendency to achieve higher results with a growing visual dictionary. These results confirm the observation of the superiority of kNN over the SVM in large-scale problems stated in [1]. The NC classifier achieves best results on $D = 2000$ for all training set sizes, however its best values are on average 13% lower than the best system (SVM or kNN). The MAP values of the SVM and the NC reveal that there is potential to improve these classifiers when involving the next to top ranking positions in the classification decision. In general the recognition accuracy decreases with an increasing number of landmarks which is not surprising. A recall value of 0.66 on the TR 45 (SVM, $D = 6000$) can be somewhat satisfying, however the best result of 0.43 on TR 900 (5NN, $D = 8000$) is less delightful. The

second experiment concentrates on the Bag-of-Phrases model. Again we report experiments in combination with the three classifiers and the four training set sizes. The BoP model requires two parameters to be set: the visual dictionary size D and the scale factor λ. As the dimension of the image's descriptor in this representation becomes very large on already small visual dictionary sizes, we examined the two sizes 500 and 1000 resulting in the descriptor dimension of 5050 and 125250, respectively. For the scale factor we choose the values 1, 2, 4 and 6. The BoP results (Figure 2) for all classifiers and all training set sizes are on average 10% lower than the BoW results. The larger visual dictionary $(500, \lambda)$ achieves better results than the smaller one $(100, \lambda)$, especially on the SVM, whereas the scale factor influences the results slightly. In the most cases the scale factor of 2 gets best results. Due to the high-dimensional descriptor ($D \geq 500$) the recognition time is many times higher than of the BoW model.

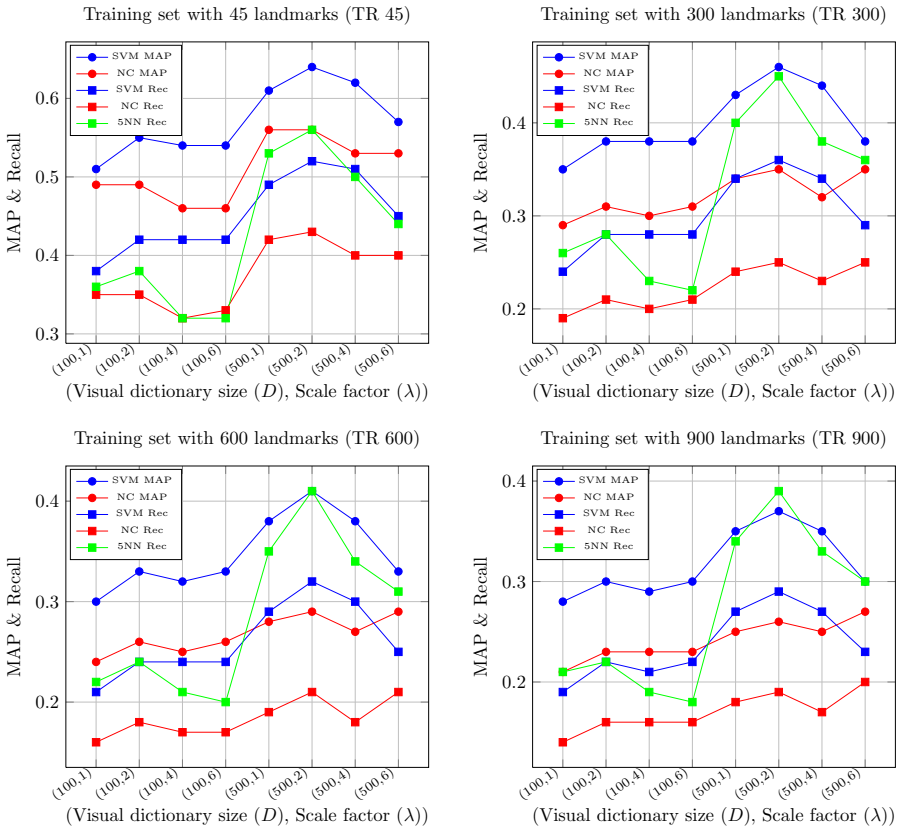

Fig. 2. Classification results of the BoP model depending on the parameters visual dictionary size D and scale factor λ (x-axis), classifier with evaluation measure (plots) and training set size (subimages)

	TR 45	TR 300	TR 600	TR 900
SVM	0.0577	0.1016	0.2710	0.5081
NC	0.0039	0.0239	0.0424	0.0612
5NN	0.1129	0.4637	0.8444	1.2232

Table 1. Average recognition time (in seconds) for the BoW model with a dictionary of size 8000 dependent on the three classifiers and the four training set sizes.

4 Summary and Future Work

To build a landmark recognition system with large number of landmarks (TR 900) supported, the Bag-of-Words model together with the kNN classifier offers a higher recognition accuracy than the SVM but on the cost of a relatively high recognition time of about 1.2 seconds per image. A solution to use kNN and to reduce the recognition time is to integrate an appropriate and efficient access structure into the system and to try to reduce the number of training images per landmark (by a compressed representation) without loosing too much relevant informations. The BoP model alone does not convince, therefore the question arises, if this model returns additional knowledge to the BoW model. In fact, some few landmarks (33% of the tested landmarks) benefit from the BoP model, others not. A detailed analysis of this and a suitable combination of both models are matters for further research beyond this work. Furthermore it would be interesting to compare our state-of-the-art approach with a system which bases on an inverted file index working directly on local features.

References

1. Deng, J., Berg, A. C., Li, K., Fei-Fei, L.: What does classifying more than 10,000 image categories tell us? In: Proceedings of the 11th European conference on Computer vision (ECCV'10), 2010
2. Gammeter, S., Bossard, L., Quack, T., Gool, L.V.: I know what you did last summer: object-level auto-annotation of holiday snaps. In: IEEE international conference on computer vision (ICCV), 2009
3. Avrithis, Y., Kalantidis, Y., Tolias, G., Spyrou, E.: Retrieving landmark and non-landmark images from community photo collections. In: Proceedings of the international conference on Multimedia (MM '10). ACM, New York, 2010
4. Zheng, Y., Zhao, M., Song, Y., Adam, H., Buddemeier, U., Bissacco, A., Brucher, F., Tat-Seng Chua, Neven, H.: Tour the world: Building a web-scale landmark recognition engine. In: Computer Vision and Pattern Recognition (CVPR), 2009
5. Philbin, J., Zisserman, A.: Object Mining Using a Matching Graph on Very Large Image Collections. In: ICVGIP, 2008
6. Crandall, D. J., Backstrom, L., Huttenlocher, D., Kleinberg, J.: Mapping the world's photos. In Proceedings of the 18th international conference on World wide web (WWW '09). ACM, New York, 2009
7. Lowe, D. G.: Distinctive Image Features from Scale-Invariant Keypoints. In: International Journal of Computer Vision, 2004
8. Zheng, Q. F., Gao, W.: Constructing visual phrases for effective and efficient object-based image retrieval. In ACM Trans. Multimedia Comput. Commun. Appl. 5, October 2008