

# The Intelligent Image

Max Jaderberg

Keble College, University of Oxford

Trinity Term 2012

## **Abstract**

In this report, I shall discuss some awesome stuff.

# Chapter 1

## Introduction

The web contains billions of images, which are often the focus of attention on the web pages that include them. However, there is almost no information about the content of these images. Images on websites are purely binary data files, occasionally with some associated meta data included in the image's HTML<sup>1</sup> code. Very little information is available about the image, let alone the objects or scenes contained within the image. The aim of this project is to create a system which automatically recognises the objects within these images, thus releasing the information within them. This is a large scale object retrieval problem.

There are a large number of applications which would benefit from having detailed information on the contents of images. With more knowledge on the objects contained within images, one can create more effective search engines, better cataloguing and classification systems, user interfaces which engage viewers more, and relevant advertising based on the content. Novel applications could also be built, for example, which retrospectively embed geographical information in the image binary by recognising where the image was photographed. The plethora of useful applications provides a great deal of motivation for this project.

The result of the project is that a query image is inputted, the objects contained within the image are recognised, and the image is returned with the recognised objects "tagged" i.e. the region of the object in the image is outlined and it can be clicked to give relevant information. In this way the standard image has been transformed into the "intelligent image" - one which knows about the contained scene and can offer up information to the consumer about it's contents.

---

<sup>1</sup>Hyper Text Markup Language <http://en.wikipedia.org/wiki/HTML>



Figure 1.1: An example “Intelligent Image” showing the various hover over states

Recognising a wide range of objects in an image is not a trivial problem. For such a system to be useful the following needed to be addressed:

- Acquire and filter enough data to create a model to perform matching on a large range of objects.
- Create a retrieval system which provides accurate matching. False positive matching needs to be avoided as much as possible.
- Ensure matching can be done quickly on a database of millions of objects.

To recognise millions of different objects, reference images are needed to create a model to match against. From the outset, Wikipedia<sup>2</sup> is used as the primary model data source. Wikipedia is a crowd-sourced online encyclopaedia with many images contained within the articles, and is considered an accurate source of information. In the system, each Wikipedia page defines an object, with the images in the article being used to provide the data to match against. The content of the Wikipedia article is used to give the user further information about the object. A web crawler was written to extract the relevant images of desired Wikipedia pages and create the image database. Filtering is also performed to ensure only useful images are included in the database. The data sources are fully explained in Chapter 3.

Object retrieval uses a method employing a bag-of-words model. This builds upon the work described in x and y, more information of which is provided in Chapter 2. The visual words in the image database from Wikipedia are precomputed. At run time, the words in the query image are computed and searched against the database of precomputed words to find the top image matches. The top matches are spatially verified, with the first verified image being the result. Subsequent improvements to the base line system

---

<sup>2</sup><http://en.wikipedia.org>

(described fully in Chapter 4) were made, including geometric improvements to the spatial verification and descriptors used to increase speed and accuracy of spatial verification, and matching improvements through query expansion using crowd-sourced data (dubbed “Turbo-boosting”) from Microsoft’s Bing<sup>3</sup> search engine. The geometric improvements and turbo-boosting is reported in Chapter 6 and Chapter 7 respectively.

Finally a website was developed to provide a front end interface for the system. A user can simply navigate to the website and upload an image. They then start the automatic tagging process, during which a realtime log of the process is displayed. After the query is complete, the intelligent image is displayed, which the user can interact with, see the names of the objects contained within the image, and click on the object to go to it’s Wikipedia page. The software architecture of the website and the backend systems is explored in Chapter 5. The result is a realtime automatic tagging system which could recognise, for example, all the buildings in a tourist’s photo album of London in seconds.

The aim of the project is to work towards recognising every object on Wikipedia, however due to time restraints a subset of objects was used for development and testing of the project. The subset chosen were the pages that appear on the Wikipedia page “List of Structures in London”<sup>4</sup>.

---

<sup>3</sup><http://www.bing.com>

<sup>4</sup>[http://en.wikipedia.org/wiki/List\\_of\\_structures\\_in\\_London](http://en.wikipedia.org/wiki/List_of_structures_in_London)

## **Chapter 2**

# **Background**

## Chapter 3

# Data

There are three main datasets used by the application: the images from Wikipedia used to build the database of objects (Section 3.1), the images from Microsoft Bing used for the Turbo-boosting (Section 3.2) and the images from Google Images used for validation and testing (Section 3.3). All images that are used are resized so that their larger dimension does not exceed 1000 pixels to reduce storage space and provide homogeneity. This chapter describes the various datasets and how they are acquired. Table 3.1 gives an overview of the data used.

