VISUAL AND GEOGRAPHICAL DATA FUSION TO CLASSIFY LANDMARKS IN GEO-TAGGED IMAGES

Mark Hughes, Noel E. O'Connor

CLARITY Centre for Sensor Web Technologies
Dublin City University
Ireland

Gareth F. Jones

Centre for Next Generation Localisation
Dublin City University
Ireland

ABSTRACT

High level semantic image recognition and classification is a challenging task and currently is a very active research domain. Computers struggle with the high level task of identifying objects and scenes within digital images accurately in unconstrained environments. In this paper, we present experiments that aim to overcome the limitations of computer vision algorithms by combining them with novel contextual based features to describe geo-tagged imagery. We adopt a machine learning based algorithm with the aim of classifying classes of geographical landmarks within digital images. We use community contributed image sets downloaded from Flickr and provide a thorough investigation, the results of which are presented in an evaluation section.

1. INTRODUCTION

Creating an algorithm to automatically recognise high-level semantic concepts, scenes or objects within digital imagery based on image content alone is a very challenging task and currently one of the most sought after goals in image retrieval. Computers excel at low-level processing of image data such that a query in an image retrieval system like 'retrieve all images within an image collection containing 30% blue pixels, 40% red pixels and 30% green pixels' is a trivial task. Humans rarely, if ever carry out search queries in this manner. A human is far more likely to provide a high level semantic query like 'retrieve all images of Paris Hilton's dog' or 'Eiffel Tower on Sunny Day'. These types of high level semantic queries are more challenging for a computer algorithm based on analysing pixels alone.

There is currently a limitation in the ability of computer algorithms to achieve this goal, a phenomenon known in the literature as 'The Semantic Gap' [1]. To circumvent the constraints of the semantic gap, we combine state of the art computer vision techniques with novel geographical contextual features with the aim of creating a machine learning based approach to high level semantic image classification. We focus on the classification of images containing commonly photographed landmarks.

The focus is on landmarks 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 450,000 images, Flickr search for Empire State returns over 370,000 images (June 2011)). Geographical features and landmarks have long been one of the most commonly photographed objects that tourists capture and commonly search for in image retrieval systems. Sanderson and Kohler [2] claim that almost one fifth of all web search engine queries had some geographical relationship, while Gan et al. [3] claimed that one in eight web queries contained the actual name of a specific location. For this study, we specifically focus on landmarks located within the Paris metropolitan region.

In recent years, as the extraction and representation of image features became more reliable, several techniques were developed to classify low-level semantic information from an image. Combinations of global low-level image features can be combined with classification techniques to infer basic information about the content of an image. For example, in the absence of EXIF information, colour based image features can be useful to determine whether an image was taken during the day or at night [4]. Successful low-level classification of semantics allowed for image retrieval systems to organise and return images based on more humanistic queries. For example, instead of returning images consisting of 30% red pixels and 70% blue pixels, a system could now be queried to retrieve images containing a sunset, snow covered landscape or perhaps a seascape scene. These types of semantics are getting closer to the types of queries that humans might make to a retrieval system.

Several successful classification methods have also been developed to recognise a variety of other low-level semantics such as the recognition of a cityscape (urban) or landscape (rural) scene [5]. Szummer and Picard [6] combined colour histograms with texture features to train a nearest neighbour classifier to recognise whether an image was taken indoors or outdoors. Vailaya et al. [7] trained a k-Nearest Neighbour classifier to group images cityscape, landscape, forest, mountain and sunset classes among others.

While low-level semantic classfiers have been shown to work well, there still is a significant gap when it comes to automatic high level semantic understanding of an image. In this paper, we aim to address this gap through the use of machine learning algorithms and additional contextual information that is available within geo-tagged images.

We hypothesise that it is possible to improve upon visual features alone to classify landmarks by combining them with geographical features extracted from associated geo-tags and publicly available geographical datasets. Accurate semantic classification can benefit image retrieval systems, particularly those based on tag based approaches, such as Flickr, due to issues with noisy human defined metadata [8].

2. LANDMARK CLASS CLASSIFICATION

Understanding the content of a scene depicted within an image is one of the core goals of computer vision. The aim is to convert the pixel data contained within an image into one or more high level semantic descriptions of the scene or event that is displayed in an image. A high level semantic description could be described as a detailed and meaningful representation of the content of an image, which would be relevant to a human observer (or perhaps a description that could be converted by a computer so that it would be relevant to a human observer). High-level semantic image classification is still a very open problem in the computer vision community, particularly in unconstrained environments. In recent years however, much progress has been made in image classification at a lower semantic level, such as the ability to classify images into different categories of scenes.

Following a machine learning approach, it was hypothesised that it would be possible to classify an image of a landmark into one of a finite number of visually distinct categories. The motivation behind this is that if accurate semantically relevant groupings of landmark images could be achieved, it would be possible more accurately annotate landmark images in a text based image retrieval system such as Flickr. Eight different classes of landmarks were chosen that had a high representation within the corpus, and could be suitable for classification using machine learning algorithms. These were:

- **Artwork** The artwork class is defined as images (that contain a painting or drawing) taken inside an art gallery or museum. From an informal empirical study of the Paris corpus, it is evident that many Flickr users commonly photograph paintings, and several well known pieces of art could be considered landmarks.

- **Bridge** Another very commonly photographed landmark is a bridge. Many iconic bridges span the river Seine and many of the canals that flow through the Parisian region, and due to their unique visual appearance and photogenicity, large numbers appear in the training corpus. In this work, a bridge is defined as a man-made object that spans across a body of water, a road or a railway track.

