Investigation of Image Models for Landmark Classification

Mark Hughes, Gareth J. F. Jones
Centre for Digital Video Processing
Dublin City University
Dublin 9, Ireland
mhughes@computing.dcu.ie

Noel E. O'Connor
Clarity Centre for Sensor Web Technologies
Dublin City University
Dublin 9, Ireland
oconnorn@eeng.dcu.ie

Abstract

One commonly used approach to scene localisation and landmark recognition is to match an input image against a large annotated database of images using local image features. However a problem exists with these approaches with memory constraints and the processing time required to compare high dimensional image feature vectors in a very large scale database.

In this paper we investigate a new landmark classification technique which takes advantage of the fact that there is considerable overlap in visually similar images of landmarks in any large public photo repository. A large number of images containing landmarks are clustered into visually similar clusters. Classification models are then implemented and trained based on global histograms of interest point features from these clusters to create models which can be used for robust real-time accurate classification of images containing these landmarks. We also investigate different techniques for the creation of these classification models to ascertain how best to guarantee a high level of robustness, accuracy and speed.

1. Introduction

With the arrival of commercial digital cameras at increasingly low cost,average consumers are able to capture and store large volumes of high-quality digital imagery quickly and cheaply. This is creating a significant challenge regarding how to efficiently organise and retrieve these images. Several online image databases such as Flickr [1] now exist which can store a user's personal photo collections and allow them to search their own and others' public images from within these collections using textual queries based on manually assigned tags.

A problem that persists is that users may not spend the required time to create rich consistent image tags leading to reduced retrieval effectiveness. Ideally a solution is required to create methods to automatically create image captions or enhance existing ones. Several techniques exist to classify low-mid level semantics about an image such as whether an image contains faces [4] or large buildings [10]. We aim to provide a much higher level semantic annotation describing the actual landmark contained within an image.

Our focus is on landmarks within images due to the significant contribution that they make to a large scale public photo repository such as Flickr (eg. Flickr search for eiffel tower returns over 310,000 images, Flickr search for empire state returns over 230,000 images). Landmarks also tend to have a unique visual appearance which leads to high discrimination values between images of landmarks.

A commonly used approach to identifying landmarks within images is to match sets of interest points to a large dataset of annotated landmark images using point to point matching [8]. The main disadvantage of this technique is that it is very processor intensive and with a large dataset would be infeasible. We propose a technique based on classification models which have small memory footprints and provide a very fast technique of classifiying a landmark.

Our approach to the classification of these landmarks is based upon the photographing behaviour of users on large scale photo-sharing websites. Users tend to visit similar destinations and landmarks. Users also tend to take images of these landmarks from a small number of photogenic locations which leads to multiple clusters of visually similar images of popular landmarks that reflect their different viewpoints. Based on this premise we intend to take advantage of this overlap by reducing the search space in a large scale dataset by clustering visually similar images thus creating more robust means of classifying an image using classification models.

In this paper we first introduce previous work in the field. In section 3 our motivation and the approach that we use for this problem is outlined. The results of experiments carried out are detailed in section 4. The final section describes how we plan to scale up these experiments by testing this approach on a very large scale dataset.

2. Previous Work

Image classification is the subject of a large amount of current research activity. In this section we review existing work which is particularly relevant to our research. Interest point detection is a term used in the computer vision community and refers to the detection of salient interest points in images for subsequent processing. Today, the main application of interest points is to detect points/regions in the image domain that are likely candidates to be useful for image matching. Ideally interest point detectors should be invariant to scale, rotation and affine-invariant to a certain degree. The most important attribute of an interest point detection algorithm is repeatability, ie. will the algorithm find the same interest point in similar images? Several different algorithms exist that meet all of these criteria. Two of the most common ones used today are the Scale Invariant Feature Transform (SIFT) [9] and the Speeded Up Robust Features (SURF) [3] algorithms.

There is some existing work on scene classification using Support Vector Machine (SVM) models combined with interest point features. Ayers et al. [2] have fused SIFT keypoints with SVM models for home interior classification. They use histograms of SIFT features as inputs to the SVMs with relatively accurate classification of rooms in a home.

We are also interested in visual bag of words features. These have been used successfully in the past for generic object classification. Tirilly et al. [13] use visual word histograms combined with language models for the generic classification of objects within images. They classify with reasonable success the presence of objects such as airplanes, motorbikes and guitars within images.

