Segmentation of Natural Images Using Hierarchical and Syntactic Methods

Paul Stefan Williams and Michael Alder

Centre for Intelligent Information Processing Systems
Department of Electrical & Electronic Engineering
The University of Western Australia
Nedlands W.A. 6907, AUSTRALIA
phone: +61-8-9380-3856 fax: +61-8-9380-1104
williams@ciips.ee.uwa.edu.au mike@maths.uwa.edu.au

Abstract. This paper examines the problem of image segmentation using hierarchical and syntactic methods. A bottom up approach takes low level feature vectors and combines them to form higher level objects corresponding to disjoint regions of the image with homogeneous characteristics. This transformation is known as *The UpWrite*. A *DownWrite* process is also introduced which reconstructs an image using only the higher level representation. This not only provides an insight into the effectiveness of the representation, but also outlines its weaknesses.
The results of the UpWrite, or segmentation, and the DownWrite are illustrated from a database of 1000 Corel Photo-CD images. Finally a simple classification scheme is presented to distinguish between predefined image classes such as Fields, Brown Bears and Elephants. The classification results indicate the effectiveness of the approach for use in content based image retrieval (CBIR).
keywords: Image Segmentation, Content Based Image Retrieval (CBIR), Syntactic Pattern Recognition, Image Recognition

1 Introduction

As the quantity and distribution of digital images grow the need for automatic image retrieval increases. Currently, systems such as IBM's Query By Image Content (QBIC) allow automatic retrieval based on simple characteristics and distributions of colour and texture [1][2]. Although useful, these simple primitives do not consider structural or spatial relationships and in general fail to capture the meaningful content of the image. The addition of spatial location, size and spatial extents of colour sets leads to more desirable results as illustrated by a sample sunset query demonstrated using VisualSEEk [3]. This does however require the user to perform searches through a subjective choice of colours and explicit spatial locations.

A more general and reusable technique which addresses a family of more specific problems is the use of body plans. These have been designed to perform segmentation and recognition in complex environments, as illustrated by

locating horses and naked people in natural scenes [4][5]. Although heavily based on structural and statistical information this approach is not fully automated, making use of *a priori* information in the form of specific models. In [4], one specific classifier is used for horses and another for naked people.

Ideally a hybrid combination of these methods is required which combines structural and spatial information with low level intensity, colour and textural characteristics to form a fully automated CBIR system. This system should make few assumptions about the image type and be robust with some level of invariance to contrast, brightness, noise, rotation, scale and translation.

2 Hierarchical Syntactic Approach

For the purpose of segmentation and CBIR, an image is considered to posses a hierarchical structure containing sub images or patterns with some underlying syntactic relationship. For example, an image of a large sailing boat may be decomposed into objects corresponding to sky, the boat and the ocean. The boat can then be further decomposed into its hull, masts and sails. Such hierarchical structure is usually depicted using a tree and can be expressed by some underlying grammar as discussed by Fu [6] a pioneer in syntactic pattern recognition. One of this most attractive point in this approach is that the relationships between objects can be expressed in much the same way as in human language and communication; *the image consists of a boat on the ocean with sky in the background. The boat has two white sails, two masks and a brown hull.*

2.1 The UpWrite

The *UpWrite* is a concept or general approach for automatically extracting hierarchical structure from an image or collection of primitives. Performed bottom up, the UpWrite can basically be summarised by the following three steps.

- *Localisation.* Initial local feature extraction involving a choice of scale.
- *Amalgamation.* A method of combining feature vectors or low level objects into a higher level object based on relevant criteria.
- *Representation.* Higher level objects are presented as a point in a Euclidean space $\mathbf{R}^N$. The representation should effectively characterise the object for classification or recognition and may include a discrete label.

This process can be repeated transforming a number of objects from a low to higher level within the hierarchy. For example, in binary images pixels can be combine into line segments, in turn into strokes and higher level objects such as circles, ellipses, rectangles [7][8] through multiple stages of localisation, amalgamation and representation. In principle the UpWrite will transform a collection of low level primitives into a single high level representation, or a point in $\mathbf{R}^N$, which can be used for training and classification.

