MULTI-SCALE EDGE DETECTION AND OBJECT EXTRACTION FOR IMAGE RETRIEVAL

Miguel Ferreira, Serkan Kiranyaz and Moncef Gabbouj
Tampere University of Technology, Finland {miguel.ferreira, serkan.kiranyaz, moncef.gabbouj}@tut.fi

ABSTRACT

A new scheme for boundary-based object extraction and description of still images, through multi-scale edge detection, is proposed in this paper. Boundary-based methods try to extract closed contours from individual edge pixels through edge-linking. Our approach is based on a connected structure of edge pixels as the initial edge-linking elements. These connected structures, the *sub-segments*, are extracted from the Canny edge map of an image. Multiple simplification-scales are derived from applying iterations of the Bilateral filter to the image, providing extra information about the relative importance of each *sub-segment*. Edge-linking towards contour closure is achieved through perceptually-driven minimum cost search. Furthermore, a shape-based description vector is derived from the extracted contours, and retrieval results are obtained via the integration of the whole scheme into MUVIS framework.

1. INTRODUCTION

The content description of still-images is a complex issue which involves the detection and integration of different low-level features in a well-structured manner, with respect to the human perception and visual system. From different features that may characterize a visual scene, the shape of objects is one of the most important for semantic analysis. The topics of image contour extraction and of shape description are covered separately in literature. Our approach in object extraction can be grouped under the boundary-based methods, such as the method described in [4]. Physiological evidence suggests that the human visual system (HVS) tends to respond to gradients, or differences, between homogeneous- regions, rather than to the interior of the regions themselves according to a process known as *Lateral Inhibition* [3]. Deriving from this consideration, an edge-detector based on the Canny edge algorithm [2] was implemented to detect the location of those boundaries in a given image. Ideally, one would like to extract automatically from an image all edges belonging to object boundaries. However, no matter what edge detector is used, its output tends to have some missing edges (which might also arise from occlusion) and/or distracting edges arising from noise, texture or illumination in the image. In this paper we propose a perceptually-motivated method to link edges together aiming to extract semantic object contours. In order to complete the broken object contours, a graph-based search algorithm is implemented over the field of edge *sub-segments* (SSs) attempting to find a minimum-cost closed path, allowing *virtual-links* to be established between nearby SSs. The psychological grouping rules known as *Gestalt* laws [7] provide a ground upon which the cost, or inversely, the saliency of a certain arrangement of SSs is based. Hence, certain *Gestalt* laws such as *proximity*, *good continuation* and *closure* are integrated in the search. Furthermore, a hierarchy of the initial SSs is created based on their respective scale information. Scale can be understood, in this context, as different stages of image simplification, through iterations of the non-linear Bilateral filter [6] over the original image. The analysis of the edges which "survive" at different stages of simplification, allows us to infer about the relative importance of each initial SS. This scale information is incorporated in the SS grouping search as an additional term in the cost function, influencing the outcome of the search in the favor of the most-relevant edges for object boundary extraction. Thus a set of closed contours is achieved by applying one search for a certain number of most important SSs (say N_{SS}), starting from the most relevant towards the least one. Each individual closed contour is then subject to shape description via a centroid-based angular-moment histogram. The individual contour descriptions are finally collected in an image feature vector (FV) which is used to calculate image distances in a given retrieval query.

The remaining of this paper is organized as follows: In section II, the initial pre-processing steps on image and subsequent edge detection with SS formation are explained. Section III presents the Bilateral filter and its role in the SS scale-"weight" calculation. Section IV sets the search space and search algorithm with respective *pruning* conditions and cost-function definition. In section V description of individual object shapes, and integration into a content-based multimedia retrieval system (MUVIS) is addressed. Experimental results are shown in section VI, and finally section VII draws some conclusions from the proposed overall scheme, which is shown in Figure 1.

2. SUB-SEGMENT FORMATION OVER EDGES

The individual images, arising either from a still-image or from a video stream, are first converted to the YUV format and only the Y (luminance) component is used in our further analysis. An image re-sampling step, with pre-defined minimum and maximum limits for the height and width, is performed in order to apply a standard approach. This also reduces complexity in the analysis, by keeping the image within reasonable dimensions. The Canny edge algorithm was used to detect the Y-edges, which accepts one smoothing factor parameter σ and two threshold parameters (high, thr_{high} and low, thr_{low}). From the Canny edge output, a set of coordinates is obtained, signaling edge pixel positions in the 2-D image space. This output is thinned using a simple mask, so that each edge pixel has at most two edge pixel neighbors (in the 8-neighborhood definition). The thinning operation is necessary for getting one-pixel thick edge SSs, and it is also responsible for the separation of joining edge *branches* (T-junctions). A s*ub-segment* (SS) is a structure containing an ordered series of 8-connected edge pixels. This structure can either have two *endpoints*, or can close itself in a loop (contour). The *Gestalt* laws of grouping are expressed in terms of these SSs, rather than based on individual edge pixels.

