

# MonuAnno: Automatic Annotation of Georeferenced Landmarks Images

Adrian Popescu  
TELECOM Bretagne  
Technopôle Brest Iroise  
29238 Plouzané, France  
+33229001435

adrian.popescu@telecom-  
bretagne.eu

Pierre-Alain Moëllic  
CEA LIST  
18 route du Panorama  
92265 Fontenay aux Roses, France  
+33146549619

pierre-alain.moellic@cea.fr

## ABSTRACT

Uploading tourist photographs is a popular activity on photo sharing platforms. The manual annotation of these images is a tedious process and the users often upload their images with no associated textual information. Automating the annotation process has received a lot of attention but the problem remains a hard one, especially when dealing with large and heterogeneous databases. Here we focus on landmarks images, very frequent among tourism pictures, and propose a new automatic technique for annotating this type of pictures. Our system, called MonuAnno, relies on the joint exploitation of localization information and of image content analysis in an efficient and scalable framework. The annotation is performed using a two steps  $k$  Nearest Neighbors ( $k$ -NN). First, only neighboring landmarks of a new unlabeled georeferenced image will be considered as potential annotations and the image will be attributed to the landmark that is visually closest. Then, we introduce a verification step that eliminates false positives (images taken near a landmark that represent something else). The technique was tested on Web images and the results show that the precision of the labeling process in MonuAnno exceeds 80%, when annotating around 50% of the images in the test set.

## Categories and Subject Descriptors

### H.3.1 Content Analysis and Indexing

### General Terms

Algorithms, Experimentation.

### Keywords

Image annotation, geographic gazetteer, georeferenced images, landmarks,  $k$ -NN, bags of visual words, Flickr, Panoramio.

## 1. INTRODUCTION

Landmarks are geographically situated objects or small areas and their localization information is useful when one wants to

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

CIVR '09, July 8-10, 2009 Santorini, GR

Copyright © 2009 ACM 978-1-60558-480-5/09/07... \$5.00

automatically label picture content. The main intuition supporting the use of spatial context when annotating georeferenced images can be formulated as follows: a spatially situated photograph can only represent neighboring objects and it is pointless to try to annotate it with distant place names. With the spread of geographically aware devices it becomes possible to exploit localization information in image annotation tasks. Early experiments in this direction were described in [4], where the authors manually constitute a database of georeferenced pictures containing 101 objects representing landmarks in Singapore. They perform an object recognition task, showing that the use of localization information improves the quality of the results. Tools like ZoneTag [2] propose an annotation of pictures with the place names surrounding a location but, to the best of our knowledge, no existing system performs an automated annotation of landmarks images from large scale and noisy georeferenced picture databases like Flickr.

Pictures can be described using a variety of descriptors which can be separated into two main classes according to type of content representation they propose: global (color or texture) and local (“bags of visual words” [20]). The basic idea of “bags of visual words” is to produce a visual vocabulary built after an unsupervised quantification of a set of patches extracted from images using local features such as SIFT descriptors [12]. New images are indexed by comparing their content to that of the vocabulary and by producing a histogram of occurrence of the vocabulary patches in the images to index. Global descriptors are simpler to compute than local features, but these last proved to be efficient in clustering landmark images [10], a task which is close to ours.

In this paper, we describe MonuAnno, a framework for annotating unlabeled images with neighboring place names that are represented by a sufficient number of georeferenced pictures on photo sharing platforms. First, MonuAnno extracts a list of landmarks (and of their geographic coordinates) from a large scale geographical gazetteer (see [8] for a definition). Then, for each element of the list, queries with landmark names, limited to a radius around the landmark coordinates, are launched in Flickr and Panoramio. MonuAnno downloads up to 500 georeferenced images for each landmark and indexes these images using the visual vocabulary.

Given an unlabeled georeferenced photo, MonuAnno compares the image to pictures of neighboring landmarks in order to determine if it represents one of these landmarks and if so, which

one. First, we employ a k-NN to select the object that is visually closest to the one depicted in the image. Second, we compare the image to those of the landmark that is visually closest and to those in a *Falses* class (a pool of irrelevant images). We annotate the picture with the landmark name only if the number of nearest neighbors depicting the landmark selected after the first step is higher than a threshold. This second phase is necessary for eliminating false positives (images taken near the *Saint Sulpice Church* in *Paris* that do not represent this landmark).

The remainder of this paper is structured as follows: section 2 reviews related work; section 3 details the data model and the problem definition; section 4 describes the constitution of the reference corpus and the indexing process; section 5 presents the classification process in more detail; the final section 6 details experiments for validating our approach.

## 2. RELATED WORK

Georeferenced images are interesting in a large number of applications and they constitute the subject of an important research effort, mostly in the image retrieval community. [1] introduced World Explorer, a tool presenting geographic tags situated on a map and associated Flickr photos. This is the first system for retrieving georeferenced images we know of that employs a large scale geographic database associated to a map-based interface. In [17], we extended the system in [1] and proposed two new functions: a thematic exploration of the world and a content based image retrieval. We exploited an enriched version of Geonames [6] to display tags and the user was able to select interest topics he or she wants to visualize, for example only *castles* and *churches* in a region. If a displayed image is selected, the system proposes a CBIR among the pictures representing the place name. In [17], the CBIR was limited to specific objects, that generally have a stable visual appearance, while here the same type of limitation is applied in an image labeling task.