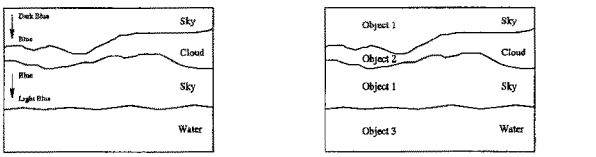
2.2 The DownWrite

Along with the UpWrite, a DownWrite process involves taking a higher level object and reconstructing a set of lower level primitives. Since the UpWrite process only extracts some features from the 'image' the DownWrite will not be a perfect reconstruction. For example, if two people were told an image consisted of a ship on the ocean their interpretation of the actual image would probably vary considerably. If an UpWrite algorithm recognised a ship on the ocean the DownWrite would hopefully produce an image which a human would agree to be a ship on the ocean. Consequently the DownWrite provides an insight into the effectiveness of the UpWrite and higher level object representation. Unlike the UpWrite the DownWrite is not unique and in general involves making decisions.

2.3 General Approach

To perform image segmentation low level feature vectors describing intensity, colour and textural characteristics are extracted from an image and amalgamated to form spatially disjoint homogeneous regions. These are represented using low order geometric moments, thus completing the first level of UpWrite. A DownWrite is illustrated using the models of these regions to gain an insight into both the effectiveness and limitations of the higher level representation. Furthermore, regions are combined based on mutual prediction of extracted image characteristics illustrating a potential second UpWrite. This enables spatially disjoint and occluded objects to be identified as depicted by Figure 1.

In summary the aim is to automatically segment the image hierarchically based on structure inferred at various levels of analysis. The remainder of this paper outlines the method of feature extraction, the first UpWrite or initial segmentation algorithm, one possible corresponding DownWrite and higher level processing to combine disjoint regions. Since at this stage the entire image is not transformed into a single point, a feature vector will be constructed from the object models and used for classification.



(a) Image content. (b) Ideal object segmentation.

Fig. 1. This example illustrates a common pattern recognition problem known as occlusion.

3 Feature Extraction

The method of feature extraction involves obtaining correlations between neighbouring low level image characteristics as described in [9] [10] [11]. In short this involves taking a square mask consisting of a set of $\mathbf{k}^2$ adjacent $\mathbf{n}$ by $\mathbf{n}$ pixel windows and making an appropriate measurement in each window. The resolution and complexity are governed by the chosen *window size* ($\mathbf{n}$) and *block size* ($\mathbf{k}$). Feature vectors are formed by moving the entire mask across the image one pixel at a time and recording each window measurement along with the mask position $(\mathbf{x}, \mathbf{y})$.

Using single window measurements of average pixel intensity and standard deviation is extremely effective in detecting anomalies in painted steel [9], and segmenting newspaper images into regions of text, picture, background and edges [10], respectively. For segmentation of natural images both of these and a further two average colour window measurements are used. In summary, the feature extraction process makes use of four window measurements as outlined by the following map.

$$\mathbf{I}[\mathbf{r}, \mathbf{g}, \mathbf{b}, \mathbf{x}, \mathbf{y}] \longmapsto \mathbf{P}[\mathbf{i}_1, \sigma_1, \mathbf{cx}_1, \mathbf{cy}_1, \mathbf{i}_2, ..., \mathbf{i}_{\mathbf{k}^2}, \sigma_{\mathbf{k}^2}, \mathbf{cx}_{\mathbf{k}^2}, \mathbf{cy}_{\mathbf{k}^2}, \mathbf{x}, \mathbf{y}]$$

