Research White Paper

WHP 177

August 2009

Report on CLARET

CLARET – CLAssification and RETrieval of Images

Krystian Mikolajczyk (University of Surrey)

and Denise Bland (BBC Research & Development)

BRITISH BROADCASTING CORPORATION

Report on CLARET

Krystian Mikolajczyk (University of Surrey) and Denise Bland (BBC Research & Development)

Abstract

This White Paper describes the operation of CLARET, an image CLAssification and RETrieval system. One method of identifying image content is implemented and used in two different scenarios. The first scenario is object classification from five learned classes (pedestrians, cars, motorbikes, bicycles and rocket propelled grenades). Images are analysed in terms of the learned classes resulting in a confidence factor that the object class is present in the image. This scenario can be used to automatically generate keyword metadata from images. The second scenario is image retrieval (search) which visually orders images according to their similarity to a selected image. This is also known as 'query by example'. Query by example can be used to search through large image archives or to prioritise incoming UGC (User Generated Content). CLARET is a collaboration project between the University Surrey and BBC R&D.

Additional key words: object, recognition, search, index, metadata

Report on CLARET

Krystian Mikolajczyk (University of Surrey) and Denise Bland (BBC Research & Development)

White Papers are distributed freely on request.

Authorisation of the Head of Research is required for publication.

Report on CLARET

Krystian Mikolajczyk (University of Surrey) and Denise Bland (BBC Research & Development)

1 Introduction

CLARET is an image CLAssification and RETrieval system. Classification is identifying a specific object such as an airplane or a car in an image and requires prior training of the system on each type of object to be recognised. Retrieval searches for similar images to the query image.

News Interactive's User Generated Hub (UGH) receives hundreds of images per week from the public, but a major incident very quickly increases the number of images received to an unworkable amount. During the London bombings of July 7th 2005 hundreds of images were received in a very short space of time. The first pictures of the incident on the BBC's web site were from the public. During the UK floods in 2007 thousands of images were received over a period of a few days. Opening up the BBC to User Generated Content (UGC) creates a practical problem in sorting the large volumes that can be submitted in a short space of time. The material is often topical and must be dealt with quickly. This report details our work in object recognition of still image content. This work is a collaboration project between the University of Surrey and BBC Research & Development.

CLARET has implemented a method of simultaneous recognition and localisation of multiple object classes in image content. CLARET analyses image stills, however key frames from a video can also be analysed as a series of still images. The recognition method is based on appearance clusters built from edge based features. The appearance clusters are shared amongst several object classes and are represented in a hierarchical tree structure. A probabilistic model allows for detection of multiple different objects in the same image.

CLARET's object recognition method is used in two different scenarios. The first scenario is object classification from image categories for pedestrians, cars, motorbikes, bicycles and rocket propelled grenades. The image categories are learned through a training process and an image is analysed in terms of the learned categories. If the resulting confidence factor for a category being present in an image is above a pre-determined threshold, the category is judged to be present in the image and identified by a bounding box. This scenario can automatically generate keyword metadata for efficient searching and leverage of archive content. The second scenario is image retrieval (search) based on an image's similarity to a selected image; also known as 'query by example'. Query by example can be used to search through large image archives or to prioritise incoming UGC.

Classification requires training for each image category. A mathematical model of the image category is built from hundreds of images. Training with non-segmented images where the object is not explicitly identified within the image requires minimal manual intervention compared to segmented training images. A classification training model built from approximately a hundred non-segmented images for each image category of pedestrians, cars, motorbikes, bicycles and rocket propelled grenades was produced by the University of Surrey. Retrieval calculates individual parameters for each image in the search area and uses over five hundred non-segmented images of landscape and urban images supplied by BBC Research & Development.

This paper describes the training procedure and classification method used by both classification and retrieval in section 2. CLARET's results are presented in section 3 and where possible they are compared to published figures. The conclusion is in section 4 and references are given in section 5.

2 Method

In this section we describe the approach taken by classification and retrieval. We first briefly discuss the training procedure and then the classification method.

2.1 Training

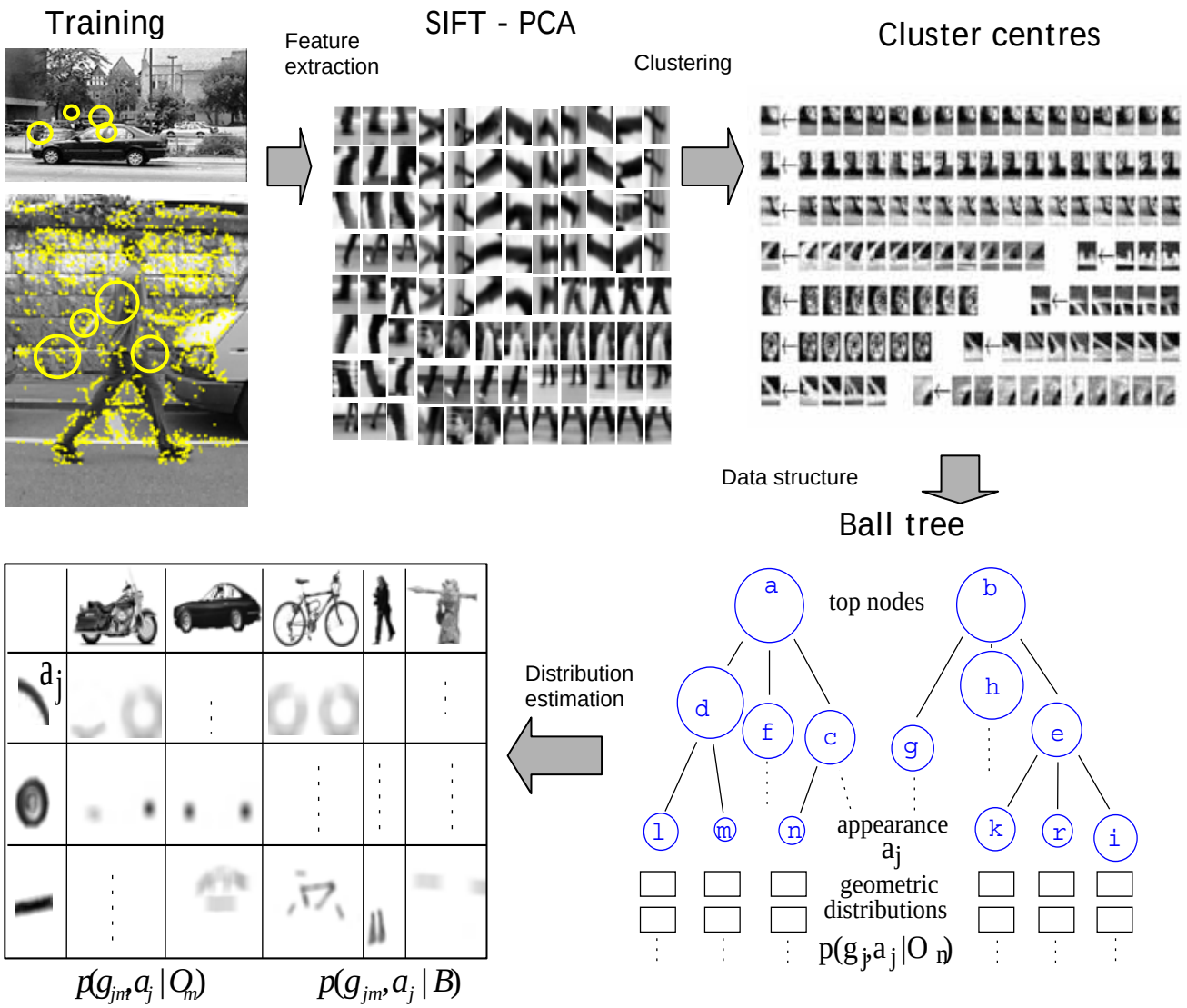


Figure 1 Training of image category models

Training of image category models for the different classes is illustrated in Figure 1 Training of image category models. The classification training category consists of pedestrians, cars, motorbikes, bicycles and rocket propelled grenades (as an interest from news). We assume that annotated image examples are given to the input of the system for each image category. The annotations assert that an object of a specific category is present.