[13] was one of the first attempts to enhance georeferenced content based image retrieval using geographic information. The authors use a dataset of around 10000 images and showed that the use of localization information, in addition to low level picture descriptors, enhances the retrieval process. The annotation of geotagged images is mentioned in [13], but left for future work.

While [13] exploits only global image descriptors, the use of local features seems more appropriate for landmarks images because landmarks generally have a stable appearance that is well captured using local descriptors. The adaptation of “bags of words”, a popular method in natural language processing, to images (“bags of visual words” [20]) endows the image processing community with new and powerful methods to build robust images descriptions. Hörster et al. [9] analyzed the use of bags of visual words for image retrieval within large and heterogeneous image datasets (Flickr) with LDA modeling and showed that this type of description is effective. For the creation of the vocabulary, [9] merge the results of multiple K-Means computed on different subsets of a large-scale collection. We computed several K-Means (using a parallel MPI implementation) on the same pictures dataset, with random initializations, and retained the best result defined as the partition with the optimal intra-cluster distance. Our approach is computationally more complex but finds an optimal (or nearly optimal) configuration of the codebook.

[10] introduced a method for clustering georeferenced Flickr images using multimodal information. Basically, a cluster is well ranked if it has a good visual coherence, a good connectivity and if its images temporally diverse, geographically localized and taken by different users. Global image descriptors are used for clustering, while local descriptors (SIFT) are only exploited for determining the cluster connectivity. We took a different approach and decided to use only local image descriptors because they are more efficient for the description of landmark images. [10] combines a visual processing and image metadata for finding representative landmark images. The photos used in [10] are already labeled with landmark names and the authors do not try to label new images. Our purpose is different because we use georeferenced landmark photos in order to annotate unlabeled images.

[7] automatically builds a database containing around six million pictures which is subsequently exploited for inferring image coordinates based on an image content analysis. We build the list of landmarks in our corpus using Gazetiki [18] and query Flickr and Panoramio [14] with both keywords and image coordinates to obtain a reference database. The main difference between the two approaches to constituting a reference database comes from the fact that we concentrate on specific objects (i.e. *Saint Sulpice Church*) while [7] uses general keywords (i.e. *Paris*). This choice is determined by the envisioned application: recognizing landmarks in our case and inferring geographic coordinates for non georeferenced images in [7].

[3] describes a framework for labeling collections of personal photographs based on a model combining metadata (GPS and time) and visual descriptors (color and SIFTs) to generate event and scene descriptions. The model discriminates 11 scenes and 11 events, a classification complexity that is well handled by SVMs. Their algorithm first annotates only photos with high confidence scores and then propagates labels to other images considering their proximity to initially annotated pictures. [3] annotates tourism images with a reduced set of generic scene and events names whereas we focus on labeling photos with specific landmark names, belonging to a list of around 5000 different elements.

ZoneTag [2] is a mobile application for uploading tagged images from a camera phone to Flickr. In addition to the tags introduced by the user, ZoneTag proposes location and event related tags based on the users' tagging history and the image coordinates and time stamp. ZoneTag does not perform any image processing and relies only on contextual information in order to propose nearby landmark names. Also, the annotation process in ZoneTag is semiautomatic because the system only proposes tags and it is up to the user to choose relevant labels.

Most existing methods for (semi)automatic image annotation combine image processing and machine learning techniques [5]. From a set of labeled pictures, the methods try to learn a correspondence between low-level features and semantic labels. Generally, this correspondence is hard to establish because of the diversity of the visual representation of concepts. As an example, try imagining what would be the low-level representation of *Paris* or *Europe*. These classes are too general to have a coherent visual representation and it is easier to attach them to georeferenced pictures using spatial reasoning. On the contrary, specific concepts (i.e. *Notre Dame* or *Saint Sulpice Church*) have a coherent visual representation and, provided that they are indexed with appropriate low-level descriptors, they can be well

represented at a visual level. With [3], we advocate that, while particularizing the annotation problem to specific domains, the joint use of contextual information and of specialized concepts in order to improve the efficiency of the automatic annotation.

In [4], the authors introduce SnapToTell, a framework for annotating georeferenced pictures. The image corpus, covering 101 landmarks in Singapore (5278 images), is manually constituted by the authors and does not contain any noise. Each object in the database is photographed from different points of view and at different distances. The first version of the system only exploits global image descriptors (color or texture). In a later work [11], the same group investigates the introduction of local descriptors in the annotation framework and reports that the annotation is improved, passing from 88% to 92%. While interesting, the work in [4] assumes that SnapToTell will only have to annotate representative images and does not treat the case when a picture near a landmark represents an object that doesn't belong to the reference database. MonuAnno is built using an open world model and handles non-representative pictures taken near a landmark. A second important difference with [4] is that the reference corpus used in our system is automatically constituted and contains pictures from Panoramio and Flickr. Third, SnapToTell is limited to landmarks in Singapore whereas the geographic gazetteer we employ [18] covers most regions of the world. The only condition for a landmark to be included in our annotation framework is for it to be represented by enough images in georeferenced corpuses like Flickr and Panoramio.

