FROM SALIENT FEATURES TO SCENE DESCRIPTION

Timor Kadir[1], Paola Hobson[2] and Michael Brady[1]

[1]Department of Engineering Science,
University of Oxford,
Oxford, UK.

[2]Motorola Labs,
Jays Close,
Basingstoke, UK.

ABSTRACT

In this paper, we discuss an image modelling method that can capture and represent a number of commonly used image characteristics such as lines, blobs, statistical and structural textures. We develop the concepts underlying the Kadir and Brady feature detection algorithm [2], namely feature space unpredictability and spatial localisation, to characterise, what is termed, the *semi-local* predictability of features. Many previous approaches to image description choose a particular representation a-priori. In contrast, the proposed approach aims to detect the presence of particular types of region. We suggest that the result is a richer description of the image from which semantic scene content can be extracted. Preliminary results are presented.

1. INTRODUCTION

Images can contain a wide variety of feature types, such as lines, blobs and textures, from which vision algorithms commonly analyse a small subset for subsequent analysis. This process of *feature selection* is typically performed for reasons of computational tractability and to facilitate inference about the scene content of interest. The latter is only possible under a specific set of imaging conditions and where the feature is related to the scene content of interest. The choice of feature set tends to be, in practice, application dependent. The system designer chooses ahead of time, usually implicitly or more rarely explicitly, a set of features, knowing the application. This approach contributes to the brittleness of vision systems.

An alternative is to learn from a training set the features necessary to discriminate between a set of objects or object classes e.g. [6]; however even with such an approach some decision about the features and their representation must be made a-priori.

In this paper we suggest an alternative approach. We propose that image content can be organized into a taxonomy of generic types – lines, blobs, statistical and structural texture. We also argue that the information that encodes a region's type is embedded within the image and that this may be determined by analyzing its statistics, in particular its spatial predictability. This idea presupposes that knowledge of a region's type is useful for subsequent processing; we support this notion with some demonstrations. The benefit of this approach is that images can be characterised for use in a range of applications where a large degree of prior knowledge normally required for a recognition task may not be available e.g. in consumer image management on mobile devices, where the user's collection of pictures comes from many possible domains, and it is not feasible to pre-load all expected object models.

2. MOTIVATION

As an illustrative example, consider the images in Figure 1 which contains many types of feature which might be used for inference about the scene. In the left image, the frames of the windows of the buildings may be labelled as lines, the boundaries between the buildings and the sky as edges. The sky and car park surface might be labelled as piecewise smooth intensity regions and the brick walls as textured regions --- the distant one as a statistical texture and the close one on the left edge as a structural texture. Clearly, with sufficient zoom the same scene components would map onto quite different feature types. The image on the right shows the same scene but at a larger focal length. In this new image the individual bricks of the yellow wall are observable as separate entities; hence a structural texture model might be more appropriate in this case. New individual features can also be observed inside the windows whereas at the original focal length they had mostly appeared as regions of constant intensity.

The main point here is that conventional approaches typically assume a certain dominant image model, say texture, and proceed to model and discriminate regions based on this model. The technique described in this paper aims to extract possible region types directly from the image from which appropriate modelling decisions can be made as part of the learning and inference process.

Figure 1 : Two images of the same scene at different focal lengths. Each contains a variety of feature types - lines, blobs, structural and statistical textures. The distant wall appears as a statistical texture in the left image, but may be considered a structural texture in the right image where the individual bricks are visible.

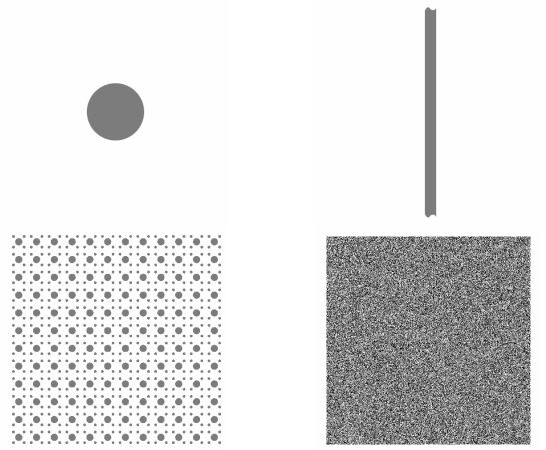

Figure 2 : Idealised examples of four image region classes: Blob, Line, Structural Texture and Statistical Texture.

3. IMAGE MODEL TAXONOMY

The key idea of the proposed approach is that image regions can be categorized into different types by analyzing their local statistics. We illustrate the concept by considering five different region classes: piecewise constant regions, blobs, lines, structural textures and statistical textures. See Figure 2 for examples. Evidently, other ontologies are possible: the one used in this article is based on our introspection.

