# EXTENT: INFERRING IMAGE METADATA FROM CONTEXT AND CONTENT

*Chang-Ming Tsai[1], Arun Qamra[1], Edward Y. Chang[2], and Yuan-Fang Wang[1]*

Department of Electrical & Computer Engineering[2] and Computer Science[1], UC Santa Barbara

**ABSTRACT**

*We present EXTENT, an image annotation system that combines the context and content information to annotate images with metadata that cannot be reliably inferred from either the context or the content alone. EXTENT first applies contextual information for restricting the search scope in an image database and reducing the complexity of ensuing content analysis. It can then afford to use more expensive (hence more robust) algorithms for performing content analysis within the restricted database scope. Our experiments show that effectively combining content with context information can infer metadata with high accuracy.*

## 1. INTRODUCTION

Recognizing objects (landmarks or people) in a photo remains to be a very challenging computer-vision research problem. However, with available contextual information, such as time, location, and a person's social network, recognizing objects among a much limited set of candidates is not as daunting a task. For example, given the location information that a photograph was taken at downtown Santa Barbara, the landmarks observed in the photo are likely to be certain famous architectures in the area. Given the fact that a photo depicts a birthday party of a person, the attendees are most likely to be his/her friends and family. With candidate objects being limited, not only does matching become easier, but also the matching time becomes shorter. Hence, we can afford to employ more elaborate matching techniques to improve accuracy.

Recently, technology advance in mobile phones and digital cameras has created rapid growth in on-line picture sharing. Most digital cameras today assign a photo a meaningless number as its filename. To help users better organize their photos, it would be desirable to provide useful metadata such as time (*when*), people (*who*), location (*where*), landmarks (*what*), and event (inferred based on when, who, where, and what). Providing the *when* and *where* information is relatively straightforward, as all cameras provide time information, and most picture phones can infer (rough) location from GPS or CellID information. However, providing the *what* and *who* metadata must rely on content analysis. Therefore, we had proposed EXTENT [1]: a combined cont**ext** and cont**ent** analysis system for annotating images with metadata.

In this work, we focus on landmark recognition (one important aspect of the *what* metadata). Given the (rough) location where a photo was taken, EXTENT generates a list of landmarks that are likely to be in the photo. EXTENT then extracts *scale-, illumination-,* and *viewpoint-invariant* features from the photo to match with those of the database images of the landmarks. The major research challenge lies in that the matching must be insensitive to variation in time of the day and viewing direction. In other words, a landmark taken under different lighting conditions and from different viewing angles should be correctly recognized as the same.

Some recent works [2][8][9][10] enable digital images to be annotated with the metadata of spatial context. Nevertheless, these methods suffer from some drawbacks. First, in [2][8][10], GPS devices are used to infer the location where a photo is taken. However, the objects in a photo may be far away from the camera [11], not at the GPS coordinate. For instance, one can take a picture of the *Bay Bridge* from many different locations in the cities of Berkeley, *Oakland*, and *San Francisco*. The GPS information often cannot definitively identify landmarks in a photo. Furthermore, a landmark in a photo may be occluded and can hence be deemed not important. For instance, a ship may occlude a bridge; a person may occlude a building, etc. Also, a person can be inside or outside of a landmark. Therefore, a robust image annotation system must perform content analysis to precisely name the landmarks.

The remainder of this paper is organized as follows. In section 2, related work on image context inference is reviewed. In section 3, we propose EXTENT. In section 4, selected experimental results are reported. In section 5, we provide a concluding remark.

## 2. RELATED WORK

We have discussed several context–based image annotation systems in Section 1 to motivate this work.

*0-7803-9332-5/05/$20.00 ©2005 IEEE*

Here, we briefly review some representative work in content-based object recognition.

To compare and match objects in images, template matching is a widely adopted technique. Nevertheless, template matching cannot deal with occlusion or scaling very well. It is also very sensitive to changes in illumination. To remedy these problems, Lowe [6][7] proposed one of the most robust local descriptors: Scale Invariant Feature Transform (or SIFT), which extracts distinctive and invariant local features. SIFT [6] consists of four processing stages: 1. scale-space extrema detection, 2. keypoint localization, 3. orientation assignment, and 4. keypoint description. (Details are presented in Section 3.2.) Features generated by SIFT have been shown to be robust in matching in the presence of affine distortion, change of 3D viewpoint, addition of noise, and variation in illumination [6].