- **Building Facade** A building facade is a category containing the main facade of a large building. If there is no notable facade, for example in the case of an office block or a skyscraper, the facade is considered to be any side of the building.

- **Fountain** A fountain is defined as a man made object that sprays or pours water either into the air or into a man made reservoir. Although originally used for human water consumption purposes, today fountains are mainly used for ornamental purposes.

- **Monument** The category of monument is quite nebulous and can refer to a large number of objects. In this work, a monument is considered a man-made structure that does not have a use as a dwelling place (such as a building) and does not contain a large statue or sculpture. Some examples of monuments in the image corpus are; the 'Eiffel Tower' and the 'Arc de Triomphe'.

- **Church:** A Church is defined as a place where a Christian might practice their religion, such as a church, cathedral or a chapel. This category is concerned solely with images that were taken outside of the structure.

- **Church(Indoor):** The church indoor is defined as an image that was photographed inside a Christian place of worship. These commonly include close up images of stained glass windows, church ornaments and altars.

- **Statue** A statue is defined as a sculpture that usually represents a person or historical event. Additionally, a sculpture within an art gallery or museum will fall into this category.

- **Other** Any landmark that does not fall into one of the above categories is defined in this class.

Although there is a large amount of variance in intra class visual similarity within each of these categories, many different landmarks within a class share some basic characteristics. Take, for example, the class 'Church', which includes churches and cathedrals, among others. In many cases, a human observer could quickly recognise a church as being a church irrespective of the size of the landmark or the architecture style, as illustrated in Figure 1. Whether a church was built in the Gothic style, such as the famous Notre Dame Cathedral, or in a more modern style such as the Sagrada Familia in Barcelona, many humans could identify from visual recognition that these structures are places of worship. This recognition could be based on knowledge obtained in their lifetime using the visual style of other visually similar places

Fig. 1. *An example illustrating many different examples in a landmark category 'Church'. Although there is a lot of visual intra class variation, it will still be possible, based on visual information alone for many human observers to quickly classify all of these images as either being churches, chapels, cathedrals or mosques.*

of worship, which could be considered analogous to supervised learning. It is logical, therefore, to assume that these two structures share enough characteristics visually, for a human observer to predict the category of both structures without heterogeneous knowledge. It is based on this premise, that a suite of classification models was implemented with the aim of grouping landmarks into a finite set of categories.

3. VISUAL SEMANTIC CLASSIFIER

We determine landmark classification as a pattern recognition problem and adopt a supervised learning approach. Given pattern x extracted from an image i, the aim is to obtain a probability measure, which indicates whether a semantic landmark class is present in image i. We use Support Vector Machines (SVM) to obtain this probability measure based on the RBF kernel function.

3.1. Support Vector Machines

A SVM is a learning algorithm originally developed by Vapnik [9] that can perform input/output mappings from labelled examples and can choose a balanced capacity for each decision function. SVMs have been widely used in many different research genres and are highly regarded for scaling well with high dimensional data [10].

3.1.1. Linear Classification

The main aim of an SVM is to seperate classes of data with the use of a hyperplane. The general equation for a hyperplane H is

$$H = w \cdot x_i + b >= 1 \text{ where } y_i = +1$$

and

$$H = w \cdot x_i + b <= -1 \text{ where } y_i = -1$$

where x is an input point (a vector) lying on the hyperplane, w is a set of weights (also a vector) and b is a constant. H_1 and H_2 are two hyperplanes, that are parellel to H where

$$H_1 = w \cdot x + b = 1$$

and

$$H_2 = w \cdot x + b = -1$$

The points that lie along the hyperplanes H_1 and H_2 are the closest points to the the hyperplane H and are called the support vectors. The support vectors are the critical elements of the training set as they are the input features that would influence the position of the dividing hyperplane decision if removed from the dataset. Distance $d+$ is defined as the distance from H to the closest positive point, while distance $d-$ is defined as the distance from H to the closest negative point. The margin of the separating hyperplane is defined as $d-$ + $d+$. This margin can be calculated as $2/||w||$.

The main aim of SVMs is to create a hyperplane with as large a margin as possible, i.e. optimise w and b so that $2/||w||$ is maximised, which is the equivalent to minimising $\frac{1}{2}||w||^2$. A maximum margin hyperplane ensures a higher certainty level of correct classification, as points located near the decision plane represent unpredictable classification decisions. A classifier with a maximum margin will make much fewer of these low certainty decisions. This provides a slight margin of error within the classification procedure. A noisy variable will not cause a classification error.

In this work, a multi-class SVM model, based on the one versus all paradigm, was trained to classify images into one of these nine categories.

3.1.2. Radial Basis Function Kernel

In situations where complex data is not linearly separable, it might be possible that a transformation of this data into a higher dimensional space could result in a linearly separable model, where the linear based SVM approach described above could then be applied. The function behind this transformation is referred to in the literature as the kernel function. Several different kernel functions were evaluated but it was the RBF kernel that performed best for this task. This function takes a parameter called gamma (g) that defines how the influence of each support vector. A large gamma value will enable a support vector to have a stronger influence over a larger area, which in turn can lead to a smaller number of support vectors in each classifier. With stronger influence over larger areas, fewer support vectors are required to define a boundary. In this work, optimal values for g were defined through *k-fold cross validation*. The RBF kernel is formally defined as:

$$K(x, y) = exp(-g||x - y||^2)$$

4. IMAGE FEATURES

As inputs into our SVM model, we use a two of state of the art computer vision features:

4.1. Edge Histogram Descriptor

One of the most commonly used texture features in the MPEG7 standard is the Edge Histogram Descriptor (EHD). The EHD is a global based feature vector containing spatially organised histograms of edge orientations detected within an image. It is based on the measurements of four directional edges (vertical, horizontal, 45° and 135°) and one non directional edge.

