Object Fingerprints for Content Analysis with Applications to Street Landmark Localization

Wen Wu and Jie Yang
School of Computer Science, Carnegie Mellon University
5000 Forbes Ave, Pittsburgh, PA 15213, U.S.
{wenwu,jie.yang}@cs.cmu.edu

ABSTRACT

An object can be a basic unit for multimedia content analysis. Besides similarity among common objects, each object has its own *unique* characteristics which we cannot find in other surrounding objects in multimedia data. We call such unique characteristics *object fingerprints*. In this paper, we propose a novel approach to *extract* and *match* object fingerprints for multimedia content analysis. In particular, we focus on the problem of street landmark localization from images. Instead of modeling and matching a street landmark as a whole, our proposed approach extracts the landmark's object fingerprints in a given image and match to a new image or video in order to localize the landmark. We formulate matching the landmark's object fingerprints as a classification problem solved by a cascade of 1NN classifiers. We develop a street landmark localization system that combines salient region detection, segmentation, and object fingerprint extraction techniques for the purpose. To evaluate, we have compiled a novel dataset which consists of 15 U.S. street landmarks' images and videos. Our experiments on this dataset show superior performance to state-of-the-art recognition algorithms [20, 33]. The proposed approach can also be well generalized to other objects of interest and content analysis tasks. We demonstrate the feasibility through the application of our approach to refine web image search results and obtained encouraging results.

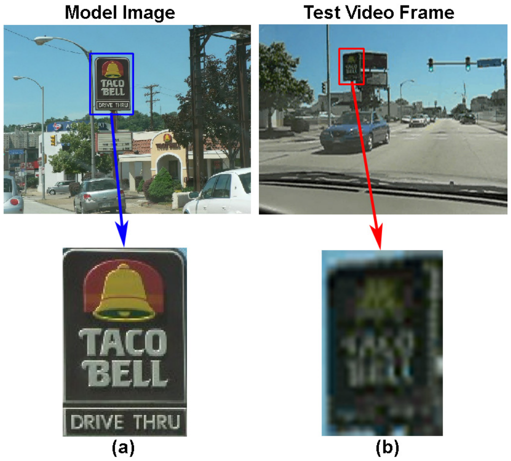

Figure 1: An example from our street landmark dataset. Model image resolution is 1280×960, while test video frame's is 320×240. Two images are rescaled to be equal size. In this example, it is even hard for human eyes to recognize the Taco Bell sign from clutter in (b) with knowledge of (a). But our approach is able to localize the sign in video with object fingerprints which this paper proposes.

1. INTRODUCTION

The goal of multimedia content analysis is to extract semantic meanings from the multimedia data such as a video sequence with an accompanying audio track. Unlike a text document, we do not have "key words" in understanding a multimedia document. In order to analyze video data, we need some explicit structures. An object is something material that may be perceived by the senses and it can be a basic unit at the lowest level of such a structure. In video analysis, many algorithms can be largely categorized in *detectors*, such as face detector [34], text detector [32], and person detector [16], etc. Simple detectors can form a more complex detector, e.g., a face detector and a shot classifier can be combined as an anchor detector. Therefore object detection plays an important role in multimedia content analysis.

Object detection determines whether or not the object is present, and, if present, determines the location and scale of each object in an given image. The challenge of this task is largely because of the infinite combinations of scale, orientation, viewpoint, lighting, and many other factors. To model such variations, a machine learning algorithm usually requires extracting a number of features at every pixel based on the statistics of the surrounding pixel values. For example, we could train a good face detector using a large number of data [34]. For some content analysis applications, however, we might not have enough training data. For example, we frequently must deal with only one query image. In this research, we are interested in object recognition and localization with very limited training data (e.g., one image per object for training). An application of such a problem is street landmark localization which is important to location-aware multimedia applications and in-car navigation tasks.

In daily life, a street landmark means anything that is easily recognizable, such as a monument, a building, a store sign, or other structure [2]. In this research, we refer to street landmarks as objects that can be used for giving directions for driving navigation. In an urban environment, commonly used street landmarks include gas stations, fast food restaurants, and churches, etc. The research is a part of our efforts in developing landmark-based navigation technologies for drivers. Current GPS based navigation systems provide turn-by-turn instructions, e.g., turn left in 50 feet. However, human drivers often use landmarks for helping navigation. For example, we tell people to turn left after a BP gas station and then make a right at Starbucks. The application scenario is as follows: John is going to visit his sister Linda in another city, to which John is new. To help John, Linda sends him some street landmark images of her city on his route to her place. John will drive a car with a video camera that can capture the scene in front of his car. We would like to build a system that could help John to automatically recognize these landmarks from the video sequence. We assume that only one image per landmark is available for training in this task. In addition, the training images and the test video sequences always have different quality and resolution. An example from such a scenario is shown in Fig.1.

Although street landmark localization from video is an object recognition problem, many object recognition algorithms [23, 34, 27] may not work well for this application because of the special characteristics of the problem. A fundamental challenge is lack of training data. Other challenges include rigid object recognition with different viewpoints, scale and illumination changes, partial occlusion, and mismatching resolutions between the model and testing images, which also prohibit us from directly applying Content-Based Image Retrieval algorithms [4, 18] and sub-image retrieval methods [13] to this task. In order to address these challenges, we propose a unified approach for learning a model based on a given landmark image such that the system can recognize the landmark from a new street scene image. Our main insight is to identify a set of features that are unique to the landmark which we call fingerprints, and consider finding any fingerprints in a new image as recognition of the landmark.

We believe both appearance and context information of the landmark are crucial for landmark recognition. However we hypothesize that we do not need all the information to model the entire landmark in order to recognize it. Rather, we believe that the recognition of the landmark's fingerprints such as the star shape in the center of BP sign (Fig.2(a)), suffices to provide a robust recognition of the landmark. The contributions of this paper are:

- We propose a new landmark recognition approach by mining and matching landmarks fingerprints. Appearance and bag of features are surprisingly useful in general object recognition [29]. We combine them to extract unique fingerprints of a given landmark. Given a new image containing the landmark, our approach not only recognizes the landmark, but also estimates its scale and pose.
- Sliding window scanning is commonly used in object recognition, however, more researchers have advocated the use of image segmentation to get better spatial support for recognition. Inspired by [21], we use multiple segmentations to extract landmark patches.
- We conduct extensive experiments to test our approach. We compile a new and challenging street landmark dataset of 15 U.S. street landmarks. Our approach demonstrates promising results on this dataset.
- We demonstrate the proposed approach can be generalized to other content analysis tasks through the application of our object fingerprint-based approach to refine web image search results and achieve improved ranking outcomes.

