Content-based retrieval of visual information

Oerlemans, A.A.J.

Chapter 6

Learning and Visual Concept Detection

Visual concept detection is the automated detection of image semantics. Detecting image semantics is particularly useful in content based retrieval because it allows us to annotate media with semantically meaningful information. One computationally efficient approach toward subimage annotation is to focus on regions which are considered salient. The novel contribution in this chapter is using the regions found from the maximization of distinctiveness (MOD) saliency approach for automatic visual concept detection. We present results based on real images and compare nearest neighbor classification, support vector machines and a neural network.

6.1 Introduction

Bridging the gap between low level features such as color histograms and semantic descriptions such as 'trees' is highly useful for searching image databases and digital libraries. From the panels of CIVR and MIR conferences, it has been said repeatedly that one of the primary goals of the community is automatic annotation of images and video. It would be ideal to have a program which receives as input an unknown image, analyzes the pictorial content of the image and then outputs a set of keywords. In this chapter we present our ongoing work in detecting visual concepts toward automatic image annotation.

The basis of visual concept detection is the need for automatically describing the content of images. For image retrieval tasks, these generated descriptions of images can be very useful. For example, the following image in Figure 6.1 could be classified as containing buildings and trees:

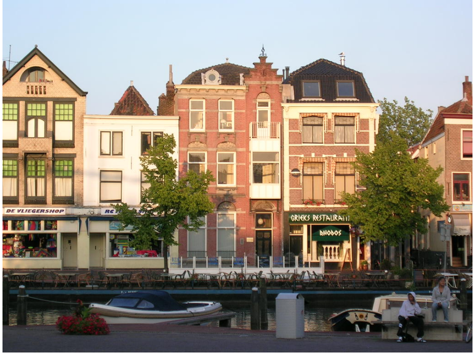
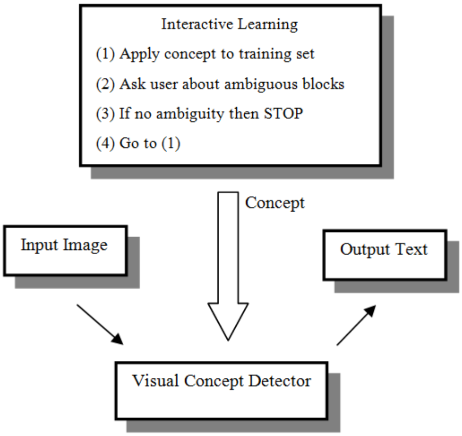


Figure 6.1: An image which could be labeled with trees, buildings and sky.

With this annotation, it is possible to search for these images using keywords. A high level overview of our system frame work is shown in Figure 6.2. In our system, a visual concept is learned interactively by our program using positive and negative responses from the user. Once the visual concept has been learned, it is applied to an unknown image and automatically outputs descriptive text.



Interactive Learning

(1) Apply concept to training set

(2) Ask user about ambiguous blocks

(3) If no ambiguity then STOP

(4) Go to (1)

Figure 6.2: The overall visual concept detection system.

For detecting visual concepts in images, we have compared the following methods: nearest neighbor classification, support vector machines and a neural network. We also use several approaches for segmenting images into sub-images using wavelet-based salient points and maximization of distinctiveness (MOD) based interest points, of which the visual concepts are determined.

6.2 Related work

In the recent years, visual concept detection or automated image annotation has become a very active field of research. As the amount of visual information available increases, the need for new methods for searching it also increases. For example, Srikanth et al. [84] describe an image annotation method that uses an ontology for linking annotation words. They measure the retrieval performance when annotation words are connected to other words in a hierarchy. Retrieval accuracy improves using this knowledge.

Other closely related work in the research literature for visual concept detection is face detection in complex images. Representative examples would be the neural network work by Rowley and Kanade [66] or the information theoretic approach by Lew and Huijsmans [39] in which the authors detect the specific visual concept of a human face. An excellent survey of the work done in face detection is found in Yang, et al. [95].

6.3 Maximization Of Distinctiveness (MOD)

Recently, we have proposed the Maximization of Distinctiveness paradigm [56] for detecting interest points in images. These interest points have the property that they are optimized for visual content matching using feature vectors. The selected points have the highest distinctiveness with respect to certain features, in a local region around the selected point.

For computing the MOD interest point, each pixel in an image is first assigned a distinctiveness value that is based on the dissimilarity of the point to all pixels in a neighborhood around it. The dissimilarity for a pixel is estimated as the inverse of the similarity value for the closest matching point in the neighborhood, when the features are considered that will be used for matching. The dissimilarity values are combined to form a distinctiveness map of an image and the local maxima in this distinctiveness map will yield the MOD interest points. The details of this method are described in chapter 5 of this thesis.

6.4 Detecting visual concepts

We used two steps in detecting visual concepts in images: Visual concept description and visual concept matching.

First, the visual concepts need to be described. This can be done by selecting positive and negative examples and to use a classification method which uses these positive and negative examples to create a general model of the concept.

For each of the positive and negative examples, a number of feature vectors is extracted. We have used an HSV based color feature, color moments and a texture feature created by Ojala [62], which is invariant under grayscale variations and rotation.

The second step for visual concept detection is to match each image to the visual concepts. Images are compared to the generalized model and are classified as either matching or non-matching for each visual concept. The list of matching concepts is then an annotation for the image.

For describing the visual concepts, we have used the interactively selected lists of positive and negative images.

For matching images with visual concepts, we have looked at three different methods: Nearest neighbor classification, support vector machines and a neural network. These methods are explained in detail in the next three sections. We expect classification using SVMs or the neural network to outperform the nearest neighbor method because of the generalization capabilities of these machine learning techniques, but we included the nearest neighbor method as a benchmark.

6.4.1 Classifiers

For our experiments, we have used three different classifiers. The details of each of these classifier types can be found in chapter 3 of this thesis. Here we only describe the parameters used (if any) for each of the classifiers.

First, as a benchmark, we used a k-nearest neighbor classifier that uses the label of the closest positive of negative example of a concept as a classification.

Secondly, we have used a neural network with one variable sized hidden layer and one output unit. In most experiments, the hidden layer consisted of 25 units.

The third classification technique we have used is the support vector machine (SVM). In these experiments, we have chosen to use a third order polynomial kernel, because we assume the feature vectors are not linearly separable.

6.5 Experiments

We have tested our system on a dataset of tourist-like images of the cities of Leiden and Amsterdam and several other images. This section shows the user interface of the program, the results of applying concepts to a few of these images and also the results of an application of visual concept detection to face detection.

Figure 6.3 shows the images that were selected from the tourist database for detecting trees, buildings and blue sky. For the beach and face concepts, we have selected images from the web.

Figure 6.4 shows the interface of our visual concept detection program. Users can add concepts and assign image regions to the positive or negative examples for each concept.

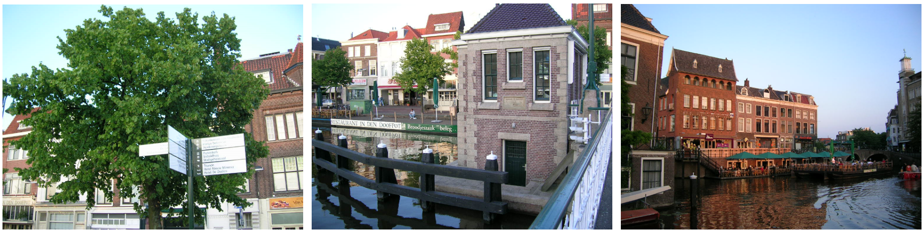
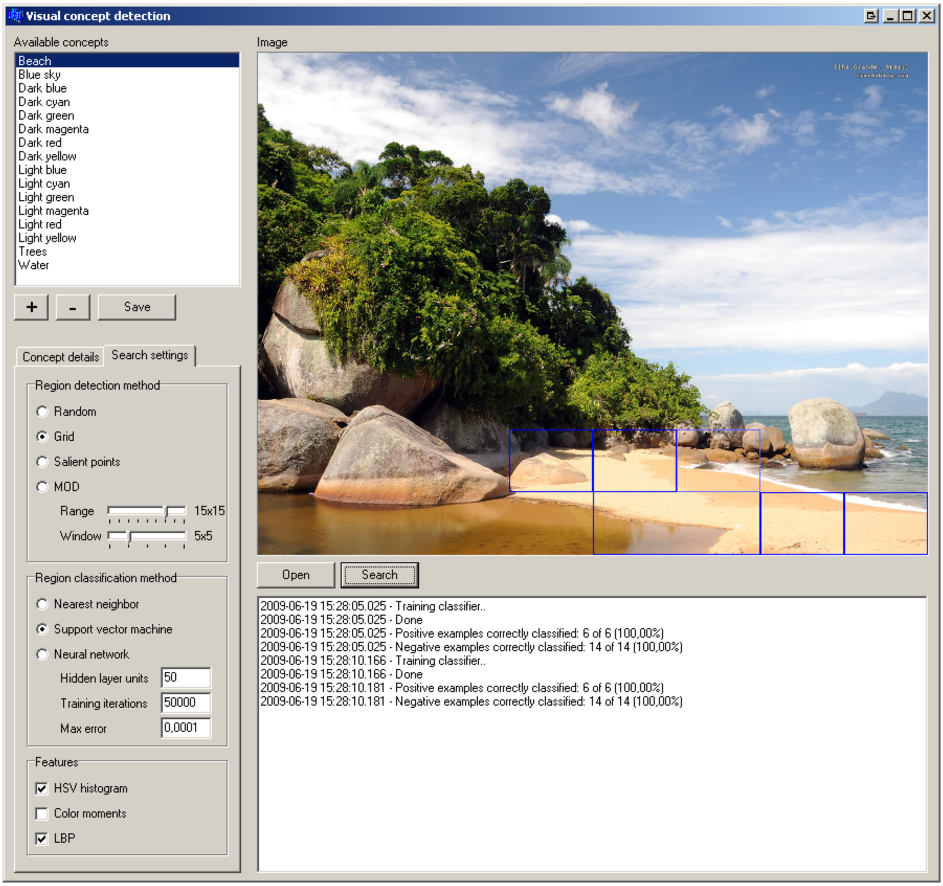


Figure 6.3: Images selected from the Tourist Database.