To train CLARET with an image category, first a feature extraction method is applied to all the input images in that category. We use a scale invariant detector, which extracts regions of high signal change from the image. Each circular region of change is then represented with a scale invariant feature transform (SIFT) histogram of 128 dimensions. SIFT combines a scale invariant region detector and a descriptor based on the gradient distribution in the detected regions. The descriptor is represented by a 3D histogram of gradient locations and orientations where contributions to the location and orientation bins are weighted by the gradient magnitude. Standard Principal Component Analysis (PCA) is applied to reduce the dimensionality of the 128 dimension SIFT to a

40 dimension PCA (SIFT-PCA). More details on feature extraction can be found in [Mikolajczyk PAMI 2005] (see References).

A clustering method is applied to the extracted 40 dimension PCA feature descriptors. The clustering method performs vector quantization and builds a hierarchical ball-tree data structure for fast similarity searching. The clustering method and the hierarchical ball-tree data structure are described in [Leibe BMVC 2006] (see References).

Finally, probability distributions that a feature prototype belongs to a given image category is estimated from positive image examples (images that contain the object) as:

$$p(g_{jm}, a_j \mid O_m)$$

And from negative image examples (background images that do not contain the object) as:

$$p(g_{jm}, a_j \mid B)$$

Where
O_m is for an object, which is a positive image category
B is for background, which is a negative image category
a_j is an appearance cluster of local image structure that occurs on different categories
g_{jm} is the corresponding geometric location of a_j with respect to the object centre.

2.2 Classification

Classification of a query image consists of matching the features extracted from the query image to the features of the trained model. Figure 2 Classification of a query image illustrates the classification process.

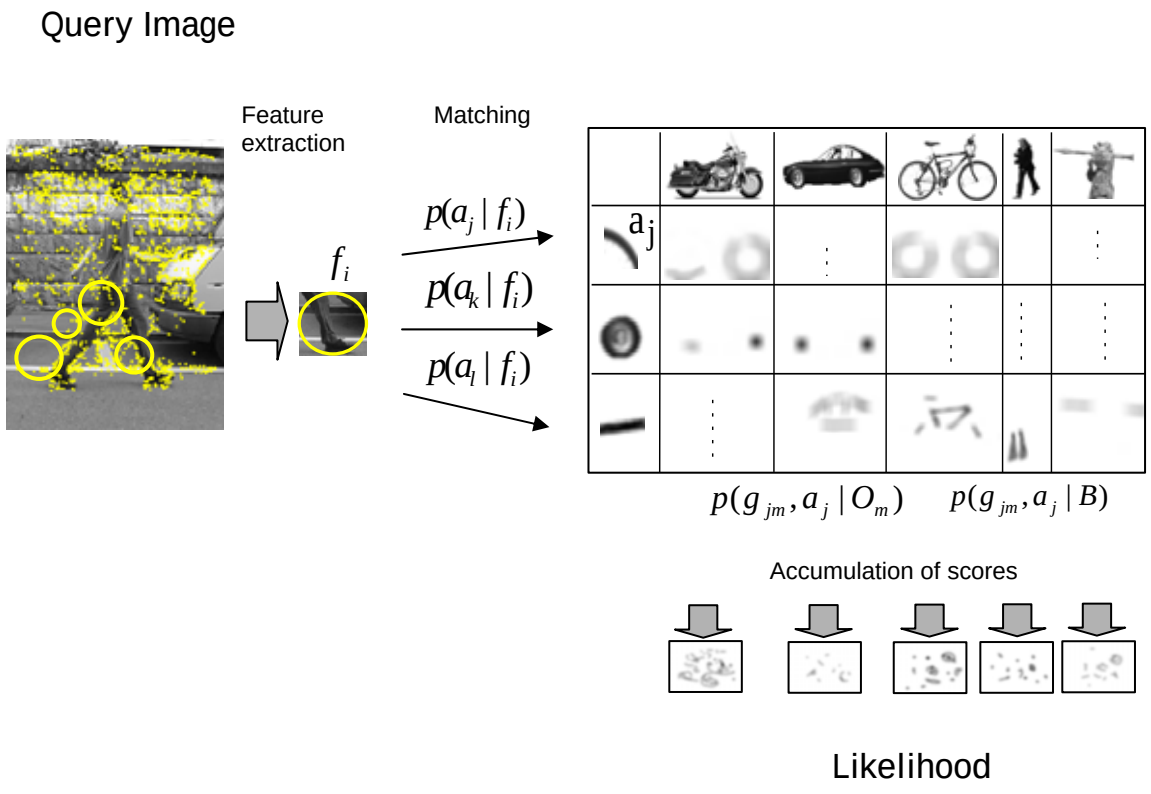


Figure 2 Classification of a query image

Given a query image, local features are extracted. Each feature f_i is then matched to the trained cluster prototypes a_j. Cluster prototypes that are similar to the query image indicate the possible object categories for this feature and its position within the image:

$$p(g_j, a_j \mid f_i, O_m) p(a_j \mid f_i)$$

The scores from different cluster prototypes are accumulated along the object category columns in likelihood maps for each feature:

$$p(G, A \mid f_i, O_m) = \sum_j p(g_j, a_j \mid f_i, O_m) p(a_j \mid f_i)$$

The scores from different query features are multiplied for each object category:

$$p(G, A \mid O_m) = \prod_i p(G, A \mid f_i, O_m)$$

The same operation is done for the probabilities estimated on background non-category examples termed B. The final score is the likelihood ratio given in equation 1 and compared with empirically chosen threshold.

$$\frac{p(G, A \mid O_m)}{p(G, A \mid B)} \geq Threshold$$

Equation 1 Classification decision

Each point on the likelihood map represents a ratio of probabilities that a given object category is present at this position to the probability that the non-category (background) is present (cf equation 1). Classification decision is based on a comparison of this value with an experimentally chosen threshold. Details for this procedure can be found in [Mikolajczyk CVPR 2006] (see References).

2.3 Retrieval

In addition to classification's extracted features from local clusters of corners and edges, retrieval extracted features also include the distribution of colour and a histogram of gradients (texture).

3 Results

3.1 Classification

3.1.1 Classification Test Data and Evaluation Criteria

We evaluate the performance of CLARET's classification system on 4 object classes from a street scenario, namely pedestrians, cars, motor-bikes and bicycles. We introduce one more challenging category which is an RPG shooter (rocket propelled grenade launcher). All training and test data except RPG come from the Pascal collection [Pascal] (see References). We use 600 pedestrian training images, which is a subset of the data used in [Dalal CVPR 2005] (see References). The test set consists of 84 images with 149 pedestrians.

Note that 20 images in this set contain only upper-bodies, which cannot be detected by the current implementation of our approach due to the object centre falling outside the image, thus outside the limits of the likelihood function. Nonetheless, we use this set in order to be directly comparable with the performance reported by [Dalal CVPR 2005] for which the results are reported in the Pascal report.

The multi-scale car test data of 108 images with 139 cars was used in [Fritz ICCV 05] (see References). We used 50 side views of cars for training. The 115 motorbike test images were previously used in [Fritz ICCV 05]. For training we used 150 motorbike examples from the 101 Caltech set [Caltech101] (see References). The bicycle test data from Pascal contains 113 images with 123 bicycles. 100 side views of bicycles were used for training.

A dummy RPG was used to prepare the training images and some of the test images. The training set contains 104 images of various RPG-shooter poses in 5 different clothes on a uniform background. The test data contains 40 images collected from feature movies, TV news as well as taken from real street scenes. Finally, 600 background images are taken from the set in [Dalal CVPR 2005] for training.

We adopt the evaluation criteria used in Pascal challenge. Detection is considered correct, if the area of overlap between the predicted bounding box B_d and ground truth (known object) bounding box B_{gt} exceeds 50% by the formula

$$\frac{B_d \cap B_{gt}}{B_d \cup B_{gt}} \geq 0.5$$

Thus, our results are directly comparable with [Dalal CVPR 2005, Fritz ICCV 05, Pascal].