2. RELATED WORK

To achieve robust object recognition from images and videos, several approaches have been proposed [20, 7, 22] to exploit object appearance and geometrical information. Meantime, object detection has made tremendous progress over past years in various directions: real-time deformable object detection [24] and application of geometric scene context for detection [9]. Location recognition has also been recently studied using local feature descriptors in the context of mobile phone cameras [19] and in the context of a moving camera on a car [27]. Most of these pioneer works, however, require modeling of the *whole* object, and recognition is performed by matching the learned object model to a new image. In contrast, our approach represents an object by a set of object fingerprints and does matching of fingerprints instead of the *whole* object.

On the other hand, recognizing a street landmark from a moving camera is a difficult task because of combined effects of egomotion, blur, constantly changing illuminations and restricted image quality. In addition, only one model image is provided which invalidates many successful object recognition techniques, such as algorithms requiring a certain amount of training data. While face recognition from a single image has been extensively studied [34] and detection of pedestrians and vehicles has been successfully demonstrated [17], little attention has been focused on recognition of street landmarks, a large class of street objects.

Road signs are an important and special class of street landmarks and have attracted much attention in various fields in the past decade [23, 5]. Road signs provide drivers with important and necessary information about the road and navigation. They are designed to be easily recognizable by human eyes through high-contrast patterns and colors. Recent research has demonstrated promising road sign detection results [32]. However, to recognize and localize general street landmarks, road sign recognition methods may not work well because general street landmarks have very diverse patterns and usually do not contain text.

Feature phrases can effectively improve image indexing, retrieval, and recognition [35, 36]. A bag of SIFT features for an image has been successfully used in a text retrieval approach for object matching in videos [29]. Other directions have been also pursued for object search in videos [11, 16]. Some recent research has utilized local features and SIFT descriptors to match place-of-interest images or logos [14, 26], among which the EXTENT system was proposed for automated photograph annotation [25] and focused on historic buildings, a different problem from ours. Very recently, in the data mining community, researchers have started investigating frequent spatial patterns in images [15] for clustering images, but not for recognition.

One close effort to ours is Ferencz, et al. [8] which has only one image for training but ample data to train a category classifier. While the two approaches agree in general princi-

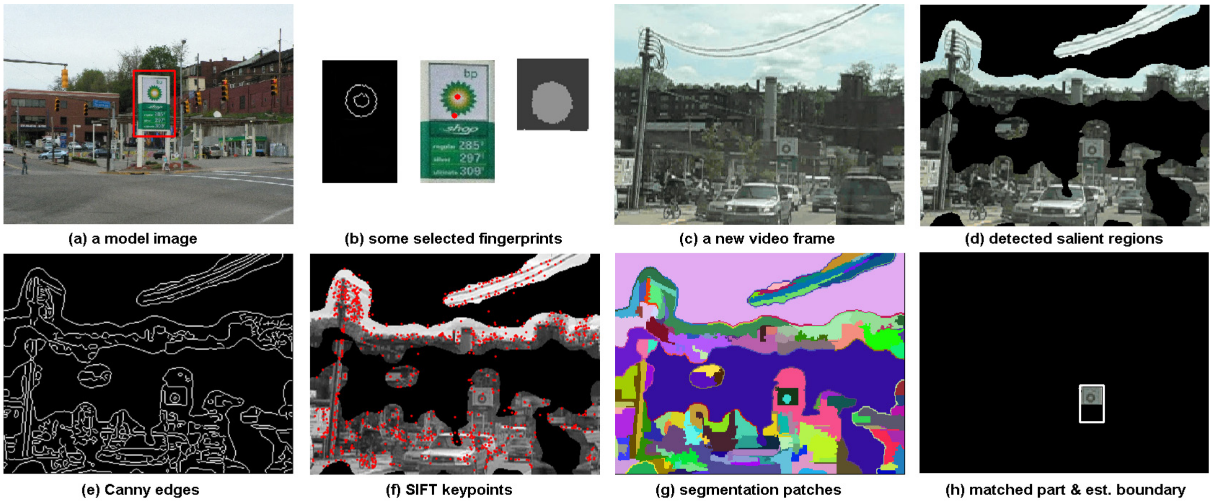

Figure 2: Given a model image of the object that is marked in a red box (a), our method first extracts salient features to select object fingerprints. Here three bigram fingerprints of the BP sign are shown in (b). For a new video frame (c), our method first detects salient regions [9] (d) and extracts SIFT keypoints [17] Canny edges and segmentation patches [5] to obtain (e)-(g) in which it matches the selected object fingerprints. (h) shows the matched object part and estimated object boundary.

ple, our approach provides a different solution to recognize street landmarks from one training image, while [8] focuses on car and face instances. Furthermore, our approach applies multiple segmentations to extract patches, which has been shown to give better spatial support for recognition than regular patches [21], which are used by [8].

3. SELECTING OBJECT FINGERPRINTS FOR CONTENT ANALYSIS

We highlight in this section key ideas of our object fingerprint-based content analysis approach applied to street landmark localization. Fig.2 provides an overview of core steps of our approach. Let us first define how recognizing a seen landmark from an input image can be formulated as a object fingerprint matching problem. During training, we construct a set $\Delta = \delta_1, ..., \delta_M$ of M object fingerprints lying on the landmark. Object fingerprints of an object are defined as a list of selected feature phrases such as bigram edges, bigram patches, and triplet keypoints which are unique to the object. At runtime, given an input image Fig.2(c), we first detect salient regions by removing backgrounds as in Fig.2(d) and then decompose the image into a set of features Fig.2(e)-(g) from which we want to decide whether or not they contain the selected object fingerprints. In other words, by representing the object through M object fingerprints we transform the landmark recognition problem to a fingerprint matching problem Fig.2(h). The underlying assumption of our approach is each selected object fingerprint represents the object identity.

