Content-based retrieval of visual information

Oerlemans, A.A.J.

Chapter 5

Interest Points Based on Maximization of Distinctiveness

Interest or salient points are typically meaningful points within an image which can be used for a wide variety of image understanding tasks. In this chapter we present a novel algorithm for detecting interest points within images. The new technique is based on finding the locations in an image which exhibit local distinctiveness. We evaluate our algorithm on the Corel stock photography test set in the context of content based image retrieval from large databases and provide quantitative comparisons to the well known SIFT interest point and Harris corner detectors as a benchmark.

5.1 Introduction

In a typical content-based image retrieval [40] task, image features are compared for matching images. When the image features are close, it is assumed the images are similar. These features can be computed globally (over the entire image) or locally (over small parts of the image). For locally computed image features, it is necessary to determine which image points should be used for describing the image content. These image points are called interest points and various methods exist to select these points.

We introduce a novel method of computing interest points based on local uniqueness and evaluate the effectiveness.

5.2 Related work

Many interest point detectors are available [1] [40] [44] [43] [51] [52] [89], and depending on the application, different performance measures can be chosen. Arguably, the original interest point detector was created by Moravec [52] who needed to find extremely computationally efficient methods for performing real time robotic navigation. In the 70s, it was impossible to perform real time video analysis on a mobile computer so his necessity led to the invention of interest points. In current times, there are now other data intensive tasks, one of which is content based image retrieval from large databases typically measured in the thousands to millions of images. In this image retrieval context, it is again important to have information efficient descriptors to perform content based searches in a user acceptable response time.

A good overview is given in Sebe, et al. [72] and also Schmid, et al. [71]. In these works, it is clear that one of the best performing interest or salient point detectors is the Harris corner detector. The Harris corner detector [23] is an interest point detector that is invariant to rotation, scale and illumination changes. It uses the auto-correlation function for comparing a small part of an image to the area around it. SIFT [44] [43] features are invariant to changes in scale and rotation. Trujillo and Olague [89] use genetic programming to detect salient points.

There are a wide variety of methods of evaluating different interest point detectors. Schmid et al. [71] use two evaluation criteria for interest points: repeatability rate and information content. The former criterion determines the stability of the interest point under various transformations. The latter is a measure of the distribution of the feature values for those interest points. A distribution that is spread out indicates more information content. Sebe, et al. [80] suggest a good measure is using the information content as measured by the average information content of all messages or the entropy. Tian et al. [88] use the retrieval accuracy in a content based image retrieval task to evaluate their wavelet-based salient point method. So far we have briefly discussed what different methods exist. In the next section we will discuss why different methods are interesting and explain the fundamental motivation behind our own interest/salient point paradigm.

5.3 Maximization Of Distinctiveness (MOD)

In Moravecs [52] original interest operator from 1979, the main intuition was to use points which had high x and y gradients which in principle would be distinctive just as there are typically far fewer edge pixels than non-edge pixels within an image.

Nearly a decade later, Harris [23] came up with a robust method for detecting corners which also had high x and y gradients and are an intuitive method for

salient point detection. The usage of Moravecs and Harris work as salient points was intuitive but also in our opinion adhoc.

Recently, Lowe [44] proposed the Scale Invariant Feature Transform (SIFT) method which focuses on looking through the neighboring scales to find extrema in the difference of the Gaussian which is an approximation to the Laplacian of the Gaussian. The fundamental notion was to find scale invariant interest points on the assumption that scale invariance was important to good features:

L = set of extrema in the Laplacian of the Gaussian
S = set of stable points in L
H = set of higher contrast points in S (see page 98 of [44])
P = set of H where edge responses are eliminated.

In addition to P, one or more orientations are assigned to each element of P and a descriptor is computed using the gradient magnitudes and orientations for each level of the image pyramid where the orientations are adjusted for the assigned orientation.

While we agree that the scale invariance is a useful aspect of a good feature, it depends on the particular context. In many areas such as texture classification, image retrieval, video retrieval, stereo matching, and motion estimation contexts, the scale is assumed to be very similar between the correspondences or between the query and the results images. For example, in texture classification, one does not want to match a fine grain texture with a coarse grain texture. In summary, there are many areas where the scale is important and where the variation of scale is beneficial in reducing candidates and maximizing the accuracy of a matching algorithm.