We use precision-recall criteria where the precision is

$$precision = \frac{\#\,true\ \ positive\ \ \ returned\ \ images}{\#\,all\ \ true\ and\ \ \ false\ \ returned\ \ images}$$

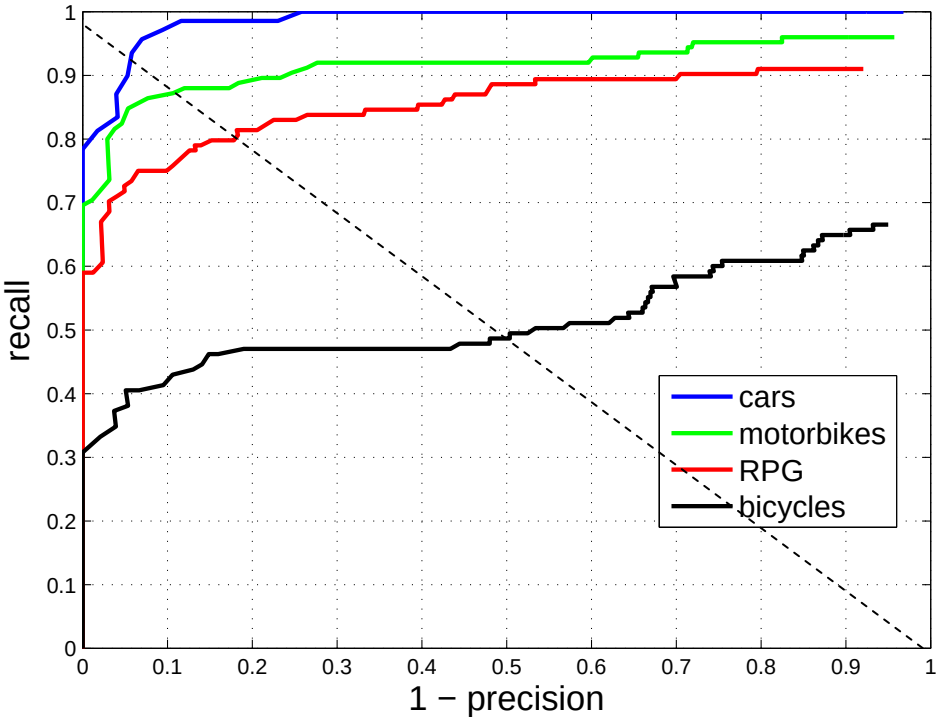
$$1 - precision = \frac{\#\,false\ \ positive\ \ \ returned\ \ images}{\#\,all\ \ true\ and\ \ \ false\ \ returned\ \ images}$$

and recall is

$$recall = \frac{\#\,true\ \ positve\ \ returned\ \ images}{\#\,ground\ \ \ truth}$$

3.1.2 Classification Results

Figure 3 Category detection results shows precision-recall curves for cars, motorbikes, bicycles, and RPG object classes. The equal error rate (EER) is shown by the dashed line and ideally where the results cross the EER line should be as high as possible. This ensures that the number of true positive images is at a maximum and the number of false positive images is at a minimum. The best recognition score is obtained for cars 95% EER (equal error rate) performance compares favourably to 87.8% in [Fritz ICCV 05]. Initial recall for our method with 0 false positives is 78%. The motorbike test data is more challenging and contains many out of plane rotations as well as occlusions. Our EER performance for this data is 88% with initial recall of 70% compared to 81% EER and 40% initial recall in [Fritz ICCV 05]. The performance for bicycles is lower since the test images contain many different viewpoints and partial occlusion. EER score is 50%. However the detector trained on side views demonstrates some tolerance to viewpoint changes (cf. figure 5) due to the rotation invariance of the model. The score for RPG class is 81%. The confusion factor between pedestrians and RPGs is low, but some false positives occur on upper bodies of pedestrians.



- Cars 95% EER
- Motorbikes 88% EER
- RPG 82% EER
- Bicycles 50% EER

Figure 3 Category detection results

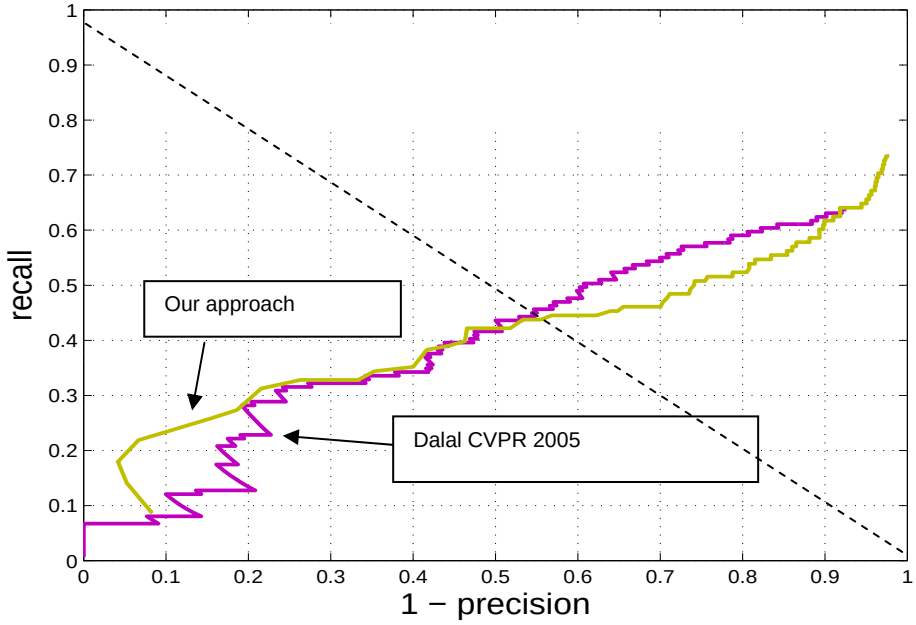


Figure 4 Pedestrian detection results

Finally the lowest performance of 45% EER is on pedestrian images in Figure 4 Pedestrian detection results. However a state of the art pedestrian detector from [Dalal CVPR 2005] obtains a similar EER score. 2416 of positive and 12180 of negative examples were used to train [Dalal CVPR 2005] detector compared to our 600 positive and 600 negative. Both curves are displayed in figure 4 above.

We also observe that the number of appearance clusters grows sub-linearly with increasing number of object classes as more clusters are shared between different categories. The multi-class model uses approximately 75% of the number of clusters in all individual detectors with the same maximum size of appearance clusters. This indicates that the method can scale to a large

number of classes since the complexity reduces compared with individual detectors. Another consequence of sharing clusters is that the required amount of training data is reduced. In particular, wheel based vehicles seem to benefit from the training data of each other. Additional experiments have to be carried out to quantify this.

Figure 5 shows detection examples on different categories. The results are displayed for recognition threshold set to EER on the pedestrian database. Note the difficulty of the recognition task in presence of occlusion on motorbike examples and RPG, various object scales for pedestrians and cars, different viewpoints for bicycles, multiple object classes occluding each other and in-plane rotations.