Variations of object scale and viewpoint are challenging issues in this problem. Keypoints have been shown to be effective for matching object within certain scale changes, while shape and appearance features can tolerate more scale variations. By combining the two, we may achieve robust feature combinations. In other recognition tasks, there is usually a large set of training data available. However, in our task of street landmark localization, it is impractical to obtain many images per street landmark. Most likely, we have only one image per landmark. In order to achieve robustness against viewpoint, scale, and illumination changes, we synthesize many images of the landmark using a simple rendering technique to extract robust features. The remainder of the paper is as follows.

Section 4 describes selecting object fingerprints from a single object view; Section 5 presents object localization in a new view; Section 6 presents experimental results; finally, Section 7 gives discussions and future directions.

4. SELECTING OBJECT FINGERPRINTS

We describe here how to extract object fingerprints from a *single* view of an object. We assume that only one object exists in the view. The object location is provided by ground truth or another algorithm. The task of learning is to start with this single view of the object, synthesize different views of the object, extract low-level features, and identify salient feature phrases and select object fingerprints. The feature phrases in our work include bigram and triplet features.

4.1 Augmenting Single Object View

The availability of object model views for content analysis training is usually limited, and this issue pushes us to rely on rendering new views of the object based on some geometric assumptions. We further extract different types of image features in these synthesized views as depicted in Fig.3. This method allows us to easily determine stable features from noises and perspective distortions, which in turn helps to make matching in low quality video frames robust to changes of viewpoint and illumination and clutter.

We focus on street landmark objects in this study. Since most street landmarks are rigid bodies and planar in the real world, a new view can be generated by warping the model view of the street landmark using an affine transformation which approximates homography. The affine transformation can be decomposed as: $H = R_\theta R_\phi^{-1} S R_\phi$, where R_θ and R_ϕ

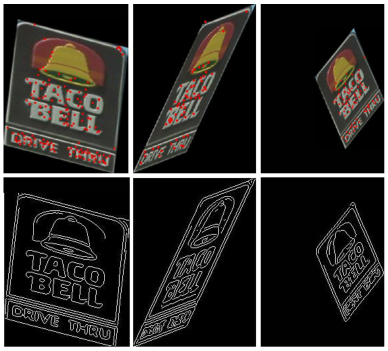

Figure 3: Synthesized views of the Taco Bell sign and extracted keypoints and edges for these views.

are two rotation matrices respectively parameterized by the angles θ and ϕ, and $S = diag[\lambda_1, \lambda_2]$ is a scaling matrix. We use a random sampling of the affine transformation space, the angles θ and ϕ changing in the range $[-\frac{1}{4}\pi, +\frac{1}{4}\pi]$, and the scales λ_1 and λ_2 changing within the range $[0.15, 2.0]$. We set $[-\frac{1}{4}\pi, +\frac{1}{4}\pi]$ as angle change range due to no visibility beyond this range. We use a much large scale range because of two reasons: 1) to select the most stable scale-invariant features and 2) to later recognize the street landmark at different distances. Therefore, we obtain a set of synthesized landmark object views, K, including the model view.

4.2 Extracting Features

Local image features are basic representation blocks of an object in images and videos. The spirit of our approach is to identify the object's unique fingerprints, and works independently of any particular choice of feature detectors and descriptors. In this work, we extract local keypoints, edges, and patches. We choose the SIFT detector for its robustness and rapidness [20]. Number of scales per octave is set to be 3 and $DoG\_thresh = 0.02$. We choose Canny edge detector to extract edges for its simplicity. Edges whose length is greater than 20 pixels and loops whose length is greater than 30 are selected.

It has been demonstrated that using multiple different segmentations of the same image can improve recognition accuracy [21]. This can be implemented if computational complexity and speed are not concerns during training. We choose the three most popular segmentation algorithms, Normalized Cuts [28], Mean-Shift [3] and the FH algorithm [6], to generate the "soup of segments" for the given object image. For Normalized Cuts, we generate 7 different segmentations per landmark by varying the number of segments $k = 5, 10, 15, 20, 25, 30, 40$. For the Mean-Shift segmentation, we get 9 segmentations by setting $min\_region = size(landmark)/15$ and varying $spatial\_band = 5, 7, 9$ and $color\_band = 7, 14, 21$. For the FH algorithm, we get 12 segmentations by setting $\sigma = 0.5, 1, 1.5, 2$, $k = 200, 500, 1000$, and $min\_region = size(landmark)/15$. From these 28 segmentations, we let the user determine the most satisfactory one as the landmark patch representation. Obviously, we cannot afford the strategy of multiple segmentations on every video frame; in contrast, we use one segmentation algorithm with a single setting as described in Section 5.

We also apply SIFT, Canny detector, and Mean-Shift segmentation algorithms on the synthesized set, K. For each feature, f_i, which is detected in the original view we define a missing score to measure what percentage of the feature is not detected in K.

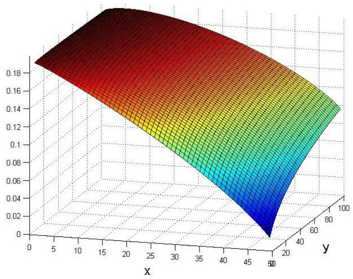

Figure 4: The minimal of conditional entropy $H(O|\delta_j)$ indicates the most salient feature phrase with respect to the existence of O. It is a function of x and y, the number of times δ_j occurs on O and at other locations. The figure shows $H(O|\delta_j)$ is minimized when δ_j occurs on every instance of O and does not at any other locations.

$$\Gamma(f_i) = 1 - \frac{C_j}{C}, \qquad (1)$$

where C_j is the number of synthesized views in which f_i is present and C is the total number of synthesized views. $\Gamma(f_i)$ is smaller, thus better.

4.3 Finding Salient Feature Phrases

The task of this section is to select the salient feature phrases (which appear in the model view) about the existence of the object. This becomes extremely important when conditions of the new object view differ greatly from the model view and only signature information of the object is preserved. We randomly select a set of 1000 natural scene images from public datasets (denoted by S) as the feature selection database.