Ke and Sukthankar recently proposed a modified version of SIFT, called the PCA-SIFT [5]. PCA-SIFT uses the Principle Components Analysis (PCA) to normalize the gradient patch based on the output of SIFT. The result is shown to be robust to image deformations [4].

Despite the success of SIFT and PCA-SIFT, these techniques can break down when the number of candidate objects to be matched is large. EXTENT successfully brings the candidate object set down to a small, manageable subset, and thereby substantially improves the matching accuracy and efficiency.

## 3. THE EXTENT SYSTEM

While the full architecture of EXTENT is detailed in [1], this section presents the part for landmark annotating. We assume that photos were taken with temporal (*when*) and spatial (*where*) information. The EXTENT system uses these contexts to first choose possible landmark candidates, and then uses the content of the photos to create invariant features for matching against a database of landmarks.

Two important issues of landmark recognition in the EXTENT system are lighting and viewing-angle variation. Since the photos of a landmark can be taken in many different times and from many different perspectives, EXTENT uses SIFT's image features, which are insensitive to lighting variations and changes in viewing directions. Moreover, to incorporate the context knowledge into the recognition process, EXTENT applies an intelligent coarse-to-fine search: First context information is analyzed for reducing the search range in the landmark database; then the feature extraction and matching algorithms are applied to locate suitable matches, and finally the context information is referenced again to filter those ambiguous matches. The procedure of EXTENT is discussed below.

### 3.1. Generating the list of candidate landmarks

The candidate landmarks are selected from the database of landmarks in two ways. First, they are chosen based on the spatial metadata, such as GPS or CellID information. For example, if the CellID indicates that a photo was taken on the Stanford campus, the candidate landmarks, if one exists in the photo, can be the *Hoover Tower*, the *Memorial Church*, and several other landmarks on the campus. The second method is to review the previously annotated photos. If some previous photos were taken at about the same time as the current photo, the landmark information of the previous photos is also used to choose the candidate landmarks. For example, if the landmark in photos taken a couple minutes ago was identified as the *Gates* building, we add the landmarks near the *Gates* building to the candidate list.

The second method can be treated as a complementary method for the first one. That is, if at some time the CellID information is not present or ambiguous, we can still infer the spatial information by using previous spatial and temporal information. (We do not use the second method in our experiments.)

### 3.2 Constructing robust features

In order to compensate for the lighting variation, view-angle change and image noise, we apply a robust feature extraction methods, the SIFT algorithm [6], for constructing invariant features from photos. The SIFT features are created by the following four steps:

1. Potentially interesting image features, called keypoints in the SIFT algorithm, are identified in the scale-space using the difference-of-Gaussian images. The detected feature points are invariant to scale and orientation, because the search area covers all image scales, and the difference-of-Gaussian images produce stable image features against image rotation.
2. The detailed location of each keypoint is calculated by fitting a 3D quadratic model to the neighboring regions of the keypoints. For extracting stable features, some tests are performed to eliminate those points that lie on an edge with poor localization.
3. To achieve orientation independent results, the principal gradient direction is used to rotate and align

the dominant direction of each keypoint. An orientation histogram is then computed using image gradient information in the neighboring regions of the keypoints.

4. Finally, a local image descriptor is formed, by collecting the normalized gradient information around the keypoints. The keypoint descriptors are designed to avoid boundary effects, and they are composed of 128-element feature vectors.

The SIFT algorithm extracts dense features that are distinctive and invariant, and hence, they are ideal for landmark detection and matching in a database [6]. Furthermore, the algorithm can be applied efficiently by processing a landmark database off line.

### 3.3. Matching features in a landmark database

For each detected keypoint in a query photo, we perform a nearest-neighbor search to locate similar features in a landmark database. Since there is no known algorithm that can identify the exact neighbors of points efficiently in a high dimensional feature space, we follow the approximate method in [6], the best-bin-first algorithm, to match the keypoints in the database. Then we cluster the keypoints from the same photos in the database to determine the possible landmarks in the query photo.