Colour	Object
green	motorbike
yellow	pedestrian
black	bicycle
blue	car
red	RPG

Table 1 Category detection key

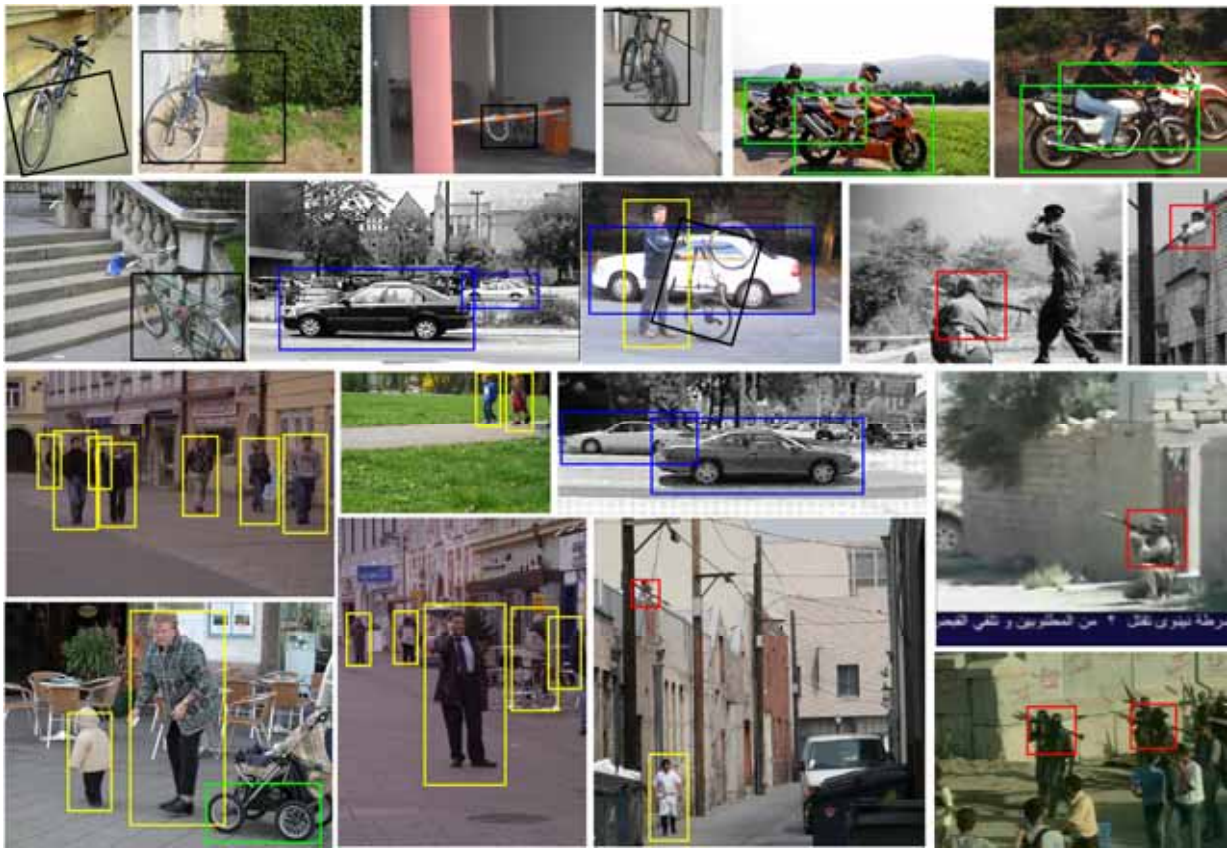


Figure 5 Examples of category detections.

3.2 Retrieval

3.2.1 Retrieval Test Data and Evaluation Criteria

We evaluated the performance of CLARET's retrieval system by measuring the ranking of visually similar images that are successfully returned. The retrieval test data consists of 230 landscape images and 253 urban images taken from 'Elvis', the BBC's stills database. The images were manually sorted according to their visual content into one of the 12 groups forming the ground truth or background. The 12 groups consist of 6 urban scenes (Big Ben, Broadcasting House, brown buildings, London skyscrapers, white skyscrapers and Canary Wharf) and 6 landscape scenes (glacier, lavender, mountains, yellow field, starfish and sunflower). A group consists of four or more similar images. The ground truth images are shown in Figure 6 Ground truth for Big Ben, Broadcasting, Buildings, Glacier, Lavender and London and in Figure 7 Ground truth for Mountains, Skyscraper, Yellow Field, Starfish, sunflower and Canary Wharf. Each image in the ground truth was selected and the similar image retrieval process initiated. The first twenty returned images out of the 483 test images were judged on their similarity to the selected image by their presence/absence in the correct ground truth group.

Three measures of precision were calculated for each image in the ground truth group. A figure of precision for the first returned image is computed as 100% if the first return image is in the correct ground truth group shown in figures 6 and 7. A figure of precision for the first and second returned image is computed as 100% if the first and second returned images are in the correct ground truth group shown in figures 6 and 7, and 50% if just the first or second image is in the correct ground truth group. A figure of precision for the first, second and third returned images is computed as 100% if the first, second and third returned images are in the correct ground truth group shown in figures 6 and 7, 67% if just two of the first three images are in the correct ground truth group, and 33% if just one of the first three images are in the correct ground truth group. Three measures of recall were calculated for each image in the ground truth group. The percentage of the first five, ten and twenty images that were in the correct ground truth group shown in figures 6 and 7 is taken as the recall.

Our approach to retrieval is compared to 3 other techniques, namely scale invariant feature transform (SIFT) wavelet and colour histogram. SIFT transforms the image data into scale invariant key points relative to local features. This is a popular technique used to perform matching between different views of the same object or scene. SIFT features are invariant to image scale and rotation, and they robustly match across a range of change in viewpoint, addition of noise and change in illumination. Wavelet time-frequency-space analysis is a popular transform used in image compression and in Retrievr, a search engine for images based entirely on visual input. Retrievr searches through photographs stored in Yahoo's Flickr database by using hand-drawn doodles as its search criteria. Colour histogram is a popular feature used in commercial image retrieval systems; IBM's Marvel (Multimedia Analysis and Retrieval Engine) extracts a number of features including colour histogram.

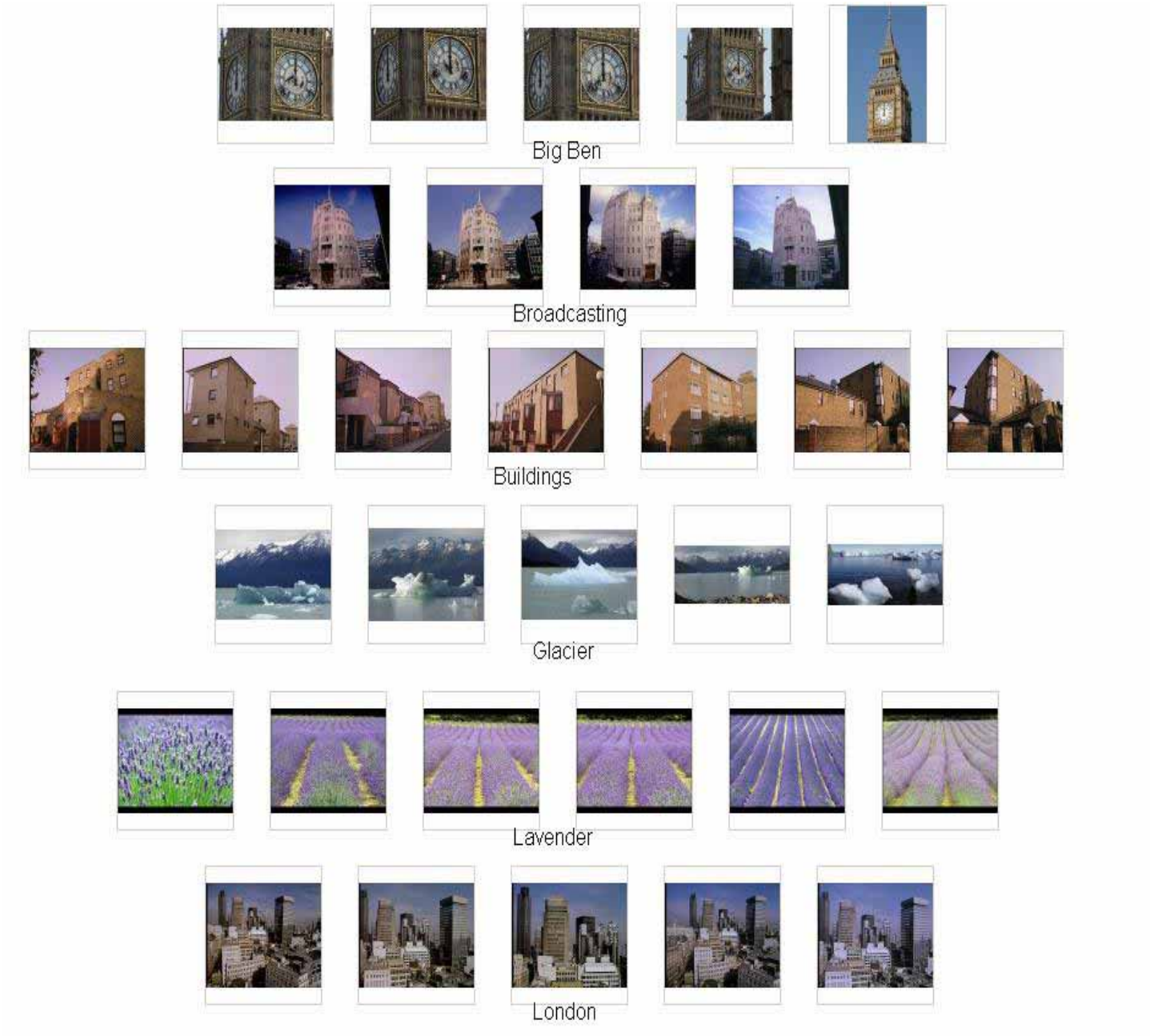


Figure 6 Ground truth for Big Ben, Broadcasting, Buildings, Glacier, Lavender and London