For every image in S, we apply the same procedure to extract SIFT keypoints, edges, and segmentation patches, from which we determine the salient feature phrases about O. Although this process seems computational heavy, however, we only need to perform feature extraction once, cluster extracted features offline, and store constructed vocabularies (one for each feature type) in memory when a new model view comes. Intuitively, we want to find feature phrases which occur on the landmark, but rarely or never occur anywhere else. This intuition can be formally implemented by the information gain criterion and is commonly used for feature selection [27].

Information gain $I(A|B)$ is a measure of how much uncertainty is removed from a distribution by adding some additional information. It is defined based on the entropy $H(A)$ and conditional entropy $I(A|B) = H(A) - H(A|B)$. In our problem, information gain $I(O|\delta_j)$ is defined with respect to the existence of the object O and a particular feature phrase δ_j. O is a binary variable that is true when the landmark is present and δ_j is a binary variable that is true when the feature phrase δ_j is present. Therefore, the information gain of feature phrase δ_j at presence of the object O is:

$$I(O|\delta_j) = H(O) - H(O|\delta_j). \qquad (2)$$

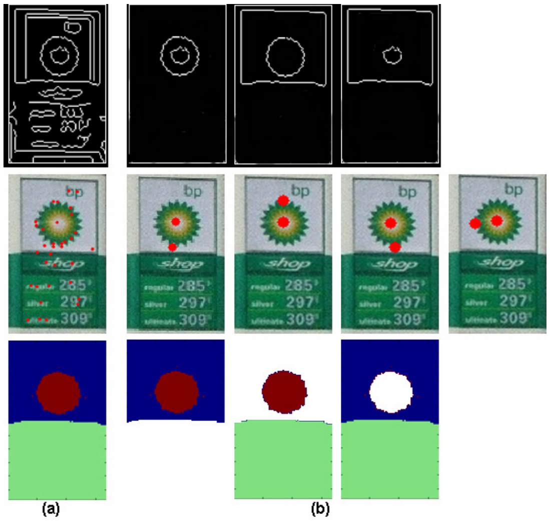

Figure 5: Extracted three types of features (a) and select bigram object fingerprints for each type (b).

Recall that our goal is to find those feature phrases on O which maximizing this information gain value. Since the entropy $H(O)$ is constant across all features on the object, then maximizing $I(O|\delta_j)$ leads to minimize the conditional entropy $H(O|\delta_j)$. We can compute $H(O|\delta_j)$ from four terms: N,N_O, $N_{\delta_j O}$, and $N_{\delta_j \overline{O}}$. The first two terms are constant: N is the total number of images; N_O is the number of object instances. The last two terms vary with feature phrase: $N_{\delta_j O}$ is the number of times δ_j occurs on O. $N_{\delta_j \overline{O}}$ is the number times f_i occurs on other instances. Based on the definition of the conditional entropy, $H(O|\delta_j)$ depends on six probabilities and essentially is a function of N,N_O, $N_{\delta_j O}$ and $N_{\delta_j \overline{O}}$. For simplicity, we substitute x and y for $N_{\delta_j O}$ and $N_{\delta_j \overline{O}}$ and then we have,

$$\begin{aligned} & H(O|\delta_j) = \\ & -\frac{x+y}{N}\left(\frac{x}{x+y}log\left(\frac{x}{x+y}\right) + \frac{y}{x+y}log\left(\frac{y}{x+y}\right)\right) \\ & -\frac{N-x-y}{N}\left(\frac{N_O - x}{N-x-y}log\left(\frac{N_O - x}{N-x-y}\right)\right. \\ & \left.+\frac{N-N_O-y}{N-x-y}log\left(\frac{N-N_O-y}{N-x-y}\right)\right). \end{aligned} \qquad (3)$$

This equation shows $H(O|\delta_j)$ is determined by a function of x and y as shown in Fig.4. This leads to fast processing of a new model view since we compute x on the new model view and y by looking through the vocabularies.

4.4 Selecting Object Fingerprints

We adopt bigram and triplet features as the feature phrases, which are our matching primitives. We allow combination of features of *different* types to form a feature phrase. By combining two or three features into a set, we increase their discriminative power over unigram features. However, the above definition of the conditional entropy only considers saliency of a bigram or triplet while ignoring the robustness of them to noises. To achieve the optimal trade-off between selecting salient and stable features, we propose to rank extracted feature phrases by using the following value.

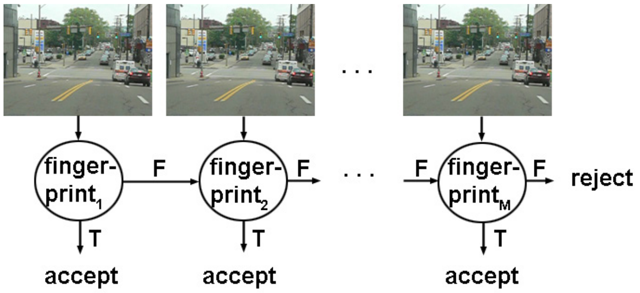

Figure 6: A cascade of selected object fingerprints using 1NN classifiers. The most salient and robust object fingerprints are selected to construct the cascade. During localization, once an object fingerprint is matched, the object will be localized.

$$Q(\delta_j) = (1-\mu)\cdot H(O|\delta_j) + \mu \cdot \Gamma(\delta_j), \qquad (4)$$

where $\Gamma(\delta_j) = \sum_{i=1}^{n} \Gamma(f_i)$, $f_i \in \delta_j$, is the sum of missing scores of each feature in δ_j, $n = 2$ for bigram and $n = 3$ for triplet. μ is a weighting factor. By computing $Q(\delta_j)$ for extracted feature phrases on the object, we rank them in ascending order and keep top M to form the set of object fingerprints, Δ, for O. Fig.5 depicts the selected bigram fingerprints for BP sign. These M fingerprints are used to build the acceptance cascade to recognize O at runtime as shown in Fig.6.

5. LOCALIZATION IN A NEW VIEW

Given a new view, J, our task is to detect whether the object exists, localize where the object is, and estimate its boundary. J can be a new image or video frame. We go through the following steps: removing less-information regions (e.g., sky and ground), extracting features, matching the selected object fingerprints to extracted features, and finally estimating the pose and boundary of the object if it exists.