### 3.4. Solving the ambiguity of matching

The previous step may produce more than one candidate landmarks. We can make the final decision by utilizing the context information again. In step 1, we use the context information to choose some candidate places. In this step, we just need one solution. Therefore, we choose the landmark that is closest to the CellID area, GPS position, or the previous recognized landmarks.

We have discussed the four main steps of the EXTENT system. The first and last steps use the contexts of photos, including spatial and temporal contexts, for narrowing down the searching ranges. The second and third steps use the photo content to perform a robust database search. Combing context and content information, EXTENT is a much more robust system compared to those use only context or content.

## 4. EXPERIMENTAL RESULTS

### 4.1. Test-bed

We have used two datasets to evaluate EXTENT. The first one, referred henceforth as the Towers dataset, was created by collecting 1,000 architecture images from various websites on the Internet. This set contained 50 tower images: 5 images each for 10 different landmark towers (such as the *Eiffel Tower*, *Tower of Pisa* etc). The rest were other landmarks from all over the world. These images were taken at different times, by different people, and under varying lighting conditions and viewing angles. They were taken at various resolutions and aspect ratios (ranging from 180×317 to 1384×1752). All of them were converted to the JPEG format before performing feature extraction. We used this dataset to evaluate the effectiveness of SIFT feature extraction for landmark detection.

The second dataset (referred hereafter as the *Stanford* dataset) was obtained from Mor Naaman [8]. The dataset was constructed by collecting photographs taken by visitors to the Stanford Visitor Center. All photographs were taken in the Stanford campus and were annotated with GPS information. From this dataset, we used a subset containing about 1,000 images. To evaluate the EXTENT system, we selected photographs with GPS coordinates around the *Memorial Church* and *Hoover Tower* (two important landmarks on the Stanford campus). All images were rescaled to 320×240, before performing SIFT feature extraction. Sample images from this dataset are shown in Figure 1.

Also, we kept a separate set of sample images for each landmark. These images were used as queries to determine the presence of a landmark.

**Figure 1.** Stanford Dataset Samples

**Table 1**. Matching Accuracy for the Towers dataset

| Tower | **A** | **B** | **C** | **D** | **E** | **F** |
|---|---|---|---|---|---|---|
| Accuracy (%) | 80 | 80 | 40 | 100 | 60 | 100 |
| Tower | **G** | **H** | **I** | **J** | **All** | |
| Accuracy (%) | 100 | 60 | 100 | 60 | 78 | |

### 4.2. Results

We used the *Towers* dataset to evaluate the effectiveness of the SIFT feature extraction for recognizing landmarks. This experiment was performed without using any contextual information. We used the 50 tower images in this experiments and performed leave-one-out cross validation. So each image was queried against the

remaining 49 images and a label was assigned to the query image based on the best match found in the dataset of the remaining 49, and compared against the ground truth. The average annotation accuracy achieved for each of the towers is listed in Table 1. The average annotation accuracy is 78%. Notice that we only have four sample images for every tower to compare with. We believe that with more representations (more photos of a landmark taken from different angles and lighting conditions), the result can be improved. The results show that SIFT is a promising feature-extraction method. SIFT's major drawback is its high comparison cost. In our second experiment, we used contextual information to improve both search efficiency and accuracy.

The second set of experiments used the *Stanford* dataset. First, we used spatial context information to narrow the pool of candidate landmarks. (Again, GPS coordinates of each image can tell us what landmarks are in the vicinity of the imaging location.) Each image in the dataset was individually processed and matched with sample images that contain candidate landmarks. We used a distance threshold to separate likely matches from non-matches. If the distance (computed using the SIFT features) between the query image and a landmark sample was within the threshold, the landmark would be a possible match. If no possible match was found (after comparing with all landmark samples), we concluded that the image contained no landmark. Otherwise, the best match was used to annotate the image. The value of distance threshold was computed offline using the sample images and a set of non-sample images (images not containing the tower and the church). This set is referred to as the training set. Evaluations were done with various thresholds to determine the optimal value, which was found to be 10.7.