Available concepts

Beach
Blue sky
Dark blue
Dark cyan
Dark green
Dark magenta
Dark red
Dark yellow
Light blue
Light cyan
Light green
Light magenta
Light red
Light yellow
Trees
Water

Region detection method

○ Random
● Grid
○ Salient points
○ MOD
Range 15x15
Window 5x5

Region classification method

○ Nearest neighbor
● Support vector machine
○ Neural network
Hidden layer units 50
Training iterations 50000
Max error 0,0001

Features

☑ HSV histogram
□ Color moments
☑ LBP

```
2009-06-19 15:28:05.025 - Training classifier..
2009-06-19 15:28:05.025 - Done
2009-06-19 15:28:05.025 - Positive examples correctly classified: 6 of 6 (100,00%)
2009-06-19 15:28:05.025 - Negative examples correctly classified: 14 of 14 (100,00%)
2009-06-19 15:28:10.166 - Training classifier..
2009-06-19 15:28:10.166 - Done
2009-06-19 15:28:10.181 - Positive examples correctly classified: 6 of 6 (100,00%)
2009-06-19 15:28:10.181 - Negative examples correctly classified: 14 of 14 (100,00%)
```

Figure 6.4: The user interface of the Visual Concept Detection program.

6.5.1 Tree detection

This section shows where the concept 'Tree' was detected in some of the Tourist Database images. Figures 6.5 to 6.7 contain the examples.

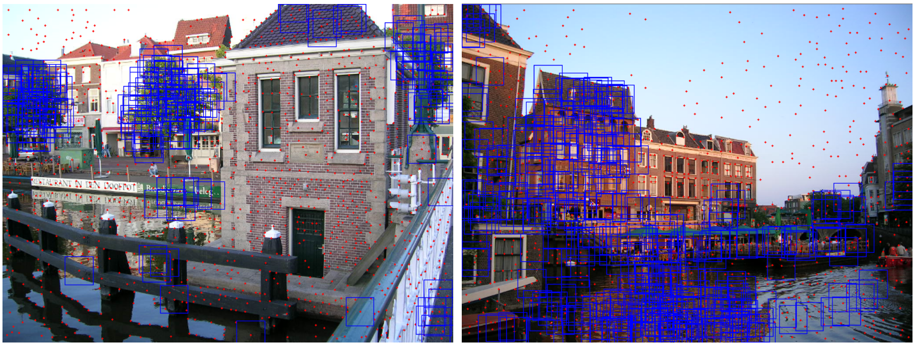


Figure 6.5: Tree detection with nearest neighbor classification.

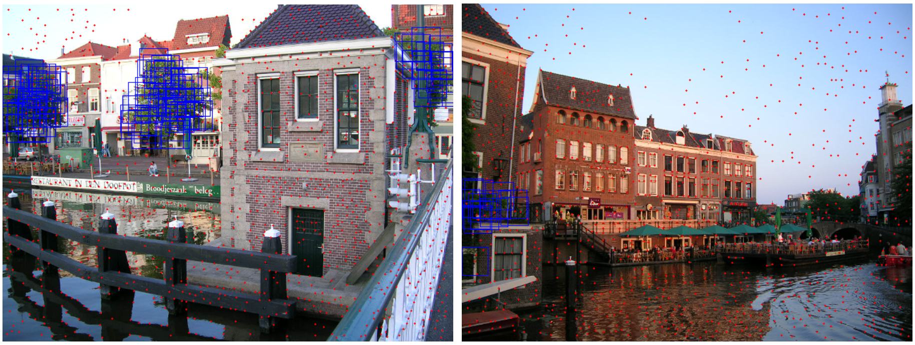


Figure 6.6: Tree detection with SVM classification.

6.5.2 Building detection

This section shows where the concept 'Building' was detected in some of the Tourist Database images. All examples are contained in figures 6.8 to 6.10.

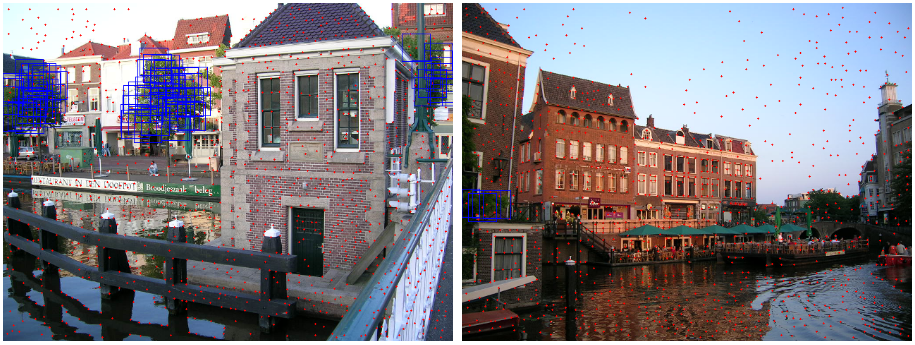


Figure 6.7: Tree detection with neural network classification.

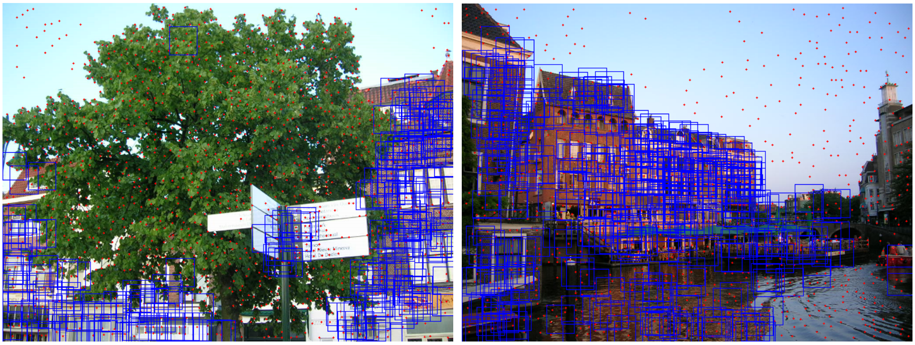


Figure 6.8: Building detection with nearest neighbor classification.

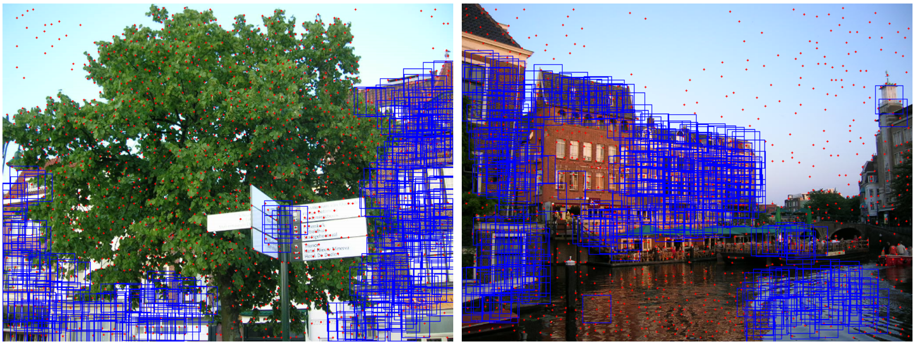


Figure 6.9: Building detection with SVM classification.

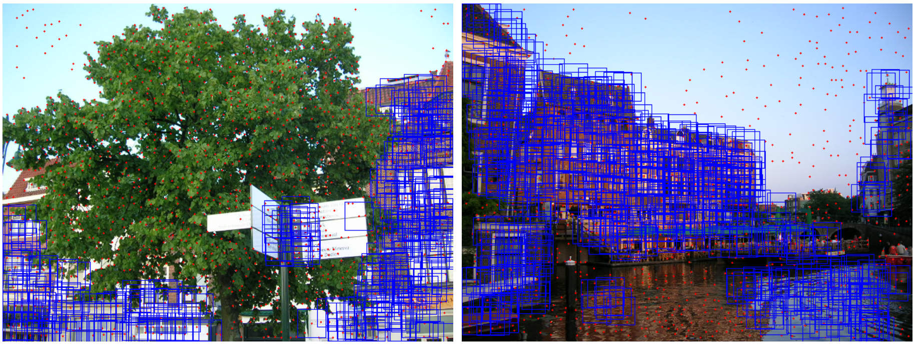


Figure 6.10: Building detection with neural network classification.