5.1 Saliency Detection and Feature Extraction

During training, the object location and boundary is given by ground truth while for J we have no prior knowledge whether the landmark exists and where it is. Instead, to avoid greedy feature extraction on the whole image, we rely on a simple method for visual saliency detection [10]. By analyzing the log-spectrum of the input image, the method extracts the spectral residual of an image in the spectral domain and constructs the corresponding saliency map in the spatial domain (Fig.2(d)).

We follow a similar procedure to extract image features as we do during training. We apply a SIFT detector to extract local keypoints which are represented by 128-dimension vectors. Canny edge detector ($\sigma = 1$) is applied to extract edge fragments. For patch features, we do not apply multiple segmentations due to computation concern, but only apply the FH algorithm [6] by letting $\sigma = 0.5, k = 250$ and $min\_region = 50$. This empirical setting was learned from our experience with three segmentation algorithms. Thus, we obtain three feature maps for J, as shown in Fig.2(e)-(g).

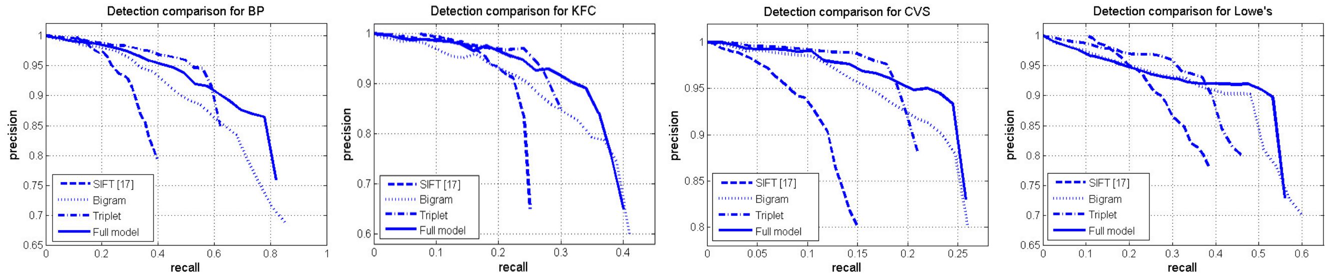

Figure 7: Precision recall curves using Lowe's algorithm [17], bigram fingerprints, triplets and full model.

5.2 Localization via Object Fingerprints

For street landmark localization, we propose an acceptance cascade of M 1-Nearest Neighbor (1NN) classifiers in Fig.6. Different from face or human detection using rejection cascades which combine a number of weak classifiers, our object fingerprints-based cascade model recognizes and localizes the object once verification occurs at any cascade level. At each cascade level, we match the object fingerprint, δ_j, to the image, J, via a pictorial model [7].

$$L^*_{\delta_j} = \arg\min_{L_{\delta_j}} (\sum_{i=1}^{n} m_i(l_i) + \sum_{(f_i,f_k)\in E,\delta_j} d_{ik}(l_i, l_k)), \qquad (5)$$

where $L = (l_1, .., l_n)$ specifies an candidate matching feature phrase in J to δ_j, where l_i specifies the matching location of feature f_i, $f_i \in \delta_j$. $n = 2$ for a bigram or $n = 3$ for a triplet. $(f_i, f_k) \in E, \delta_j$ indicates f_i and f_k are connected in δ_j. $m_i(l_i)$ measures the degree of mismatch when feature f_i is matched to the feature at location l_i in J. $d_{ik}(l_i, l_k)$ measures the deformation of the model when f_i is matched to l_i and f_k matched to l_k. We define $m_i(l_i)$ as Shape Context cost [1] for edge and patch features and $1 - cosine(a, b)$ distance metric for keypoint features (SIFT vectors). For $d_{ik}(l_i, l_k)$, we choose either feature in the bigram, or the middle feature in the triplet, as a reference point, say f_i, and its matched position in J, l_i. We then define $d_{ik}(l_i, l_k)$ as the L_2 distance from l_k to f_k. For triplets, the second term in Eq.(5) is sum of deformation costs on two edges. Since the pictorial model in our case is a two-node or three-node tree, an efficient search algorithm [7] is applied to obtain the minimal-distance match to δ_j in given J.

If the mismatch distance from δ_j to the minimal-distance matching feature phrase in J (the first term in Eq.(5)) is smaller than the sum of features' mismatch thresholds, the found feature phrase will be confirmed to match to δ_j and the object is recognized. The mismatch thresholds for keypoints (ψ_k), edges (ψ_e) and patches (ψ_p) are learned from the synthesized view set K.

The pose of the landmark can be estimated by projecting back to the synthesized landmark views which give rise to the minimal residual on the matched fingerprints. In order to estimate the landmark boundary, we can either re-scale and overlay the boundary of the synthesized view obtained from the last step, or compute the landmark image height from its image position, 3D height (obtained from the training image), and the viewpoint using the method introduced in [9]. We use the first method in this work.

6. EXPERIMENTS

Due to the lack of established research in street landmark recognition, it is difficult to obtain a standard dataset to compare our approach with. Thus, we compile a new dataset of 15 U.S. street landmarks including 4 categories: 1) gas station, 2) fast food restaurant, 3) pharmacy and 4) store (Fig.8). For each landmark, we have a single model image and several test videos. We collected videos from a moving vehicle, so it's a real-world dataset. For each test video, a human expert labels video frames which contain the street landmark and those which do not. Precision, recall and F_1 are evaluated on the frame base. We consider a recognition correct if the matched object fingerprint is within the landmark region. The number of cascade levels are set at $M = 20$. We used the UCLA implementation of SIFT [1],which are very similar to Lowe's [20]. The vocabulary size of each feature channel for clustering in training is set at 1000, as suggested in [31].