Overall annotation accuracy was computed by averaging the annotation accuracies for images that contain the *Hoover Tower*, those that contain the *Memorial Church*, and those that contain neither. Overall annotation accuracy is plotted for varying numbers of sample/base images (denoted as $k$) for each landmark. Figure 2 shows that without contextual information, SIFT features can achieve 90% annotation accuracy for all $k$ values. With contextual information, the annotation accuracy improves to 93%. Due to the space limitation, we do not plot of the result of the *Memory Church* query. Its accuracy without and with contextual information is 73% and 92%, respectively. The *Memory Church* is more challenging to be identified because many buildings look similar to it. Once the number of candidates can be narrowed down to a handful based on the spatial information, the accuracy rises to a satisfactory level.

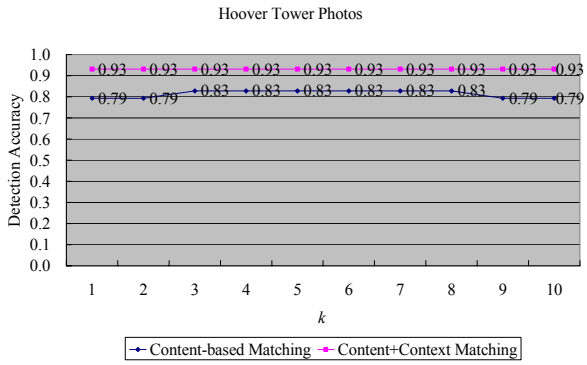


**Figure 2**. Hoover Tower Annotation Accuracy

## 5. CONCLUSIONS

In this paper, we presented the landmark annotation of the EXTENT system. Context information was first used to narrow down the search to a small pool of candidate landmarks. Next SIFT features were used to analyze content and locate landmarks. We demonstrated experimentally on two different data sets that the system achieved high annotation accuracy.

## 6. REFERENCES


[1] Chang, E. "Extent: Fusing Context, Content, and Semantic Ontology for Photo Annotation, US Patent, March 2005 (submitted).

[2] Diomidis, D. S., "Position-annotated photographs: a geotemporal web," in *IEEE Pervasive Computing*, Vol. 2(2), pp 72-79, 2003.

[3] Dey, A. K., "Understanding and Using Context," in *Personal and Ubiquitous Computing Journal*, Vol. 5, Issue 1, pp 4-7, 2001.

[4] Ke, Y., Sukthankar, R., Huston, L., "Efficient Near-duplicate Detection and Sub-image Retrieval," in *Proc. of 12th ACM International Conference on Multimedia*, pp 869-876, 2004.

[5] Ke,Y. and Sukthankar, R., "PCA-SIFT: A more distinctive representation for local image descriptors," in *Proceedings of IEEE Computer Vision and Pattern Recognition*, 2004.

[6] Lowe, D. G. , "Distinctive image features from scale-invariant keypoints," in *International Journal of Computer Vision*, 2004.

[7] Lowe, D. G., "Object recognition from local scale-invariant features," in *Proceedings of International Conference on Computer Vision*, pp 1150–1157, 1999.

[8] Naaman, M., Paepcke, A., and Garcia-Molina, H., "From Where to What: Metadata Sharing for Digital Photographs with Geographic Coordinates," in *Proc. of 10th International Conference on Cooperative Information Systems (CoopIS 2003)*, pp 196-217, 2003.

[9] Sarvas, R., Herrarte, E., Wilhelm, A., and Davis, M., "Metadata Creation System for Mobile Images," in *Proceedings of the Second International Conference on Mobile Systems, Applications, and Services (MobiSys2004)*, pp 36-48, 2004.

[10] Toyama, K., Logan, R., and Roseway, A., "Geographic Location Tags on Digital Images," in *Proc. of 11th Annual ACM International Conference on Multimedia,* pp 156-166, 2003.

[11] Davis, M., King, S., Good, N., and Sarvas, R., "From Context to Content: Leveraging Context to Infer Media Metadata," in *Proc. of 12th Annual ACM International Conference on Multimedia*, 2004.