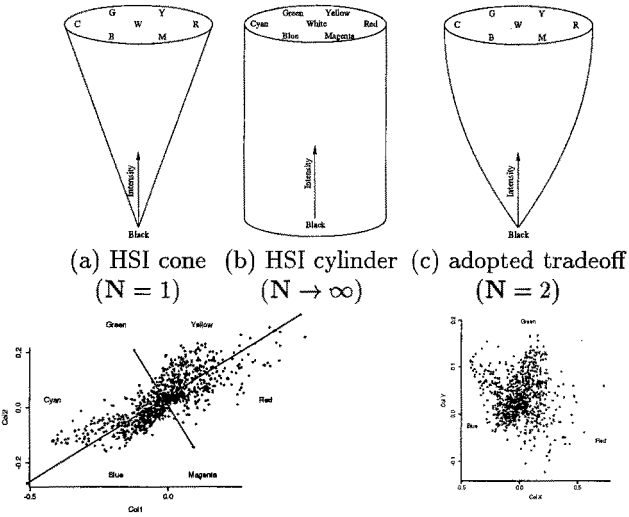
3.1 Colour Measurements

Since RGB colour representation is not well suited to image analysis, other colour spaces such as CIE-XYZ and HSI are typically used. While the CIE-XYZ colour space does provide an equation for *perceptual difference* between colours it requires a white reference and is highly nonlinear ruling out averaging and statistical techniques [12]. The more commonly used HSI cylindrical space [3] suffers from noise sensitivity causing very dark (black looking) colours to be assigned a fully saturated colour. The alternative HSI conical space [13][14] prevents this by adding a linear dependency between intensity and maximum saturation, hence unfairly distinguishing between the same perceived colour from mid to high range intensities.

In order to obtain a tradeoff between the HSI cylindrical and conical spaces a tradeoff is used whereby the saturation is limited by $1-(1-\mathit{intensity})^{\mathbf{N}}$ as seen in Figure 2(a)-(c). Furthermore the hue saturation plane is rotated and scaled to account for the distribution and correlation between colours in natural images apparent in Figure 2(d). The final two colour measurements used in feature vectors correspond to the Cartesian coordinates of this modified HSI space.

4 Segmentation Algorithm: The UpWrite

As introduced in Section 2 the UpWrite involves localisation, amalgamation and representation. The localisation or feature extraction phase uses the methods outlined in Section 3 to form a set of feature vectors $\mathbf{P}$ in $\mathbf{R}^{4\mathbf{k}^2+2}$. The implementation uses a block size $\mathbf{k} = 3$, window size $\mathbf{n} = 6$ pixels and a rotation of the



(a) HSI cone ($N = 1$) (b) HSI cylinder ($N \to \infty$) (c) adopted tradeoff ($N = 2$)

(d) Hue-saturation plane. (e) Adopted colour measurements.

Fig. 2. HSI colour spaces and distributions in the hue-saturation plane for natural images.

0.545 radians in the hue-saturation plane (see Figure 2). Furthermore window measurements are scaled by 1, 4, 415 and 1225 to allow a Euclidean metric to be used.

As outlined in [10] in regions where there is little local change corresponding feature vectors $\mathbf{p} \in \mathbf{P}$ will be close to the vector $\mathbf{1} = [1, 1, \ldots, 1]$, regardless of dimension[1]. Hence, a suitable initial estimate of local change is given by the orthogonal distance from this vector to $\mathbf{p}$, i.e. *local change* $= ||\mathbf{p} - (||proj_{\mathbf{1}}\mathbf{p}||)\mathbf{1}||$. Using this measure of local change the segmentation algorithm is performed as follows[2].

- Threshold the local change image and identify all disjoint regions using a *blobify* algorithm. The initial threshold will allow a very low tolerance to local change allowing more homogeneous regions to be identified first. For implementation an initial threshold of 5 is used which is increased in steps of 2 until it exceeds 80.
- Model sufficiently large regions for this threshold. Regions exceeding 400 pixels are modelled.
- Label spatial positions for the modelled regions hence excluding these from further modelling.
- Repeat these steps with an increased tolerance to local change.

[1] Ignoring the $\mathbf{x}, \mathbf{y}$ elements of the feature vector.

[2] Refer to [11] for more comprehensive details of algorithm stages.

Regions are modelled using these steps.