Quack et al. [19] downloaded a database of around 200 000 georeferenced Flickr images from 9 urban areas and clustered them using local image descriptors in order to discover place names and events. The authors state that their method is also suited for image tagging and provide some examples but no evaluation is presented.

### 3. DATA MODEL AND PROBLEM STATEMENT

We first describe the data model used in this paper, pointing its main features. Second, we define the research problem that is approached in this work.

Each landmark  $\ell$  in the dataset in MonuAnno is described by the tuple:

$$\ell = (tag^\ell, type^\ell, pic^\ell, coord^\ell)$$

Where:

- $tag^\ell$  – the name of the landmark;
- $type^\ell$  – the parent concept of the landmark;
- $pic^\ell = \{I_i^\ell\}_{i \in NI}$  – the set of images describing the landmark;
- $coord^\ell$  – are the coordinates (latitude and longitude) of the place where the picture was taken;

The tags define the set of salient geographic objects which are annotated using MonuAnno. The saliency is determined using a popularity measure over a set of a georeferenced photographic

corpus [18]. Each  $tag$  has an associated parent concept ( $type$ ) which facilitates an automatic annotation expansion with this parent concept using the inheritance relation. Each landmark is represented by its set of photos, which ideally capture diverse representations of the object, allowing its automatic recognition. The photos have an associated location, which is used during the corpus constitution and the annotation phases in order to restrain the search space.

We use Internet to constitute a reference database and there is no guarantee concerning the relevance of the photographs in the dataset. In particular  $pic^\ell$  will contain a certain amount of noise (shots which do not depict  $\ell$  but are annotated with its name). The location of the photos is also approximate but usually precise enough for our purposes.

Our research problem can be formulated as follows: given an unlabeled georeferenced image, decide if it depicts a neighboring landmark and, if so, label the picture with that landmark name. The photographs taken near a landmark but representing something that is not described the reference database should not be annotated. We model a real world situation and, if the combination of visual and contextual image description is efficient, the proposed solution can be implemented at a world scale.

## 4. AUTOMATIC CONSTITUTION OF A REFERENCE IMAGE CORPUS

When wanting to annotate thousands of different objects, it is necessary to constitute a large-scale reference database. First, we present the data sources we used, the way the landmarks list was built and some details concerning the obtained database. Second, we describe the indexing reference corpus.

### 4.1 Reference Corpus Constitution

#### 4.1.1 Data Sources

There are two image sources we used to build the reference corpus:

- **Panoramio**, a website for sharing georeferenced pictures, contains around 10 million images. The validation of photos relevance by other users constitutes a useful particularity of this platform compared to other sites, such as Flickr, because the amount of noisy images is greatly reduced. Panoramio data are available via an API which provides: the title and localization of the images, the name of the user who uploaded it and a link towards the image itself. In order to obtain the images for a landmark, it is necessary to match its name against the image title.
- **Flickr**, with over 90 million georeferenced images, is the largest source of localized pictures. Flickr images representing a landmark are generally noisier than Panoramio photos but volume of pictures in Flickr is more around nine times higher. The data is also available via an API which, given a landmark name and coordinates, provides images matching this information.

#### 4.1.2 Landmarks Selection

The annotation method described here relies upon the use of a reference corpus. New images of an object can be annotated only if the object is represented by a sufficient number of images in the

reference corpus. A landmark is selected only if it represents specific entity and it is represented by a sufficient number of images.

In the geographic domain, there is a direct relation between the degree of generality of a concept and the possibility to automatically annotate images representing it. This relation is explained by the visual coherence of image sets associated to concepts. For instance, the representation of *France* is very diverse and recognize automatically while those of the *Saint Sulpice Church* or of the *Centre Pompidou* are by far more coherent and easier to recognize. In other, the georeferenced images representing general concepts like *Paris* or *France* can be reliably annotated using tools like ZoneTag [2], based uniquely on the image coordinates that are matched against spatial footprints of the respective geographic areas. MonuAnno focuses on specific objects like *bridges*, *churches* or *squares* and we retain only the elements belonging to such parent classes from a geographical gazetteer. This restriction is possible since we dispose of a conceptually organized geographic database, which categorizes specific entities in more general classes [2].