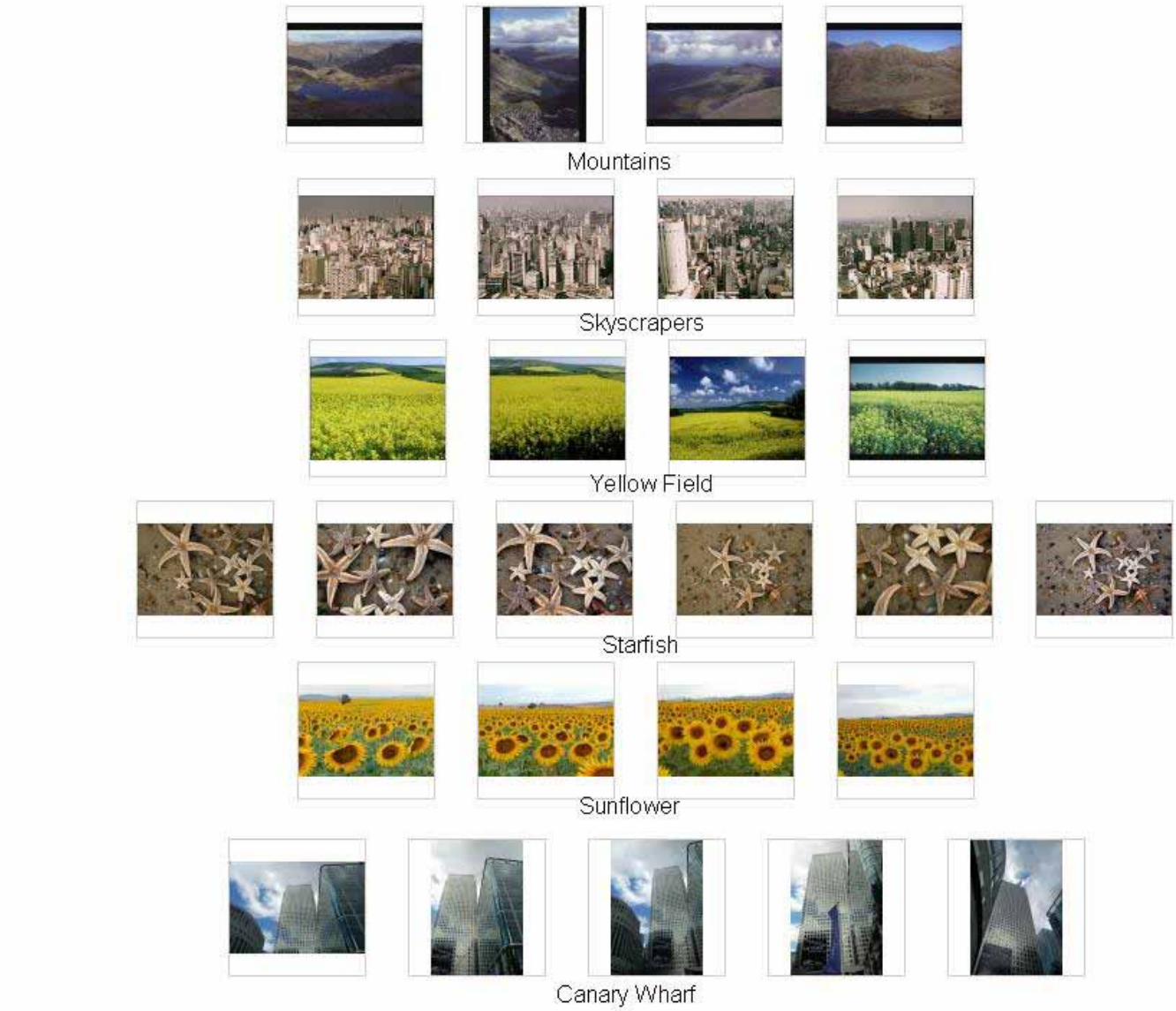


Figure 7 Ground truth for Mountains, Skyscraper, Yellow Field, Starfish, sunflower and Canary Wharf

3.2.2 Retrieval Results

Surrey	PRECISION				RECALL			
	1st	1st, 2nd	1st,2nd,3rd	**Ave**	First 5	First 10	First 20	**Ave**
Big Ben	80%	80%	80%	**80%**	60%	60%	65%	**62%**
Broadcasting	50%	38%	33%	**40%**	42%	59%	67%	**56%**
Buildings	71%	50%	43%	**55%**	31%	53%	62%	**48%**
Glacier	60%	30%	27%	**39%**	20%	20%	25%	**22%**
Lavender	83%	67%	50%	**67%**	33%	53%	63%	**50%**
London	100%	100%	100%	**100%**	100%	100%	100%	**100%**
Mountains	50%	50%	42%	**47%**	42%	50%	50%	**47%**
Skyscrapers	100%	100%	92%	**97%**	100%	100%	100%	**100%**
Yellow Field	25%	25%	25%	**25%**	33%	50%	58%	**47%**
Starfish	100%	100%	83%	**94%**	65%	83%	100%	**83%**
Sunflower	100%	100%	92%	**97%**	100%	100%	100%	**100%**
Canary Wharf	60%	70%	60%	**63%**	55%	60%	60%	**58%**
Total Ave				**67%**				**64%**

Table 2 Surrey Retrieval Results

SIFT	PRECISION				RECALL			
	1st	1st, 2nd	1st,2nd,3rd	**Ave**	First 5	First 10	First 20	**Ave**
Big Ben	100%	100%	100%	**100%**	100%	100%	100%	**100%**
Broadcasting	75%	75%	50%	**67%**	50%	50%	58%	**53%**
Buildings	0%	0%	10%	**3%**	7%	12%	19%	**13%**
Glacier	80%	80%	80%	**80%**	65%	65%	65%	**65%**
Lavender	50%	50%	33%	**44%**	20%	20%	20%	**20%**
London	100%	100%	100%	**100%**	100%	100%	100%	**100%**
Mountains	0%	0%	0%	**0%**	0%	17%	17%	**11%**
Skyscrapers	0%	0%	0%	**0%**	0%	0%	0%	**0%**
Yellow Field	0%	0%	8%	**3%**	8%	17%	17%	**14%**
Starfish	100%	100%	100%	**100%**	100%	100%	100%	**100%**
Sunflower	67%	33%	22%	**41%**	22%	22%	56%	**33%**
Canary Wharf	80%	90%	87%	**86%**	80%	90%	90%	**87%**
Total Ave				**52%**				**50%**

Table 3 SIFT Retrieval Results

Wavelet	PRECISION				RECALL			
	1st	1st, 2nd	1st,2nd,3rd	**Ave**	First 5	First 10	First 20	**Ave**
Big Ben	80%	80%	60%	**73%**	45%	45%	50%	**47%**
Broadcasting	50%	25%	25%	**33%**	25%	33%	58%	**39%**
Buildings	71%	50%	38%	**53%**	21%	29%	36%	**29%**
Glacier	80%	40%	27%	**49%**	25%	35%	45%	**35%**
Lavender	100%	100%	94%	**98%**	80%	93%	97%	**90%**
London	100%	80%	60%	**80%**	65%	85%	100%	**83%**
Mountains	0%	0%	17%	**6%**	42%	50%	75%	**56%**
Skyscrapers	25%	13%	8%	**15%**	17%	42%	50%	**36%**
Yellow Field	25%	25%	17%	**22%**	17%	17%	17%	**17%**
Starfish	100%	100%	72%	**91%**	50%	57%	73%	**60%**
Sunflower	67%	67%	56%	**63%**	67%	78%	78%	**74%**
Canary Wharf	80%	70%	53%	**68%**	60%	60%	60%	**60%**
Total Ave				**54%**				**52%**