When extracting image features to represent objects/landmarks within an image, it is preferable to have some division of the image into meaningful regions that are relevant to the actual objects/landmarks depicted. Once this division has been calculated, features can be calculated for each region, which allows for the inclusion of local information to be embedded into these features.

Traditionally, global based features were calculated based on the whole content of an image. The main disadvantage of this is that all geometrical information is disregarded regarding the layout of the extracted features.

Block based segmentation, is the process of partitioning the image into blocks or regions, each one a predetermined size, and calculated in a defined manner. Each of these blocks is then treated as separate entity for the purposes of feature extraction and the geometrical information regarding the regions location and relationship to other regions can be preserved in the feature descriptor. This information can provide a weak form of geometric consistency when comparing and matching features from multiple images. The MPEG7 edge histogram feature utilises a block based segmentation scheme.

To calculate the feature, firstly the image is partitioned into 4×4 (16) equal sized sub images, the width and height of each block is $W/4$ and $H/4$ respectively, where W and H represent the overall width and height of the image. Each of these sub images is then treated as a separate entity. Irrespective of the size of the image, each of these blocks are further divided into 1100 smaller block. Each of these smaller sub blocks are then processed with a suite of 5 oriented edge detectors (0, 45, 90, 145 and non-directional). The sub block is then marked as the orientation that had the maximum edge strength outputted from these edge detectors, if above a threshold, if not, the block is disregarded. For each original larger sub block (16 in total) the average numbers of edges in each orientation is histogramed into 5 features. As this process is repeated for each larger sub block, this gives a total of 80 values, to create the global EHD (see Figure 2 for example).

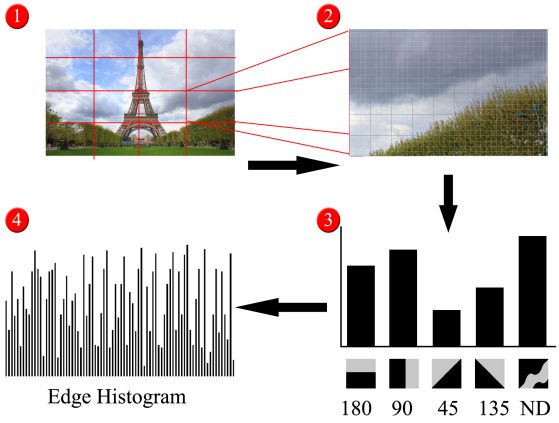

Fig. 2. *An illustration displaying the process of extracting an edge histogram feature from an image. Firstly the image is split into 16 sub images (1), followed by the further block segmentation of each of these sub blocks to 1100 much smaller blocks (2). A histogram is created for each large sub block, containing 5 values (3). All of these smaller histograms are merged into one global histogram (4).*

4.2. Visual Bag Of Words

Bag of words (BOW) models have been used in document classification successfully in the past [11]. A BOW model is a technique where a document is represented as an unordered collection of words that are then used to classify a document based on these representations. Visual bag of words (VBOW) features are based upon the same basic premise, however the bag of words is replaced by a bag of descriptions of image patches. These image patches can be identified from a sample set of images using a variety of approaches such as dense sampling, random sampling or using an interest point detection algorithm, in this work we use the SURF algorithm [12]. Descriptor vectors are then processed for each of these image patches. A collection of these descriptors is referred to as a visual vocabulary or a codebook.

Once a codebook is created, the VBOW approach provides an efficient method to quantise large numbers of image descriptors. Each image is represented as a bag of visual words that are created based on the presence of visually similar image descriptions of salient regions in an image and contained within the visual vocabulary [13].

There are several steps involved in creating a VBOW model:

- Local image feature descriptions are extracted from each image or from a subset of images within the dataset.
- These image features are then quantised into a visual vocabulary using a k-means clustering algorithm with k being the vocabulary size of the dictionary.
- Using this vocabulary each image can then be represented by a global histogram value that is calculated by

comparing each image feature to every feature in the dictionary and a vote is counted for the entry in the dictionary that has the smallest distance from the image feature. The histogram forms a vector where the number of possible words is the length of the feature vector.

This VBOW model effectively quantises large numbers of image features into a single feature vector while retaining a high level of discrimination. A VBOW histogram is an orderless image feature, in that the order of feature values is not determined in advance and has little or no impact of classification/matching accuracy. We experiment with three values for k, 1024, 2048 and 4096.

4.2.1. Soft Assignmentof Visual Word Features

Traditionally in the VBOW model, image features are assigned to their closest neighbour in the vocabulary and only their closest neighbour. This assignment process is referred to as 'hard assignment' and can be formally defined as:

$$Hist(i) = \frac{1}{n}\sum_{i=1}^{n}\begin{cases}1, & \text{if } i = argmin(Dist(v, p_j)\\ 0, & \text{otherwise}\end{cases}$$

where n is the number of interest points extracted from an image, p_j represents an interest point j extracted from an image, and $Dist(v, p_j)$ represents the distance between a vocabulary word i and an interest point j. This hard assignment model has many disadvantages in that for each input feature, it's similarity is only considered for the closest neighbour in feature space. This model disregards all other features that could be also quite similar, some of these might only be marginally further away in feature space than the nearest neighbour. Clearly this approach is not ideal as relevant information that can aid discrimination of each feature is simply disregarded (see Figure 3).