It is necessary to know which of the objects selected in the first step are salient enough to be automatically annotated. In [18], we introduced a simple and efficient measure for ranking entities based on the information in a georeferenced image corpus, like Panoramio. The measure combines the total number of retrieved Panoramio images (a classical term frequency measure) and the number of different users (a community-based relevance assessment) having uploaded those images. We can reasonably suppose that, if a geographic object was photographed by several people, it is more representative than an object photographed by one person only, and should thus be ranked higher. The final list of objects annotated in MonuAnno contains only those objects having a rank superior to an empirical threshold, established at 50.

### 4.1.3 The Database

The final list of landmarks contains over 5000 items from different regions of the world and we decided to download a maximum of 500 images per object. This volume of photos should enable a diverse representation of the objects, including various points of view, different moments of the day and meteorological conditions. Given that Panoramio photos are generally more representative, we privilege this data source over Flickr, which is used only when there are not enough Panoramio photos representing an object. As underlined in [7], downloading of a large number of images from a photo sharing platform is a long process because the number of queries per day is limited. The constitution of the reference database for the current version of MonuAnno (5000 different landmarks) took around a month to complete but the download can be performed faster (for instance Flickr imposes a limit of 1 query per second, allowing for around 80000 images to be downloaded each day).

## 4.2 Reference Corpus Indexing

### 4.2.1 Bag of features descriptor and visual vocabulary

Our annotation problem is ultimately a classification one and a bag of features approach was shown to be efficient for object categorization tasks [15]. The SIFT descriptor [12] captures local properties (here, around Harris Laplace interest points) of represented objects and complements traditional approaches to image indexing, which provide a global characterization of the

content. We computed a visual vocabulary considering a selection of Flickr pictures that stand for different geographic categories. This way, we ensure that a high variety of local patches will be included in the visual vocabulary. We analyzed around 5000 images and extracted a maximum of 1000 Harris-Laplace keypoints per image described by SIFT descriptors. Then, we constituted a 5000 size codebook using the K-Means algorithm. To overcome the initialization dependency of the K-Means algorithm we computed several K-Means with random initializations and retained the best result defined as the partition with the optimal intra-clustered distance. We disposed of a cluster for processing the visual vocabulary and the use of a parallel implementation of the K-Means<sup>1</sup> we could afford to apply the K-Means on the whole dataset. We designate our codebook by  $W$  and note each of its visual word by  $w_i$ . One day was necessary for finding an optimal configuration of a visual vocabulary of 5000 codewords.

### 4.2.2 Image Indexing

An image is characterized according to  $W$  with a 5000-bins histogram  $h_i^b$ . We extract up to 1000 keypoints with the Harris-Laplace detector and match each point with its nearest  $w_i$  from  $W$ . Then, each bin of  $h_i^b$  can be seen as the frequency of the visual word  $w_i$  in the image. Inspired by text mining approaches, we compute the similarity between two images using the Cosine distance.

The indexing of a 600x400 pixels image takes 5 seconds in average and we decided to reduce the size of the images in order for their larger side to measure 250 pixels. The indexing time is reduced to less than 1.5 seconds on average. The reference database needs to be pre-indexed but the computational load does not represent a problem since we dispose of a 100 processors cluster. An index file is created for each landmark in the database in order to optimize the access to image indexes during the annotation phase. The index of a single image occupies 19 kB of memory space and the required memory space is of 19 GB for a million images, a size that is not problematic given the current storage capacities of personal computers.

## 5. IMAGE ANNOTATION

Automatic classification methods perform well when evaluating a reduced number of classes. Consequently, it is essential to reduce the complexity of the problem as much as possible. The reference corpus in MonuAnno contains 5000 different objects, each one represented by a large set of images, and we decided to run the experiments presented in this paper using a simple but flexible classification method, a k-NN, that scales up easily and does not implicate a training step. As we mentioned, there is no guarantee that a georeferenced photo depicts a neighboring landmark and we introduce a step which allows the detection of false positives elicited after the forced classification into a neighboring class by comparing the image to a supplementary class which contains only noisy images. With minor modifications, the annotation technique can make use of more complex classification methods, like SVMs. Note however that SVMs would be applied to a noisy dataset and this could deteriorate their performance. On the contrary, k-NN doesn't imply a learning phase and the influence

---

<sup>1</sup> <http://www.ece.northwestern.edu/~wkliao/Kmeans/index.html>

of noisy images should be smaller compared to the use of SVMs because only parts of the referenced database are used when annotating a particular photo.

