

# Perceptual Grouping for Image Retrieval and Classification\*

Qasim Iqbal and J. K. Aggarwal  
Computer and Vision Research Center  
Department of Electrical and Computer Engineering  
The University of Texas at Austin  
Austin, Texas 78712, USA.  
{qasim, aggarwaljk}@mail.utexas.edu

## Extended Abstract

### 1 Introduction

In this paper we use perceptual grouping for image structure extraction for both image retrieval and classification. Image retrieval refers to the retrieval of images similar to a given query image, whereas image classification denotes the classification of all images in a database into a known set of classes. The developed approach is invariant to reflection, rotation and translation of image data. In addition, segmentation and detailed object representation is not required. For recent advances in the field of content-based access of images, refer to [1].

The human visual system can detect many classes of patterns and statistically significant arrangements of image elements. Perceptual grouping refers to the human visual ability to extract significant image relations from lower-level primitive image features without any knowledge of the image content. The grouping process hierarchically groups these relations continuously until a meaningful semantic representation is achieved that may be used by a higher-level reasoning process. The grouping principles proposed by Gestalt psychologists embodied such concepts as grouping by *proximity*, *similarity*, *continuation*, *closure*, and *symmetry* [2]. It has been noted that many of the perceptually salient image properties identified by the Gestalt psychologists (in their study of perceptual grouping of lower-level primitive image events into meaningful higher-level structures), such as collinearity, parallelism, and good continuation, are viewpoint invariant [3]. Quantitative analysis of image structure can also help to detect changes in image structure in a sequence of images [4].

### 2 Structure extraction via perceptual grouping – Feature selection

We extract the following features hierarchically in an unconstrained environment, i.e., with no constraints on the viewing angle and depth, using the approach detailed in [5]: *line segments*, *longer linear lines*, *coterminations*, *“L” junctions*, *“U” junctions*, *parallel lines*, *parallel groups*, *“significant” parallel groups* (Figures 1(a) - (f)). As an enhancement to that approach, we also extract closed figures comprised of *polygons* (Figure 1(g)). Perceptual grouping rules of similarity, continuity, parallelism and closure are used to extract these features.

Polygons are closed figures formed by non-parallel lines. A polygon is a significant image relation. According to the *closure* rule of perceptual grouping, human vision tends to complete curves to form enclosed regions [2]. Extracting closed figures corresponds to this feature of human vision. Polygons are non-accidental image relationships, since the coterminations forming them are non-accidental.

---

\*This work was supported in part by the Army Research Office under contracts DAAD19-00-1-0044, DAAG55-98-1-0230 and DAAD19-99-1-0012 (Johns Hopkins University subcontract agreement 8905-48168).

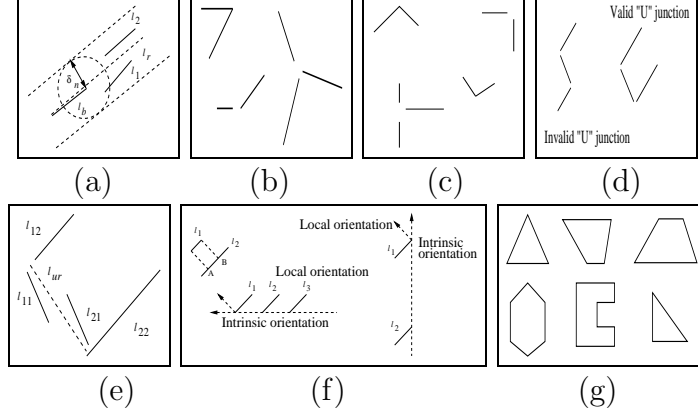


Figure 1: Visualization of the groupings. (a) Longer linear line (b) Coterminals (c) “L” junctions (d) “U” junction (e) “U” junction (f) Parallel groups (g) Polygons

Hence, polygons represent significant structure in an image. Elements of graph theory [6] are employed to extract polygons from an image using the cotermination graph. The underlying idea is to take advantage of the one-to-one correspondence between the closed figures comprised of line segments and the circuits in the graph. The connected components of the graph are found, and each sub-graph corresponding to each component is processed separately. The weight of a spanning tree is the sum of the weights of all the branches in the tree. We search for the maximal spanning tree, which may be found by slightly altering the minimal spanning tree algorithm to incorporate the vertices resulting in maximal-weight spanning tree [6]. The maximal spanning tree is employed to extract the fundamental circuits. Each fundamental circuit represents a closed figure in the image, where edges on this circuit correspond to line segments on the closed figure.