One method to atone for this shortcoming, is to utilise an approach to feature assignment based on the similarity of features to each vocabulary word, called soft assignment. This is where each input can be assigned to k bins in a histogram, where k that represent it's nearest neighbours in feature space. Usually it is desirable to set the amount of the value assigned to each to be directly proportional to how close the input feature is to each of its neighbours. This ranking based feature essentially calculates the proportionality of the score assigned to the k nearest neighbour bins, by using the position of each bin in a ranked list (of length k) of nearest neighbours to designate the score. The ranking based function that we use in this work can be formally defined as:

$$Hist(k) += \frac{1}{2^i - 1}$$

where i is the position of the bin k in the ranked list of nearest neighbours.

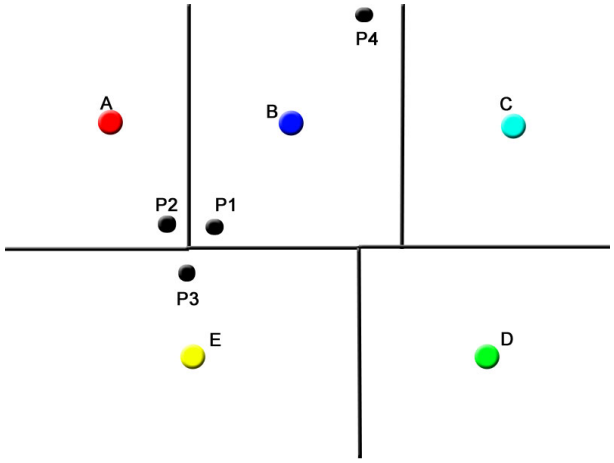

Fig. 3. *An illustration outlining the advantages of the soft assignment of visual word features. This diagram presents a hypothetical partition of a visual word vocabulary containing 5 visual word features A-E, and four feature points to be assigned to a visual word cluster center, P1-P4. It can be clearly seen that points P1, P2 and P3 are quite close together in feature space, however, using a hard assignment approach would not take into account the similarity between these features and they would never be matched. P1 would only be associated with the visual word B, while points P2 and P3 would only be associated with the visual words A and E respectively. Using hard assignment the only point to be matched to P1 would in fact be P4, even though P2 and P3 are closer in feature space. Using soft assignment the points P1, P2 and P3 would be assigned to each the visual words A, B and E albeit with different weights, which are calculated based on the distance from the visual words. This would allow these features to matched as they are closer in feature space.*

5. DATASETS

A collection of training and testing images was collected for this purpose. The training collection was gathered from two sources. The first source was the SUN image dataset [14], which is a large scale collection of images categorised into 899 scene categories. Of these 899 scenes, 7 were deemed useful for the purposes of this work. These included: bridge, building facade, church (outdoor), church (indoor), fountain and statue.

The other source used to gather data for the training set was the Flickr API. For 8 of the 9 semantic categories, the Flickr API was queried using the category name as the query text. All retrieved images were manually analysed and if they conformed to the category class, they were added to the training data. In total the training collection consisted of 3886 images:

- Artwork - 246 images

- Bridge - 562 images
- Building - 625 images
- Church - 480 images
- Church (Indoor) - 616 images
- Fountain - 709 images
- Monument - 185 images
- Other - 155 images
- Statue - 308 images

6. LANDMARK CLASS CLASSIFICATION WITH COMMUNITY CREATED GEOGRAPHICAL DATA

Visual information can be useful when classifying low-level semantic information from digital images [6], however it is more difficult to infer high level semantics. From results in Figure 4, it is evident that global based image features alone are insufficient for accurate classification across all semantic classes. To overcome this, it is hypothesised that utilising geographical contextual information will help to bridge this 'Semantic Gap'. There now exist rich geographical databases, accessible online, that contain high level semantic information describing a specific region. In this section, it is proposed that by fusing visual and geographical information, it would be possible to classify an image into a high level semantic landmark category with a higher degree of accuracy than if using visual information alone.

In recent years, there has been a surge in the creation and dissemination of information on-line by large communities of contributors. One particular type of information accessible online includes geographical data. Large numbers of websites have recently been created that enable for the creation of large scale semantic databases describing geographical locations.

In this work, a database containing geographical points of interest (POI) was created. This consisted of a number of objects referenced by geographical location, which were harvested from two online sources. A technique was proposed to classify an image into one of the 9 landmark categories based on the objects stored in this POI dataset.

6.1. Open Street Map

One example of an online geographical community is Open Street Map. Open Street Map is an online repository where community contributors upload the spatial coordinates of a wide range of geographical entities, along with semantic data describing these entities. With a large community of users, these present a very valuable resource for research communities across several fields.

Human contributors can upload map data which is represented by lists of waypoints. Each waypoint contains latitude and longitude coordinates. Users can also upload geographical objects, otherwise known as 'Points of Interest' (POI), and assign them a location. OpenStreetMap has a strict set of guidelines to ensure that uploaded data is accurate. Each uploaded POI can be assigned one of a finite number of feature classes dependent on the use and attributes of the feature. In this work 8 different feature classes were selected to coincide with the set of semantic classes desired to be classified in this work. These feature classes were:

- Bridge
- Building
- Fountain
- Gallery
- Monument
- Museum
- Place of Worship
- Statue

All of these feature classes located in the Paris region were downloaded and stored in the POI dataset.