6.5.3 Sky detection

This section shows where the concept 'Sky' was detected in some of the Tourist Database images. All examples are contained in figures 6.11 to 6.13.

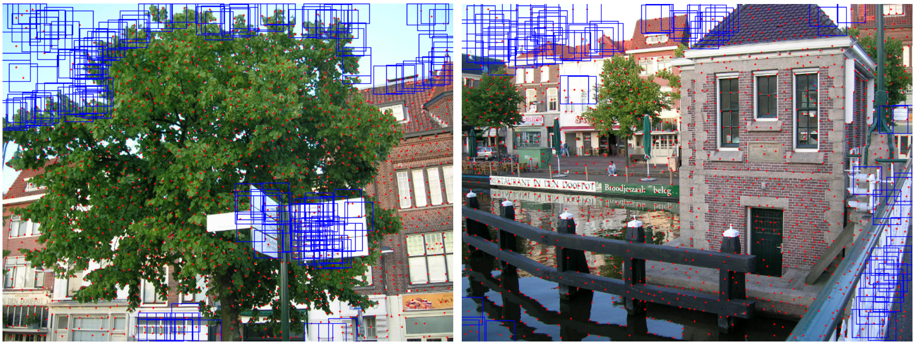


Figure 6.11: Sky detection with nearest neighbor classification.

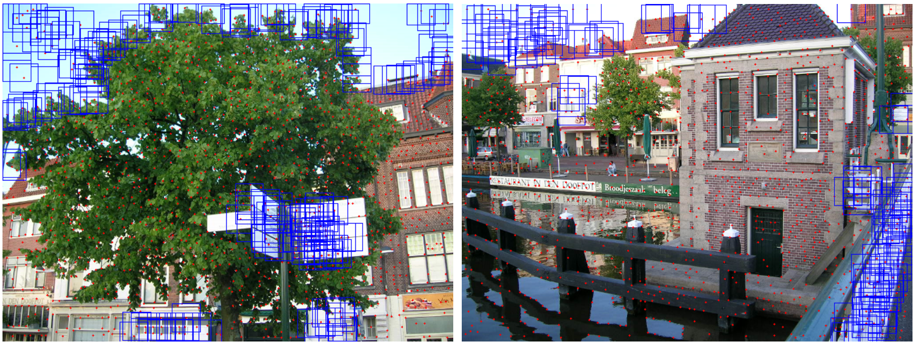


Figure 6.12: Sky detection with SVM classification.

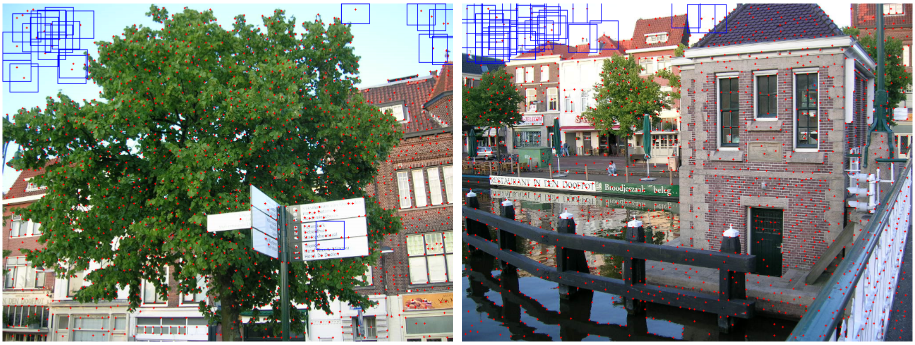


Figure 6.13: Sky detection with neural network classification.

6.5.4 Beach classification

This section shows the detection of the 'Beach' concept on an image from the web. Figure 6.14 shows the example.

6.5.5 Face detection

Figure 6.15 shows the results of applying a 'Face' concept to an image. In this case, the training image set and test set were very small, but we present this result as an example of other uses of the visual concept detection.

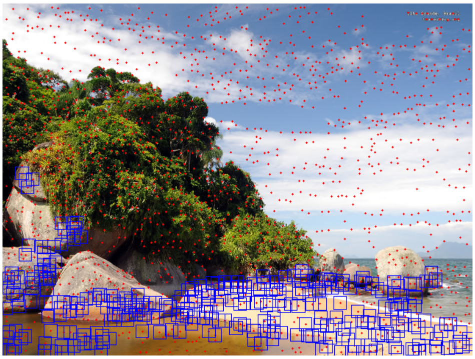


Figure 6.14: Beach detection with neural network classification.

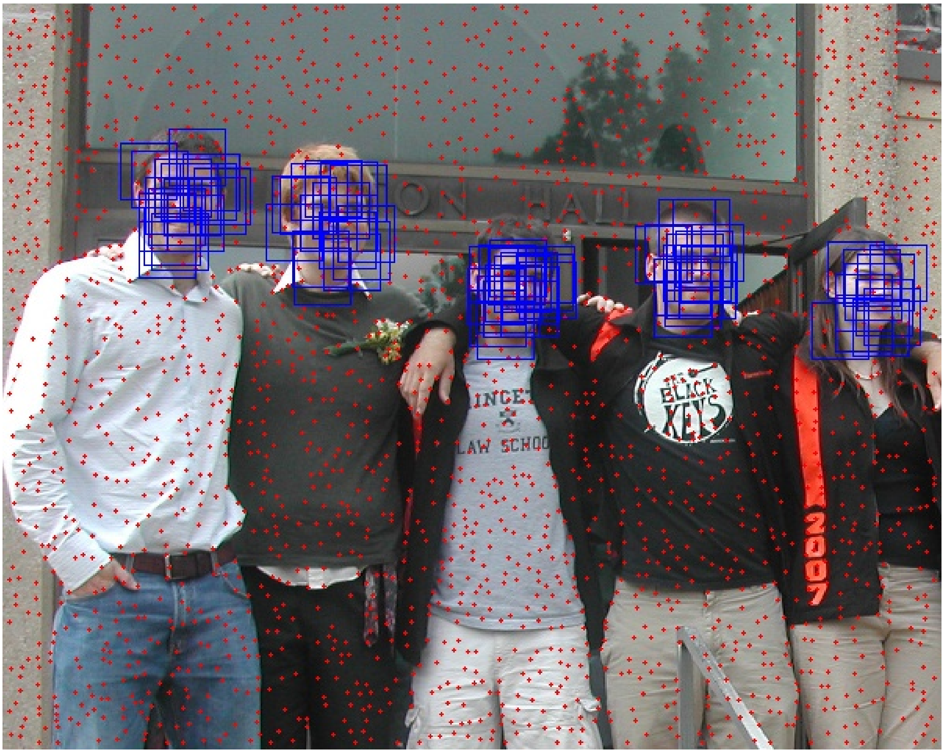


Figure 6.15: Output of the visual concept detector for the 'face' concept.

6.6 Experiments on MIRFLICKR-25000 dataset

The MIRFLICKR-25000 dataset was presented at the MIR conference held in 2008 [31]. The dataset contains 25000 images that were retrieved from the Flickr website. All original user annotations are available, as well as the EXIF metadata. Also, a ground truth is supplied for a large number of visual concepts, from general to very specific topics. For example, the concept 'people' is annotated, but also 'baby'.

The ten general topics of the dataset annotations are:

- Animals
- Indoor
- Night
- People
- Plant life
- Sky
- Structures
- Sunset
- Transport
- Water

We have used these general annotations in our experiments. We have compared the classification accuracy of SIFT [44] features with a support vector machine (SVM) as a classifier to MOD interest points with both an SVM and a neural network as classifiers with a basic set of features.

For the MOD based classification, we used the EWF, an Extended Wavelet Feature set which combines the wavelet [85] [86] [92] representation of a grayscale version of the image region [73], with the HSV Histogram [74] and Local Binary Patterns [60] [75].

For a comparative benchmark, SIFT features with an SVM classifier have shown to be a good classifier for visual concepts [3], so we have chosen to use that as a baseline method.

The SVM classifiers were constructed using a radial basis function (RBF) kernel function, which has been suggested as the optimal kernel in our context [6].

For our tests, we have selected a set of 50 training images for each concept. The training set contained 25 positive examples and 25 negative examples. Each positive example was manually segmented into regions that contain the concept. Only interest points within these regions were used in the tests.

For the test set for each concept, we have randomly selected 5000 images from the dataset, of which 50% were positively labeled and 50% were negatively labeled, although not all concepts have 2500 positive examples in the dataset. In those cases, more negatives were added.