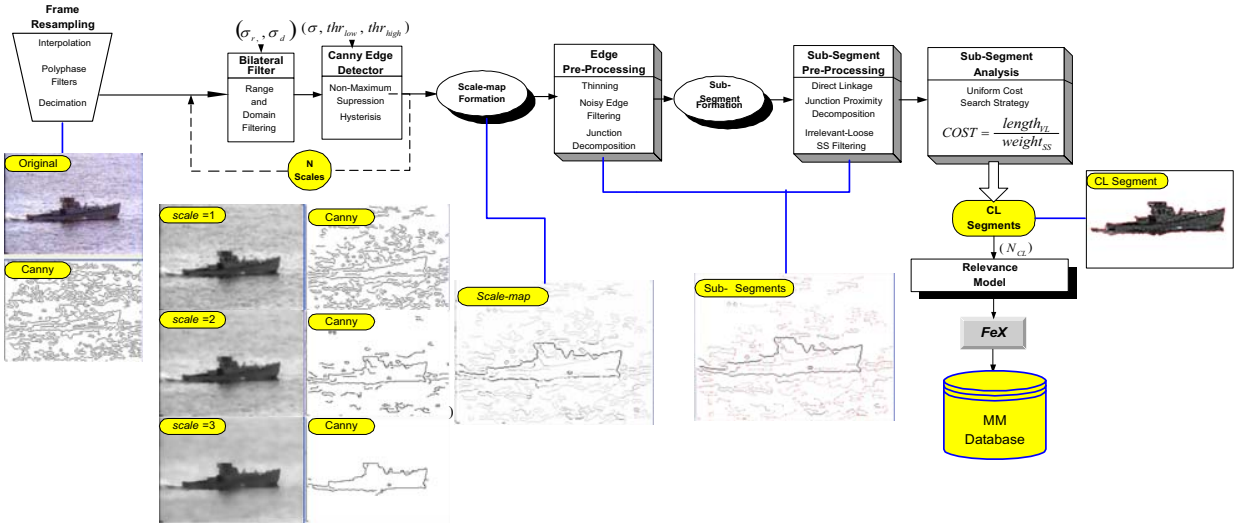

Figure 1: Overview of the proposed scheme

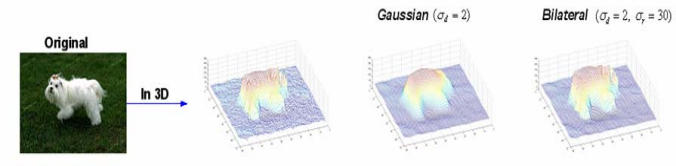

Figure 2: Gaussian vs. Bilateral filtering in 3D.

3. BILATERAL FILTERING AND WEIGHT ASSIGNMENT

The non-linear Bilateral filter was first proposed in [6] as a faster and simpler substitute for anisotropic diffusion [1], with the aim to preserve strong edges, whilst removing image details. When linear Gaussian filtering is applied to smooth out image noise and details, also important edges become smeared (i.e. edges become weaker), and in consequence their location might change as well. In scale-space theory, a solid scale interpretation is given in terms of applying varying Gaussian spatial filter widths (σ) to the image. However, resulting from the smearing of important edges, the problem of tracking boundary information across different scales is an obstacle. In the present work we implement a scale formulation which keeps fixed locations for "surviving" edges by using the Bilateral filter. A similar scale formulation using anisotropic diffusion is discussed in [5]. If I_{in} is an input image frame, then we can obtain the Bilateral filtered output, I_{out}, according to the following expression:

$$I_{out}(\vec{x}) = \int I_{in}(\vec{x}).\exp\left\{-\frac{(\vec{x}-\vec{\varepsilon})^2}{2\sigma_d^{\,2}}\right\}\exp\left\{-\frac{(I_{in}(\vec{x})-I_{in}(\vec{\varepsilon}))^2}{2\sigma_r^{\,2}}\right\}d\vec{\varepsilon} \qquad (1)$$