Table 4 Wavelet Retrieval Results

YIQ	PRECISION				RECALL			
	1st	1st, 2nd	1st,2nd,3rd	**Ave**	First 5	First 10	First 20	**Ave**
Big Ben	80%	80%	67%	**76%**	55%	55%	60%	**57%**
Broadcasting	0%	0%	8%	**3%**	8%	8%	17%	**11%**
Buildings	57%	50%	38%	**48%**	21%	33%	52%	**36%**
Glacier	40%	20%	13%	**24%**	15%	15%	20%	**17%**
Lavender	100%	92%	89%	**94%**	63%	77%	83%	**74%**
London	100%	90%	87%	**92%**	70%	85%	85%	**80%**
Mountains	75%	38%	42%	**51%**	42%	50%	50%	**47%**
Skyscrapers	50%	50%	58%	**53%**	58%	83%	92%	**78%**
Yellow Field	50%	25%	25%	**33%**	25%	50%	50%	**42%**
Starfish	100%	58%	61%	**73%**	37%	43%	60%	**47%**
Sunflower	100%	100%	89%	**96%**	100%	100%	100%	**100%**
Canary Wharf	100%	90%	67%	**86%**	65%	75%	100%	**80%**
Total Ave				**61%**				**56%**

Table 5 Colour Histogram Results

Colour	Algorithm
pink	SIFT
red	Wavelet
green	colour histogram
blue	Surrey algorithm