In the case of georeferenced images, the use of spatial proximity between the image and candidate objects is a simple and intuitive manner to reduce the classification to a search space of manageable dimensions. Given an unlabeled georeferenced image, MonuAnno considers only objects within a range of a few hundred meters or a kilometer and the classifier will have to decide among one or two dozens of candidate classes (at most). The limitation of the search space using the spatial context of unlabeled georeferenced images is central to the scaling up of MonuAnno. An image taken in the *Central Paris* will only be compared to neighboring landmarks (*Centre Pompidou, Saint Eustache Church, Notre Dame de Paris, Sainte Chapelle, Hôtel de Ville* etc.) regardless of the number of landmarks in other regions in the world. The presence or absence of landmarks in *Moscow* or *Beijing* in the reference database does not affect the complexity of the classification because these landmarks are too far from *Central Paris* and will not be considered. Inversely, landmarks in *Paris* will not influence the annotation of unlabeled photos taken in *Moscow* or *Beijing*. The complexity of the annotation process will increase only when new objects in the area around an unlabeled image will be added to the reference database. This happens only if supplementary objects will be represented by a sufficient number of images in Flickr and Panoramio. However, the number of landmarks in a radius of the order of one kilometer around any point in the world will remain manageable. The annotation problem reaches its maximal complexity for regions containing a lot of landmarks, associated to known tourist destinations like *Paris* or *London*. In these cases, the current reference database used in MonuAnno contains around 50 different landmarks. For most of the regions where landmarks appear, the classification problem is much simpler and the number of candidate objects within a given distance from the photo coordinates is reduced to few objects.

The annotation is performed in two main steps: classification and checking, both using a k-NN approach. The algorithm first restricts the set of candidate landmarks, attributes the candidate image to one of the neighboring classes and then evaluates if this image really represents that object by also comparing it to the *Falses* class.

### 5.1 STEP1 - classification

The algorithm first selects only those landmarks found within a maximal distance (*maxDist*) from the georeferenced unlabeled photo. The complexity of the classification task is correlated with *maxDist*, a higher value of this parameter determining the examination of more candidate landmarks. The Evaluation section presents results for several *maxDist* and shows that the radius has an important impact on the results.

The picture to label is indexed using the same “bag of visual words” approach employed for the reference database and compared to the sets of images associated to neighboring landmarks. Similarly to images in the reference database, the size of unlabeled images is reduced to a maximum of 250 pixels per side and the indexing process takes around 1.5 seconds. The 5000 landmarks in MonuAnno are stored in a MySQL table and, given the coordinates of the unlabeled images, the selection of nearby landmarks takes less than 0.1 seconds. The algorithm computes

the cumulated sum of visual distances for the 5-NN pictures of each landmark and the image is temporarily annotated with the landmark name having the minimum cumulated distance. This step forces each tested picture into a neighboring class, annotating even the images that do not represent landmarks. The entire process takes 2 seconds in average but, for the moment, we focused rather on proving the concept than on the optimization of the indexing time.

The corpus contains a variable number of images for each class and we tested whether the classification results are improved if a balanced comparison is performed by retaining the same number of images for each neighboring landmark. This test implied the reduction of the number of pictures for each object considering the poorest represented landmark. The obtained results were worse compared to an unbalanced classification and this indicates that the number of images per object has an important influence on the classification process.

### 5.2 STEP2 – classification checking

The second step of the annotation procedure consists of a balanced classification implying the images of the landmark retained in the classification step and *Falses*, a pool of irrelevant images. First, the number of irrelevant pictures is adapted to the number of photos representing the landmark and the algorithm computes the visual distances between the test images and the elements of the two pools, which are temporarily stored in an array. This array is ranked so as to favor images that are visually similar to the unlabeled image and the algorithm counts the number of irrelevant images among the 10-NN. The number of neighbors (k=10) was empirically fixed at ten after testing both smaller and greater values.

The higher the number of neighboring irrelevant pictures (*maxFalse*) the more relaxed the filtering of potentially irrelevant images is. When *maxFalse* has a small value, the precision of the annotation is high but the percentage of annotated images is low. Inversely, a higher value of *maxFalse* determines a high percentage of annotated images accompanied by a precision loss. The influence of *maxFalse* on the annotation process is studied in the Evaluation section.

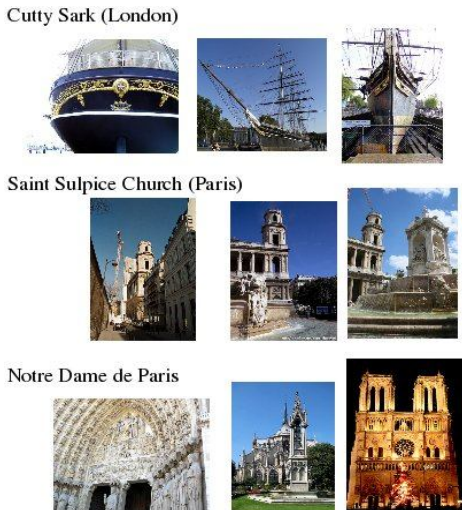
## 6. RESULTS AND EVALUATION

We evaluated MonuAnno on four urban areas with a high concentration of landmarks. The choice of *Paris* (48 landmarks), *London* (55 landmarks), *New York* (42 landmarks) and *San Francisco* (33 landmarks) insures diversity among the evaluated landmarks both in terms of inclusion into more generic geographic classes (*churches, palaces, museums, bridges, squares* etc.) and of visual appearance. As we mentioned, the only condition for a landmark to be included in MonuAnno is to be sufficiently well represented in Panoramio. Dense areas like the four chosen cities represent the most difficult case of annotation of georeferenced images because the classification must be performed on a high number of different landmarks. The 178 landmarks in the four cities are represented by over 30000 images in the reference dataset.