	Source	# Images	# Classes
Model images	Wikipedia	3963	732
Turbo-boosting images	Microsoft Bing	18273	732
Validation images	Google	701	294

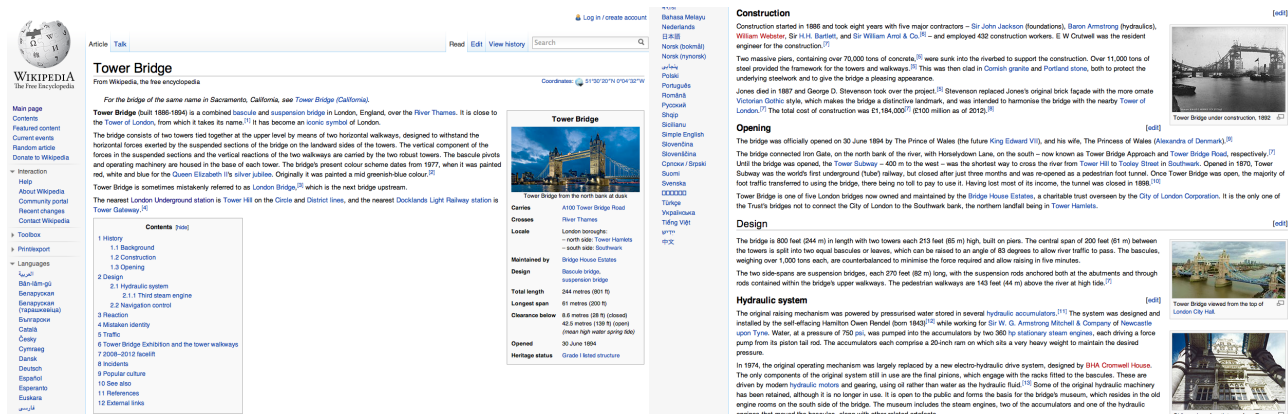
Table 3.1: A summary of the datasets.

### 3.1 Model Images

The model comprises of a dataset of images that depict the objects that are to be able to be recognised.

Each page of Wikipedia that contains images represents an object which can be matched. The database of images which is used to build the model is simply created by visiting each page on Wikipedia for the

objects desired and downloading the relevant images contained on the web page, labelling those images as being associated with the object. A script automates this process of building the model database.



(a)

(b)

Figure 3.1: The Wikipedia page for “Tower Bridge”. Note the images contained are those used in the model to represent this object.

To automate the downloading of object images from Wikipedia, Python<sup>1</sup> is used. Wikipedia offers a public application programming interface (API) over HTTP to access its data, as it is built on the MediaWiki framework<sup>2</sup>. However it is cumbersome and not easy to consume. Instead, a web crawler was written to explore Wikipedia pages and extract the relevant images.

A crawler object in Python finds all the images and notes the URLs of them for subsequent download. Firstly, the HTML of the Wikipedia page must be downloaded, as it appears to a web browser (an example of a web browser being Google Chrome). However, Wikipedia does not allow crawlers and automated bots to access its web pages. To overcome this, the HTTP header<sup>3</sup> of the crawler is edited to emulate that of a browser. This is implemented using the urllib2 library<sup>4</sup>. The code shown in Listing 3.1 shows an example of how to read the HTML of the main Wikipedia homepage. The HTML document for each Wikipedia page is parsed using the BeautifulSoup library<sup>5</sup>. All the anchor elements are found and stored for further

<sup>1</sup><http://www.python.org>

<sup>2</sup>The MediaWiki framework was originally developed for Wikipedia and provides an API over HTTP as standard. <http://www.mediawiki.org/wiki/API> provides documentation for the API.

<sup>3</sup><http://www.w3.org/Protocols/rfc2616/rfc2616.html> describes the Hyper Text Transport Protocol and the various header fields.

<sup>4</sup><http://docs.python.org/library/urllib2.html>

<sup>5</sup><http://www.crummy.com/software/BeautifulSoup/>



crawling. The images contained within the HTML are also found by looking within the part of the HTML document that is unique to the specific Wikipedia page (see Listing 3.2).

