©Springer-Verlag. http://www.springer.de/comp/lncs/index.html

Salient Regions for Query by Image Content

Jonathon S. Hare and Paul H. Lewis

Intelligence, Agents, Multimedia Group,
School of Electronics and Computer Science,
University of Southampton,
Southampton, SO17 1BJ,
United Kingdom
{jsh02r, phl}@ecs.soton.ac.uk

Abstract. Much previous work on image retrieval has used global features such as colour and texture to describe the content of the image. However, these global features are insufficient to accurately describe the image content when different parts of the image have different characteristics. This paper discusses how this problem can be circumvented by using salient interest points and compares and contrasts an extension to previous work in which the concept of scale is incorporated into the selection of salient regions to select the areas of the image that are most interesting and generate local descriptors to describe the image characteristics in that region. The paper describes and contrasts two such salient region descriptors and compares them through their repeatability rate under a range of common image transforms. Finally, the paper goes on to investigate the performance of one of the salient region detectors in an image retrieval situation.

1 Introduction

Much previous work in the field of content based retrieval has been based around the concepts of using global descriptors to describe the content of the image. More recently researchers have begun to realise that global descriptors are not neccessarily good when it comes to describing the actual objects within the images and their associated semantics. Two approaches have grown from this realisation; firstly approaches have been developed whereby the image is segmented into multiple regions, and separate descriptors are built for each region; and secondly, the use of salient points has been suggested.

The first approach has been demonstrated to work [1], although it has a large problem - that of how to perform the segmentation. Over the years many techniques for performing image segmentation have been suggested, although none really solve the problem of linking the segmented region to the actual object that is being described. Indeed, this shows that the non-naive segmentation problem is not just a bottom-up image processing problem, but also a top-down problem that requires knowledge of the true object, before it can be successfully segmented.

©Springer-Verlag. http://www.springer.de/comp/lncs/index.html

The second approach avoids the problem of segmentation altogether by choosing to describe the image and its contents in an altogether different way. The use of saliency in computer vision has become quite widespread in recent years. Saliency is often used to provide the basis for a visual attention mechanism that reduces the need for computational resources [2]. Historically, saliency was described by the term 'interest point detectors', but use of the term 'saliency' has come about from the large amount of psychology-based work on selective visual attention. By using salient points within an image, it is possible to derive a compact image description based around the local attributes of the salient points. A number of different methods for finding salient points have been suggested, from the simple Harris' & Stephens [3] corner detector, to wavelet based approaches [4, 5, 6], to methods based around image entropy [7, 8]. Many previous approaches to using salient points have generated feature-vectors from pixel data in fixed-sized regions around the salient point, usually a 3x3 or 9x9 pixel neighbourhood centred on the point [5], although some of the modern state-of-the-art detectors find affine invariant regions and generate descriptors from within the region [9, 10, 11]. This paper compares and contrasts an extension to previous work in which the concept of scale is used in the selection of salient points (or rather salient regions), and the pixel content of the entire region content to build the feature vector of the local descriptor.

2 Salient Regions

2.1 Scale Saliency

The Scale-Saliency algorithm developed by Kadir and Brady [8, 7] was based on earlier work by Gilles [12]. Gilles investigated salient local image patches or 'icons' to match and register two images (specifically aerial reconnaissance images). Gilles suggested that by extracting locally salient features from the pair of images and matching these, it would be possible to estimate the global transform between the two images. Gilles defined saliency in terms of local signal complexity or unpredictability. More specifically, he suggested the use of Shannon Entropy of local attributes to estimate the saliency. Basically, image segments with flatter intensity histogram distributions[1] tend to have higher signal complexity and thus higher entropy. Gilles method only worked at a single scale, and picked single salient points, rather than salient regions.

Kadir and Brady modified Gilles original algorithm to make it perform well on images other than those from aerial reconnaissance imagery. Essentially they changed the algorithm so that it detected salient regions at multiple scales by looking for self-similarity across scales. The modified algorithm located circular patches of the original image that were considered salient. The size of the patch was determined automatically by the multi-scale additions to Gilles algorithm.