6.2. GeoNames

Another online resource that contains accessible geographical data is the GeoNames repository located at geonames.org. GeoNames is an online geographical repository that contains over 10 million geographically mapped location names, along with 7.5 million geographical features. These features are split into 9 feature classes, which are then split into 645 feature types. Of these feature types, 5 were deemed relevant to the set of landmark classes outlined in section 6.3. Each feature type is associated with a set of metadata, including geographical coordinates, a code representing the country, and the name of the geographical feature. Each of these feature types is considered to be a POI. GeoNames data has been gathered from many reputable sources, including the United States Geological Survey, Netherlands Statistics Office, and the French National Institute of Statistics and Economic Studies, it is therefore expected that this data is quite accurate.

Using the publicly available API, all geographical features located within Paris associated with a set of feature classes was retrieved and stored in a database. This set of feature classes included:

- Bridge
- Building
- Church
- Monument

- Museum

It must be noted that the GeoNames data collection is by no means a comprehensive list for each geographical feature. For several of the features retrieved, the data was quite sparse. For example, for the class Church, only 15 geographical features were found. It must be noted that the majority of geographical features that populate the dataset tend to be well known landmarks, which could be beneficial for this work as these are the objects that users are most likely to visit and photograph. All of these feature classes located in the Paris region were downloaded and stored in the POI dataset. In total, the OpenStreetMap and GeoNames data combined comprised of 1235 POIs.

7. CLASSIFICATION USING GEOGRAPHICAL DATA

To analyse the effectiveness of community geographical data to classify landmark classes, the test collection of images was processed based on a nearest neighbour scheme. Each test image within our collection was gathered from Flickr and was geo-tagged within the Paris metropolitan region. The location information from each test image was extracted and all POIs within a radius of 250 metres were retrieved from the POI database. Retrieved features were then ranked according to geographical distance, using the Haversine formula, with the shortest distance ranked at the top. This top ranked feature was then assigned to the test image.

It is assumed that the POIs 'Gallery' and 'Museum' might be useful to classify the semantic class 'Artwork', due to the likelihood of pieces of art appearing in both of these locations. If the closest POI to an 'Artwork' test image is 'Museum' or 'Gallery' then this images is marked as being correctly classified. Similarly for the semantic class 'Statue', it is assumed that there is a correlation with the POI class 'Museum'.

7.1. Fusion of Visual and Geographical Features for Semantic Classification

In this section, experiments were carried out that fused the visual and geographical data. Two fusion approaches were implemented, one based on the presence of a POI in the vicinity of a test image and the other based on the distance between a test image and nearby associated POIs.

7.1.1. Presence of POI Approach

The first fusion technique was based on combining the output values from the SVM classifier with a static value to represent whether a landmark class was present in the POI database. There was no weighted measure applied to this value, and all landmark classes detected within a spatial radius had this value added to its corresponding output from the classifier.

A minor change was made to the libSVM library to output an array of confidence measures C, with a value representing each landmark class $C_1....C_n$ (where n is the number of landmark classes). If the presence of a landmark class was found in the database within a spatial radius of a test image (defined to be 250 metres), a value v was added to c_i, where i is the associated landmark class. Therefore if a POI was discovered in the database associated with the landmark class i then C_i becomes $C_i + v$.

The values in C are normalised into the range 0-1. A value for v is selected based on the maximum value in C, ie. $v = argmax(C)$. Several variations of this calculation were evaluated, some providing a weighted bias towards the visual data and others providing a weighted bias towards the geographical data. Three weighted variations of $v = w \times argmax(C)$ were evaluated, where w is equal to:

- 2 (denoted as weight 1 in the evaluation)
- $\frac{1}{2}$ (denoted as weight 2 in the evaluation)
- $\frac{1}{4}$ (denoted as weight 3 in the evaluation)

A value of 2 for w weights the metric in favour of the geographical data. Values of $\frac{1}{2}$ and $\frac{1}{4}$ for w weight the metric in favour of the visual data.

7.1.2. Weighted Distance Approach

Similarly to the first approach, the second method added a value to relevant confidence measures outputted by the SVM model. The weight of these values was determined by the distance from a POI to the test image. Landmark classes that were nearby had a higher weight assigned to them than those they were located further away. As with above, an array of confidence measures C with a value representing each landmark class $C_1....C_n$ was output from the SVM model.

If the presence of a landmark class i was found in the database within a spatial radius of a test image, a value v was added to C_i. The value v is determined by calculating the distance, denoted as $dist$, between i and a test image t, that was calculated using the Haversine formula. Therefore for each POI class that was located within the geographical radius C_i becomes $C_i + (1 - dist(t, i))$ where $dist(t, i)$ is normalised into the range 0 - 1. Four weighted variations of the metric $C_i = C_i + w(1 - dist(t, i))$ were evaluated, where w is equal to

- 1 (denoted as weight 1 in the evaluation)
- $\frac{1}{2}$ (denoted as weight 2 in the evaluation)
- $\frac{1}{4}$ (denoted as weight 3 in the evaluation)

- $\frac{1}{8}$ (denoted as weight 4 in the evaluation)

8. EVALUATION

8.1. Visual Semantic Classification Evaluation

To evaluate the classifier, a test collection of images was collected. All of these images were retrieved from Flickr using their corresponding landmark class as the query text. In total for each landmark class 100 images was collected. Each of these images contained geographical data and had been photographed in the Paris region.