Listing 3.1: The code used to gain access to Wikipedia's content using a crawler by emulating a browser.

```
## Python
import urllib2
# Emulate the user agent as that of a browser
user_agent = "Mozilla/5.0 (Macintosh; U; Intel Mac OS X; en-US; rv:1.8.1.7)
             Gecko/2007091417 Firefox/2.0.0.7"
headers = {"User-Agent": user_agent}
# Request the webpage
req = urllib2.Request("http://en.wikipedia.org", headers=headers)
resp=urllib2.urlopen(req)
# Read the HTML of the response
html = resp.read()
```

Listing 3.2: Parsing the Wikipedia article HTML document to find the relevant images.

```
## Python
def _get_content_body(self, soup):
    main_content = soup.find('div', {"class": "mw-content-ltr"})
    if main_content is None:
        return None
    # remove navboxes
    navboxes = main_content.findAll('table', {'class': 'navbox'})
    [navbox.extract() for navbox in navboxes]
    return main_content

def _get_image_links(self, soup):
    return soup.findAll('a', {'class': 'image'})

# Parse the HTML
soup = BeautifulSoup(html)
# Get the main article body
soup = _get_content_body(soup)
# Get a list of image links
image_links = _get_image_links(soup)
```

The output of the crawler is a CSV file of the image URLs and the object class the images belong to. The object class is simply named from the URL of the Wikipedia page (for example all images appearing on [http://en.wikipedia.org/wiki/Tower\\_Bridge](http://en.wikipedia.org/wiki/Tower_Bridge) will have class `Tower_Bridge`).

The images mentioned in the CSV file produced by the crawler are then downloaded to local storage. Each image that appears on Wikipedia has a “file page” which displays the image along with properties and metadata on the image<sup>6</sup>. This “file page” is visited for each image, and the page is parsed to extract the storage URL of the image as well as its file format and original size. As the application resizes all images that exceed 1000 pixels to 1000 pixels, it is a waste of time and storage space to download the original

---

<sup>6</sup>For an example see [http://en.wikipedia.org/wiki/File:Tower\\_bridge\\_London\\_Twilight\\_-\\_November\\_2006.jpg](http://en.wikipedia.org/wiki/File:Tower_bridge_London_Twilight_-_November_2006.jpg)

image and later resize it. Instead, Wikipedia's inbuilt thumbnail engine is exploited, which resizes the image on Wikipedia's servers and allows you to download a thumbnail of a user selected width<sup>7</sup>. Therefore if the original image on Wikipedia exceeds 1000 pixels, the 1000 pixel thumbnail version is downloaded instead. The images downloaded are saved in a folder named after it's class. The result is a directory containing a folder for each class, within which are the images for that class.

This process creates a structured dataset of model images from the Wikipedia pages visited. For the List of Structures in London dataset used, there were 732 classes which had 3963 images associated with them (on average 6 images per class).

## 3.2 Turbo-boosting Images

To achieve effective turbo-boosting, many additional images are needed to supplement the model images acquired from Wikipedia. Microsoft Bing is used as the source of the turbo-boosting images. For each class, 25 additional images are downloaded to boost that class.

Bing offers a public API that can be used to perform image searches programmatically. After obtaining an application ID from Bing for authentication, complex search requests can be made over HTTP, with the results returned in JavaScript Object Notation (JSON) format<sup>8</sup>.

Bing image search takes a number of keywords - the query - and returns a list of images from web pages related to the query. Further filters can be applied to further narrow down the search to the most relevant images. This is shown in Figure 3.2 on the website version of Bing. The "Style" and "Size" filters are especially useful in this application, as all images should be photographs for the List of Structures in London dataset, and large images are preferable so as to include as much detail as possible. Setting these filters precludes many instances of graphics and logos which are not suitable for turbo-boosting.

All the parameters that appear in the web interface for Bing image search can be replicated in the API request with query parameters. A MATLAB script is used to consume the API and download the images. Listing 3.3 shows the URL used for the API call to get the search results for the images for a particular

---

<sup>7</sup><http://www.algorithm.co.il/blogs/programming/wikipedia-images/> describes how this is exploited

<sup>8</sup>JSON is an alternative to XML for representing structured data. <http://www.json.org/> provides more information.