[1] Kadir and Brady [8] note that the method is not limited to the intensity histogram and that it is equally possible to use a histogram from a different descriptor, such as colour or edge strength.

©Springer-Verlag. http://www.springer.de/comp/lncs/index.html

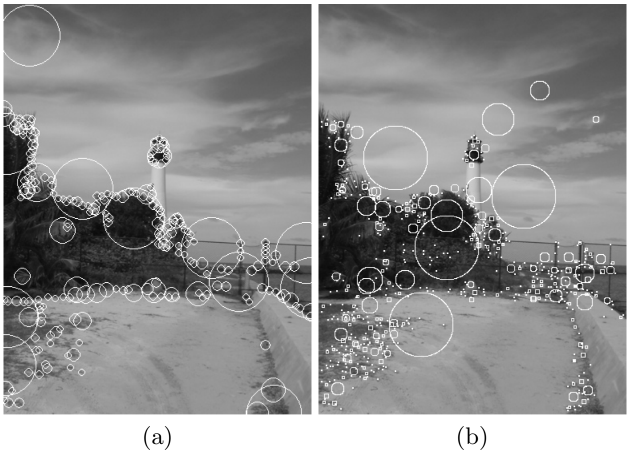


Fig. 1. (a) Salient regions found by the Scale-Saliency algorithm; (b) Salient regions found by from peaks in a difference-of-Gaussian pyramid

In addition Kadir and Brady developed a simple clustering algorithm to group together features within the $\mathbb{R}^3$ space that have similar x and y location, and scale. Figure 1(a) illustrates the results of applying the algorithm to an image.

2.2 Peaks in the difference-of-Gaussian pyramid

We take the idea of using peaks in a difference-of-Gaussian pyramid from the work of Lowe [13, 14] on object recognition using keypoints. Lowe has shown that by searching a difference-of-Gaussian pyramid for local peaks, both spatially and across scale, it is possible to select points robust to a range of projective transformations. The difference-of-Gaussian closely approximates the scale-normalised Laplacian-of-Gaussian [15, 13], $\sigma^2\nabla^2 G$. Mikolajczyk [16] showed that the minima and maxima of $\sigma^2\nabla^2 G$ produced the most stable interest points when compared to a range of other operators. Figure 1(b) illustrates the results of finding peaks in a difference-of-Gaussian pyramid.

2.3 Comparison of Salient Region Methods

Both of the methods for selecting salient regions described above are quite similar. For example, when the response of a difference-of-Gaussian filter is large, we would also expect the entropy taken over the same area as the filter to be large. Note that the converse is not always true though - high entropy does not necessarily mean that there would be a large difference of gaussian response. This is illustrated in Figure 2.

One problem with entropy is that it is very sensitive to noise. This is especially so at small scales, where there are relatively few pixels to sample and estimate the probability density function from, in order to estimate the entropy.

©Springer-Verlag. http://www.springer.de/comp/lncs/index.html

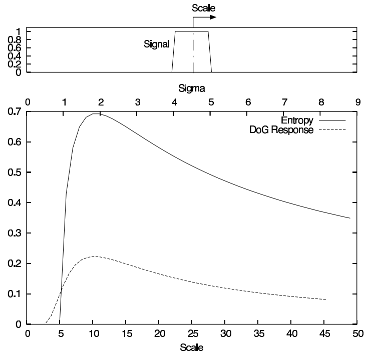


Fig. 2. Entropy and difference-of-Gaussian (ratio of $\sigma's = 1 : 1.6$, smaller σ is shown on the top x-axis) response versus scale to a one-dimensional signal as illustrated in the top diagram. The centre of the DoG and Entropy mask are kept at a constant position relative to the signal (shown by the dashed line). The graph illustrates how the response functions behave in a similar manner across scale-space