where the parameters σ_d and σ_r define one Gaussian in the spatial domain and one in the intensity domain of the image, respectively. As shown in Figure 2, the stronger edges remain while the texture and noise details are reduced. This "cartooning"-effect can be controlled by both σ-parameters and the number of iterations performed. If the σ's are kept fixed, and successive iterations of the filter are applied to an image, then we may consider the resulting simplification steps as different scales of the image. Then more iterations are performed to achieve higher scales. If the original edge positions are extracted from the lowest scale, one can assign scale-"weights" to them by adding a constant value to surviving edges at each successive scale.. Thus a simple scale-edge-map can be built, providing both edge locations and respective "importance" values, as shown in Figure 1, where darker pixel intensities represent higher scales. Using the previously formed connected structures, the total SS weight is calculated by the sum of its individual pixel weights.

4. SUB-SEGMENT ANALYSIS

In order to find the desired closed contours via SS merging, a search space is defined with respective states, possible transitions and a cost function. However, some pre-processing steps need to be performed beforehand, in order to achieve better SSs for merging. Also, the *virtual-link* (VL) structure is defined, as a straight line SS connection between two endpoints, with an allowed maximum length of VL_{thr}.

4.1. Sub-segment pre-processing

From the initial list of SSs, with assigned weights, only the N_{SS} most relevant are taken. Also, SSs with 3 pixels or less are considered as noise and therefore removed. A SS *relevancy measure* is simply given by its total weight, since it carries information about the number of pixels and their scale. Following this initial filtering, three steps are performed in the given order:

1) Connect single possible links; if an endpoint of a given SS may only establish a *virtual-link* to one other endpoint, of the same or different SS, then this link is immediately created and the SS(s) merged, resulting in a single and longer SS.

2) Break sub-segments by nearby endpoints; in many cases, an endpoint can lie very close to a middle point of another SS. Perceptually, this endpoint "breaks" that SS in two at the location of closest distance, and an implicit visual link is formed between the endpoint and the body of the SS. Since our model only assumes VLs between endpoints, additional endpoints need to be created at these locations.

3) Remove sub-segments with "loose" endpoints; an endpoint is "loose" when no VL can be established to any other endpoint. In this case, the SS to which the endpoint belongs is removed from

the search, since it cannot be part of a closed contour. These steps are illustrated in Figure 3.

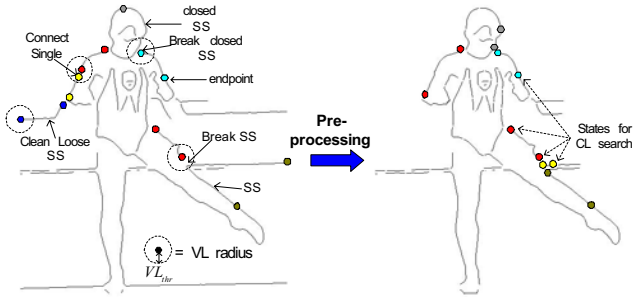

Figure 3: Pre-processing of SSs and final states for SS merging with the VL structure.

4.2. Search for closed-loops

A state in the proposed state space is simply a given endpoint with its location and associated SS identification. A minimum-cost path forming a closed loop (CL) among states is desired (*Gestalt* principle of *closure*), with the additional restrictions that it may not use twice the same SS and crossing between SSs is to be avoided. A cost function which agrees with two other *Gestalt* principles, *proximity* and *good continuation*, can be defined as:

$$C(e_1, e_2) = \frac{length_{VL}(e_1, e_2)}{weight_{SS}(SS_2)} \qquad (2)$$

The cost of virtual-linking the endpoints e_1 and e_2, with e_2 belonging to SS_2 is directly proportional to the length of the VL, and inversely proportional to the weight of SS_2, which can be decomposed in length and scale. So small VL "jumps" to more relevant SSs exhibit lower cost and are favored, complying with our perceptual judgment. The cost at a current state, following a certain previous path, is simply the sum of all transition costs in that path. A *uniform cost* search algorithm is applied to this state-space, and further techniques for enhancing time and space complexity of the algorithm are implemented using knowledge about restrictions on the desired solution. A hash table is built to keep the current minimum cost for each state, thus providing the possibility of *pruning*, not repeating higher-cost search sub-trees. Starting from the SS with highest weight, a start and end state are automatically defined using the SS's two endpoints. The minimum-cost CL is extracted for that SS, which is then removed and the search is repeated for the next SS with highest weight. At the end we obtain a list of CL segments from which the N_{CL} most important ones are used to describe the major objects in the image. The *relevancy, R(CL),* of an extracted *CL* segment is expressed by:

$$R(CL) = \frac{\sum_i weight_{SS}(SS_i)}{1 + \sqrt{cost(CL)}} \qquad (3)$$

where the SS_i are the SSs forming the CL segment.