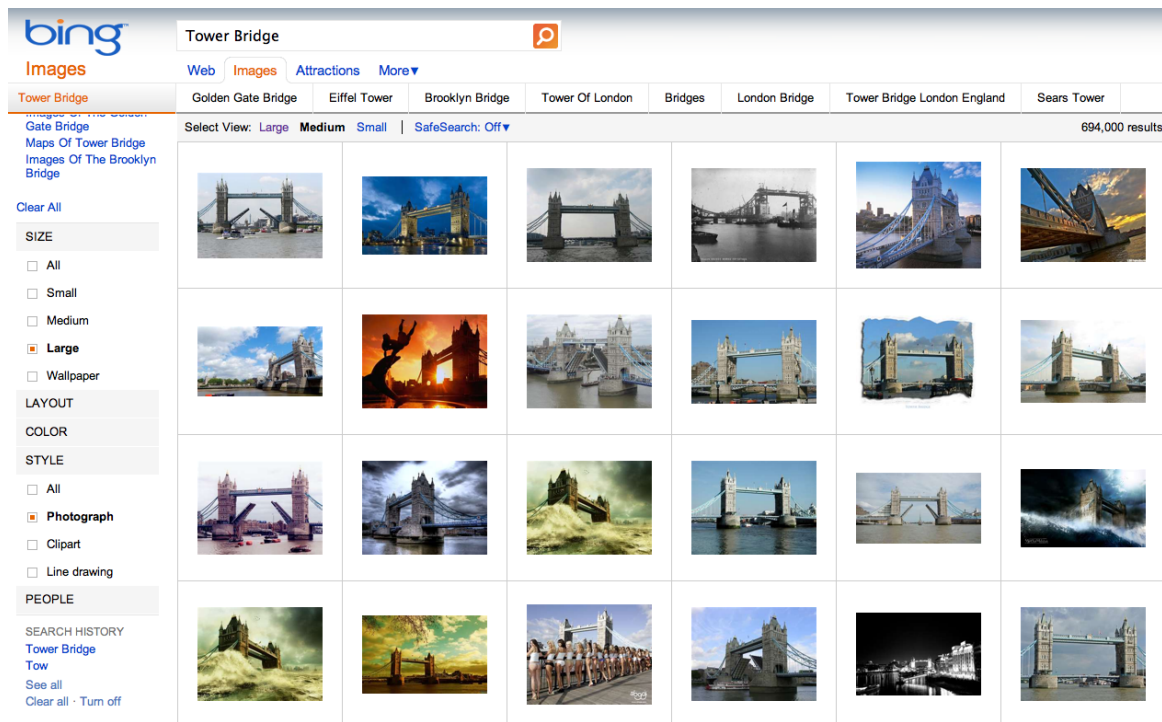


Figure 3.2: The browser interface of Bing image search which is replicated in the API query URL. Note the filters in the left column.

class. The URL encoded<sup>9</sup> class name is used as the query. For example, for the class `Tower_Bridge`, the variable `search_term` appearing in Listing 3.3 will be set to `"Tower%20Bridge"` (note `" "` is URL encoded as `"%20"`). Both the "Style" and "Size" image filters are set to "Photo" and "Large" respectively by setting the `Image.Filters` parameter.

Listing 3.3: A Bing image search API request.

```
%% MATLAB
request_url = ['http://api.bing.net/json.aspx?' ...
    '&AppId=' app_id ...
    '&Query=' search_term ...
    '&Sources=Image' ...
    '&Version=2.0' ...
    '&Adult=Strict' ...
    '&Image.Count=' nPhotos ...
    '&Image.Filters=Style:Photo+Size:Large' ...
    '&JsonType=raw' ...
];
% Read the result of the request
response = urlread(request_url);
% Parse the result from JSON to MATLAB structure form
resp_struct = parse_json(response);
```

<sup>9</sup>URL encoding ensures that all characters are in a form which can be used as a URL. See [http://www.w3schools.com/tags/ref\\_urlencode.asp](http://www.w3schools.com/tags/ref_urlencode.asp) for more information.

Using MATLAB's inbuilt `urlread` function, the search results are requested and returned in JSON format. The JSON result is then parsed and converted into a MATLAB structure object for reading. An example JSON response is shown in Listing 3.4. Each element in the array `Results` is an image result. Each image is then downloaded from its `MediaUrl` field using a modified version of the `imread` function which allows for request timeouts, as some image resources may have expired since their submission to the Bing database.