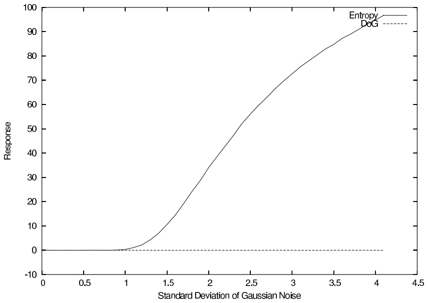


Fig. 3. Response of Entropy and difference-of-Gaussian functions to a constant signal with increasing amounts of zero-mean additive Gaussian noise. The DoG response stays stationary, whilst the Entropy response increases with noise

The difference-of-Gaussian is much less sensitive to noise due to the smoothing effect of the Gaussians. This is illustrated in Figure 3.

The remainder of this section is devoted to objectively comparing the stability of the two salient region detectors.

Repeatability. We take the measure of repeatability of interest points from Schmid *et al* [17]. The concept of repeatability is described below together with some results.

Repeatability Criterion. Repeatability is a measure of how independent an interest point detector is to the imaging conditions, i.e. camera parameters - position relative to the scene, zoom, etc. 3D points detected in one image should also be detected at aproximately the same locations in subsequent images. Given a point X in 3D space and two projection matrices, P_1 and P_2, the projections of X in two images I_1 and I_2 are given by $p_1 = P_1X$ and $p_2 = P_2X$ respectively. The point p_1, detected in image I_1, is repeated if the corresponding point p_2 is detected in image I_2. In order to estimate the repeatability, a unique relation between the points p_1 and p_2 has to be found. In the case of a planar scene, points in one image are related to points in a second image by a planar homography: $p_2 = Hp_1$.

The percentage of points that are repeated with respect to the total number of detected points is called the repeatability rate. In general, a point is not repeated

©Springer-Verlag. http://www.springer.de/comp/lncs/index.html

at exactly the same position as given by Hp_1, but in a small neighbourhood of that point. Denoting the size of the neighbourhood by ϵ , we can define the ϵ-*repeatability*. Interest points that cannot be observed in both images will corrupt the repeatability measure, thus only points in the common part of the scene are used to calculate the repeatability. The common part of the scene is defined by the homography, thus points $\tilde{p_1}$ and $\tilde{p_2}$ which lie in the common parts of images I_1 and I_2 are defined by $\{\tilde{p_1}\} = \{p_1|Hp_1 \epsilon I_2\}$ and $\{\tilde{p_2}\} = \{p_2|H^{-1}p_2 \epsilon I_1\}$. The set of point pairs $(\tilde{p_1}, \tilde{p_2})$ that correspond within an ϵ-*neighbourhood* is $D(\epsilon) = \{(\tilde{p_2}, \tilde{p_1})|dist(\tilde{p_2}, H\tilde{p_1}) < \epsilon\}$.

As the number of detected points in the two images may be different, the repeatability rate is defined as:

$$r(\epsilon) = \frac{|D(\epsilon)|}{min(|\{\tilde{p_1}\}|, |\{\tilde{p_2}\}|)}. \tag{1}$$

Repeatability Results. Using the repeatability criterion, we investigated the robustness of the two salient region descriptors to image rotation and scaling. The rotation and scaling were performed digitally, using bilinear interpolation. As a baseline, we also calculated the repeatability of the well-known Harris corner detector (using a [-2 -1 0 1 2] kernel), and an improved version of the Harris detector that calculates the derivatives more precisely by replacing the [-2 -1 0 1 2] kernel with one calculated from the derivatives of a Gaussian ($\sigma = 1.0$).