Table 6 Precision and Recall graph's key

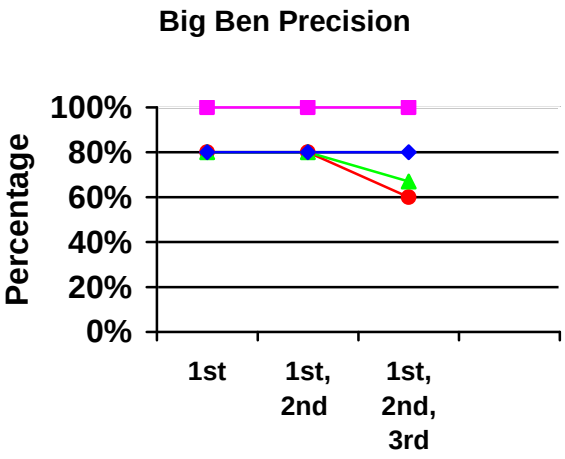


Figure 8

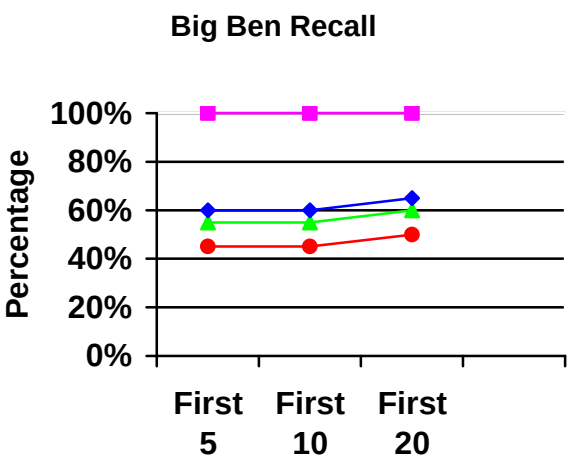


Figure 9

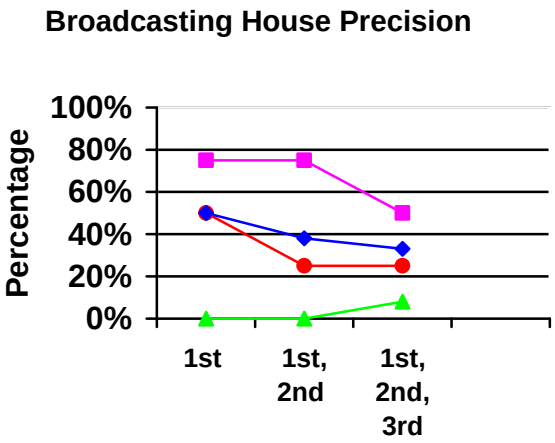


Figure 10

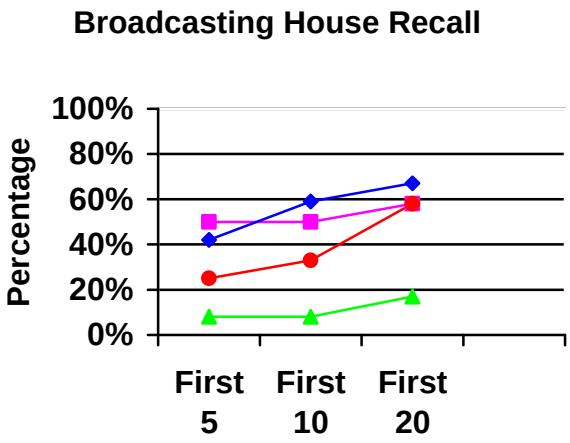


Figure 11

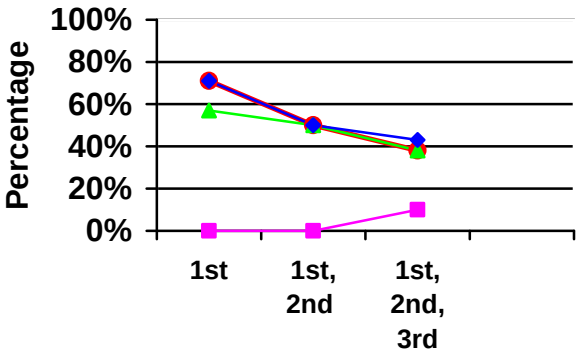


Figure 12

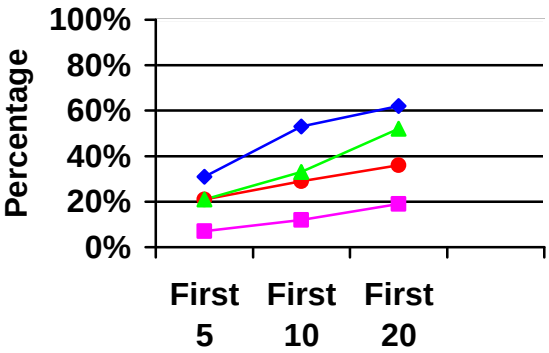


Figure 13

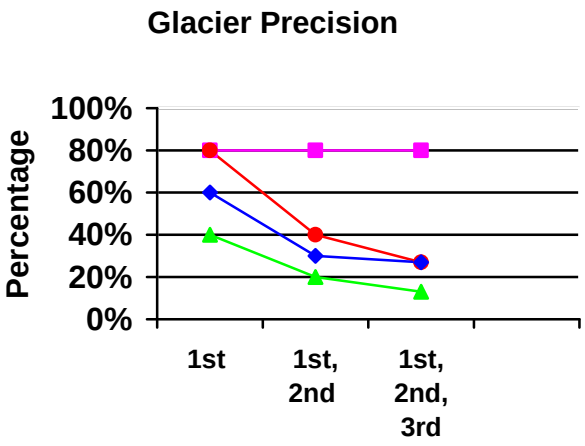


Figure 14

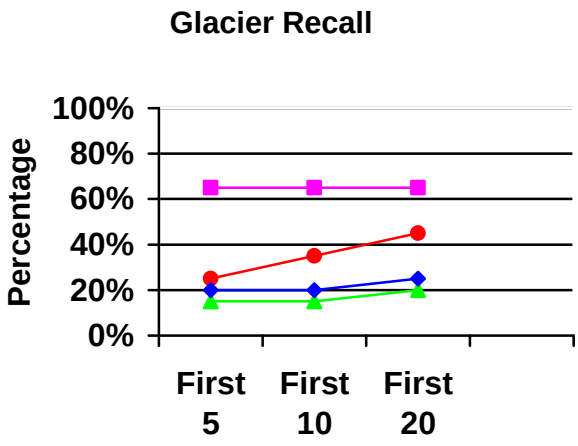


Figure 15

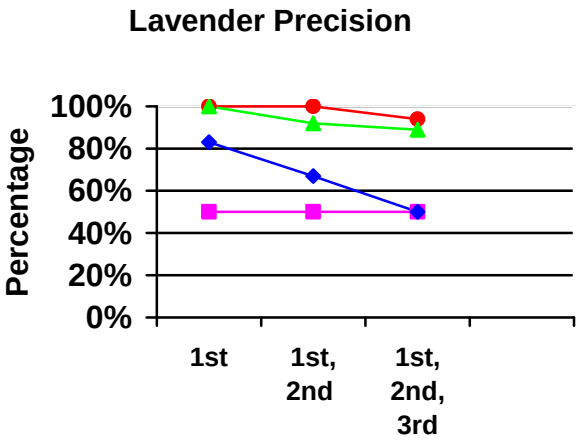


Figure 16

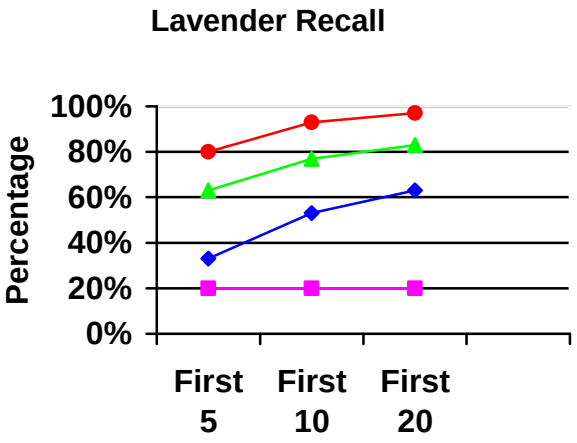


Figure 17

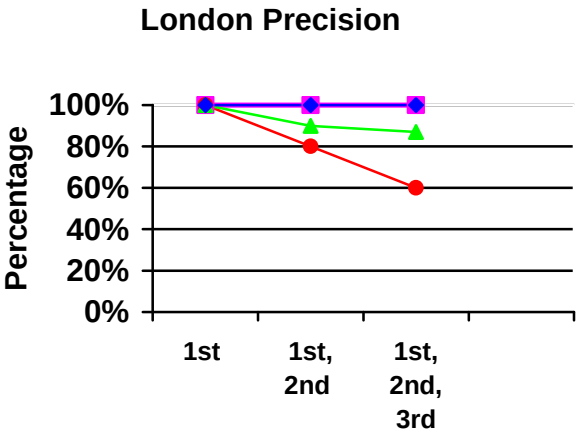


Figure 18

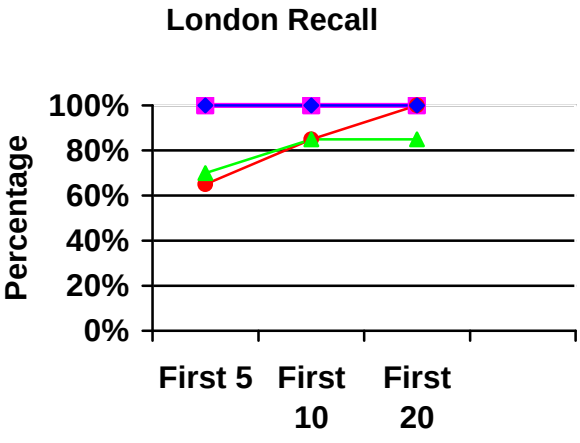


Figure 19

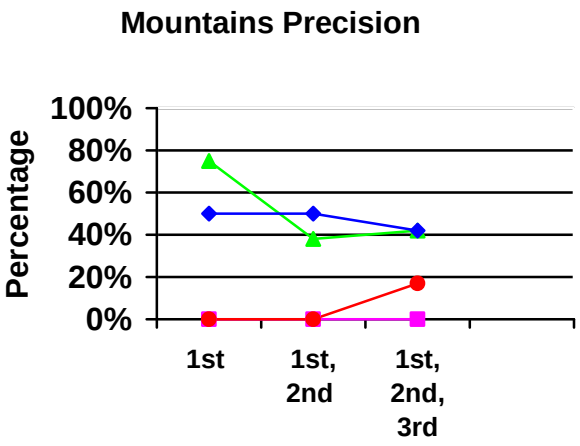


Figure 20

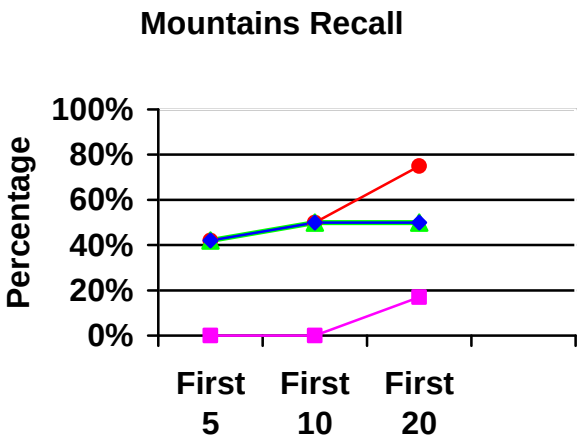


Figure 21

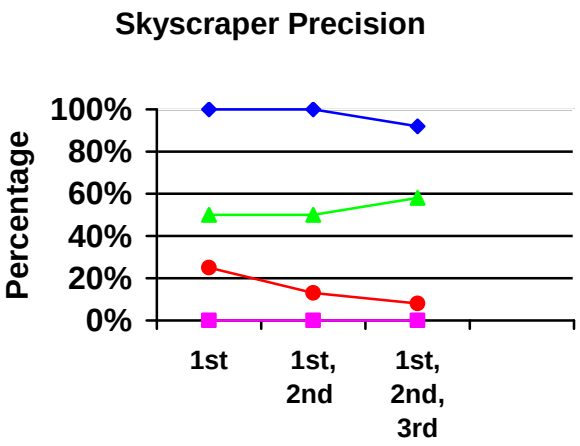


Figure 22

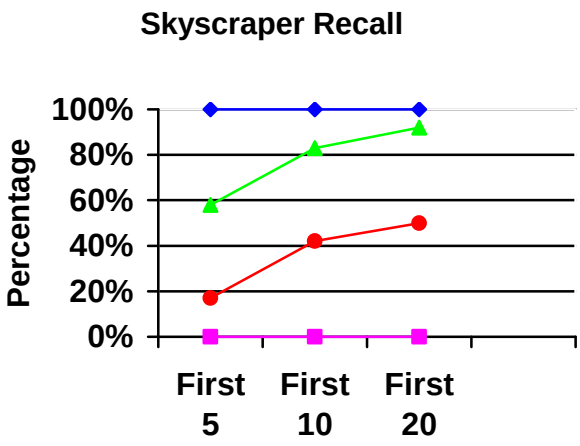


Figure 23

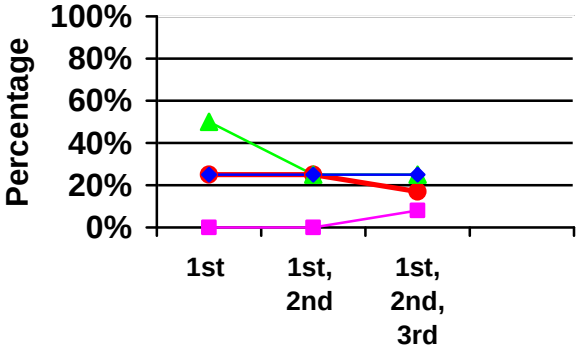


Figure 24

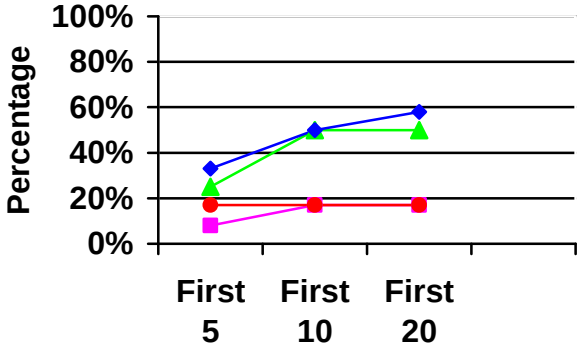


Figure 25

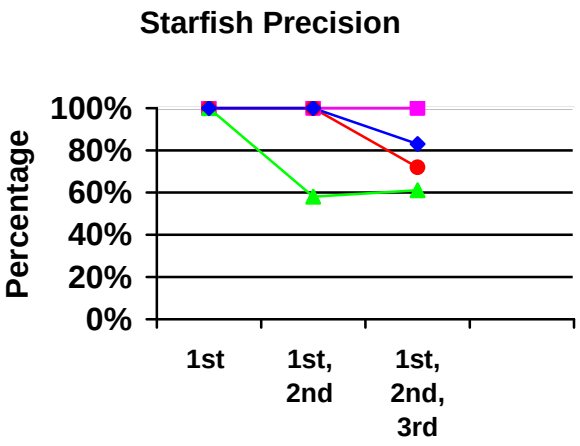


Figure 26

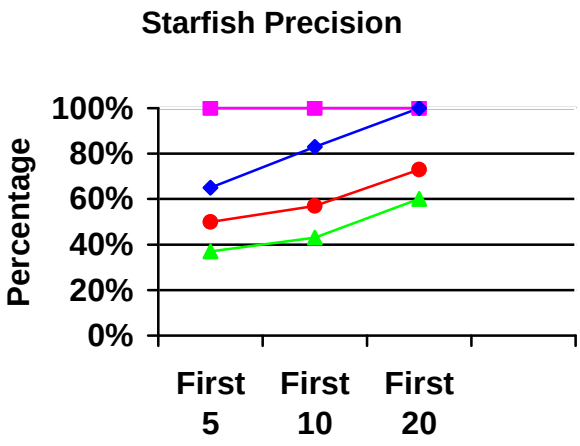


Figure 27

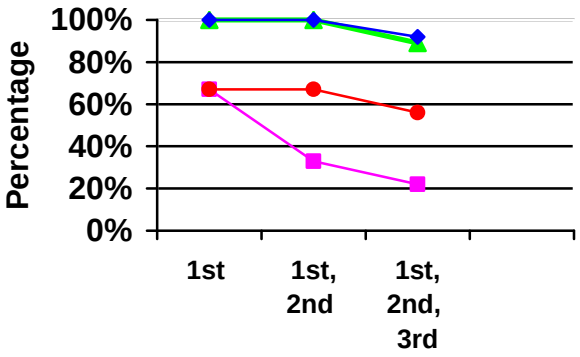


Figure 28

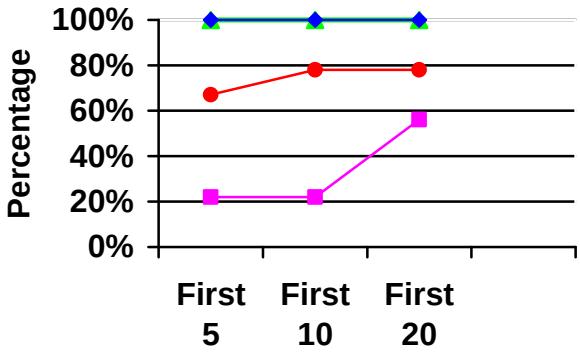


Figure 29

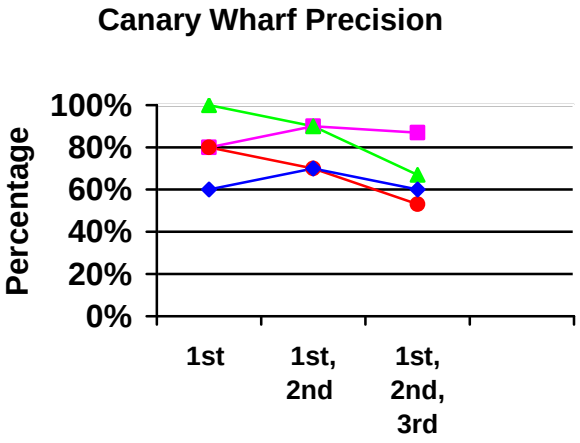


Figure 30

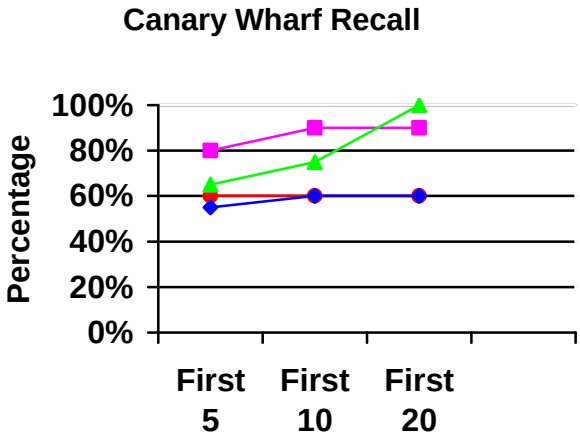


Figure 31

4 Conclusion

CLARET employs object recognition of still image content to generate keyword metadata using object classification and provides content based image searching using image retrieval.

The first scenario is object classification which analyses images in terms of the following trained classes; bicycles, motor-bikes, cars, pedestrians and rocket propelled grenades. The performance of CLARET is compared with published results from the 2005 PASCAL visual object class challenge for bicycles, motor-bikes, cars and pedestrians, as shown below:

Class	No. of training images	No. of test images	Comparison with Report	Report EER	CLARET EER	Report Initial Recall	CLARET Initial Recall
Pedestrians	600	84	Dalal CVPR 05	46%	45%	-	-
Cars	50	108	Fritz ICCV 05	87.8%	95%	-	78%
Motor-bikes	150	115	Fritz ICCV 05	81%	88%	40%	70%
Bicycles	100	113	-	-	50%	30%	-
Rocket propelled grenades	104	40	-	-	81%	-	-

Table 7 Comparison of results

Looking at the equal error rate EER where a high value is required, generally CLARET performed as well as, if not better than, the published PASCAL classifiers. CLARET performed well in the presence of occlusion on motorbike and RPG examples. CLARET also handled various object scales for pedestrians and cars, different viewpoints for bicycles and multiple object classes occluding each other and in-plane rotations. Object classification automatically generates keyword metadata for efficient searching and leverage of archived content that otherwise would not have any metadata for searching.