Kennedy et al. [6] have done some work with classifying specific landmarks and visual concepts around the New York area using low-level colour and texture features combined with Support Vector Machine classification models. Our classification method improves upon this work by using more discriminative local image features.

Some work has taken place with the organisation and displaying of images containing landmarks based on combining textual tag data, spatial data and interest point features. Crandall et al. [5] use a process to classify location information about an image. They attempt to estimate the location where an image was taken using its content (image attributes and text tags).

Kennedy et al. [7] cluster images of landmarks from a dataset of 110,000 images all taken in San Francisco which was downloaded from Flickr. Location data was first used to create initial clusters of images. Sub-clustering based on the tags which are associated with the images is then carried out with expensive comparison of interest points used as a last phase in their process. Using these clusters they then generate representative images for popular landmarks. We aim to improve upon this work by providing a means for automatically classifying new images based on visually similar clusters.

Zheng et al. [14] have been experimenting with landmark recognition on a large scale. They utilise images mined from the web and online articles which describe landmarks to create a very large dataset (21 million images). They cluster these images into clusters of visually similar landmarks based on location data and local feature matching. Once these images are clustered they use a k-nearest-neighbour classifier based on local feature matching.

Popescu et al. [12] used a geo-reference collection of 5000 landmarks worldwide. They use colour, texture and local image features to represent each image in their dataset. All of their landmarks are organised spatially and they classify a test image using spatial distance along with nearest neighbour classification. We aim to improve upon this work by using more robust support vector machine classification along with analysing the effects of affine variation within a classification model.

3. Landmark Classification

3.1 Approach

The traditional method to classify landmarks within images is to compare the interest points extracted from a test image against all images within a dataset using point to point matching [8]. However the actual matching between keypoints can be very computationally expensive, and for very large image databases would be computationally infeasible. Commonly used interest point detection methods depend on image content and image size, but will typically generate up to 1000 keypoints per image. This presents a considerable computational challenge in terms of matching two images using their individual interest points. To put it into perspective, to compare one image against all images in a 1000 image database using the SURF algorithm would require 64 million comparisons to be made. To compare one image against a database of 100,000 images would require over 6 trillion comparisons to be made, and this number would grow considerably as the size of the database grows.

A new approach is desired which does not require test images to be compared against large numbers of training images in order to get a successful match. We propose an approach which is motivated by the observation that the majority of photos containing landmarks, for example by tourists, are frequently taken from a number of limited viewpoints due to geographical constraints and photogenicity as illustrated in figure 1. In our approach we cluster multiple image views of a landmark into single view classification models based on image features extracted from

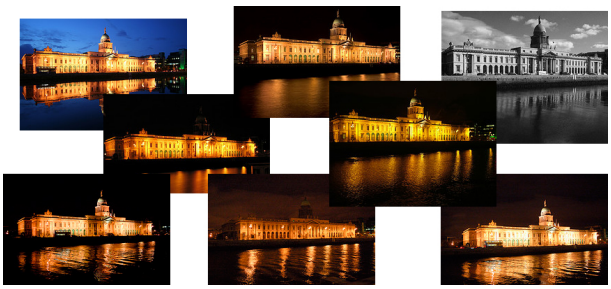

Figure 1. An example of images returned from a Flickr search using the term 'Customs House Dublin'. It is straightforward to see that all images are visually very similar and have been taken from a relatively similar position.

these images. There are two main advantages to be gained from clustering multiple image views into single classification models:

- Computational overhead: The amount of time taken to compare and classify images in a large-scale database is drastically reduced. With efficient filtering methods this classification could be theoretically done in real time in large-scale databases.
- Robustness: Increased robustness is obtained by combining features obtained under multiple imaging conditions into a single model view.

In our work we explore the hypothesis that classification models can be used to create robust methods to classify an individual landmark. We wish to ascertain which classification technique would provide the best trade-off between classification accuracy and the time taken to classify an image.

The difference in the viewpoints of each image can greatly effect the accuracy of the classification as training images which are not visually similar will add noise to the training data therefore we aim to determine the best approach in the creation of landmark classification models and what level of dis-similarity or affine variation between training images provides accurate matching in the fastest time. To determine the affine and lighting variability parameters that would provide the highest level of robustness in the selection of input images for each classification model, we perform classification experiments on 2 different datasets.