5.3.1 The MOD paradigm

In the scale dependent visual matching areas such as stereo matching or motion estimation, the typical techniques are variations of feature vector matching, of which the most popular and intuitive method would be template matching.

Unlike the SIFT method which strives to find points which are scale invariant, we strive to find points which optimize distinctiveness in matching. In matching, the most common problem to address is the one to many mapping, where one point may have many good matches. We first compute a distinctiveness of matching measure which finds the most distinctive point in a local region based on the matching method to be finally used. This means searching the neighboring region around a point, computing the dissimilarity to each point in the region and then estimating the distinctiveness as the minimum of the dissimilarity values in the region. One benefit this has beyond the local gradient interest point methods is that it can remove points which are only similar from the perspective of the matching algorithm.

We were motivated to design a new paradigm for salient point detection which would both be intuitive, adaptable to diverse image matching fields, and be centered on optimizing a criterion function.

Our fundamental notion is that we want to minimize the probability of a mismatch when we select a match based on a distortion or dissimilarity measure. Therefore, we want to select salient points which will have a lower number of similar candidates in any local region.

We assume for simplicity that the distinctiveness of a point is inversely related to the similarity of that point to the closest wrong match.

Let $D(x, y)$ represent the distinctiveness of a pixel at (x, y). Then the function we are optimizing can be expressed elegantly as

$$D(P) = \underset{R}{\operatorname{argmin}}[-Similarity(R, P)] \tag{5.1}$$

where R represents a region based on a pixel location P.

and the constellation of salient points would be

$$C = maxima\ of\ D(P) \tag{5.2}$$

This means that we select the set of pixels which are local maxima of distinctiveness with regard to the similarity function used in the matching algorithm.

One of the advantages of this paradigm is that the similarity function can be adapted to the area of computer vision or pattern recognition. It can be adapted to be rotation invariant, scale invariant, color invariant, etc. As mentioned earlier, different problem areas have different constraints.

Another advantage of the MOD paradigm is that the similarity function can be also utilize sets of imagery such as all of the frames in a video shot because R can include the region near a pixel in the video shot specified as (x, y, t) where t represents the frame number. This would mean that we could find all of the salient pixels over a video shot, not merely a single frame.

5.3.2 The special case of template matching

The implementation of interest point method for the special case of template matching is based on selecting points in the image that maximize local distinctiveness. In this case, the distinctiveness value of a point is determined by the distance to the best matching neighbor in an area surrounding that point. We determine the distance between two points by calculating the SAD of grayscale values in a square template window.

Figure 5.1: Example images from the 'aviation' class.

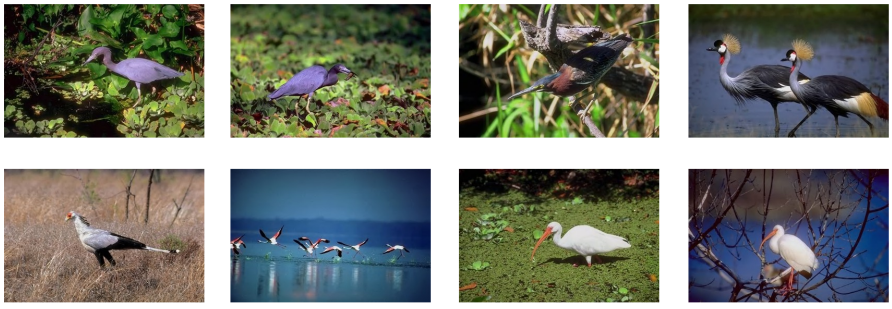

Figure 5.2: Example images from the 'wl_bird1' class.

5.3.3 Detector output

Figures 5.1 to 5.3 show example images from three classes of the Corel database.

Figure 5.3: Example images from the 'dogs' class.