Listing 3.4: A Bing image search response in JSON format. The “Results” array is truncated to one element.

```
// JSON
{
  "SearchResponse":{
    "Version":"2.0",
    "Query":{
      "SearchTerms":"Tower Bridge"
    },
    "Image":{
      "Total":780000,
      "Offset":0,
      "Results":[
        {
          "Title":"Tower Bridge - London Photo (551176) - Fanpop",
          "MediaUrl":"http://images.fanpop.com/images/image_uploads/Tower-Bridge.jpg",
          "Url":"http://www.fanpop.com/spots/london/images/551176/title/tower-bridge",
          "DisplayUrl":"http://www.fanpop.com/spots/london/images/title/tower-bridge",
          "Width":1600,
          "Height":1200,
          "FileSize":761104,
          "Thumbnail":{
            "Url":"http://ts3.mm.bing.net/images/thumbnail.aspx?q=4757...",
            "ContentType":"image/jpeg",
            "Width":160,
            "Height":120,
            "FileSize":3274
          }
        }, ...
      ]
    }
  }
}
```

The downloaded images are resized if larger than 1000 pixels and stored in a folder named after its class. As for the model images, the result is a directory containing a folder for each class, within which are the turbo-boosting images for that class.

### 3.3 Validation Images

Images are needed to test and validate the yield performance of the object recognition system. Therefore a separate dataset of images with their ground truth classes is needed. Ideally all classes would be tested and each test image should be a fair representation of the object of that class. Google image search was used first to automatically download 8 images for each class. The images were then checked manually to refine the dataset.

Google image search is very similar to Bing image search described in the previous section. However the results are markedly different, providing another set of images that are perfect for testing. Google offers an API over HTTP which can be used to search based on a text query and, as with Bing, filters can be applied. The result is returned in JSON format.

Listing 3.5: A Google image search API request.

```
%% MATLAB
request_url = ['https://ajax.googleapis.com/ajax/services/search/images?v=1.0' ...
    '&q=' search_term ...
    '&as_filetype=jpg' ...
    '&imgsz=xxlarge' ...
    '&imgtype=photo' ...
    '&rsz=8' ...
];
% Read the result of the request
response = urlread(request_url);
% Parse the result from JSON to MATLAB structure form
resp_struct = parse_json(response);
```

Listing 3.5 shows the formulation and request of a Google search API request. As with the Bing requests, the search term is the URL encoded class name. The JSON response (Listing 3.6) is parsed into a MATLAB object and the `url` field is used to download the image.

Listing 3.6: A Google image search response in JSON format. The “results” array is truncated to one element.

```
// JSON
{
  "responseData":{
    "results":[
      {
        "GsearchResultClass":"GimageSearch",
        "width":"1024",
        "height":"819",
        "imageId":"ANd9GcSTU9Qv930E3Q5kjo0h5Dcs16sYBGURxSlmw8tfUbISx6heecXcY9VKkquZ",
        "tbWidth":"150",
        "tbHeight":"120",
        "unescapedUrl":"http://www2.hiren.info/desktopwallpapers/natural/the-tower-...",
        "url":"http://www2.hiren.info/desktopwallpapers/natural/the-tower-bridge.jpg",
        "visibleUrl":"www.hiren.info",
```

```

        "title": "Desktop Wallpapers Natural Backgrounds The \u003cb\u003eTower
        Bridge\u003c/b\u003e \u003cb\u003e...\u003c/b\u003e",
        "titleNoFormatting": "Desktop Wallpapers Natural Backgrounds The Tower Bridge
        ...",
        "originalContextUrl": "http://www.hiren.info/desktop-wallpapers/natural-pic...",
        "content": "The \u003cb\u003eTower Bridge\u003c/b\u003e, London,",
        "contentNoFormatting": "The Tower Bridge, London,",
        "tbUrl": "http://t1.gstatic.com/images?q\u003dtbn:ANd9GcSTU9Qv930E3Q5kjo..."
    }, ...
], ...
"responseDetails": null,
"responseStatus": 200
}

```

As with the turbo-boosting images, the downloaded images are resized if larger than 1000 pixels stored in a folder named after its class. Again, the result is a directory containing a folder for each class, within which are the turbo-boosting images for that class.

After automatic download of potential test images for each class, the dataset was checked over manually. Images that appear in the model dataset, as well as images that do not fairly depict the class they are to test are removed from the test set .