5. CL DESCRIPTION FOR INDEXING

In MUVIS [8], feature extraction modules (*FeX*) are designed to build a fixed-size feature vector (FV) describing an image. From the extracted CL objects, a normalized histogram of the angular first moment is used as a shape feature. In Figure 4, the formation of one such histogram is represented for a CL object. The histogram has a number of bins (nP) determined by the user of the *FeX* module. The bin number also determines the angular resolution at which the shape moment is calculated. The angular first moment in a discrete CL segment is simply the sum of pixel distances, from the center of mass (CoM), within an angular section. The histograms of the most relevant N_{CL} CL segments can be grouped in an array to form the FV as shown in Figure 4.

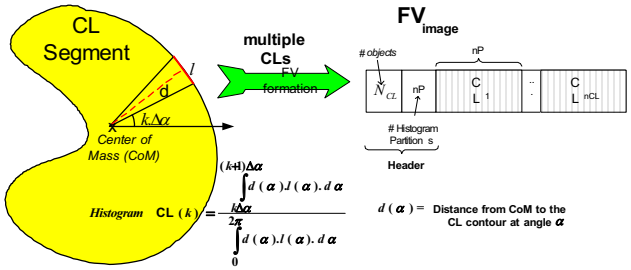

Figure 4: Feature Vector formation from CL histograms with nP partitions each.

Retrieval process in MUVIS is based on the traditional query by example (QBE) operation. The features of the query item are used for (dis-) similarity measurement among all the features of the visual items in the database. Ranking the database items according to their similarity distances yields the retrieval result. Between the query frame and a particular frame in the database, the feature vectors of all (N_{CL}) segments are used for similarity distance calculation with the following matching criteria: for each CL segment in the query frame, a "matching" CL segment in the compared frame is found. The *Euclidean* metric based minimum similarity distance is used as the matching criterion. Afterwards the total similarity distance will be the sum of matching similarity distances of all segments present in the query frame.

In order to achieve rotation invariance, the feature vector of a CL segment (angular histogram bins) is shifted one bin at a time with the other vector kept fixed and the similarity distance is calculated per step. The particular shift, which gives minimum similarity distance, is chosen for that particular CL segment of the query frame. Since this sliding operation on the histogram bins basically represents the feature vector of the rotated CL segment, the rotation invariance can therefore be achieved. The proposed shape descriptor is also translation and size invariant, since the CoM is used as a reference point and the histogram is normalized by the total CL boundary length.

6. EXPERIMENTAL RESULTS

During the experiments we used the following parameters: $\sigma = 1.5$, $thr_{low} = 0.4$ and $thr_{high} = 0.8$ (Canny detector), $\sigma_d = 2$ and $\sigma_r = 20$ for the Bilateral filter with 5 scales (1 iteration of the Bilateral filter each), and N_{CL} was set to one for the retrieval experiments in the natural image database (single CL in the final FV). For the binary (black and white) image shape database, the parameters used were the same, except from the number of scales, which was set to zero (no iteration). In order to examine the experimental results, a distinction must be made between *object extraction accuracy* and *descriptive power* of the proposed FV.

Figure 5: Two examples showing 12-best retrievals from the binary shape database. The top-left is the query image.

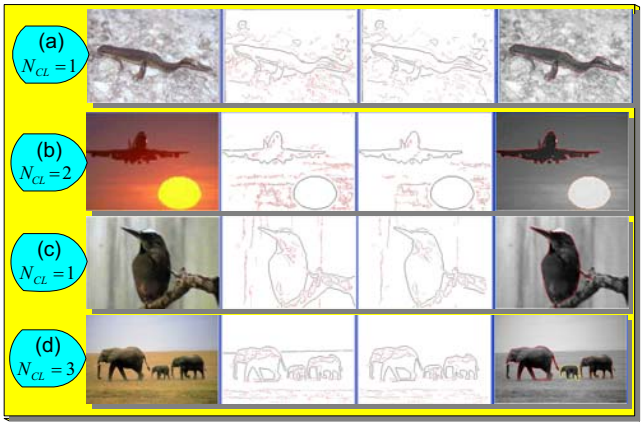

Figure 6: Typical CL extraction results. The leftmost shows the original image, the second and third columns are the *scale-maps* and SSs. The rightmost shows the CL objects.