- Calculate an initial region model by taking low order geometric moments. The region is modelled by a mean $\mu \in \mathbf{R}^6$ and symmetric 6 by 6 covariance matrix $\mathbf{C}$ using the 4 window measurements and location (x, y) for all window placements within the region.
- Perform a flood fill algorithm using the Mahalanobis distance with respect to the regions model.
  - A dynamic Mahalanobis threshold of $\max\{5.5-0.1\sigma_{\mathbf{max}}, 0\}$ is used where $\sigma_{\mathbf{max}}$ is the largest window measurement standard deviation as obtained from the covariance matrix, i.e. $\sigma_{\mathbf{max}} = \max\{\sqrt{\mathbf{C_{ii}}} : \mathbf{i} = \mathbf{1}\ldots\mathbf{4}\}$. This inhibition factor prevents regions from leaking into one another along small fingers of gradual local change.
  - After the addition of new points the regions model is refined.
- Model the final region by the the number of data points, mean μ and covariance matrix $\mathbf{C}$.

After all regions have be identified and modelled they are stored to disk. Currently no shape model is used, but rather is binary quad-tree for each region's shape is stored.

5 The DownWrite

Each region is modelled by the multi-variable Gaussian $\mathbf{G(p)}$ as defined below. The expected value of this model can be obtained by minimising the equation to the right, thus obtaining the mean. The expected value at a restricted $(\mathbf{x}, \mathbf{y})$ location can also be obtained in this way hence accounting for the modelled lighting, colour and textural variation across the region.

$$\mathbf{G(p)} = \frac{1}{\sqrt{|C|}(2\pi)^{N/2}} e^{\frac{-1}{2}(\mathbf{p}-\mu)^T\mathbf{C}^{-1}(\mathbf{p}-\mu)} \quad \mathbf{E(p)} : \min((\mathbf{p}-\mu)^T\mathbf{C}^{-1}(\mathbf{p}-\mu))$$

The DownWrite is achieved by calculating the expected value at the position specified by each pixel within the region and adding Gaussian random noise at a standard deviation determined by the model.

6 Higher Level Processing

The second UpWrite makes use of the spatially constrained estimate used by the DownWrite. The model for a region is deemed to predict another region if it can successfully estimate the characteristics or mean at the centre of the other region. If two regions successfully predict one another the regions are labelled the same.

7 Results and Classification

Results at each stage of processing using a database of 1000 Corel Photo-CD images are illustrated online [15]. Figure 3 shows the types of image classes in the database also used in [13][14].

The UpWrite, DownWrite and higher level processing are illustrated in Figure 4 on images of the categories *fields*, *brown bears* and *elephants*.[3] The regions in white are not considered to belong to sizeable homogeneous regions. In terms of large objects, the sky and field are identified in the first image, the bear and sky in the second and the elephant, ground and sky in the third. One point to note is the variation of the uppermost sky region in the elephant example which prevents the two sky regions from being combined. Similar results can be observed for the *tree* regions outlining the inadequacy of the simple region prediction method.

Fig. 3. Corel images classed as bald eagles, brown bears, fields, Canadian Rockies and sunsets.[3]

[3] All figures are available online at http://ciips.ee.uwa.edu.au/Papers/Conference_Papers/1998/03/

Fig. 4. Results of performing the UpWrite, DownWrite and region prediction algorithm.

Overall the segmentation into homogeneous regions appears to be effective. It is also encouraging that the DownWrite is visually similar to the original image highlighting the strength of the object representation.

On a single R10000 the UpWrite, DownWrite and prediction take 2 minutes, 8 seconds and less than 1 second respectively.

	Classification										
Image class	AS	PB	BB	E	T	C	BE	CR	F	D	S
Airshows	70	8	1	1	1	0	6	3	6	3	1
Polar Bears	29	43	0	3	3	0	11	3	6	0	3
Brown Bears	2	0	48	9	18	5	12	0	3	2	2
Elephants	4	2	5	64	9	5	2	1	1	6	1
Tigers	0	0	11	12	48	5	3	4	9	7	1
Cheetahs	3	1	5	7	4	58	2	1	3	13	3
Bald Eagles	5	1	5	2	5	0	70	1	1	9	1
Canadian Rockies	6	1	0	1	6	0	4	56	10	11	5
Fields	6	3	2	1	13	6	2	7	48	10	2
Deserts	0	2	8	8	12	12	7	12	10	23	6
Sunsets	2	0	2	4	4	5	3	4	2	6	68

Table 1. Classification results for 1000 Corel Photo CD images.