We constituted a test set containing 740 unlabeled Flickr pictures, 370 representing 80 objects in the four cities and 370 that are irrelevant for the retained landmarks but were included in the test dataset to reproduce situations when images taken near landmarks

depict something else. The test dataset contains a maximum of five relevant diversified images per landmark (see figure 3).

Relevant images are manually chosen while irrelevant ones are automatically selected to have significantly different coordinates in a georeferenced corpus. The tested objects belong to different geographic categories (*museums, bridges, squares* etc.), are represented by a variable number of element in the reference dataset and, as shown in figure 3, the images represent the object in different conditions and from different points of view. Ideally, MonuAnno should annotate all relevant images and filter out all irrelevant ones.



3. Positive examples in the test dataset.

We evaluate the influence of the radius around the object used for selecting neighboring landmarks, correlated to that of the number of “false” images accepted among the nearest neighbors of the image. Then we provide a detailed errors analysis, including: the contribution of the first and the second step in the algorithm, the influence of the number of images per landmark and the distribution of errors in general classes such as *monument, museum, bridge, square* etc.

### 6.1 Influence of parameters

We studied the distribution of georeferenced images around the coordinates and the results show that 50% of the images are taken within 250 meters, 72% within 500m, 83% within 1km and 90% within 2km. We attributed these four values to *maxDist* in order to assess the influence of the radius on the classification results.

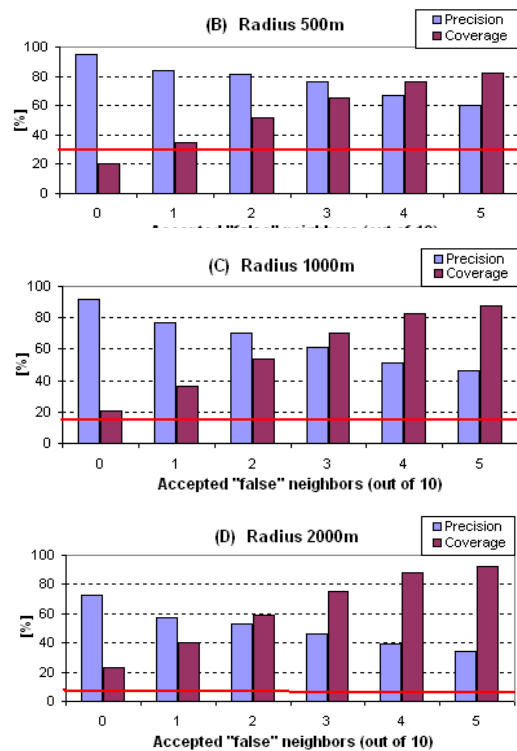
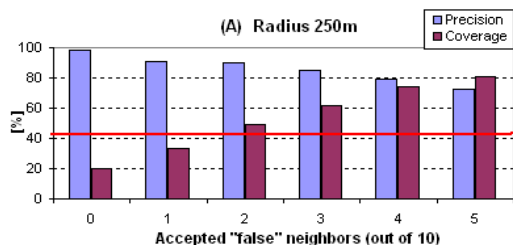


Figure 4. Annotation results for different values of the radius (*maxDist*) and of the number of “false” images accepted among the 10-NN (*maxFalse*). The red line indicates the random probability to obtain a good annotation.

The second important parameter influencing the annotation is *maxFalse*, the threshold used to eliminate false positives. For each unlabeled image, we use *Falses* images class and count the number of elements of this class appearing among the 10 visually closest images. We tested the approach on *maxFalse* values equal to 0, 1, 2, 3, 4 and 5. In figures 4 (A) to (D), we present annotation results for radiuses of 250, 500, 1000 and 2000 meters. For each distance, we plot results for different values of *maxFalse*. We represent the correlated values of precision and coverage (number of annotated images among the positive examples in the test dataset).

The radius around the coordinates of a candidate photo determines the number of candidate landmarks to test and the random probability to obtain a good annotation for an image (red line in figures 4A to 4D): 42.3% (250m); 30.4% (500m); 16.3% (1000m) and 8.2% (2000m). For all tested *maxDist* and *maxFalse* the precision of the annotation is significantly higher than the random probability. As expected, the precision and coverage score are inversely correlated and their values are influenced by both parameters in our algorithm. For a fixed radius, the precision decreases with the increase of the acceptable maximum number of neighbors from the *Falses* class.

The best precision values are obtained for the 250m (figure 4A), which corresponds to the lowest classification complexity and the worst precision for 2000m (figure 4D), corresponding to the highest classification complexity. For 500m, the annotation precision varies between 89.8% (coverage of 20.1%, *maxFalse* =

0) and 60.6% (coverage of 82.3%,  $maxFalse = 5$ ). The coverage increases for higher values of  $maxFalse$  but is accompanied by a precision loss.