## 2.1 Feature extraction

In the general form the extracted feature vector is expressed as  $\mathbf{X}_S = (\tilde{\mathbf{x}}_{S_1}, \dots, \tilde{\mathbf{x}}_{S_d})^t$ , where  $d$  is the dimensionality of the feature space and  $\tilde{\mathbf{x}}_{S_i} = \frac{\sum_j \chi_{\omega_{S_i}}(l_j)}{\sum_k \chi_{\omega_{\Theta_l}}(l_k)}$ . In this representation  $\chi$  denotes the characteristic (indicator) function,  $l$  is a longer linear line,  $\omega_{\Theta_l}$  is the set of all longer linear lines, and  $\omega_{S_i}$  is the set of all higher-level structures extracted of a particular type. It is evident that  $\tilde{\mathbf{x}}_{S_i} \in [0, 1]$  ( $i \in [1, \dots, d]$ ), i.e., the feature space is represented by a unit hypercube.

For the generation of results for both retrieval and classification, we set  $d = 3$ . We let  $\omega_{S_i}$  represent “L” junctions, “U” junctions, and “(significant) parallel groups and polygons” for  $i \in \{1, 2, 3\}$ , respectively, i.e.,  $\tilde{\mathbf{x}}_{S_i}$  represents the corresponding normalized number of lines. Hence, the feature vector extracted is expressed as  $\mathbf{X}_S = (\tilde{\mathbf{x}}_{S_1}, \tilde{\mathbf{x}}_{S_2}, \tilde{\mathbf{x}}_{S_3})^t$ , where

$$\tilde{\mathbf{x}}_{S_1} = \frac{\# \text{ of lines in “L” junctions}}{\text{Total \# of longer linear lines}} \quad (1)$$

$$\tilde{\mathbf{x}}_{S_2} = \frac{\# \text{ of lines in “U” junctions}}{\text{Total \# of longer linear lines}} \quad (2)$$

$$\tilde{\mathbf{x}}_{S_3} = \frac{\# \text{ of lines in (significant) parallel groups and polygons}}{\text{Total \# of longer linear lines}} \quad (3)$$

Detailed justification for using this feature vector is provided in [7]. In addition, elimination of *weak-edged* line segments, and lines *shorter* than a threshold help to keep background clutter to a minimum [5, 7].



Figure 2: Image retrieval: An image query – a building facade. (Databases #1 and #2.)

### 3 Euclidean isotropy of $\mathbf{X}_S$

In image analysis, the input and output are functions of  $\mathcal{R}^2$ . An appropriate notion of the isotropy of computations is the Euclidean invariance – any rotation, translation or reflection of the input should produce an identical result under these transformations, thus achieving orientation and position invariance. These image transformations are generated by the action of the planar Euclidean group (the semi-direct product of the orthogonal group and the translation group). The Euclidean group is the group of isometries of  $\mathcal{R}^2$  – mappings that preserve distances – and its action on the space of positions and directions  $\mathcal{R}^2 \times \mathcal{S}^1$ , where positions are represented using  $\mathcal{R}^2$  and directions using the unit circle  $\mathcal{S}^1$ , generates isometric geometrical objects. Indeed, it has been argued that visual computations occur on  $\mathcal{R}^2 \times \mathcal{S}^1$ , rather than on just  $\mathcal{R}^2$  [8]. In our continuing work [9] we have demonstrated that  $\mathbf{X}_S$  obtained after perceptual grouping is a representation of isotropic mapping – mapping that is invariant to the action of the Euclidean group.

### 4 Results obtained

Our image databases consist of 2660 24-bit color images. Database #1 consists of 2139 images of size adjusted to  $1024 \times 1024$  acquired from two CDs obtained from “The Visual Delights Inc.” (<http://www.visualdelights.net>). Database #2 consists of 521 images of size adjusted to  $512 \times 512$  acquired from the ground level using the Sony Digital Mavica camera.