First, we give the detailed annotation results for a few of the general concepts. We show graphs of the true positive rate compared to the true negative rate. The true positive rate is defined as the fraction of positives detected:

$$true\ positive\ rate = tp/(tp + fn) \tag{6.1}$$

Note that in a classification context, this is also known as the 'recall' value. The true negative rate is then defined as the fraction of negatives detected:

$$true\ negative\ rate = tn/(tn + fp) \tag{6.2}$$

In order to give a proper idea of the diversity and difficulty level of the concepts, we show in Figures 6.16 to 6.45 examples of interesting concepts in each subsection below along with the performance graphs and the detected salient points corresponding to the visual concept. In Figure 6.46 we show the averaged results over all of the concepts.

6.6.1 Concept 'Animals'

Figure 6.16 shows some example images for the 'Animals' concept. These images show the wide variety of images that can be classified as containing animals. Not only real animals that are clearly visible, but also hand drawn animals or parts of an animal result in the same annotation. Also note that the animal does not have to be the subject of the image, but it might also be seen in the background.

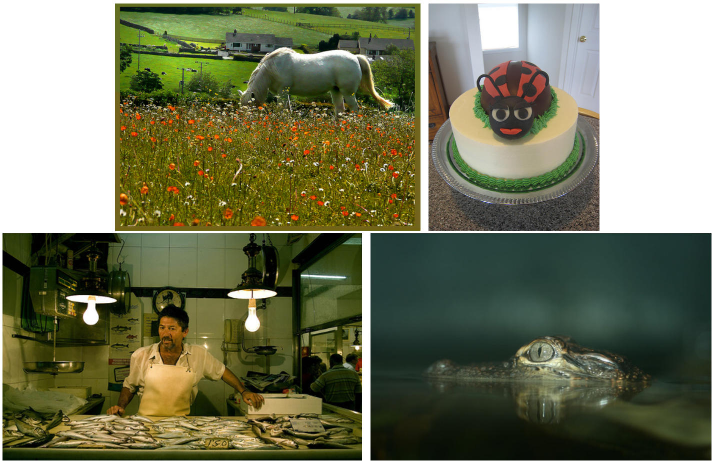


Figure 6.16: Some example images for the 'Animals' concept.

Figure 6.17 shows the classification results for the 'Animals' concept and figure 6.18 shows some detection examples of the MOD based concept detection.

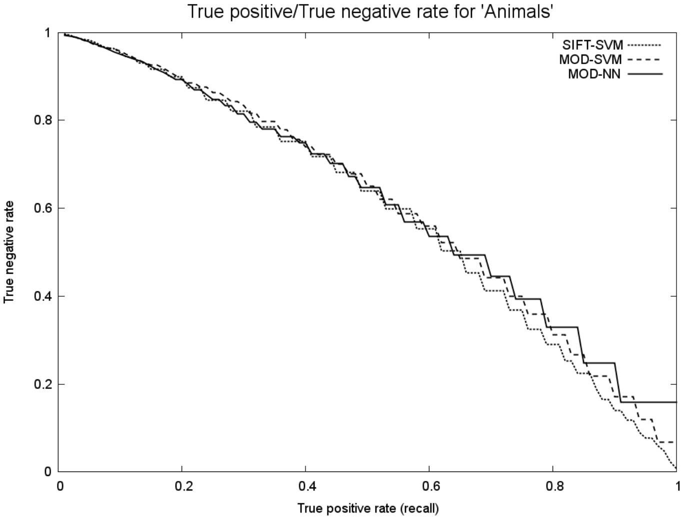


Figure 6.17: Classification results for the 'Animals' concept.

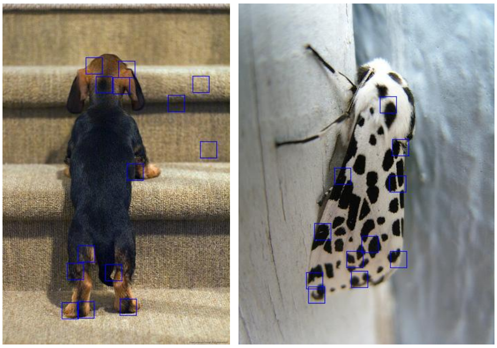


Figure 6.18: Some MOD based concept detection results for the 'Animals' concept.

As explained above, the Animals concept contains a wide variety of images and it seems that 25 training images is probably not enough to create a reliable classifier for all images with animals. As there are many more different types of animals than there are training images, a larger training set should be used.

6.6.2 Concept ’Indoor’

Figure 6.19 shows some example images for the ’Indoor’ concept. The concept ’Indoor’ contains images that were taken indoors. However, the images do not explicitly have to show a typical indoor situation like a room with a view to the outside world. Intuitively, the concept can be described as for example ’not containing sky’ or ’not containing grass’.

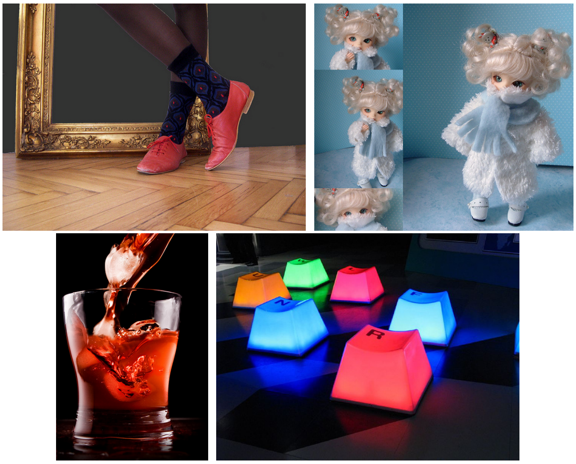


Figure 6.19: Some example images for the ’Indoor’ concept.

Figure 6.20 shows the classification results for the ’Indoor’ concept and figure 6.21 shows some detection examples of the MOD based concept detection.

The detection of the concept ’indoor’ shows to be very difficult. The real question is if there is a direct relationship between visual properties of an image region and the concept indoor. The concept cannot be pointed out in an image, although the image does get the label ’indoor’. The results show that the number of detected regions does give a hint of the probability of the image being indoor.

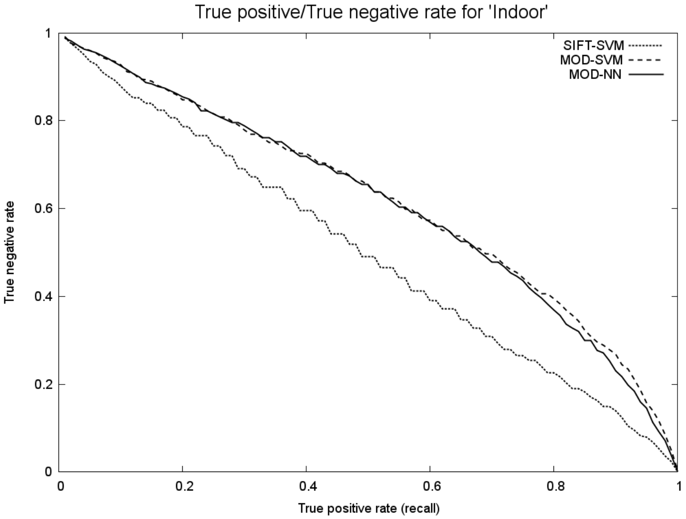


Figure 6.20: Classification results for the ’Indoor’ concept.

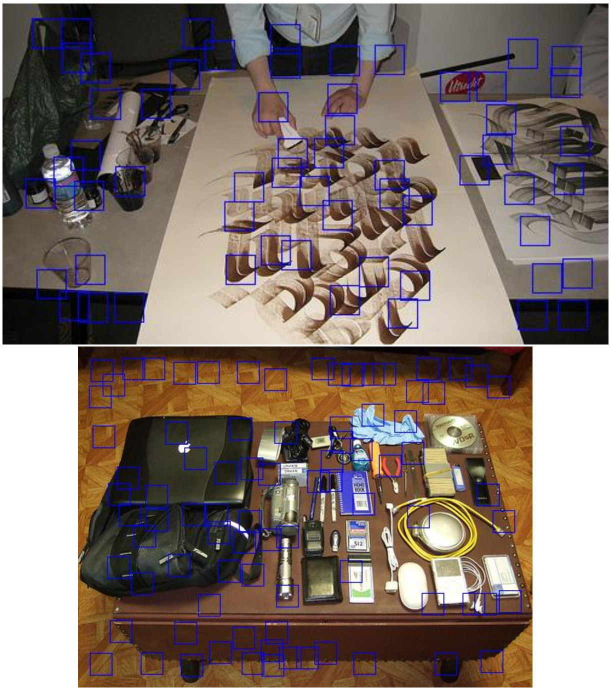


Figure 6.21: Some MOD based concept detection results for the ’Indoor’ concept.

6.6.3 Concept ’Night’

Figure 6.16 shows some example images for the ’Night’ concept. The concept ’night’ contains images that were taken at night. Although this sounds like a trivial concept to detect, because the images would usually be dark in color, also the contents are needed to determine if an image was really taken at night. Not all areas of an image need to be dark for it to be taken at night and not all images with many dark areas were taken at night.