CLARET's second scenario is image retrieval (search) based on the image's similarity to a selected image. This scenario is also called 'query by example'. For image retrieval, CLARET is trained with urban and landscape images. A ground truth is established for 12 groups containing 4 or more similar images. The 12 groups consist of 6 urban scenes (Big Ben, Broadcasting House, brown buildings, London skyscrapers, white skyscrapers and Canary Wharf) and 6 landscape scenes (glacier, lavender, mountains, yellow field, starfish and sunflower). The percentage measure for retrieval precision is based on the first 3 returned images being members of the selected group. The percentage measure for retrieval recall is based on the first 5, 10 and 20 returned images being members of the selected group. Our results are compared against SIFT, wavelet and colour histogram methods. Looking at figures 8 to 31 inclusive, the performance of retrieval (search) shows a strong correlation between type of image content and successful algorithm. SIFT based features alone perform very well when matching specific objects or scenes (as in Big Ben, London skyscraper, starfish and Canary Wharf) but perform very poorly on generic objects or scenes (buildings, mountains and white skyscrapers). Wavelet performance is reasonable across most of the images and very well on the lavender images that have a strong repeated pattern but poorly on the mountains which are structurally very different. Colour histogram performance is reasonable across all the images apart from Broadcasting House where there is a strong variation of colour. Surrey's retrieval algorithm performs reasonably well across all the different images when compared to the other methods with an average precision on 67% and an average recall of 64%.

This work will enable BBC Research & Development to accurately specify the BBC's challenging future system requirements and brings the general field of image recognition closer to everyday use.

5 References

Caltech101	http://www.vision.caltech.edu/Image_Datasets/Caltech10
Dalal CVPR 2005	N. Dalal and B. Triggs Histograms of Oriented Gradients for Human Detection. In *CVPR*, pages 886–893, 2005
Fritz ICCV 05	M. Fritz, B. Leibe, B. Caputo, and B. Schiele. Integrating Representative and Discriminant Models for Object Category Detection. In *ICCV*, pages 1363–1370, 2005
Leibe BMVC 2006	Efficient Clustering and Matching for Object Class Recognition, BMVC 2006
Mikolajczyk PAMI 2005	A Performance Evaluation of Local Descriptors, PAMI 2005.
Mikolajczyk CVPR 2006	Multiple Object Class Detection With a Generative Model, CVPR 2006
Pascal	http://www.pascal-network.org/challenges/VOC/voc2005