Figures 5.4 to 5.12 show the output of the Harris corner detector, the SIFT interest point detector and the MOD interest point detector for the same input images. Visually, the Harris detector seems to capture the structure of the objects in the image better than the MOD detector, but we will show the effect of this for the retrieval results in section 5.5.

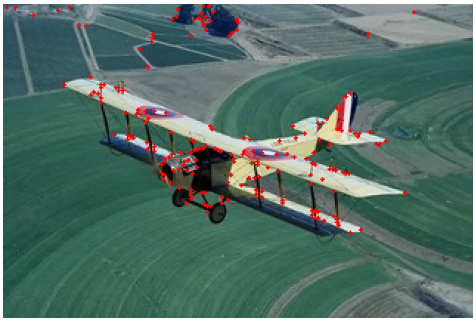

Figure 5.4: An image with the output of the Harris corner detector.

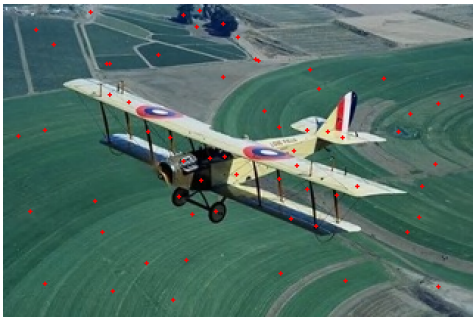

Figure 5.5: An image with the output of the SIFT interest point detector.

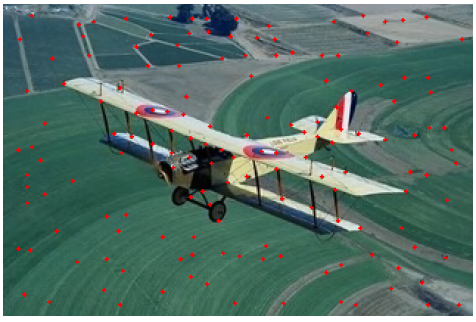

Figure 5.6: An image with the output of the MOD interest point detector.

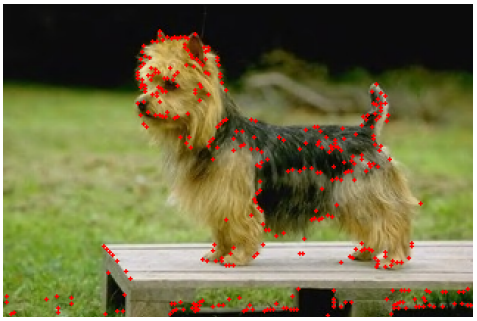

Figure 5.7: An image with the output of the Harris corner detector.

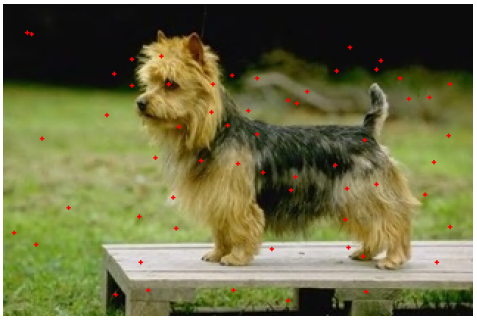

Figure 5.8: An image with the output of the SIFT interest point detector.

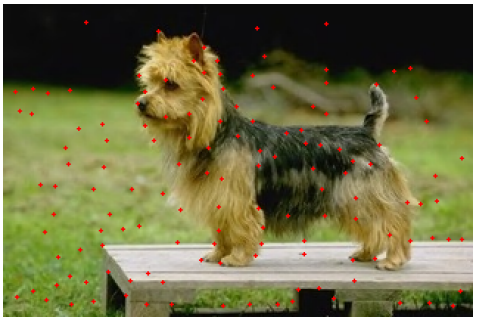

Figure 5.9: An image with the output of the MOD interest point detector.