## Chapter 4

# Baseline system

The base line system draws upon the research and literature mentioned in Chapter 2. Upon starting the project, an application was provided that includes a basic database creation and image matching process based on a dataset of structures in Oxford. This application was expanded upon to create the current system.

There are two separate processes which form the project. The first is the pre-computation process that creates the databases and data structures required for object recognition. The second is the object recognition process that takes a query image and returns the names and locations of the objects in the image. A summary of these two processes is described in Figure 4.1 and Figure 4.2 respectively.

The pre-computation process is run once on the dataset to produce the database and working data structures required for the object recognition process. For each model image, the features are detected and associated descriptors generated. This is described further in Section 4.1. A sample of the features are then used to generate the visual word vocabulary (Section 4.2). The histograms of visual words are then weighted and collected into an index ready for querying (Section 4.3). This completes the basic pre-computation process and the system is ready for use.

The object recognition process takes a query image and attempts to recognise the objects contained within the image. Firstly, the feature descriptors are computed for the detected features within the query image. The visual words are computed based on the vocabulary created during the pre-computation process, and the weighted histogram produced for the query image. A search is then performed on the index of

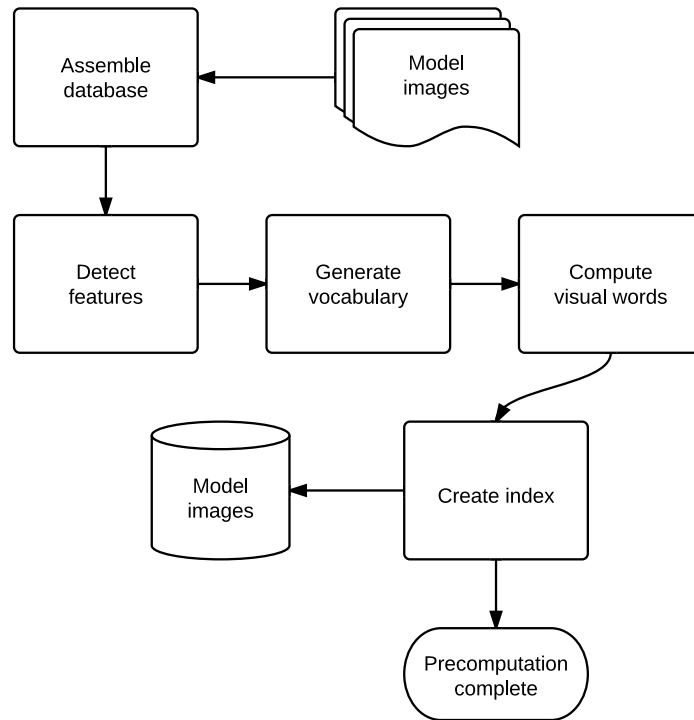


Figure 4.1: The flow diagram for the basic pre-computation process.

histograms for the model images (see Section 4.3) the output of which is a list of images based on how highly they match the query image. Going down the list of top matches by histogram, spatial verification is performed to ensure the visual words in both the query image and match image form the same shaped object (all objects are assumed rigid). This is described in Section 4.4. Once a match has been spatially verified, this object is deemed recognised and added to the list of found objects. Multiple matching is then performed by repeating this process, excluding regions of the image containing a previously recognised object (Section 4.5).

The remainder of this chapter describes further details of the parts of the processes described above.

## 4.1 Feature Detection and Description

The feature detection and description methods used are the original scale-invariant feature transform (SIFT) algorithms. The advantages of using SIFT are that the detected features and their descriptors are invariant to image translation, scaling, and rotation, partially invariant to illumination changes and robust to local geometric distortion. This is essential to be able to match the same object features across varied sources