- To test the hypothesis that classification models can be used to classify new landmark images we first implemented a set of experiments using a manually clustered dataset. This dataset is based on single viewpoint images where all training images contain a certain view of a landmark, are visually similar and are taken from a similar viewpoint.
- To test if affine and viewpoint variations greatly effect classification accuracy we created a second dataset which is based on multiple viewpoint images which is where the training images contain a large range of affine variation and a number of different lighting conditions (eg. night and day). Images within this dataset can be visually different and contain different viewpoints of a landmark such as front, back, side or the closeup of a landmark feature.

Using the single viewpoint dataset we implement a set of models trained to recognise 42 visually different landmarks from a single viewpoint. On average each landmark was represented by 63 different images. In our single viewpoint experiments we test two different classification techniques to determine which provides the best tradeoff between the accuracy of classification and the amount of time it takes to classify an image as containing a certain landmark.

Using the multiple view dataset we first train classifiers using different views and features of 10 landmarks and process a test dataset based on these 10 landmarks (Arc De-Triomphe, Colloseum, Golden Temple, Florence Cathedral, Notre Dame Cathedral, Reichstag, Rialto Bridge, Statue of Liberty, Tower Bridge, Trevi Fountain). We then cluster these images based on visual similarity to create multiple models for each landmark and process the same dataset to determine which method is more effective.

3.2 Dataset

We used the Flickr API to collect images of 42 popular landmarks located around the world. We deliberately chose a wide range of different types of landmarks such as cathedrals, statues, buildings, monuments, bridges, castles and geographical landmarks to ascertain how this approach would perform . On average there were over 50 visually similar images for each landmark. This training set consisted of images containing a single viewpoint of each landmark as illustrated in figure 2. A separate test collection consisting of 10 different images of each landmark was selected which resulted in a total of 420 test images.

To determine affine variability parameters in the model creation process, we created a second dataset of over 7000 images containing 10 of these landmarks taken from a wide range of viewpoints under a variety of lighting conditions as illustrated in figure 3. A separate test set of 1000 images was also created containing 100 images of each of these landmarks. The test set is quite challenging as no preprocessing has taken place therefor many of the landmarks are partially occluded and contain large human figures and faces.

Figure 2. Single View Models: An illustration of the some of the images used in the Statue of Liberty model. As can be seen from the images, they are all taken from a similar viewpoint and are visually similar.

Figure 3. Multiple View Models: An illustration of the some of the images used in the Arc de Triomphe model. As can be seen from the images, they are all taken from different viewpoints and under different lighting conditions.

3.3 Visual Bag of Words

Local image features have many advantages over traditional low-level global features such as those based on colour, texture and shape. Local image features are more discriminative than global features and they are also invariant to an extent to scale and affine variations. Due to the fact that they are based around small salient regions within an image they are also more robust to partial occlusion than global features.

We decided to use image features extracted using the SURF algorithm [3]. SURF descriptors are half the length (descriptor length of 64) of SIFT [9] descriptors which means less processing time while still retaining high repeatability and discrimination.

To train classification models it is necessary to organise these SURF features into a global feature vector due to the fact that varying amounts of SURF features will be extracted from different images. To organise these interest point features into discriminative global feature vectors we use the visual bag of words approach. Bag of words models were originally developed for document classification. A bag of words model is a technique where a document is represented as an unordered collection of words which are then used to classify a document based on these representations. Visual bag of words features use the same basic idea. An image is represented as a bag of visual words which are usually created based on image descriptions of salient regions within an image [13].

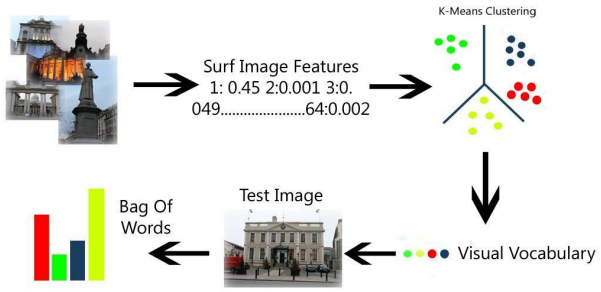

Figure 4. An illustration of the creation of visual bag of words histograms from a large dataset of images

To create a visual word histogram a visual dictionary is constructed from the training dataset. Local image feature descriptions are extracted from each image within the dataset using the SURF algorithm. These image features are then quantized into a visual vocabulary using a k-means clustering algorithm with k being the vocabulary size of the dictionary as illustrated in figure 4. During these experiments we used k values of 256, 512, 1024 and 2048. Using this vocabulary each image can then be represented by a global histogram value based on the nearest neighbour to each image feature in the dictionary.