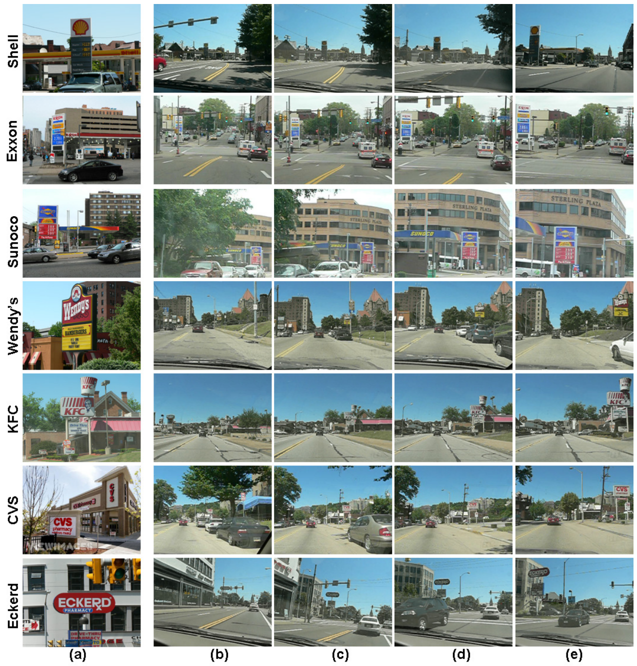

Figure 8: Example images from our street landmark dataset, which consists of 15 U.S. street landmarks. (a) training images; (b)-(e) test video images.

[1] http://vision.ucla.edu/∼vedaldi/code/sift/sift.html

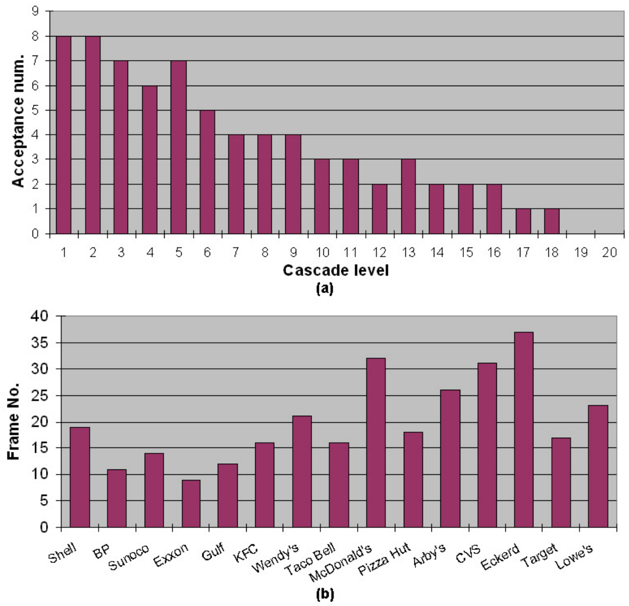

Figure 9: (a) The number of first detections at each cascade level on all videos. (b) The video frame numbers when the landmark is first recognized after averaging all videos for each landmark.

Fig.7 depicts the detection performance using precision-recall curves on four landmark examples, i.e., 1) BP, 2) KFC, 3) CVS and 4) Lowe's, which in general represents the detection patterns of general street landmarks. Each subfigure contains the precision recall curves of four detection algorithms, i.e., SIFT [20], bigram only fingerprints, triplets only, and the full model that uses both bigrams and triplets. Parameters such as $DoG\_thresh, match\_thresh, \mu, \psi_e, \psi_k, \psi_p$ are varied to obtain precision-recall curves for Lowe's algorithm and our methods.The figure shows several interesting observations. First, we observe that four methods consistently perform better on some landmarks than others. Best performance appears for the BP sign, which can be explained by its salient green star logo with high contrast pattern. However, for landmarks which mainly contain text such as CVS, four methods all perform poorly due to lack of appropriate low-level features to capture saliency of text on the landmark. Second, we can see that our three methods all perform better than SIFT on these four examples. Third, for most examples, the bigram-based method achieves higher recall while lower precision than the triplet method. This can be understood by the fact that the triplet method requires stronger verification than the bigram method. Finally, by combining both bigram and triplet feature phrases, the full model obtains the best performance in F_1 across all examples.

To further examine the proposed algorithms, Table 1 lists a quantitative comparison on a dataset of 15 U.S. street landmarks. There are about 7000 labeled video frames and images. F_1 (defined as $\frac{2pr}{p+r}$) is reported for each landmark using three algorithms, Lowe's [20], Zhang's [33][2] and our object fingerprint-based approach. Our proposed algorithm consistently performs better than two other state-of-the-art SIFT-based recognition algorithms. Overall, the average F_1 increases 32.02%, from 0.506 to 0.668. By combining both selected object fingerprints from different feature types, our proposed approach achieves the best performance for each landmark. As we can see, our approach performs very well on gas station landmarks which are generally big with salient appearance, performs well on fast food and store categories which also often contain signature logos or shapes. In contrast, pharmacy landmarks such as CVS and Eckerd only have text in their signs, which is hard to recognize by our approach and two other algorithms. In addition, pharmacy signs are much smaller than other landmarks; sometimes even hard for humans to distinguish in videos. Another advantage of our approach is to combine low level features such as edges and keypoints with mid-level features such as segmentation patches to detect landmarks in various scales, which is an essential ability for street landmark localization and other content analysis tasks. Our approach usually runs for about half a minute processing a video frame of 320 × 240 on a PC with 3.2GHz CPU and 2GB RAM. Although it is still far from real time, our approach can be applied to real-time tasks with good sampling strategy.

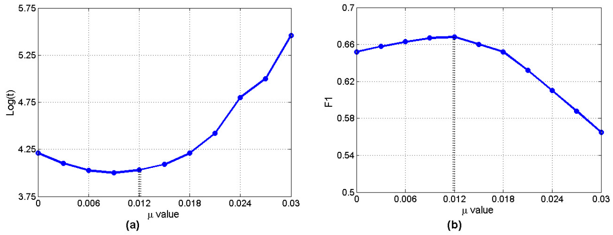

Figure 10: The effect of different μ on running time in Log scale of seconds (a) and overall performance of our approach on the dataset (b).

Fig.9(a) depicts the number of first detections at each cascade level for all test videos. As we can see, the majority of test video landmarks are first detected before the $10th$ level. This observation corroborates saliency and robustness of top fingerprints selected by our approach. Fig.9(b) shows when each landmark is usually first recognized in video. It confirms that gas station landmarks are usually detected earlier than other kinds of landmarks. Also, pharmacy signs are more difficult and take longer to detect.