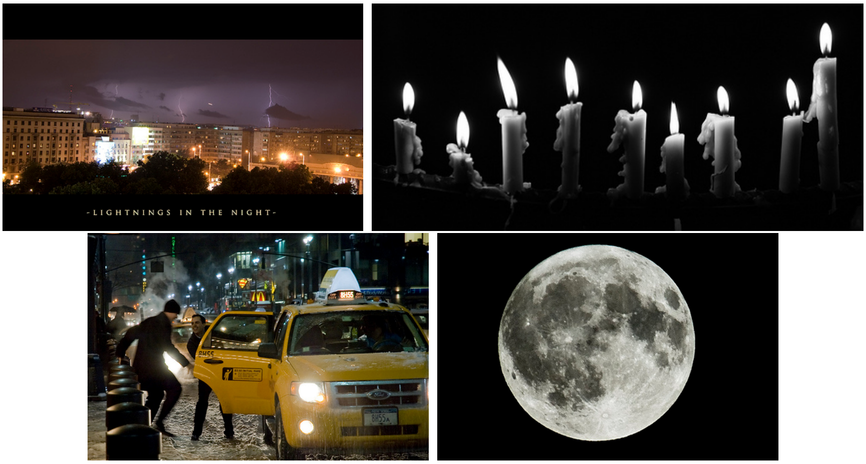


Figure 6.22: Some example images for the ’Night’ concept.

Figure 6.23 shows the classification results for the ’Night’ concept and figure 6.24 shows some detection examples of the MOD based concept detection.

The results of the MOD based detector show that many image areas that are not dark, are still classified as being part of the night concept. We anticipated this earlier when discussing what images in this concept class would look like. The retrieval performance of the MOD based methods clearly outperform the SIFT based method.

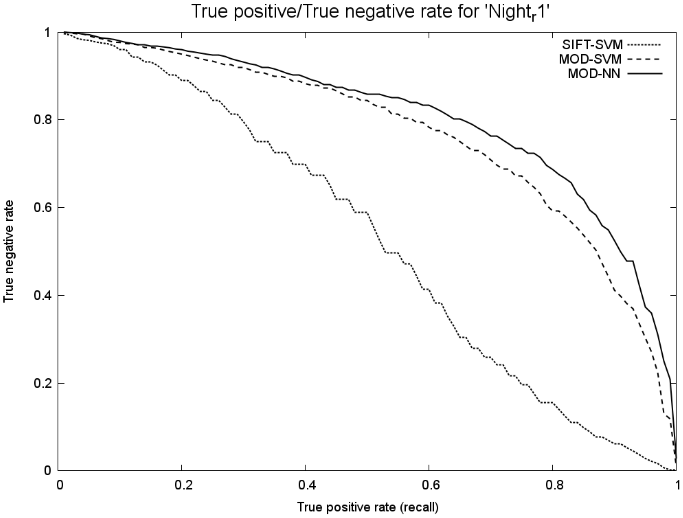


Figure 6.23: Classification results for the 'Night' concept.

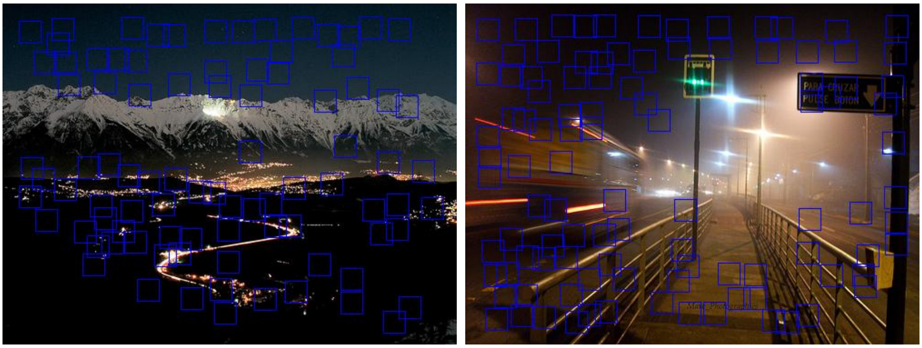


Figure 6.24: Some MOD based concept detection results for the 'Night' concept.

6.6.4 Concept 'People'

Figure 6.25 shows some example images for the 'People' concept. In the MIRFLICKR-25000 dataset, there are two different annotations for the 'People' concept. The annotation 'people' is a less strict annotation than 'people_r1', in which one or more persons are really clearly visible, probably even the subject of the picture. We have used the 'people_r1' annotation. The example images show that in some images only a face is visible, but in other images the entire person can be seen or even a large crowd of people.

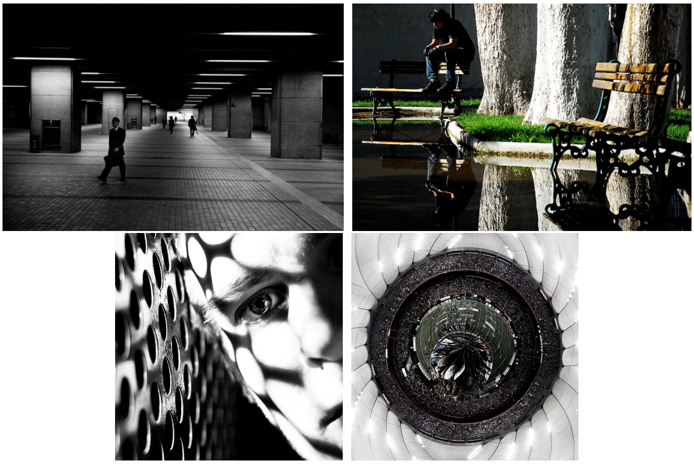


Figure 6.25: Some example images for the 'People' concept.

Figure 6.26 shows the classification results for the 'People' concept and figure 6.27 shows some detection examples of the MOD based concept detection.

Compared to the results for 'Animals', the 'People' concept detection results are more promising for the MOD method. Both the SVM and the NN classifiers slightly outperform SIFT-SVM.

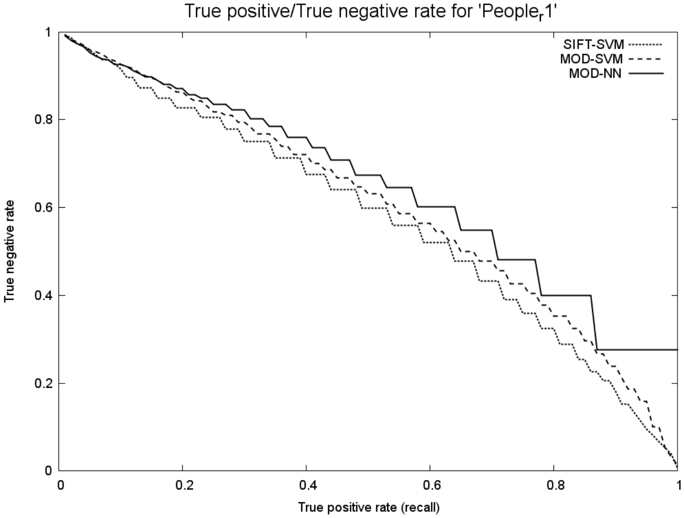


Figure 6.26: Classification results for the 'People' concept.

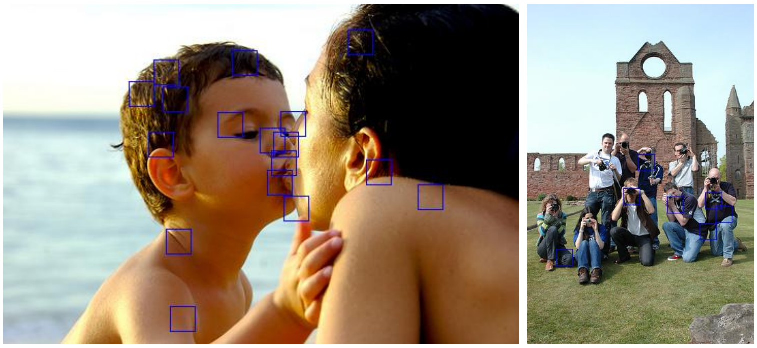


Figure 6.27: Some MOD based concept detection results for the 'People' concept.

6.6.5 Concept 'Plant life'

Figure 6.28 shows some example images for the 'Plant life' concept. The 'plant life' concept is intuitively linked to green plants, but the images again show the wide variety of images that are in the dataset. The 'Plant life' concept contains plants, trees, grass, flowers.

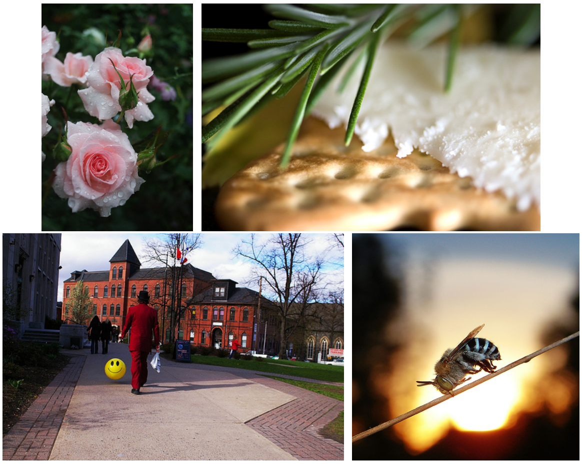


Figure 6.28: Some example images for the 'Plant life' concept.