Piecewise constant. An ideal constant intensity region consists of an infinite plane with one intensity value. It is non-localised spatially and in scale. The intensity distribution in such a perfect constant intensity region is perfectly predictable at every scale and position.

Blob. An ideal “blob” is a two dimensionally spatially constrained feature on an infinite plane. It is intended to be dual to a line in the sense that it has position but does not have a well-defined orientation. In an isotropic scale space it is perfectly constrained if it is circular, or whatever the shape of the scale-space kernel is.

Line. An ideal line, that is one with a finite width and infinite length, is a one dimensionally spatially constrained (in the perpendicular direction) feature. In both the isotropic and anisotropic scale spaces, it is partially constrained. A perfect line is unpredictable in the perpendicular direction and continuously predictable in the tangential direction. That is, it has a well-defined direction but may not have a well-defined position.

Structural Texture. An ideal structural texture consists of an infinite plane with tessellated blobs or line features. Considering the former case first, each feature is perfectly two dimensionally spatially constrained. The line feature case is similar except that the clusters are of medium saliency. Such a region is discontinuously predictable at specific positions and scales.

Statistical Texture. An ideal statistical texture consists of an infinite plane with no spatial or scale localization; it is unpredictable at all scales and locations.

The region types represent idealized patches. However, real images are scale constrained, or in the signal processing lexicon ``band limited", versions of the scene. The constraints are determined by imaging factors such as camera focal length, field of view and resolution. As such, in general real images are unlikely to map onto any one of these classes exactly. Rather, the correspondence is likely to be to a point somewhere on the continuum or space of image classes. The index or free parameter of this space is spatial predictability. For

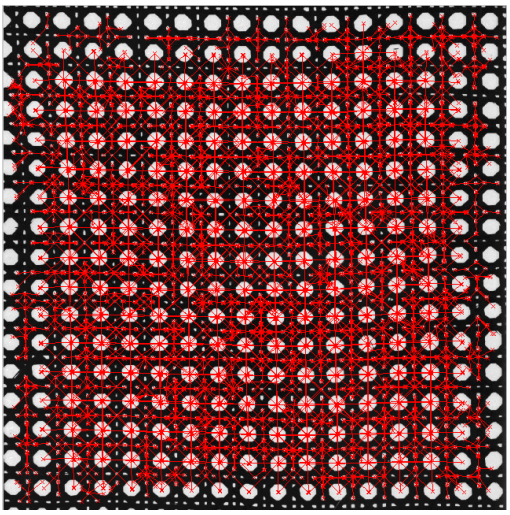

Figure 3 : Determining the texton structure of D102 Wicker Brodatz texture. The red lines indicate the salient feature matches.

example, an ellipse may be *interpreted* either as a blob or as a line depending on its aspect ratio.

Intuitively then, such regions can be distinguished by examining their spatial predictability at two principle scales: local and semi-local. The former determines the degree to which pixels are locally predictable. The latter examines the regions predictability. The question is how do we determine these quantitatively?

4. MEASURING PREDICTABILITY

Local predictability. For the local analysis we choose to use the feature detection algorithm proposed by Kadir and Brady [2,3]. This method is built upon two measures of local predictability. The first, local entropy, captures the degree to which a region's pixels are predictable in feature space. High entropy regions correspond to unpredictable pixel values and are deemed salient. The second, termed inter-scale saliency measures the predictability of a regions pixels over scale.

This method has a number of desirable properties for our application. It is invariant to translation, rotation and scale and it also performs scale selection. However, its principle benefit is that it uses a very loose definition of saliency compared to other feature detection methods such as a Laplacian blob detector. This allows it to detect a wide range of locally constrained features.

Semi-local predictability. Semi-local predictability is measured by searching for repetition of the detected features found using the Kadir and Brady detector. Similar approaches have been used for detecting repeated elements in [4,5]. In the preliminary results presented,

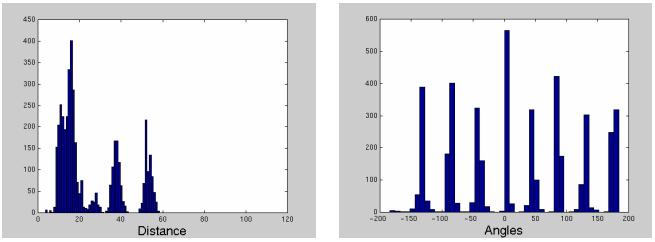

Figure 4 : The histograms of distance (left) and angles (right) for the D102 salient feature structure.

local correlation of the image patch corresponding to the feature has been used. We perform the following steps:

1. Apply Kadir and Brady algorithm to the image.
2. For all detected features do:
   a. Extract local image patch.
   b. Calculate normalised correlation in a semi-local window of size s x k, where k is some constant (3 in our implementation) and s is the feature scale found in step 1.
   c. Find peaks in correlation surface

The peaks in the correlation are the matches of the local salient icon, and capture the semi-local spatial predictability of the salient feature. The last step is not strictly necessary for linear structures since these should have a continuous correlation response. However, it can be useful for real images where the correlation will not be perfect. Applied to textures, our method resembles in part the pattern regularity technique of [1] in which global autocorrelation is used to determine the regularity of different texture. However, in our approach texture regularity is only one of the qualities of interest.