We have also conducted experiments to study the influence of μ on the performance of our approach and computational complexity at runtime as shown in Fig.10. We can see from Fig.10(b) that in our dataset the best performance appears when $\mu = 0.012$. Although this parameter is not necessarily a generalized good setting for any dataset, nor does it lead to the lowest computation complexity (Fig.10(a)), it shows that a good trade-off between selecting informative and robust features can achieve better results. For instance, $F_1 = 0.652$ when selecting the most informative fingerprints $\mu = 0$, is lower than $F_1 = 0.668$ when emphasizing robustness of selected fingerprints $\mu = 0.012$.

Organization of web image search results have been recently studied in the multimedia community [12]. To demonstrate that our object fingerprint-based approach can be generalized to other content analysis tasks, we apply our method to refine web image search results by re-ranking re-

[2] We implemented the method as in [33] without motion estimation .

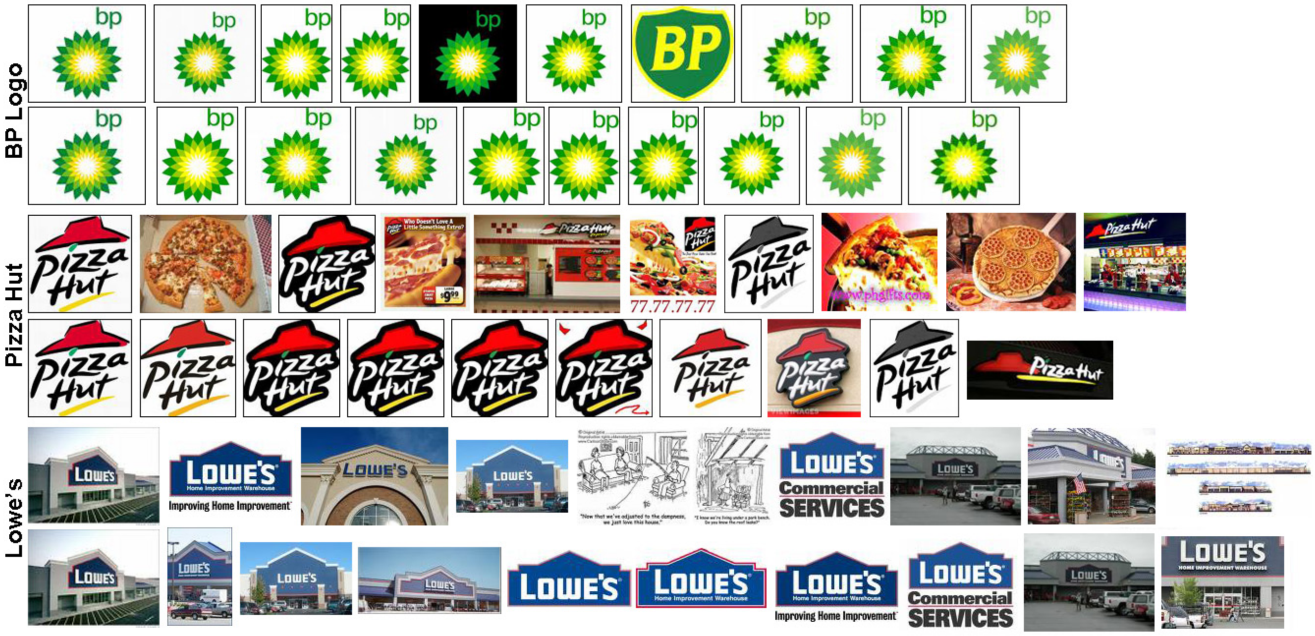

Figure 11: Refine web image search results by our approach. Google search results by 3 queries, i.e. *BP logo*, *Pizza Hut* and *Lowe's*, are shown. For each query, the first row shows top 10 images returned by Google and the second row shows refined top 10 images by our method using the first returned image as model image.

turned images. We used Google image search in this experiment. We assume that the No.1 returned image is correct and use it as the model image to re-rank the rest of the top 100 images returned by Google. Instead of quitting after acceptance, we run matching through all object fingerprints in the cascade and count the number of matched fingerprints for each image. Use the count as the ranking score and images with same counts are ranked by their sizes. Images with at least one matched fingerprint are included in the re-ranking list. Fig.11 depicts the refined top results by our approach. On the other hand, to achieve diversity of search results, our approach can also be applied to eliminate near-duplicates from search results.

	Lowe	**Zhang**	**F.P.**	**↑(%)**
Shell (404)	0.512	0.480	0.689	34.57
BP (448)	0.605	0.582	0.827	36.69
Sunoco (660)	0.619	0.581	0.761	22.94
Exxon (540)	0.633	0.605	0.756	19.43
Gulf (411)	0.625	0.587	0.776	24.16
KFC (488)	0.430	0.415	0.584	35.81
Wendy's (408)	0.448	0.429	0.608	35.71
Ta∼Bell (490)	0.667	0.637	0.815	22.19
McDona∼(519)	0.501	0.471	0.651	29.94
Piz∼Hut (492)	0.485	0.468	0.708	45.98
Arby's (475)	0.472	0.470	0.637	34.96
CVS (415)	0.316	0.304	0.419	32.59
Eckerd (484)	0.293	0.257	0.386	31.74
Target (498)	0.469	0.455	0.740	55.22
Lowe's (427)	0.512	0.472	0.669	27.73
Avg_F_1	**0.506**	**0.493**	**0.668**	**32.02**

Table 1: Performances on our dataset of 15 U.S. street landmarks in F_1 measure. Lowe uses the algorithm as in [17]; Zhang use supplementary cosine criterion as in [31]; F.P.: our approach; ↑(%) is relative improvement of F.P. over Lowe. Numbers in the first column are numbers of test images.