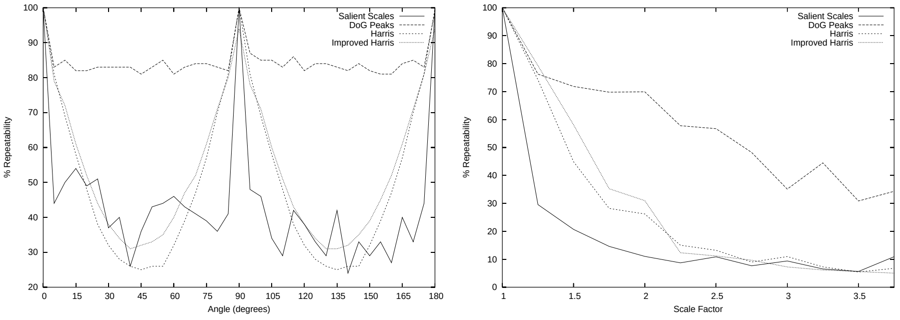
Figure 4(a) illustrates the results of repeatability against rotation angle, averaged over all of the images in the dataset, and Figure 4(b) illustrates the variation in repeatability over a range of image scales, again averaged over all the images in the dataset. The results show that the salient regions detected by finding peaks in the difference-of-Gaussian pyramid are by far the most stable to both rotation and scaling. The salient-scales algorithm performs more-or-less on a par with the Harris detector. Unfortunately, whilst the salient-scales algorithm should be robust to both scaling and rotation, in practice it is affected by discretisation of the digital raster, especially at small scales. Also, we have found that the clustering part of the salient scales algorithm does little to help its stability.

3 Query by Image Content using Salient Regions

In previous work by Sebe *et al* [5], the use of salient point detectors for content-based rerieval was shown to have better performance than when using global descriptors. In this section we describe a new metric for measuring the performance of content-based retrieval based on salient points, and illustrate it with some preliminary results that show that the performance when using salient regions is indeed better than when using global descriptors.

In order to facilitate the testing of the the use of salient regions for content-based retrieval, we have developed a system that returns the N closest matches to a given query image. The system enables queries to be made using either global descriptors or a descriptor based on salient regions. Following Sebe *et*

©Springer-Verlag. http://www.springer.de/comp/lncs/index.html



(a) (b)

Fig. 4. Repeatability rate for image rotation (a), and for scale change (b). $\epsilon = 1.5$ in both cases

al, we fix the number of salient regions to 50 per image. In the case of global descriptors, the distance between two images, I_1 and I_2, is given by the euclidean distance between the feature descriptors, $\mathbf{F_1}$ and $\mathbf{F_2}$:

$$\mathrm{D_E}(\mathbf{F_1}, \mathbf{F_2}) = |\mathbf{F_1} - \mathbf{F_2}| = \sqrt{\sum_{i=1}^{K} |\mathbf{F_{1_i}} - \mathbf{F_{2_i}}|^{\mathbf{2}}}, \tag{2}$$

where K is the number of elements in the feature descriptors. In the case of matching using salient regions, the distance between two images is given by a linear summation of the closest matching feature vector in the second image for each feature vector in the first image. Denoting the set of M feature vectors in images I_1 and I_2 as $\{\mathbf{F_1}\}$ and $\{\mathbf{F_2}\}$:

$$\mathrm{D_{salient}}(\{\mathbf{F_1}\}, \{\mathbf{F_2}\}) = \sum_{j}^{M} \min_k(\mathrm{D_E}(\{\mathbf{F_1}\}_j, \{\mathbf{F_2}\}_k)), \tag{3}$$

where $\{\mathbf{F_1}\}_j$ refers to the jth feature vector of image I_1 and $\{\mathbf{F_2}\}_k$ refers to the kth feature vector of image I_2.

3.1 Semantic Relevance

The problem with global descriptors is that they cannot fully describe all parts of an image having different characteristics. The use of salient regions aims to avoid this problem by developing descriptors that do capture the characteristics of each part of the image. Given this aim, it should not be unreasonable to expect that an image description generated from salient regions will be *better* than an image described wholy by a global descriptor. In order to test this we have developed a metric that uses semantically marked images as ground-truth against the results from our retrieval system.