3.4 Single-Viewpoint Classification

We first aim to test the hypothesis that it is possible to create robust landmark classification models based on visual word features by implementing and testing a set of 42 landmark models all trained using visually similar images from similar viewpoints. We test 2 different classification techniques.

3.4.1 Nearest Neighbour

The first classification technique that we experimented with was a nearest-neighbour classifier. We chose the nearest neighbour classifier as it performs quickly and is easy to implement. For each of the 42 landmarks we created a binary k nearest neighbour model. The visual word histograms from each image in a landmark cluster were used

as positive inputs while image features from an equal number of randomly chosen images in the dataset were used as the negative inputs for each model. We experimented with a variety of values for k and a variety of distance measures. We found that k = 5 and the Euclidean distance formula provided the highest level of classification accuracy in this task. We also experimented with a number of vocabulary sizes to determine which would provide the most desirable accuracy and speed. Each of the test images in the test collection were processed through their corresponding model and the results can be seen in table 1.

3.4.2 Support Vector Machines

SVMs are a popular learning algorithm which have been used extensively in a number of applications. The SVM is characterised by high generalisation ability, and based on the idea of finding the hyperplane that best separates two classes after mapping the training data into a higher-dimensional feature space via some kernel function. We used a linear kernel function in these experiments. A binary classifier model was trained for each of the landmarks in the dataset using the visual word histograms from each image in a landmark cluster as positive inputs and an equal amount of features from random images in the dataset as negative images. Each of the test images in the test collection were processed through their corresponding model and the results can be seen in table 1.

3.5 Multiple-Viewpoint Classification

We trained classification models to recognise 10 different landmarks. In our first experiment we trained 10 classification models using all the training data for that landmark as inputs. We then clustered all images based on visual similarity and created multiple models for each of the landmarks.

3.5.1 Multiple View Clusters

To compare the effect of differences in viewpoint in the creation of landmark classifiers we first train SVM models using a training set of images which contain a large variety of views of a landmark under different lighting conditions. Again visual word histograms were used as inputs. Due to the fact that a vocabulary sizes of 1024 and 2048 outperform smaller sizes we used these sizes in our multiple view classification process. On average over 700 images were used to train a model to recognise each landmark. We then process a test collection of 100 images per landmark the results of which can be seen in table 2.

Table 1. Single Viewpoint Test Set Landmark Classification Accuracy (Percentage of test images correctly classified)

Vocabulary Size	256	512	1024	2048
Nearest Neighbour	40%	44%	35%	24%
Support Vector Machine	91.9%	91.4%	93%	93%

3.5.2 Visually Similar Clusters

For each of the 10 landmarks in the multiple view dataset we clustered their images into a number of visually similar sub clusters. Images which have been clustered incorrectly will add noise to the training data used to train the classification models and this in turn will effect the classification accuracy therefore it is important to use a clustering method that will only organise images taken of a landmark from similar viewpoints into single view clusters.

To cluster the images we used an efficient k-d vocabulary tree approach. A vocabulary tree is a group of visual image features organised into a tree data structure for efficient matching purposes and nearest neighbour searching [11]. A k-d vocabulary tree takes multi-dimensional visual word features as inputs and splits the data into two sets based on the median dimension of largest variance within the visual words. It then repeats this process cycling through the dimensions until all points have been covered. Each leaf node in the tree represents a visual word from our vocabulary.

We used a tree with a vocabulary size of 65,000 which was generated using the features extracted from all images within the multiple view dataset. Each image feature extracted from an image was propagated down the tree to create a single feature vector for each image based on the nearest neighbours in the vocabulary. Visually similar images were then clustered based on matches between these feature vectors.

Once these clusters have been created an SVM model was created for each of the clusters. The same test dataset was then run through each of these models and if any of the models belonging to a landmark returned a positive prediction value the landmark was classed as being correctly classified. We compare the results of these 2 approaches in table 3.