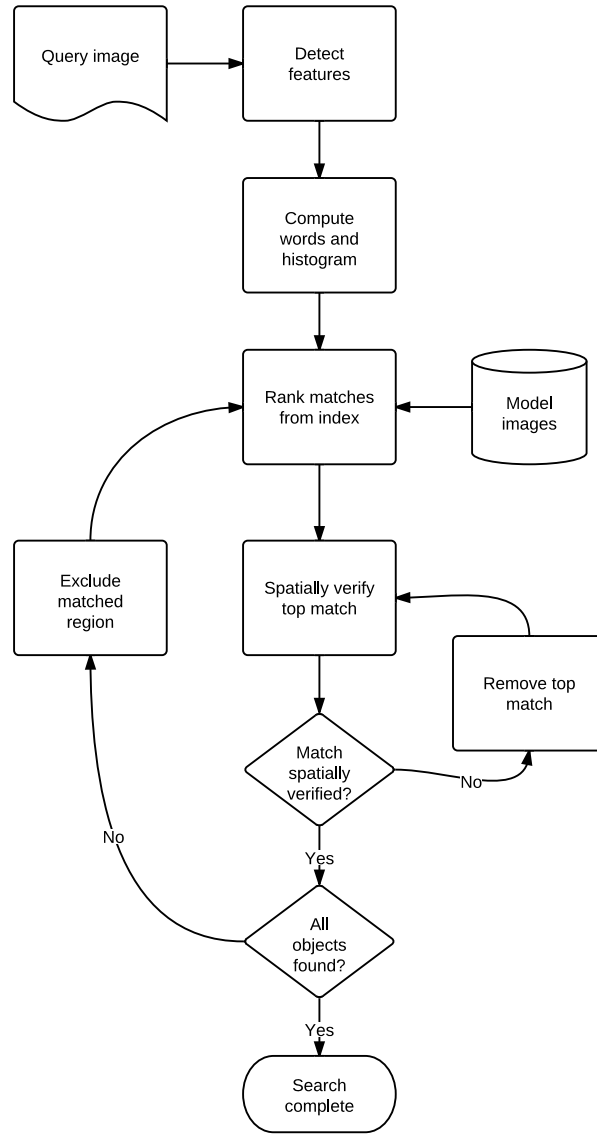


Figure 4.2: The flow diagram for the basic object recognition process.

of images.

The SIFT features are defined as maxima and minima of the result of difference of Gaussians function<sup>1</sup> applied in scale-space to a series of smoothed and resampled images. Low contrast candidate points and edge response points along an edge are discarded. These features are then described by the SIFT feature descriptor - a 128-dimensional vector.

The result of the SIFT feature detection and description algorithm are two matrices. The first is a matrix of feature points or frames, Equation 4.1, where each column describes the position  $(x_i, y_i)$  in the image, the scale  $s_i$  and orientation  $\theta_i$  for feature  $i$ . Corresponding to the frames matrix is a descriptor

<sup>1</sup>[http://en.wikipedia.org/wiki/Difference\\_of\\_Gaussians](http://en.wikipedia.org/wiki/Difference_of_Gaussians)

matrix, Equation 4.2, where each column is the 128-D vector  $\mathbf{d}_i$  that describes feature  $i$ .

$$\begin{bmatrix} x_1 & x_2 & \cdots & x_{N-1} & x_N \\ y_1 & y_2 & \cdots & y_{N-1} & y_N \\ s_1 & s_2 & \cdots & s_{N-1} & s_N \\ \theta_1 & \theta_2 & \cdots & \theta_{N-1} & \theta_N \end{bmatrix} \quad (4.1)$$

$$\begin{bmatrix} \mathbf{d}_1 & \mathbf{d}_2 & \cdots & \mathbf{d}_{N-1} & \mathbf{d}_N \end{bmatrix} \quad (4.2)$$

## 4.2 Visual Words

To avoid trying to match features in unbounded, 128-dimensional space, the SIFT features are quantised. These quantised SIFT features are known as visual words.

During the pre-computation process, the vocabulary of words is created. The vocabulary is essentially the clustering of SIFT space. The number of clusters (visual words) to be created is the vocabulary size, 100,000 for this application. Vocabulary creation is done using the approximate nearest neighbours K-means algorithm. The clustering is performed on a random sample from all the feature descriptors for the entire model images dataset. The number of features sampled was 30 times the size of the vocabulary, i.e. 3 million.

The result of the vocabulary creation is a kd-tree which can be used to get the word associated with a SIFT descriptor. Each word is assigned an ID, and the images can then be represented as a list of words, where each word is the nearest visual word to the SIFT descriptor.

## 4.3 Histograms and Index

## 4.4 Spatial Verification

## 4.5 Multiple Object Matching

## **Chapter 5**

# **Software architecture**

## **Chapter 6**

# **Geometric Improvements**

## **Chapter 7**

# **Turbo-boosting**

## **Chapter 8**

### **Summary**

# Bibliography

- [1] P. Turcot and D. G. Lowe. Better matching with fewer features: The selection of useful features in large database recognition problems In *WS-LAVD, ICCV*, 2009.