©Springer-Verlag. http://www.springer.de/comp/lncs/index.html

Table 1. Averaged Semantic Relevance for queries based on the Rank 1 result image and the closest 5 result images

	Rank 1 Result Image		Averaged Top 5 Result Images	
Feature Type	DoG Peaks	Global	DoG Peaks	Global
RGB Histogram	42.1%	37.6%	51.0%	45.6%
HSI Histogram	45.2%	36.9%	50.4%	49.6%
Mono Histogram	31.6%	36.9%	42.3%	45.0%
HU Moment	41.1%	22.6%	52.4 %	39.5%
RGB Colour Moment	33.7%	24.1%	41.9%	35.4%
HSI Color Moment	34.9%	30.2%	43.5%	40.5%

The University of Washington Ground Truth Dataset [18] contains a large number of images that have been semantically marked up. For example an image may have a number of labels describing the image content, such as “trees”, “bushes”, “clear sky”, etc. Given a query image with a set of labels, we should expect that the images returned by the retrieval system should have the same labels as the query image. Let A be the set of all labels from the query image, and B be the set of labels from a returned image. We then define the semantic relevance, R, of the query to be:

$$R = \frac{|A \cap B|}{|A|} \tag{4}$$

This implies that if all the labels in set A exist in set B then the semantic relevance will be 100%, and if only half of the labels in set A exist in set B then the semantic relevance will be 50%.

3.2 Results

We used all of the semantically marked images from the Washington dataset to form our test set. Taking each image in the test set in turn as a query, we calculated the distance to each of the other images in the test set using a range of feature types. We then calculated the semantic relevance for the rank one image (the closest image, not counting the query image), and we also calculated the averaged semantic relevance over the closest 5 images. The results of this are shown in Table 1. The table shows that the use of salient regions does indeed produce better semantic relevance than using global descriptors, although we believe that there is still scope for improvement of the semantic relevance from the salient regions. We believe that using a single feature type to describe a salient region (or indeed the whole image) is not sufficient. For example, the RGB histogram that represents a ”blue sky” semantic label may be very similar to the histogram representing the ”water” label. In our future work we hope to show it is possible to improve the semantic relevance of queries using salient regions by fusing multiple feature descriptors. Figure 5 illustrates the differences between a query based on a global RGB-Histogram descriptor, versus multiple

©Springer-Verlag. http://www.springer.de/comp/lncs/index.html

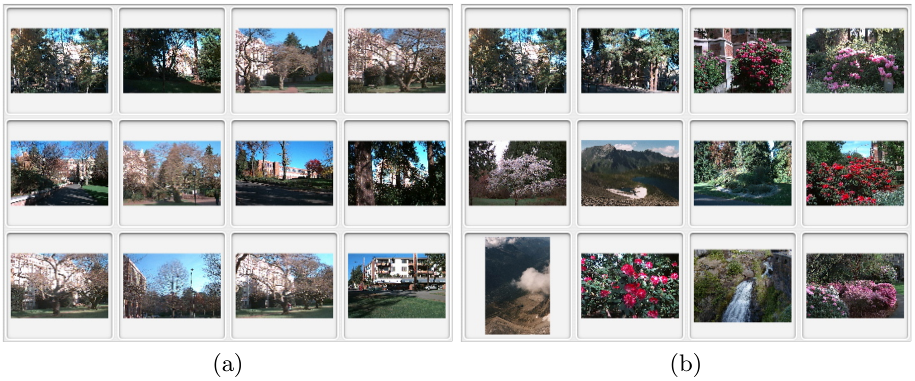


Fig. 5. Example Retrieval: (a) shows the results of a query using the Difference of Gaussian salient region method, and (b) shows the results of the same query with the Global method. In both cases, RGB Histograms are used as the feature descriptor and the first image shown is the query image

RGB-Histogram descriptors based around salient regions found from the peaks in the difference-of-Gaussian pyramid.