Figure 6.29 shows the classification results for the 'Plant life' concept and figure 6.30 shows some detection examples of the MOD based concept detection.

The graph again shows an improved detection rate for the MOD based methods, however in this case the MOD-SVM outperforms the MOD-NN method, which we have not seen for the previous two concepts.

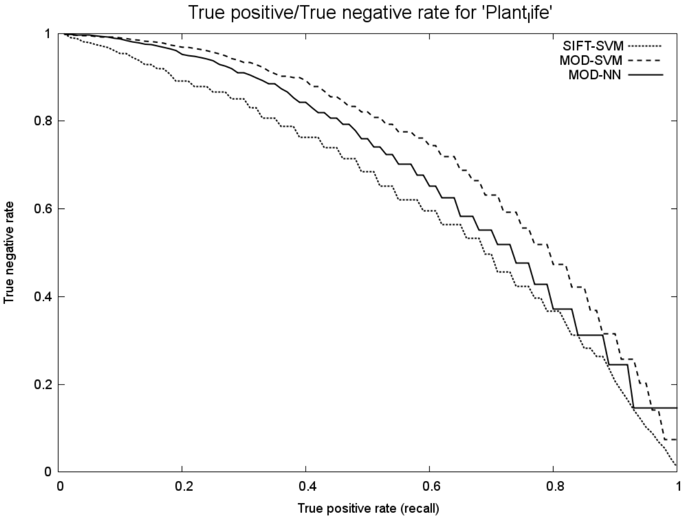


Figure 6.29: Classification results for the ’Plant life’ concept.

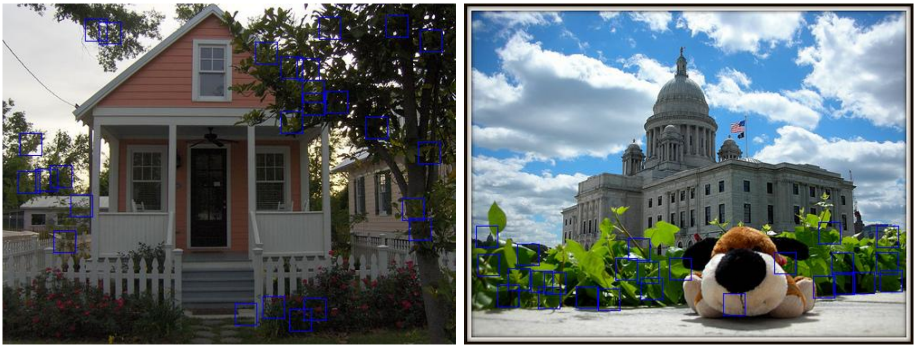


Figure 6.30: Some MOD based concept detection results for the ’Plant life’ concept.

6.6.6 Concept ’Sky’

Figure 6.31 shows some example images for the ’Sky’ concept. The ’Sky’ concept is usually regarded as an easy concept to detect. Blue sky has a very specific color and almost no texture. However, in this dataset, the sky concept can also be considered a reasonably difficult one. Cloudy skies and night skies, or just a vague notion of sky somewhere in the image are the main reasons for this.

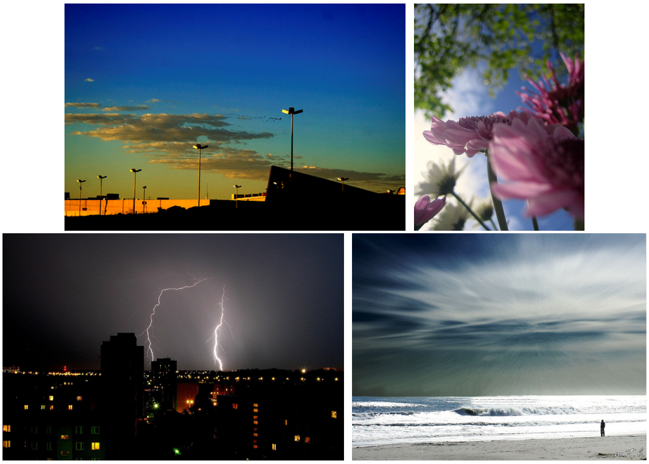


Figure 6.31: Some example images for the ’Sky’ concept.

Figure 6.32 shows the classification results for the ’Sky’ concept and figure 6.33 shows some detection examples of the MOD based concept detection.

The graph above shows the classification results for the ’sky’ concept. Again, the MOD based methods outperform the SIFT method.

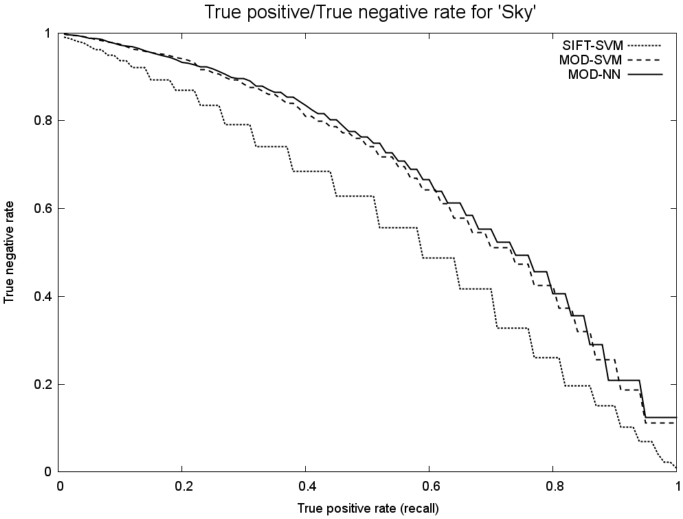


Figure 6.32: Classification results for the 'Sky' concept.

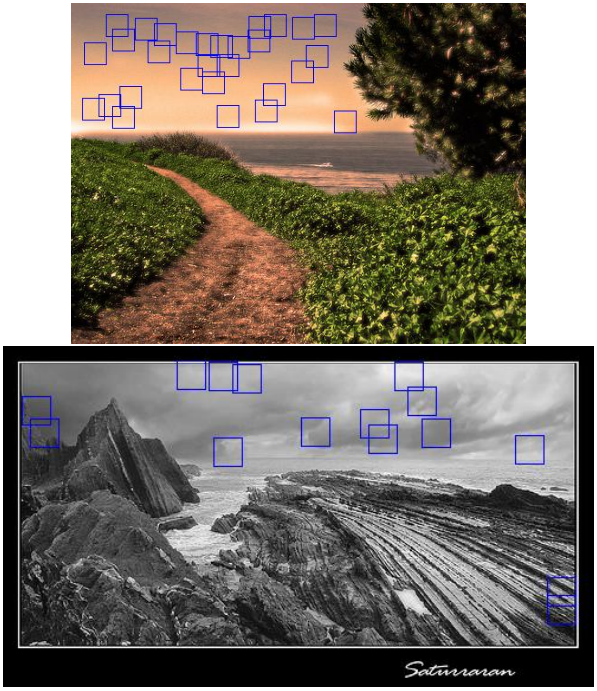


Figure 6.33: Some MOD based concept detection results for the 'Sky' concept.

6.6.7 Concept ’Structures’

Figure 6.34 shows some example images for the ’Structures’ concept. The ’Structures’ concept contains images with man-made structures on it. One can think of buildings, roads, bridges or fences.

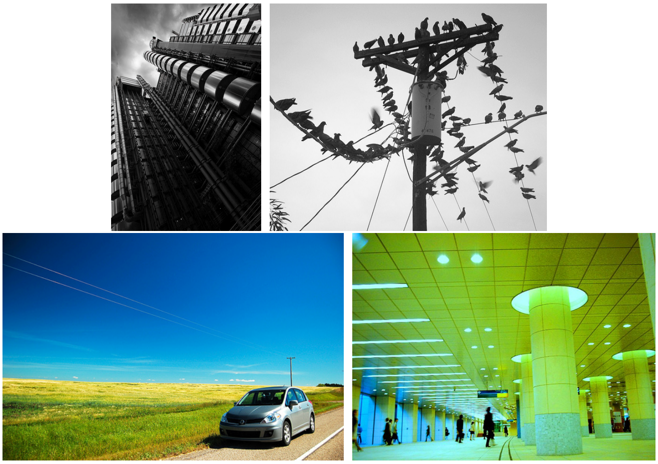


Figure 6.34: Some example images for the ’Structures’ concept.

Figure 6.35 shows the classification results for the ’Structures’ concept and figure 6.36 shows some detection examples of the MOD based concept detection.

In this case, the increase in performance is less obvious, although the MOD methods perform slightly better.

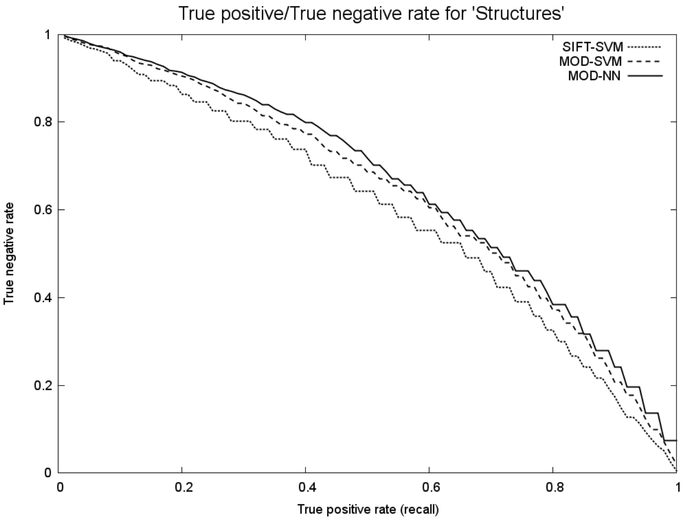


Figure 6.35: Classification results for the ’Structures’ concept.