Table 1 shows the classification of the Corel images using the closest sizeable region to the centre of each image and a k-nearest neighbours classifier. This crude classification scheme produces results of 50.5% and 55.1% for $k = 3$ and $k = 1$ respectively.

Classification rates of between 50% and 55% are also achieved in [13][14]. Results in this range are reasonably good considering a rate of 94.4% with human judgment despite obtaining 100% for all animal and *airshow* images.

8 Conclusion

The methods presented effectively use low level intensity, colour and textural characteristics along with inferred structural information to hierarchically segment images into homogeneous regions. The concept of the UpWrite and DownWrite are introduced highlighting the syntactic approach.

Segmentation results are illustrated from a database which is loosely classified by topic. A simple classification scheme demonstrates the strength of the technique on a problem which is not clear cut. Improvements and further work is required in higher level analysis, image representation, comparison and classification. The method provides a general framework for initial stages of content based image retrieval.

References

1. J. Ashley, R. Barber, M. Flickner, J. Hafner, D. Lee, W. Niblack, and D. Petkovic. Automatic and semi-automatic methods for image annotation and retrieval in QBIC. Technical Report RJ 9951-87910, IBM Almaden Research Center, April 1995.
2. P. Aigrain, H. Zhang, and D. Petkovic. Content-based representation and retrieval of visual media: A state-of-the-art review. In *Multimedia tools and applications*, volume 3, pages 179–202. Kluwer Academic Publishers, 1996.
3. J.R. Smith and S.F. Chang. Visualseek: a fully automated content-based image query system. In *ACM Multimedia '96*, November 1996.
4. D.A. Forsyth and M. Fleck. Body plans. In *Proceeding of IEEE Conference on Computer Vision and Pattern Recognition*, Puerto Rico, June 1997.
5. D.A. Forsyth M.M. Fleck and C. Bregler. Finding naked people. In *Proceedings of the Fourth European Conference on Computer Vision*, 1996.
6. K.S. Fu. *Syntactic Methods in Pattern Recognition.* Academic Press, a subsidiary of Harcourt Brace Jovanovich Publishers, 1974.
7. R.A. McLaughlin and M.D. Alder. Syntactic pattern recognition of simple shapes. In *Proceeding for ANZIIS93*, pages 5–9, Perth, December 1993.
8. R.A. McLaughlin and M.D. Alder. Technical report - the Hough Transform versus the UpWrite. Technical Report 97-02, CIIPS, Dept of EE Engineering, The University of Western Australia, http://ciips.ee.uwa.edu.au/Papers/, 1997.
9. P.S. Williams and M.D. Alder. Generic flaw detection within images. In *Proceeding for ICNN95*, volume 1, pages 141–145, Perth, November 1995.
10. P.S. Williams and M.D. Alder. Generic texture analysis applied to newspaper segmentation. In *Proceeding for ICNN96*, volume 3, pages 1664–1669, Washington DC, June 1996.
11. P.S. Williams and M.D. Alder. Image segmentation using an entropic UpWrite. Technical Report TR97-03, CIIPS, Dept of EE Engineering, The University of Western Australia, http://ciips.ee.uwa.edu.au/Papers/, 1996.
12. L. MacDonald R. Jackson and K. Freeman. *Computer Generated Color.* John Wiley and Sons, 1994.
13. H. Greenspan S. Belongie, C. Carson and J. Malik. Recognition of images in large databases using color and texture. In *CVPR '97*, 1997.
14. H. Greenspan S. Belongie, C Carson and J. Malik. Recognition of images in large databases using a learning framework. Technical Report 939, Computer Vision Group, UC Berkeley, 1997.
15. P.S. Williams. Corel images and segmentation results. Available online via "http://ciips.ee.uwa.edu.au/~williams/research/segmentation/", 1997.