4 Conclusions and Future Work

In this paper, we have illustrated the concept of using peaks in a difference-of-Gaussian pyramid to select scale-invariant salient regions. We have shown that peaks in the difference-of-Gaussian pyramid are robust to a range of transformations, and that they perform better than an alternative approach to finding salient regions based on image entropy.

We have also demonstrated the concept of using salient regions for content-based retrieval. We have introduced a new metric, which we have termed *semantic relevance*, for the measurement of the relevance of a semantically marked result image from a semantically marked query image.

Our results have shown that the use of salient regions for content-based retrieval produces better semantic relevance than global descriptors. However, we note that it should be possible to improve these results even more by the use of better feature descriptors.

As previously mentioned, our future plans are to use the fusion of multiple features to try and improve the semantic relevance. We also plan to extend our system to use a better distance metric, such as the Mahalanobis distance, D_M.

5 Acknowledgements

We are grateful to the EPSRC and Motorola UK Research Laboratory for their support of this work.

©Springer-Verlag. http://www.springer.de/comp/lncs/index.html

References

[1] Carson, C., Thomas, M., Belongie, S., Hellerstein, J.M., Malik, J.: Blobworld: Image segmentation using expectation-maximization and its application to image querying. In: Third International Conference on Visual Information Systems, Springer (1999)

[2] Itti, L., Koch, C.: Computational modelling of visual attention. Nat. Rev. Neurosci. **2** (2001) 194–203

[3] Harris, C., Stephens, M.: A combined corner and edge detector. In Mathews, M.M., ed.: Proceedings of the 4th ALVEY vision conference, University of Manchester, England (1988) 147–151

[4] Shokoufandeh, A., Marsic, I., Dickinson, S.: View-based object recognition using saliency maps. Image Vis. Comput. **17** (1999) 445–460

[5] Sebe, N., Tian, Q., Loupias, E., Lew, M., Huang, T.: Evaluation of salient point techniques. Image and Vision Computing **21** (2003) 1087–1095

[6] Sebe, N., Lew, M.S.: Comparing salient point detectors. Pattern Recognition Letters **24** (2003) 89–96

[7] Kadir, T.: Scale, Saliency and Scene Description. PhD thesis, University of Oxford, Deptartment of Engineering Science, Robotics Research Group, University of Oxford, Oxford, UK (2001)

[8] Kadir, T., Brady, M.: Saliency, scale and image description. Int. J. Comput. Vis. **45** (2001) 83–105

[9] Tuytelaars, T., Gool, L.V.: Content-based image retrieval based on local affinely invariant regions. In: Third International Conference on Visual Information Systems. (1999) 493–500

[10] Sivic, J., Zisserman, A.: Video google: A text retrieval approach to object matching in videos. In: Internaional Conference on Computer Vision. (2003)

[11] Obdrzalek, S., Matas, J.: Image retrieval using local compact dct-based representation. In: DAGM-Symposium 2003. (2003) 490–497

[12] Gilles, S.: Robust Description and Matching of Images. PhD thesis, University of Oxford (1998)

[13] Lowe, D.: Distinctive image features from scale-invariant keypoints. To appear in International Journal of Computer Vision (2004)

[14] Lowe, D.: Object recognition from local scale-invariant features. In: Proc. of the International Conference on Computer Vision ICCV, Corfu (1999) 1150–1157

[15] Marr, D.: VISION: A computational Investigation into Human Represenation and Processing of Visual Information. W. H. Freeman and Company (1982)

[16] Mikolajczyk, K.: Detection of local features invariant to affine transformations. PhD thesis, Institut National Polytechnique de Grenoble, France (2002)

[17] Schmid, C., Mohr, R., Bauckhage, C.: Evaluation of interest detectors. Int. J. Comput. Vis. **37** (2000) 151–172

[18] University of Washington: Ground truth image database. http://www.cs.washington.edu/research/imagedatabase/groundtruth/ (2004)