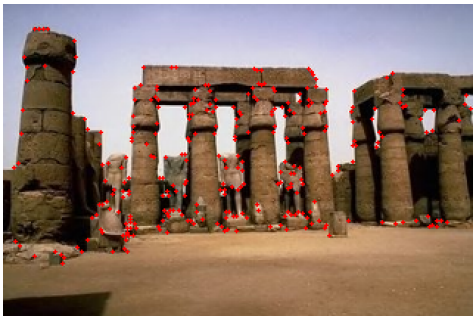

Figure 5.10: An image with the output of the Harris corner detector.

Figure 5.11: An image with the output of the SIFT interest point detector.

Figure 5.12: An image with the output of the MOD interest point detector.

5.4 Matching images

For each interest point, a feature vector is created that contains color and texture information. We have chosen color moments [87] as the color feature and local binary patterns [61] as the texture feature. Color moments are calculated for a 3x3 region around the interest point. The local binary patterns feature is calculated for a 19x19 region around the interest point.

We then use these feature vectors to compare images, by determining the best matches between interest points. The sum of all these best matches is the distance between the images.

In other words, for each point in an image I, the closest matching point in image J is searched for and this distance is used for calculating the overall distance of image I to image J. The distance between two images I and J is defined as the sum of the distance from I to J and the distance from J to I:

$$d(I,J) = \sum_x bestmatch(I_x, J) + \sum_y bestmatch(J_y, I) \tag{5.3}$$

where I_x is interest point x in image I.

5.5 Experiments and results

For testing the new salient points method, we used a subset of the Corel photo database. We selected 18 classes of images, each class containing 100 images. We have used 20 randomly selected images from each class as the query images. The results are averaged over these 20 queries.

Each image was resized to have a width of 320 pixels, to speed up computation and to make sure the calculated features are more scale-invariant. The settings for our interest point detector were set to a neighborhood size of 15 and a template size of 3, which was empirically determined by looking at the detector output.

Table 5.1 gives an overview of retrieval accuracy for the first 15 images returned, for the three interest point detectors.

Image class	Harris	SIFT	MOD
aviation	0.110	0.370	0.390
beaches	0.213	0.597	0.320
butterfly	0.140	0.157	0.307
cactus	0.127	0.303	0.257
castles	0.147	0.137	0.150
cats	0.243	0.267	0.493
dogs	0.207	0.157	0,333
horses	0.070	0.067	0.080
mammals	0.103	0.077	0.130
models	0.357	0.228	0.253
mountain	0.130	0.077	0.260
orchids	0.137	0.197	0.323
pyramids	0.170	0.373	0.627
roses	0.373	0.357	0.703
tulips	0.137	0.170	0.143
waterfall	0.210	0.170	0.383
wl_bird1	0.217	0.237	0.283
wl_fish	0.120	0.333	0.309

Table 5.1: Average precision of the three methods, using 20 query images and a result set of 15 images.

Among the best results for the MOD algorithm are 'pyramids' and 'roses', of which the results are shown in Figures 5.13 and 5.14.

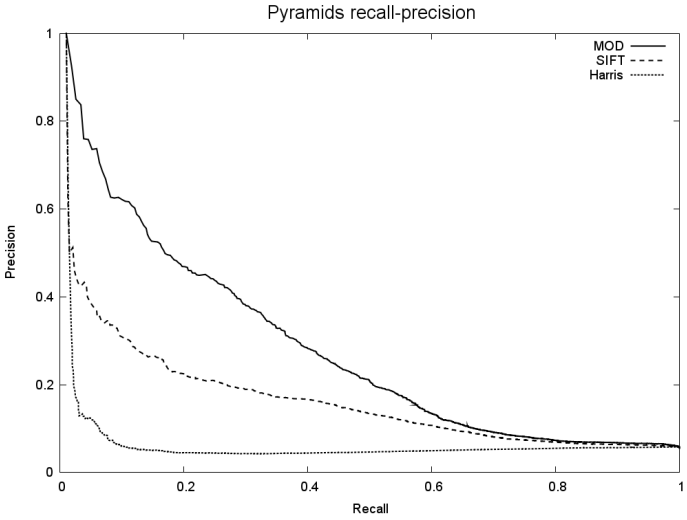

Figure 5.13: Recall-precision for the 'pyramids' class.