All of the test images were processed through the multi-class semantic classifier with a variety of different input features:

- MPEG7 Edge Histogram
- Visual Bag of Words (Hard Assignment) $k = 1024$, $k = 2048$, $k = 4096$
- Visual Bag of Words (Soft Assignment) $k = 1024$, $k = 2048$, $k = 4096$

The results of this evaluation can be seen in Figures 4 and 5. If selecting a baseline classification score based on random selection, it would be expected that a correct selection could be achieved around 11% of the time. Therefore, on average the visual classifiers performed significantly better than the baseline.

As expected, some landmark classes could be classified more successfully than others. The highest performing class was 'Church (Indoor)', which achieved an accuracy score of 88% correct. From informal inspection, the intra class visual variation in this class was deemed to be the lowest across all the classes. The class with the highest level of intra class visual variation, 'Monument' performed very poorly.

A vocabulary size of 2048 performed best for this task. Interestingly, there was a large improvement when using soft assignment as opposed to hard assignment. From these results, it is evident that visual information alone does not allow for an acceptable classification accuracy across all classes.

8.2. Geographical Fusion Evaluation

Experiments were carried out to ascertain how accurately a fusion approach (visual and geographical) would perform, using the same test collection outlined in 8.1. From the results in Figure 5, it would seem that the fusion of geographical and visual data for classifying images into semantic landmark categories improves performance slightly over using either visual or geographical features alone for a subset of the landmark classes. On average however, the fusion of visual data with geographical data hinders performance over using visual features alone. The main reason behind this is the sparsity of the geographical database. For many of the landmark classes, there was insufficient data and visual confidence measures were being decreased to the extent that other landmark features that populated the dataset were being incorrectly classified.

To illustrate this point with an example, it can be seen in Figure 5, geographical data alone works well for the concept class 'Artwork' but performs very poorly for other concepts, such as 'Building' for example. It would appear that the general poor performance of geographical data is down to the sparsity of the datasets. In the example of the concept 'Artwork', there are very few locations within the city where one would expect to find geo-tagged community images of this concept, possibly less than a dozen (restricted to museums and art galleries). From the geographical data, it can be seen that the largest museum and largest art gallery in Paris(La Louvre and the Musee D'Orsay) are included in the geographical dataset. For the concept 'Building', however, one would expect to find images in a wide variety of locations across the city. It is logical to assume that the majority of images within the test set of 'Artwork' were geotagged at one of these locations. Additionally, the significant improvement in accuracy that is garnered from the fusion approach over visual features alone for the concept 'Artwork' would imply that with a comprehensive, accurate geographical dataset, it might be possible to classify all well represented concepts with a high degree of accuracy.

Overall, while the geographical fusion approach hindered classification accuracy, we believe that this was down to the sparsity of the geographical datasets being used. In situations where the visual feature alone performed poorly and there was sufficient data available, the geographical fusion approach could be used to improve performance significantly. We believe that these experiments demonstrate promising results and are encouraged that as the density of the online community contributed datasets increases, the fusion based approach would increase in classification performance in parallel.

9. REFERENCES

[1] Arnold W. M. Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain, "Content-based image retrieval at the end of the early years," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1349–1380, 2000.

[2] Mark Sanderson and Janet Kohler, "Analyzing geographic queries," in *Proc. of the Workshop on Geographic Information Retrieval*, 2005.

[3] Qingqing Gan, Josh Attenberg, Alexander Markowetz, and Torsten Suel, "Analysis of geographic queries in