7. DISCUSSIONS AND FUTURE WORKS

Objects are basic units for multimedia content analysis. Much research has focused on exploring similarity among objects and contexts for content analysis. In this paper, we have proposed a novel approach for multimedia content analysis based on objects' unique characteristics which we call object fingerprints. In particular, we have demonstrated its performance on the problem of street landmark localization. We introduce the concept of object fingerprints to represent salient appearances of the object as well as the geometric relationships among local features. Another contribution of this work is a way to deal with object recognition and localization with very limited training data. We have focused on single landmark recognition from images. However, our approach can be well generalized to tackle other kinds of content analysis tasks. Our approach employs some state-of-the-art techniques for feature extraction, and has achieved better results than existing methods. In addition, occlusion can be handled by matching object fingerprints at various locations of the object as long as at least one object fingerprint is matched. There is still much room to improve the approach. We expect that using more advanced features would boost performance. While our approach really focuses on the *content* of the object given a model image, a sound future direction is to combine *context* with content for more robust recognition and localization. Some pioneer work in this direction has been published, using image context [30], geometric scene context [9], and image location tags [14]. Another future direction is to study the importance of various features in matched object fingerprints for different street landmarks.

8. ACKNOWLEDGEMENT

This research was partially supported by the gift grant from the General Motors, the gift grant from Microsoft, and NSF grant IIS-0534625.

9. REFERENCES

[1] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Trans. on PAMI*, 2002.

[2] G.E. Burnett. Turn right at the king's head: Drivers' requirements for route guidance information. In *PhD thesis, Loughborough University, UK*, 1998.

[3] D. Comaniciu and P. Meer. Mean shift: a robust approach toward feature space analysis. *IEEE Trans. on PAMI*, 2002.

[4] R. Datta, D. Joshi, J. Li, and J.Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys*, 2008.

[5] A. de la Escalera, L.E. Moreno, M.A. Salichs, and J.M. Armingol. Road traffic sign detection and classification. *IEEE Transactions on Industrial Electronics*, 1997.

[6] P. Felzenszwalb and D. Huttenlocher. Efficient graph-based image segmentation. *IJCV*, 2004.

[7] P.F. Felzenszwalb and D.P. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 2005.

[8] A. Ferencz, E. G. Learned-Miller, and J. Malik. Learning to locate informative features for visual identification. *IJCV: Special Issue on Learning and Vision*, 2007.

[9] D. Hoiem, A.A. Efros, and M. Hebert. Putting objects in perspective. In *CVPR*, 2006.

[10] X. Hou and L. Zhang. Saliency detection: a spectral residual approach. In *CVPR*, 2007.

[11] W.H. Hsu, L.S. Kennedy, and S.F. Chang. Video search reranking via information bottleneck principle. In *ACM Multimedia*, 2006.

[12] F. Jing, C. Wang, Y. Yao, K. Deng, L. Zhang, and W.Y. Ma. Igroup: web image search results clustering. In *ACM Multimedia*, 2006.

[13] Y. Ke, R. Sukthankar, and L. Huston. Efficient near-duplicate detection and sub-image retrieval. In *ACM Multimedia*, 2004.

[14] L. Kennedy, M. Naaman, S. Ahern, R. Nair, and T. Rattenbury. How flickr helps us make sense of the world: context and content in community-contributed media collections. In *ACM Multimedia*, 2007.

[15] S. Kim, X. Jin, and J. Han. Sparclus: Spatial relationship pattern-based hierarchical clustering. In *SIAM Int. Conf. on Data Mining*, 2008.

[16] B. Kroon, S. Boughorbel, and A. Hanjalic. Person-based search in videos. In *ACM Multimedia, Demo*, 2007.

[17] B. Leibe, N. Cornelis, K. Cornelis, and L. Van Gool. Dynamic 3d scene analysis from a moving vehicle. In *CVPR*, 2007.

[18] M.S. Lew, N. Sebe, C. Djeraba, and R. Jain. Content-based multimedia information retrieval: State of the art and challenges. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2006.

[19] Y. Li and J.H. Lim. Outdoor place recognition using compact local descriptors and multiple queries with user verification. In *ACM Multimedia*, 2007.

[20] D.G. Lowe. Distinctive image features from scale-invariant keypoints. In *IJCV*, 2004.

[21] T. Malisiewicz and A.A. Efros. Improving spatial support for objects via multiple segmentations. In *British Machine Vision Conference*, 2007.

[22] H. Noor, S.H. Mirza, Y. Sheikh, A. Jain, and M. Shah. Model generation for video-based object recognition. In *ACM Multimedia*, 2006.

[23] G. Piccioli, E. De Micheli, P. Parodi, and M. Campani. Robust method for road sign detection and recognition. *Image and Vision Computing*, 1996.

[24] J. Pilet, V. Lepetit, and P. Fua. Real-time non-rigid surface detection. In *CVPR*, 2005.

[25] A. Qamra and E. Y. Chang. Scalable landmark recognition using extent. In *Journal of Multimedia Tools and Applications*, 2007.

[26] S. Sanyal and S.H. Srinivasan. Logoseeker: a system for detecting and matching logos in natural images. In *ACM Multimedia*, 2007.

[27] G. Schindler, M. Brown, and R. Szeliski. City-scale location recognition. In *CVPR*, 2007.

[28] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Trans. on PAMI*, 2000.

[29] J. Sivic and A. Zisserman. Video google: a text retrieval approach to object matching in videos. In *Intl. Conference on Computer Vision*, 2003.

[30] A. Torralba, K.P. Murphy, W.T. Freeman, and M. Rubin. Context-based vision system for place and object recognition. In *Intl. Conference on Computer Vision*, 2003.

[31] J. Willamowski, D. Arregui, G. Csurka, C.R. Dance, and L. Fan. Categorizing nine visual classes using local appearance descriptors. In *IWLAVS*, 2004.

[32] W. Wu, X.L. Chen, and J. Yang. Incremental detection of text on road signs from video with application to a driving assistant system. *ACM Multimedia*, 2004.

[33] W. Zhang and J. Kosecka. Image based localization in urban environments. In *Intl. Symposium on 3D Data Processing, Visualization and Transmission*, 2006.

[34] W. Zhao, R. Chellappa, A. Rosenfeld, and P.J. Phillips. Face recognition: A literature survey. In *ACM Computing Surveys*, 2003.

[35] Q.F. Zheng, W.Q. Wang, and W. Gao. Effective and efficient object-based image retrieval using visual phrases. In *ACM Multimedia*, 2006.

[36] C.L. Zitnick, J. Sun, R. Szeliski, and S. Winder. Object instance recognition using triplets of feature symbols. In *Microsoft Research Technical Report*, 2007.