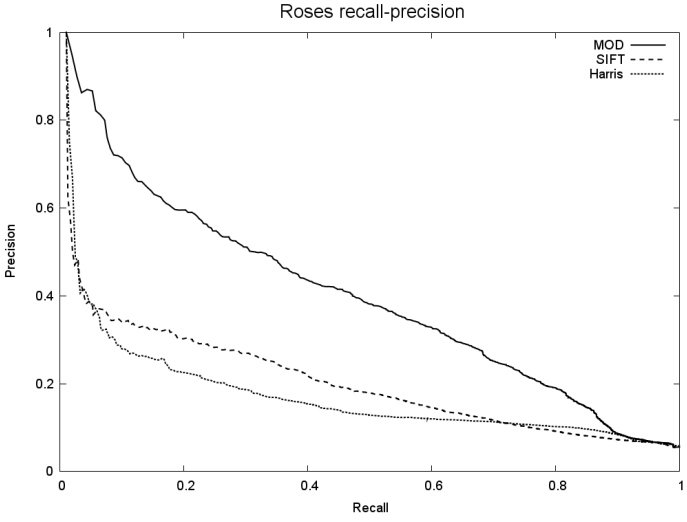

Figure 5.14: Recall-precision for the 'roses' class.

Figure 5.15 shows the overall retrieval results for our interest point method compared to the SIFT and Harris methods. It is clear that the MOD gives significantly better results in the context of image retrieval than either the SIFT method or the Harris points.

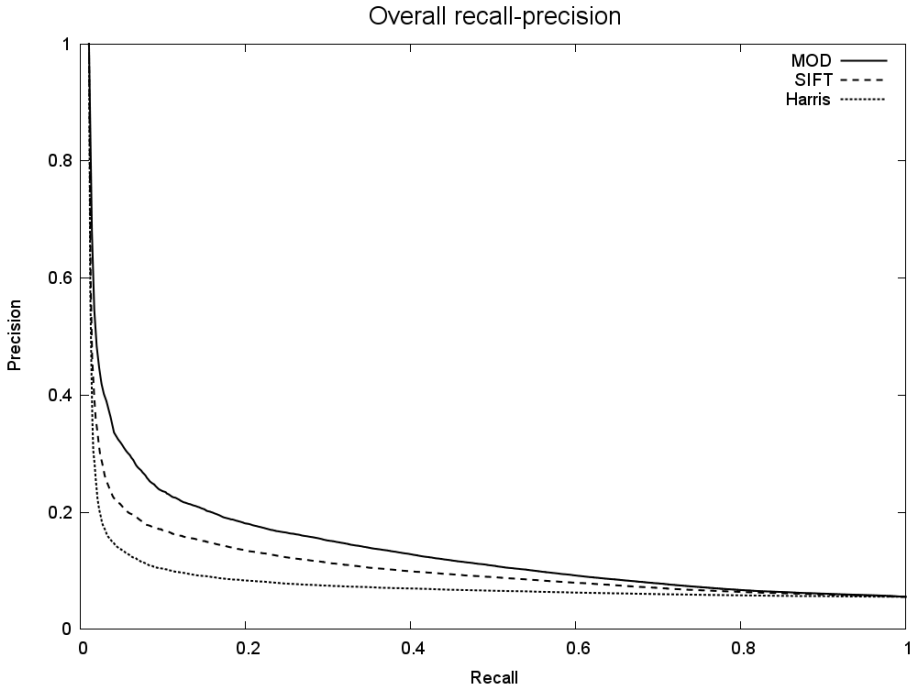

Figure 5.15: Overall retrieval recall-precision.

5.6 Discussion and conclusions

In this chapter we introduced a novel paradigm for interest point detection based on a criterion of maximization of distinctiveness. We have used two features for describing each interest point, the local binary pattern for texture and color moments for color. We compared the MOD method in a content-based image retrieval experiment to the SIFT interest point and the Harris corner detector and we have showed that it outperforms both detectors for this task.

Intuitively, our interest point detector places points at locations that are more uniform in color or texture. These points clearly are useful for content-based retrieval tasks, since these areas also contain information.