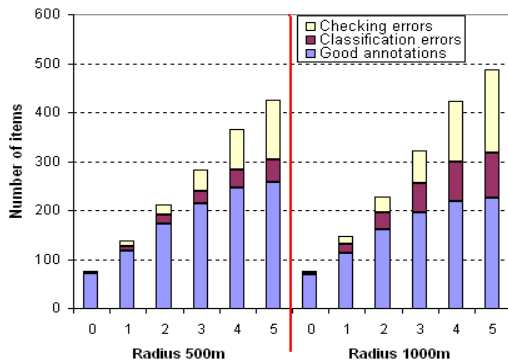
In Flickr, a manually annotated corpus, the precision of the annotation of georeferenced pictures of landmarks tops at 90% [1]. The precision of the automatic annotation approaches this value when  $maxFalse = 2$ , corresponding to an annotation of a half of the positive test images (89.3% at 250m, 81.1% at 500m). The scores for the automatic annotation are obtained in difficult cases (i.e. regions of the world with a high concentration of landmarks) and are likely to be higher for regions that contain fewer landmarks, where the classification step is performed on a small number of candidate objects.

The results presented here are very encouraging. Even if we consider a higher number of landmarks in the reference dataset, the complexity of the classification process remains manageable. MonuAnno can be easily tuned to maximize precision or coverage by modifying the values of  $maxDist$  and of  $maxFalse$ , corresponding to different application needs.

## 6.2 Errors analysis

We provide a detailed analysis of the errors in MonuAnno, which is useful for understanding the advantages and the limits of the approach.

### 6.2.1 Classification vs. checking errors



**Figure 5. Comparison of the number of errors generated by the classification (STEP 1 in the annotation algorithm) and by the checking (STEP2). The number of errors is plotted against the total number of annotated images.**

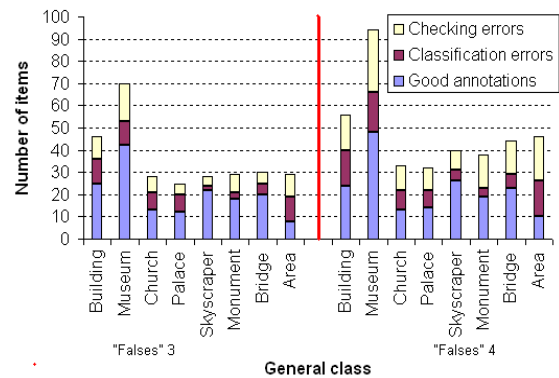
The errors in MonuAnno are caused either by the classification step (STEP 1 of the annotation algorithm) or by the verification step (STEP 2). In figure 5, we plot the number of errors and the number of well annotated images at 500m and 1000m, with  $maxFalse$  varying from 0 to 5.

The evolution of the number of errors with the variation of  $maxFalse$  (figure 5) is similar for 500m and 1000m. For small values of  $maxFalse$  (0, 1), the classification errors are dominant whereas for higher values of the same parameter (3, 4, 5), the checking errors become predominant. This situation appears because the sets of images representing a monument in the reference corpus tend to be coherent and the relaxation of the acceptability criterion for a test image allows for more false positives in the test dataset to be annotated.

### 6.2.2 Influence of the number of images per landmark

The reference corpus contains a variable number of images per landmark (between 50 and 500) and our annotation method is based on comparing a fixed number of neighbors to the image we want to label. We looked at the relation between the number of images per landmark and the number of errors and noted that the number of images per object has a negative influence on results for landmarks represented by less than 100 images in the reference database. The observed results indicate that 100 images are enough to ensure a diverse representation of a landmark and to automatically annotate it when fixing the number of neighbors in the classification step of the algorithm at 5.

### 6.2.3 Distribution of errors in semantic classes



**Figure 6. Distribution of errors in semantic classes. The errors are plotted for a radius of 1 km, when accepting respectively 3 and 4 "false" neighbors among the most similar images. The number of errors is plotted against the total number of annotated images.**

The landmark type has an important influence on the diversity of representative images. For instance, it is likely to have interior images labeled with *museum* or *arts venues* names whereas the same is not true for *statues* or *bridges* and, consequently, the annotation will be more complicated for *museums* or *venues*. We attributed each landmark in the reference corpus to one of the general classes in figure 6. The types of buildings that are well represented are presented in separated classes whereas the others are grouped in the *building* class. We also separate monuments (*statues*, *arches*); *bridges* and *areas* (*squares*, *parks*, *places* etc.). The distribution of errors is presented for a radius of 1000m and for 3 and 4 neighbors in order to have a significant number of errors (figure 6). The errors generated in the classification and checking (STEP 1 and STEP 2 of the annotation algorithm) are presented separately.

The results in figure 6 show that visually complex classes (*museums*, *church*, *areas*) generate more errors than simpler classes (*skyscrapers*, *bridges*, *monuments*). The success rate when annotating landmarks with having a significant surface (*parks*, *squares*) is low and we should consider using spatial reasoning to label image content with these types of geographic objects. We should also consider finer grained distinctions when annotating complex objects (for instance, separate interior and exterior images for complex landmarks having a significant number of interior photos).