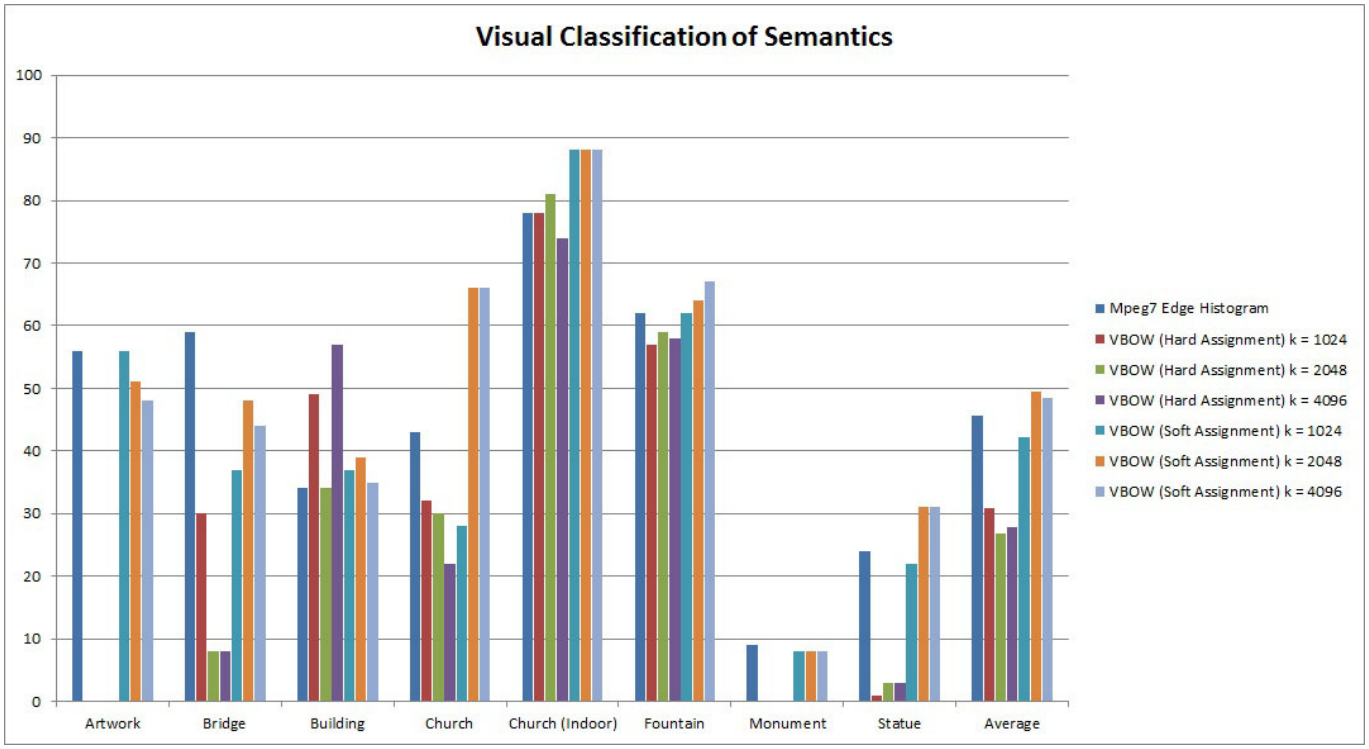

Fig. 4. *The results of the evaluation of the visual semantic classification experiments. An array of input features were evaluated, which as visible to the right of the chart*

a search engine log," in *LOCWEB '08: Proceedings of the first international workshop on Location and the web*, New York, NY, USA, 2008, pp. 49–56, ACM.

[4] Stefan Kuthan and Allan Hanbury, "Hierarchical image classification," *imageval.org 2006*, 2006.

[5] Rong Yan, Yan Liu, Rong Jin, and Alex Hauptmann, "On predicting rare classes with svm ensembles in scene classification," in *In ICASSP*, 2003, pp. 21–24.

[6] M. Szummer and R. W. Picard, "Indoor-outdoor image classification," in *Content-Based Access of Image and Video Database, 1998. Proceedings., 1998 IEEE International Workshop on*, 1998, pp. 42–51.

[7] Aditya Vailaya, Anil Jain, and Hong Jiang Zhang, "On image classification: City images vs. landscapes," *PATTERN RECOGNITION*, vol. 31, pp. 1921–1935, 1998.

[8] Mark Hughes, Gareth J. F. Jones, and Noel E. O'Connor, "A machine learning approach to determining tag relevance in geotagged flickr imagery," in *WIAMIS: The 13th International Workshop on Image Analysis for Multimedia Interactive Services*, 2012.

[9] Corinna Cortes and Vladimir Vapnik, "Support-vector networks," in *Machine Learning*, 1995, pp. 273–297.

[10] Tsau Lin and Tam Ngo, "Clustering high dimensional data using svm," in *Rough Sets, Fuzzy Sets, Data Mining and Granular Computing*, Aijun An, Jerzy Stefanowski, Sheela Ramanna, Cory Butz, Witold Pedrycz, and Guoyin Wang, Eds., vol. 4482 of *Lecture Notes in Computer Science*, pp. 256–262. Springer Berlin / Heidelberg, 2007.

[11] K. Torkkola, "Discriminative features for document classification," in *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, 2002, vol. 1, pp. 472–475 vol.1.

[12] Herbert Bay, Tinne Tuytelaars, and Van Gool L., "Surf: Speeded up robust features," *9th European Conference on Computer Vision*, pp. 404–417, 2006.

[13] Pierre Tirilly, Vincent Claveau, and Patrick 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*, New York, NY, USA, 2008, pp. 249–258, ACM.

[14] Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba, "SUN database: Large-scale scene recognition from abbey to zoo," in *2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2010, pp. 3485–3492, IEEE.