Figure 2 displays an example of image retrieval using both databases #1 and #2, where a query image is presented, and images similar to it are retrieved. Figure 2(a) shows the first 12 images retrieved using the proposed approach of extracting structure via perceptual grouping. A distance function based upon the Euclidean norm is used for this purpose. For comparison, Figure 2(b) shows the images retrieved for the same query image using a 512-bin uniformly quantized CIE Lab space color histogram. Similarly, Figure 2(c) shows the images retrieved by texture analysis using a *channel energy model* employing a bank of 16 even-symmetric Gabor filters in 4 scales and 4 orientations and measuring the energy in each channel of the CIE Lab space. (For details of the channel energy model see [10]; in this paper the original approach has been extended from grayscale to color images, and the texture analysis is done in the CIE Lab space. Refer to [9] for more

Class	T	R	C	Recall (C/T)	Precision (C/R)	Recall	Precision	Recall	Precision
S	255	235	198	77.65%	84.26%	64.31%	68.62%	48.24%	75.00%
N	140	143	114	81.43%	79.72%	70.00%	55.37%	45.71%	65.31%
I	96	113	49	51.04%	43.36%	28.13%	36.00%	58.33%	24.45%

(a) Perceptual

(b) Histogram

(c) Texture

Table 1: Image classification: Recall and precision. (Database # 2.) T = Total, R = Retrieved, C = Correct.

Class	1-100	201-300	301-400	401-500	501-521	T	M	E=M/T	E	E
S	78	78	68	24	7	255	200	78.43%	58.82%	59.22%
N	86	38	15	1	-	140	109	77.86%	42.86%	52.86%
I	47	21	9	16	3	96	45	46.88%	45.83%	34.38%

(a) Perceptual

(b) Hist.

(c) Tex.

Table 2: Image classification (Database # 2): Distribution of images *actually* belonging to a particular class in the “best matches” for that class, in intervals of 100 images (best matches), and the efficiency of the system. T = Total # of images belonging to a certain class, M = # of images that actually belong to a certain class in the first T best matches for that class, and E = Efficiency.

information.) The results obtained show the efficacy of using perceptual grouping over histogram and texture measures for the query presented.

The results for image classification were obtained using a nearest neighbor classifier and partitioning the image space into three classes, *Structure*, *Non-structure* and *Intermediate*, denoted as S, N, and I, respectively, based upon the measure of structure present in an image. A total of 30 training images are employed. Each class is represented by 10 training images. Table 1 displays performance measured in terms of *recall* and *precision*. For comparison, the results obtained for color histogram and texture analysis, as mentioned above, are also displayed. Table 2 shows the distribution of images that *actually* belong to a particular class within the “best matches” for that class, in intervals of 100 images, and the corresponding *efficiency* of the system. The best matches were obtained by sorting images based upon their distance from the training samples of each class. Again, perceptual grouping performs better than histogram and texture measures for analyzing the structural content of an image.

Space limitation precludes describing a retrieval methodology based upon a *combination* of structure, histogram and texture. For a detailed exposition of two different methods of such a combination, refer to [9, 10].

## References

- [1] A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, “Content-based image retrieval at the end of the early years,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1349 – 1380, December 2000.
- [2] David G. Lowe, *Perceptual organization and visual recognition*, Kluwer Academic publishers, 1985.
- [3] D.W. Jacobs, “What makes viewpoint invariant properties perceptually salient?: A computational perspective,” in *Perceptual Organization for Artificial Vision Systems*, K.L. Boyer, Ed., pp. 121–138. Kluwer, 2000.
- [4] Sudeep Sarkar and Kim L. Boyer, “Quantitative measures of change based on feature organization: Eigenvalues and eigenvectors,” *Computer Vision and Image Understanding*, vol. 71, no. 1, pp. 110–136, July 1998.
- [5] Qasim Iqbal and J. K. Aggarwal, “Applying perceptual grouping to content-based image retrieval: Building images,” in *IEEE International Conference on Computer Vision and Pattern Recognition, Fort Collins, Colorado*, June 1999, vol. 1, pp. 42–48.
- [6] Alan Gibbons, *Algorithmic Graph Theory*, Cambridge University Press, 1985.
- [7] Qasim Iqbal and J. K. Aggarwal, “Retrieval by classification of outdoor images containing large manmade objects using perceptual grouping,” *Pattern Recognition*, to appear.
- [8] J. Zweck and L.R. Williams, “Euclidean group invariant computation of stochastic completion fields using shifttable-twistable functions,” in *6<sup>th</sup> European Conference on Computer Vision (ECCV ’00), Dublin, Ireland*, 2000, pp. 100–116.
- [9] Qasim Iqbal and J. K. Aggarwal, “Image retrieval via isotropic and anisotropic mappings,” in *IAPR Workshop on Pattern Recognition in Information Systems, Setubal, Portugal*, July 6-8, 2001, to appear. <http://amazon.ece.utexas.edu/qasim/papers.htm>
- [10] Qasim Iqbal and J. K. Aggarwal, “Lower-level and higher-level approaches to content-based image retrieval,” in *IEEE Southwest Symposium on Image Analysis and Interpretation*, April 2000, pp. 197–201.