## 7. CONCLUSION AND FUTURE WORK

We introduced a method for automatically annotating georeferenced landmark images. The main contribution of this paper is a scalable approach for labeling landmarks images, based on a two step k-NN. The algorithm handles a real-world situation, namely that when a picture represents or not a neighboring landmarks. Our annotation method is applied on an automatically constituted photographic corpus and we show that it is robust to noise. The presented evaluations show that the labeling precision reaches values similar to those of manual annotations when annotating half of the positive examples in the test dataset. We evaluate the influence of the search radius around unlabeled image coordinates, of the number of irrelevant images admitted among the nearest neighbors and the typology of errors (classification vs. checking; type of represented semantic classes). The evaluation was performed on four major cities, with a high concentration of landmarks, and the presented results are better for regions that contain fewer landmarks. Note that, since the search radius (*maxDist*) limits the search space around a new image, the complexity of the annotation would remain the same if more regions (or even the entire world) would be considered in the evaluation.

In the future, we plan to compare the performances of the k-NN based classification to a SVM approach, similar to that in [4]. We also plan to add loose geometric constraints to the content-based similarity process ([16], [20]). Finally, we plan on testing an annotation procedure that integrates the spatial distribution of individual photos. It would be interesting to limit the search space using annotated photos that are taken in the immediate proximity of the image to annotate. This limitation would focus the classification process on a comparison of the new image to annotated picture representing similar points of view on landmarks and is likely to further decrease the complexity of the classification task.

## 8. REFERENCES

- [1] Ahem, S., Naaman, M., Nair, R. and Yang, J. 2007. World Explorer: Visualizing Aggregate Data from Unstructured Text in georeferenced Collections. In *Proc. of JCDL 2007* (Vancouver, Canada, June 2007).
- [2] Ames, M., Naaman, M. Why We Tag: Motivation for Annotation in Mobile and Online Media. In *Proc of SIGCHI 2007* (San Jose, CA, USA, 2007).
- [3] Cao, L., Luo, J., Huang, T. S. Annotating Photo Collections by Label Propagation According to Multiple Similarity Cues. In *Proc. of ACM MM 2008* (Vancouver, Canada, 2008).
- [4] Chevallet, J.-P., Lim, J.-H., Leong, M.-K. Object Identification and Retrieval from Efficient Image Matching. Snap2Tell with STOIC dataset. In *Proc. of AIRS* (Jeju Island, Korea, 2005).
- [5] Datta, R., Joshi, D., Li, J., Wang, J. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys*. 40, 2008.
- [6] Geonames - <http://geonames.org>
- [7] Hays, J., Efros, A. IM2GPS: estimating geographic information from a single image. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2008.
- [8] Hill, L. L., Frew, J. and Zheng, Q. Geographic names – the implementation of a gazetteer in a georeferenced digital library. *CNRI D-Lib Magazine* (January, 1999).
- [9] Hörster, E., Leinhart, R., Slaney, M. Image Retrieval on Large-Scale Image Databases. In *Proc of ACM CIVR* (Amsterdam, The Netherlands, 2007).
- [10] Kennedy, L., Naaman, M. Generating diverse and representative image search results for landmarks. In *Proc. of WWW 2008* (April 2008, Beijing, China).
- [11] Lim, J.-H., Li, Y., You, Y., Chevallet, J.-P. Scene Recognition with Camera Phones for Tourist Information Access. In *Proc. of IEEE ICME* (Beijing, China, 2007).
- [12] Lowe D., "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, 60, 2 (2004), pp. 91-110.
- [13] O'Hare N., Gurrin, C., Smeaton A. F., Jones G. F. G. 2005. Combination of content analysis and context features for digital photograph retrieval. In *Proc. of EWIMT 2005*.
- [14] Panoramio – <http://panoramio.com>
- [15] Pascal Visual Object Classes <http://pascalvin.ecs.soton.ac.uk/challenges/VOC/>
- [16] Philbin, J. , Chum, O. , Isard, M. , Sivic, J. and Zisserman, A. Object retrieval with large vocabularies and fast spatial matching, *Proc. of the ICCV 2007*.
- [17] Popescu, A., Moëllic P.-A., Kanellos, I. ThemExplorer: Finding and Browsing geo-referenced Images. In *Proc. of Content Based Multimedia Indexing Workshop* (London, UK, June 2008).
- [18] Popescu, A., Grefenstette, G., Moëllic P.-A. Gazetiki: Automatic Creation of a Geographical Gazetteer. In *Proc. of ACM/IEEE JCDL* (Pittsburgh, PA, USA, June 2008).
- [19] Quack, T, Leibe, B., van Gool, L. World-Scale Mining of Objects and Events from Community Photo Collections, In *Proc. of ACM CIVR'08*.
- [20] Sivic, J, Zisserman, A. Video Google : Efficient Visual Search of Videos, Towards Category-Level Object Recognition, LNCS, pp 127-144, Springer, 2006.