4 Results

As can be seen from the results in table 1 it is possible to train an SVM model to recognise certain landmarks from a single viewpoint with a high level of accuracy. The SVM method outperforms the basic nearest neighbour method by

Table 2. Single Viewpoint Test Set Landmark Classification Speed (Average Time per image in milliseconds)

Vocabulary Size	256	512	1024	2048
Nearest Neighbour	120	380	440	720
Support Vector Machine	110	200	390	900

Table 3. Multiple Viewpoint Test Set Landmark Classification Accuracy

Vocabulary Size	1024	2048
Multiple View Classification (All images)	65%	71.6%
Single View Classification (Clustered Images)	90.3%	91.5%

a significant margin. Interestingly the vocabulary size used does not seem to effect the classification accuracy by a large degree. The differences in accuracy between a vocabulary size of 256 and 2048 is just 1.1%, however classifying an image using a vocab size of 256 takes only 11% of the time required using a size of 2048. Therefore it seems that using a vocabulary size of 256 is most desirable due to the tradeoff between accuracy and speed.

As can be seen from the results in table 3 landmark classification models perform significantly better when all inputs images are visually similar. Multiple views of landmarks add noise to the training data. As can be seen from the results in table 3 the largest vocabulary size of 2048 outperformed the smaller size. Even using a very challenging test collection of images our classifiers classified just over 90% of the test images correctly.

5 Future Work

We are currently in the process of implementing our technique on a very large scale image set (1 million images) containing landmark images from around the world which has been crawled from the online photo repository Flickr. It is hypothesised that using spatial filtering techniques combined with efficient image similarity metrics will allow for the creation of a large scale classification system based on SVM models. We are currently implementing automated techniques to efficiently cluster very large numbers of landmark images. Using these clusters we will train a spatially organised databases of SVM models which will provide a very efficient and accurate technique for large scale landmark recognition.

6 Acknowledgements

This material is based upon work supported by the European Community in the TRIPOD (FP6 cr n 045335) project (www.projecttripod.org)

References

[1] Flickr. `http://www.flickr.com` .

[2] B. Ayers and M. Boutell. Home interior classification using sift keypoint histograms. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 0:1–6, 2007.

[3] H. Bay, T. Tuytelaars, and V. G. L. Surf: Speeded up robust features. *9th European Conference on Computer Vision*, pages 404–417, 2006.

[4] S. Cooray and N. O'Connor. A hybrid technique for face detection in color images. In *AVSS - International Conference on Advanced Video and Signal based Surveillance*, pages 253–258, 2005.

[5] D. J. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg. Mapping the world's photos. In *WWW '09: Proceedings of the 18th international conference on World wide web*, pages 761–770. ACM, 2009.

[6] L. S. Kennedy, S.-F. Chang, and I. V. Kozintsev. To search or to label?: predicting the performance of search-based automatic image classifiers. In *MIR '06: Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, pages 249–258, 2006.

[7] L. S. Kennedy and M. Naaman. Generating diverse and representative image search results for landmarks. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 297–306, 2008.

[8] J. Liu, M. Wan, and J. Zhang. Monocular robot navigation using invariant natural features. In *Intelligent Control and Automation, 2008. WCICA 2008. 7th World Congress on*, pages 5733–5738, June 2008.

[9] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60:91–110, 2004.

[10] J. Malobabic, H. le Borgne, N. Murphy, and N. O'Connor. Detecting the presence of large buildings in natural images. In *CBMI 2005 - 4th International Workshop on Content-Based Multimedia Indexing*, 2005.

[11] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 2:2161–2168, 2006.

[12] A. Popescu and P.-A. Moellic. Monuanno: Automatic annotation of geo-referenced landmark images. *ACM International Conference on Image and Video Retrieval*, 2009.

[13] P. Tirilly, V. Claveau, and P. Gros. Language modeling for bag-of-visual words image categorization. In *CIVR '08: Proceedings of the 2008 international conference on Content-based image and video retrieval*, pages 249–258, 2008.

[14] Y. S. Yan-Tao Zheng, Ming Zhao. Tour the world: building a web scale landmark recognition engine. In *Proceedings of International Conference on Computer Vision and Pattern Recognition*, 2009.