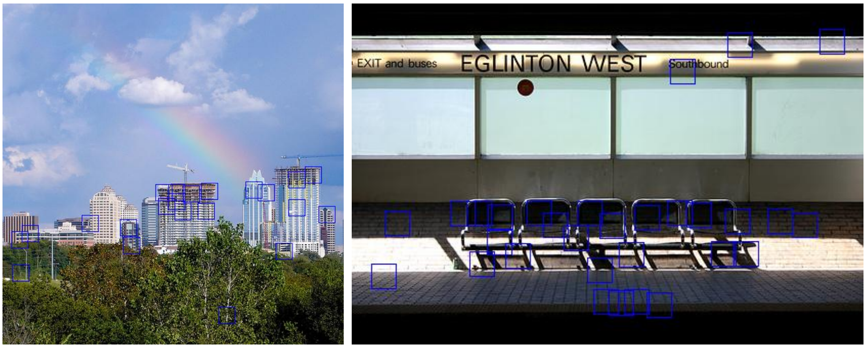


Figure 6.36: Some MOD based concept detection results for the ’Structures’ concept.

6.6.8 Concept ’Sunset’

Figure 6.37 shows some example images for the ’Sunset’ concept. The ’Sunset’ concept is a typical color-based concept. Usually the colors red, orange or dark blue can be found and one would expect higher classification accuracy because of the more obvious relation between image features and the concept.

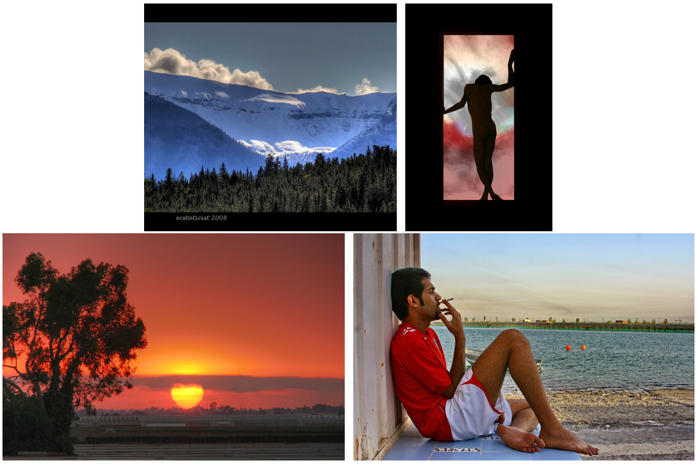


Figure 6.37: Some example images for the ’Sunset’ concept.

Figure 6.38 shows the classification results for the ’Sunset’ concept and figure 6.39 shows some detection examples of the MOD based concept detection.

The test set for the sunset concept did not contain 50% positively labeled images, as there are not enough positive images in the MIRFLICKR dataset. So, if the accuracy would be plotted in a graph (not visible here), it would flatten around 0.58, which is the percentage of negatively labeled images. For very high thresholds, all images will be classified as not containing the concept, so for 58% of the images this will be true, instead of the expected 50% as with the other tested concepts.

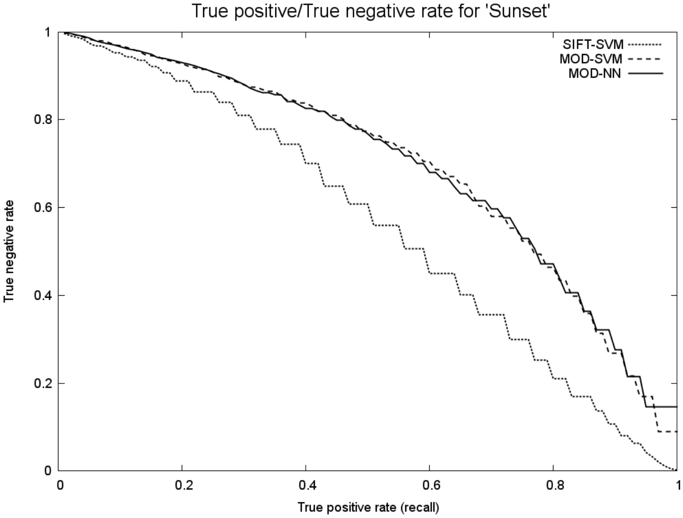


Figure 6.38: Classification results for the 'Sunset' concept.

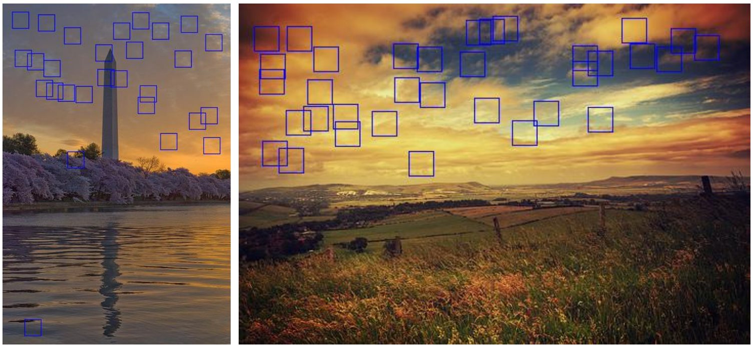


Figure 6.39: Some MOD based concept detection results for the 'Sunset' concept.

6.6.9 Concept ’Transport’

Figure 6.40 shows some example images for the ’Transport’ concept. The ’Transport’ concept covers all kinds of transport, such as cars, boats, bicycles. Like with other concepts, often only partial views of a concept are visible in the images. A few common shapes exist among the cars and bicycles, for example the tires. These are round and black. For boats, water is usually present around them.

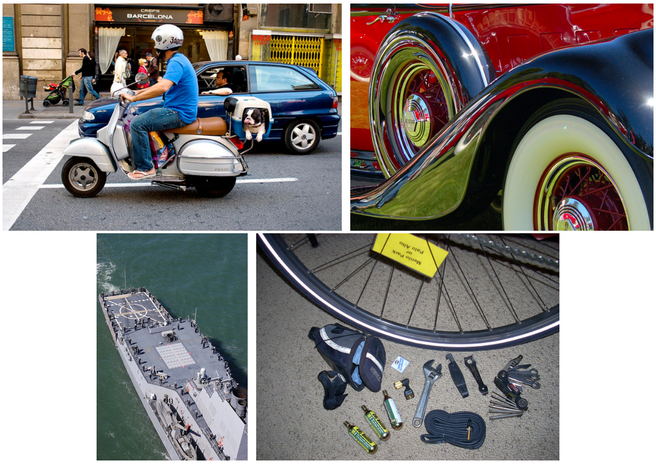


Figure 6.40: Some example images for the ’Transport’ concept.

Figure 6.41 shows the classification results for the ’Transport’ concept and figure 6.42 shows some detection examples of the MOD based concept detection.

Just like the concept indoor, the transport concept is a difficult concept to grasp with visual descriptors. Many different objects and situations relate to the word transport and as such, a high retrieval performance was not expected from our classifiers with the limited training set. Still, the MOD based methods perform better than the SIFT based method.

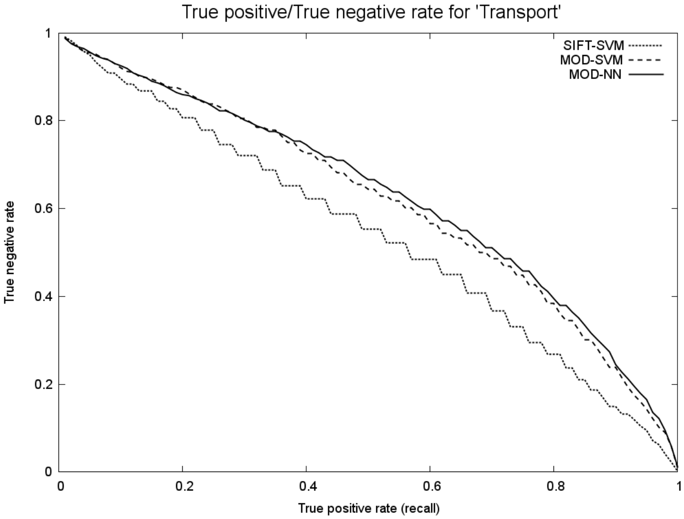


Figure 6.41: Classification results for the 'Transport' concept.