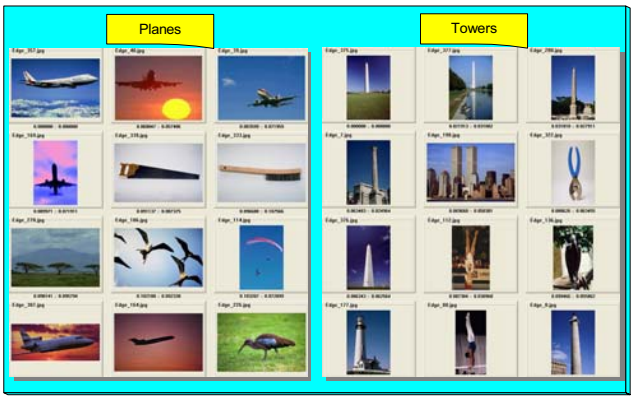

Figure 7: Two examples showing 12-best retrievals from the natural image database. The top-left is the query image.

In the binary shape database, which consists of 1400 images grouped in 70 different shape classes, the latter can be directly tested via QBE retrieval operations, since the extraction accuracy for the most relevant CL segment (i.e. the object) is around 100% as expected (due to clean edges and the absence of texture or noise). Two typical retrieval results showing the discriminative power of the proposed feature (angular-moment histogram) are shown in Figure 5, where size and rotation invariance is also observed. To test object extraction accuracy, many natural images have been processed, and promising results were obtained such as those shown in Figure 6 where N_{CL} CL segments are extracted from each image. Red, orange and yellow colors are used to represent the 1st, 2nd and 3rd most relevant CL segment. Finally, the retrieval results in a natural image database, which contains 400 images with various content, are shown in Figure 7.

The possible usefulness of this method for extraction and description of dominant shapes in an image is observed in the retrieval results. In the natural image database, coarse shape extraction and classification is obtained in the examples shown. Indexing both databases on a Pentium-4 processor with 1.8 GHz and 1024 MB of memory, lead to an average duration of 9.82 seconds per image in the natural database, and 1.21 seconds per image in the binary database.

7. CONCLUSIONS

The proposed method is, to the authors' knowledge, one of the first attempts at performing both object extraction and shape description of still images in an integrated and fully automatic way. Although many difficulties remain, such as parameter dependency, promising results can be obtained from a completely generic database, as shown in the examples given. A new scale definition is proposed by using iterations of the Bilateral filter over the image. Edges at different scales of the image are extracted, and this information is used in the CL object extraction procedure. Input parameters are used by the algorithm to adapt to specific database domains. The CL extraction is still dependent on the quality of the edge detection for multiple scales, although Canny edge and Bilateral filtering are used in the current work, future work will focus on more efficient ways to detect and to form SSs over edge pixels. The features extracted from the overall object extraction scheme provide an effective indexing and retrieval performance, as assessed via subjective query results

8. REFERENCES

[1] D. Barash, "A Fundamental Relationship between Bilateral Filtering, Adaptive Smoothing, and the Nonlinear Diffusion Equation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(6): 844-847, 2002.

[2] J. Canny, "A Computational Approach to Edge Detection", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 8, no. 6, pp. 679-698, November 1986.

[3] H. K. Hartline, "Inhibition of Activity of Visual Receptors by Illuminating Nearby Retinal Areas in the Limulus Eye", *Federation Proc.*, 8: 69, 1949.

[4] S. Mahamud, L. R. Williams, K. K. Thornber, and K. Xu, "Segmentation of multiple salient closed contours from real images", *IEEE Trans. on Pattern Analysis and Machine Intelligence,* 25(4):433-444, 2003.

[5] P. Perona and J. Malik "Scale-Space and Edge Detection Using Anisotropic Diffusion", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.12, No. 7, pp. 629639, 1990.

[6] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images", In *Proc. of the Sixth International Conference on Computer Vision*, Bombay, India, January 1998.

[7] M. Wertheimer, "Laws of organization in perceptual forms", Partial translation in W.B. Ellis, editor, *A Sourcebook of Gestalt Psychology*, pages 71-88, NY, Harcourt, Brace and Company, 1938.

[8] http://muvis.cs.tut.fi/