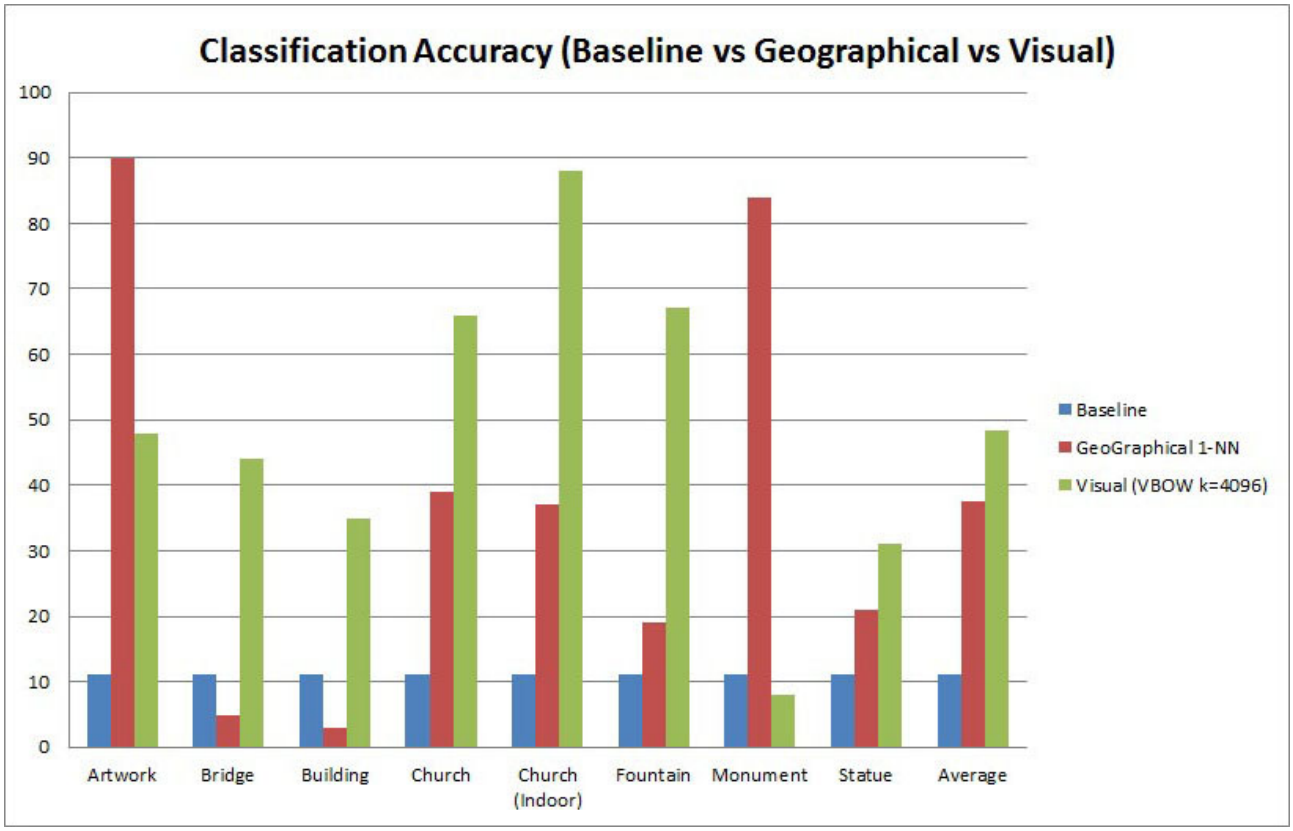

Fig. 5. *A chart comparing the classification accuracy of geographical information and visual information when classifying images into semantic landmark classes. Both are compared against the expected baseline.*

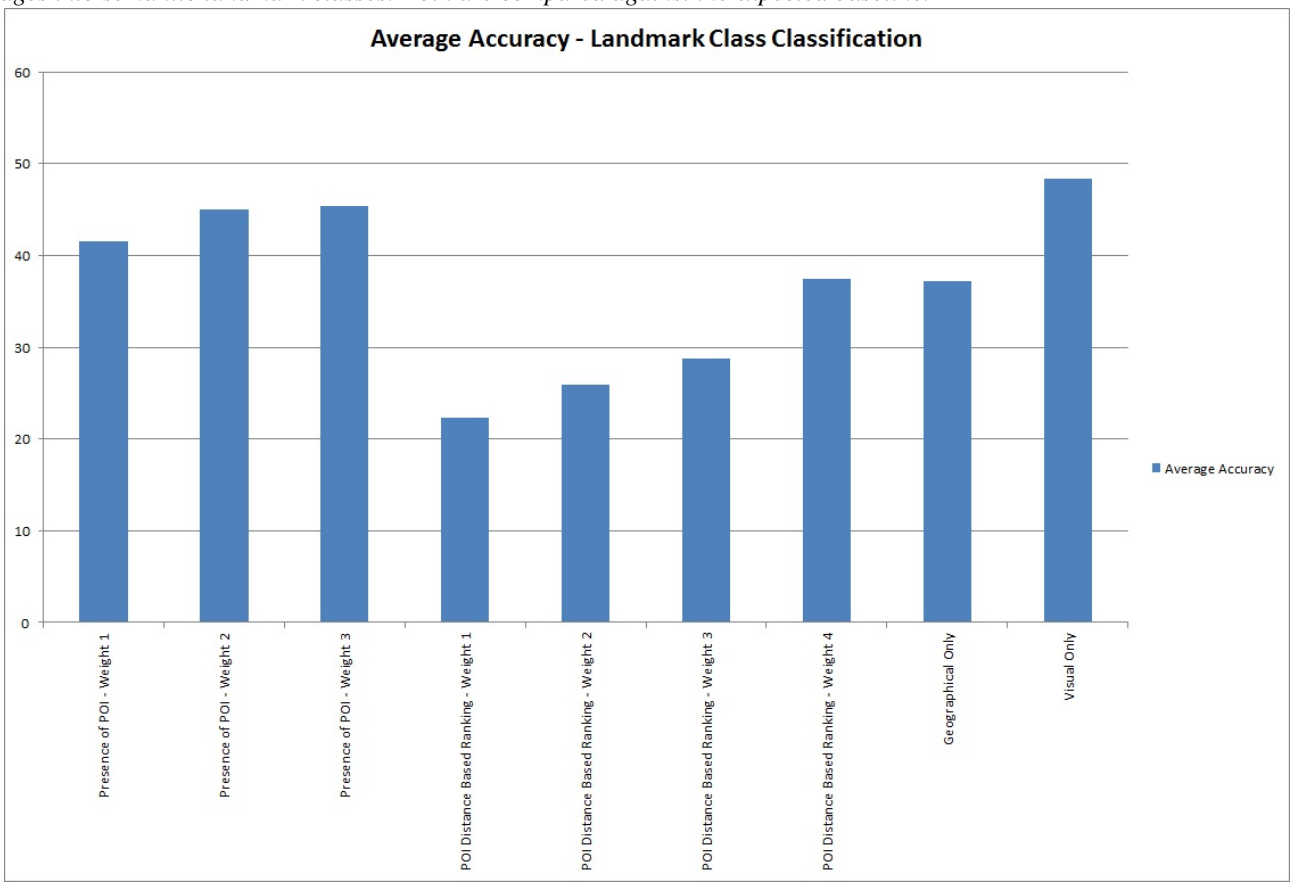

Fig. 6. *A chart comparing the classification accuracy of hybrid approaches to landmark classification against approaches based on geographical and visual information*