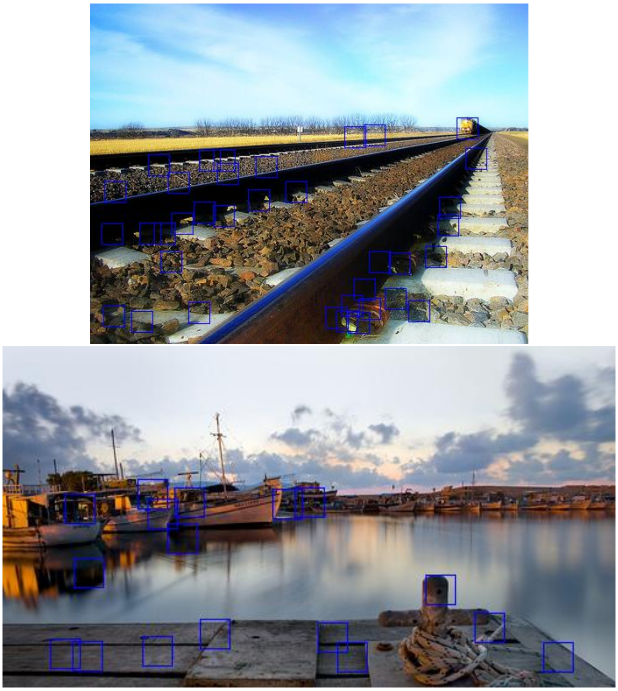


Figure 6.42: Some MOD based concept detection results for the 'Transport' concept.

6.6.10 Concept 'Water'

Figure 6.43 shows some example images for the 'Water' concept. The concept 'Water' refers to more than what one would first expect: not just water like a river or an ocean, but also rain or an aquarium are covered by this concept. The standard 'blue and ripples' description is clearly not sufficient.

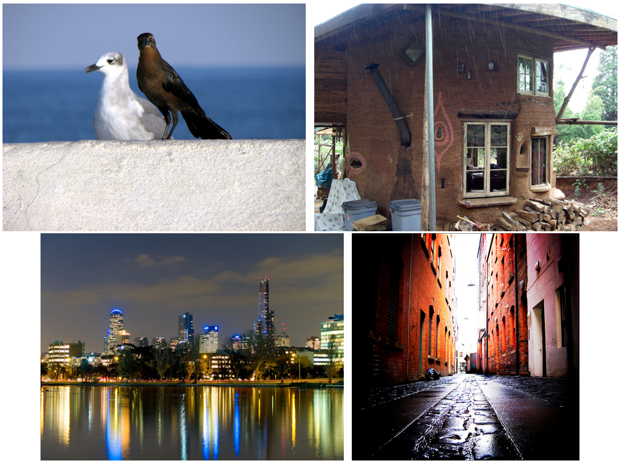


Figure 6.43: Some example images for the 'Water' concept.

Figure 6.44 shows the classification results for the 'Water' concept and figure 6.45 shows some detection examples of the MOD based concept detection.

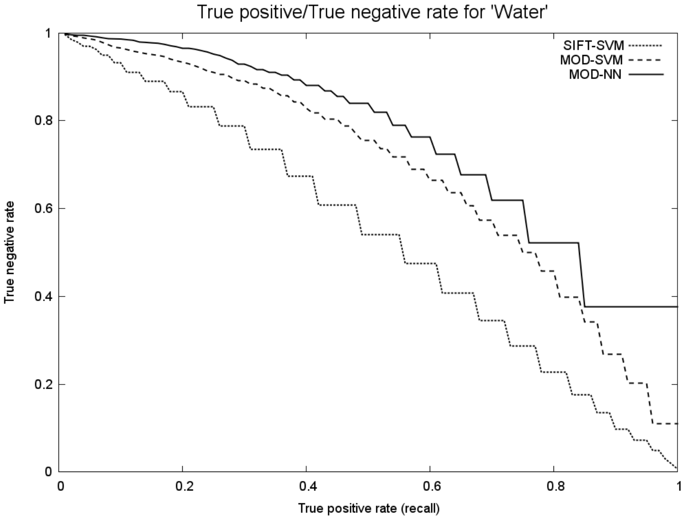


Figure 6.44: Classification results for the 'Water' concept.

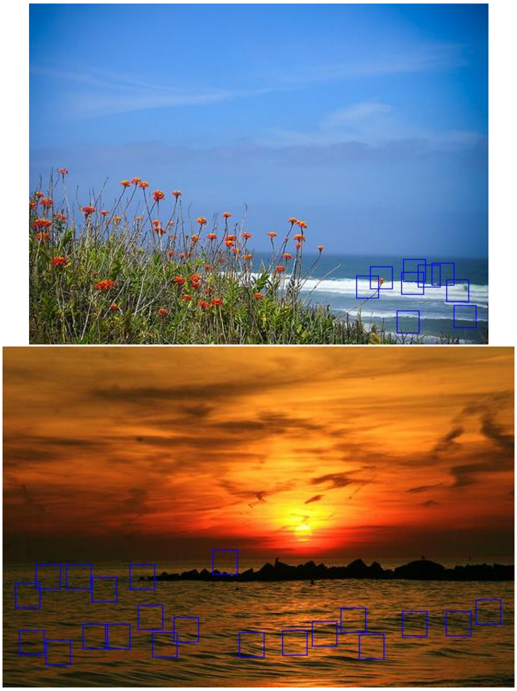


Figure 6.45: Some MOD based concept detection results for the 'Water' concept.

6.6.11 Overall results

Figure 6.46 shows the overall Recall versus True Negative rate. The MOD based methods outperform the SIFT based method, but the distinction between SVM and a neural network for the MOD based methods is not obvious. The neural network slightly outperforms the support vector machine approach, especially for high recall values.

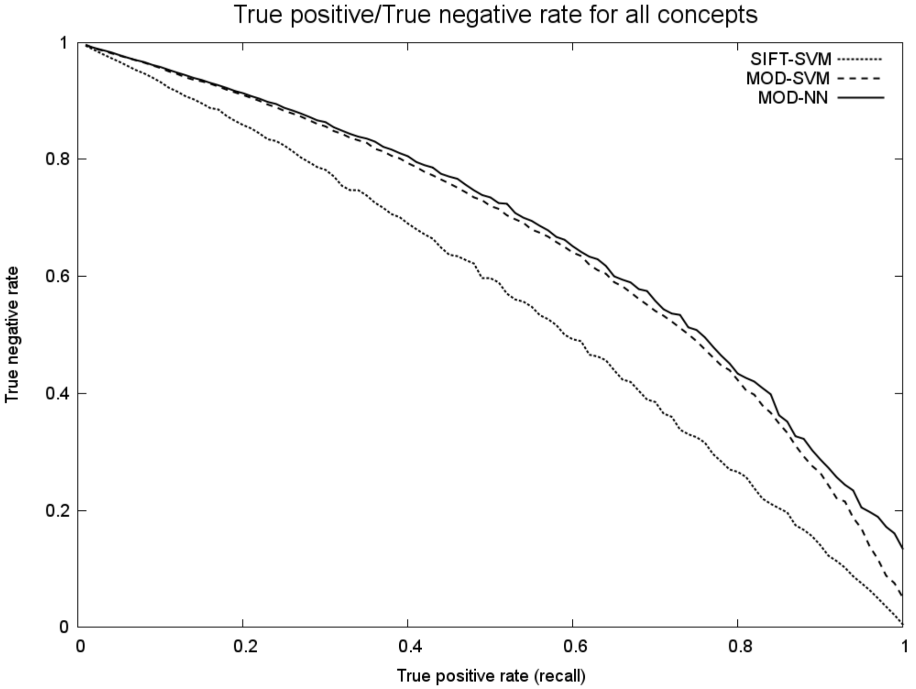


Figure 6.46: Average performance over all concepts.

6.7 Discussion, conclusions and future work

The results from the different concepts and classifiers show the promising results that can be obtained when using the MOD interest points as a hint on where to look for concepts within images. Especially the combination of a more advanced classifier like the support vector machine or a neural network with the MOD regions turns out to give very interesting results.

As an example, detecting the 'tree' concept, as shown in Figure 6.9, with a support vector machine yields a correct classification of a small image region, that we did not expect to be easily detected.

The other results on the Leiden-Amsterdam database show that the ’building’ concept detection results in some errors. The nature of this concept is visually more complex than a ’sky’ concept, so we expected more difficulties with detecting the concept, which can be clearly seen in Figure 6.10.

For the MIRFLICKR experiments, our research has focused on a new method of interest point detection for sub-image visual concept detection. Our experiments have shown that the MOD based classifiers with a few standard features outperform the SIFT based classifier with SIFT feature descriptors, a method that has been recently shown to be very effective for visual concept detection.

The results of the experiments also indicate that for the far majority of the concepts, the neural network based classifiers have equal or greater performance than the SVM based classifiers. On average over all the concepts, the neural network has a 7.4% improvement, although the main contribution comes from the highest recall values, where the neural network clearly outperforms SVM.

In our results we compare relative detection rates for the same training and test sets between MOD and SIFT salient points. We expect that the absolute detection rates can be significantly improved by using more training images, which we intend on pursuing in the future.

Also for future research, we would like to improve our MOD based concept detector by using other features. In the current tests, only three basic features were used and we expect that for example a scale-invariant feature would benefit the detection of concepts that vary in size.