6. PRELIMINARY RESULTS

In the first experiment we apply our algorithm to the D102 wicker Brodatz texture with the aim of identifying the symmetric structure of the salient features. The results are shown in Figure 3 where the structure has been highlighted in red. Both the micro and macro texton structures can be clearly seen. In Figure 4 the statistics for the structure are shown in the form of histograms. The left graph shows the distribution of distances between feature matches, thus representing the magnitude of the vector translations. The first peak corresponds to the translation of the micro textons whilst the latter two to the macro textons. The graph on the right shows the distribution of angles and clearly matches with the diagonal and orthogonal regularity of the texture.

In the second experiment, we demonstrate how the spatial distribution information may be used to distinguish between linear structures and textures. For the purposes of this experiment, linear structures are defined to have in-line matches, that is, the angles between all the

Figure 3 : Misbahah (prayer beads) on coarse gravel. Distinguishing between linear structure (yellow) and irregular texture (green).

local symmetries (or matches) are in-line. Textures, on the other hand, exhibit a broad range of angles. We test this by measuring the variance of the one sided angles.

Figure 5 shows the results of this experiment. The texture, in this case coarse gravel-tarmac, is indicated in green, while the (curvi)linear structure, the wooden Misbahah, are marked highlighted in yellow. Overall, both regions have been correctly identified. These are encouraging results given the few assumptions that have been made about the image content. There are a few notable exceptions where the method has failed. One is where the curvature of the bead structure is high. This is because the in-line angle model is only a piecewise linear approximation of curvilinear structure. A slightly more sophisticated model could solve this problem.

7. APPLICATIONS

The approaches described in this paper are particularly interesting in emerging multimedia applications, where consumers are actively embracing digital image and video technology for leisure applications. Users are increasingly frustrated by the need to manage their growing content collections, which may be augmented by images and video acquired from their mobile phones.

Unlike many image analysis approaches that search for specific features or objects with a prior knowledge model, the method described in this paper allows more general analysis of the image content to be achieved, which enables it to be more flexible in coping with the wide range of content that consumers normally deal with. This is especially important for mobile imaging applications, where users may create content relating to just about any domain e.g. family, sports, holidays, events etc, and so need tools to later analyse their large content collections to form albums or to search for specific content to show to friends, family etc. In such cases, users may have an expectation about an application's ability to assist in automatic album classification, or assigning a new image to an existing collection. It is unfeasible to preload onto a mobile device object models covering all possible objects that the mobile unit is likely to encounter.

The approach described in this paper enables meaningful image classification to take place, such that images can associate with a signature, based on the spatial relationship between the identified regions. Such signatures can then be used for downstream processing such as clustering of images for album composition, and retrieval, without any high level decisions being made by the system about specific image content. The method may also be used as a preliminary step in finding objects for later classification using classical object recognition techniques, where training and prior knowledge base construction can feasibly take place.

8. CONCLUSION

A novel approach for image modelling different kinds of image content has been presented. The key idea is that a region's type may be determined from the image directly and can be used to further enhance latter processing steps. Preliminary results demonstrating the principles of the approach were presented. For future work, we propose that optimal models be selected using MDL.

9. REFERENCES

[1] D Chetverikov. *Pattern regularity as a visual key.* Image and Vision Computing, 18(12) 2000. Pages 975-985.

[2] T. Kadir and J.M. Brady, *Scale, Saliency and Scene Description*, International Journal of Computer Vision, 45(2), 2001. Pages 83-105,.

[3] T. Kadir, A. Zisserman and J.M. Brady, *An Affine Invariant Salient Region Detector*. European Conference on Computer Vision 2004. Pages 228 – 241.

[4] T. Leung and J. Malik. *Detecting, localizing and grouping repeated scene elements from an image.* In Proc. European Conference Computer Vision 1996. Pages 546-555.

[5] F. Schaffalitzky and A. Zisserman. *Geometric grouping of repeated elements within images.* In Proc. British Machine Vision Conference 1998, Pages 13-22.

[6] P. Viola and M. Jones. *Rapid object detection using a boosted cascade of simple features*. In Proc. Computer Vision and Pattern Recognition. 2001. Pages 511–518.

ACKNOWLEDGEMENTS

The work contained in this paper was carried out under the Motorola University Partners in Research programme.