Contour Segment Matching by Integrating Intra and Inter Shape Cues of Objects

Ishani Chakraborty
http://research.rutgers.edu/~ishanic
Ahmed Elgammal
http://www.cs.rutgers.edu/~elgammal

Dept. of Computer Science,
Rutgers, State University of New Jersey
Piscataway, NJ
USA

Abstract

In this paper we propose an algorithm for contour-based object detection in cluttered images. Contour of an object shape is approximated as a set of line segments and object detection is framed as matching contour segments of an image (i.e.,an *edge image*) to a boundary model of an object (i.e., a *line drawing*). Local shape is abstracted as a group of k-adjacent segments. We use a *multi-level* shape description (with different k's) to capture complexity variations in local shape. Between images, shape descriptors are matched to give *inter-shape correspondences* and within images the underlying segment grouping enforces *intra-shape contextual constraints*. We use an efficient relaxation labeling approach that integrates these shape cues to qualify a contour match. To this end, we propose a novel framework that solves the problem of object detection as a contour segments correspondence problem. We then demonstrate the efficacy of the method for detecting various objects in cluttered images by comparing them to simple line drawings.

1 Introduction

Shape-based methods are a natural choice for color and texture invariant object detection. In recent years, a large body of research has focused on contour based techniques for shape representation. Most of the methods can be broadly classified as point-based approaches eg., [1] or boundary-curve based approaches eg., [2, 7].

In general, curve based shape representation has a natural advantage over point based approaches in terms of exploiting locality. This is because spatial neighborhood of a point is always limited to a localized region around the point, in terms of radius, patch size etc. An arrangement of connected curves, on the other hand, may emanate from a spatially localized region or it can be a spatially extended set of long segments associated only at their termination points. Hence, a set of boundary curves naturally handles scale variation better than a set of points in representing shape.

We perform object detection by framing it as a correspondence problem between contours segments in an input image and an object model. The model is a line drawing that consists of a small number of strokes defining the boundary contour of an object. In the input image, we identify an instance of the object category in a cluttered environment by searching for contour segments in a similar topology as that of the model. Hence, contour segments in the input image that match those of the line drawing delineate the object shape out of the cluttered background.

BMVC 2009 doi:10.5244/C.23.27

Figure 1: Overview of our approach. Left: Input image. (b) Left Center: Line segments (in white) extracted to form an *edge image*. (c)Right Center: Contour segments detected by *inter-shape* matching only (d)Right: Contour segments detected by combining *inter-shape* correspondences and and *intra-shape* contextual constraints. This is the output of our framework.

An important consideration in any shape-based method, including ours, is the extent of spatial context considered in the representation. A single global representation of shape is usually based on medial axis transformation [13] or a polynomial curve that defines the entire contour [10]. Such a representation provides a succinct description of the object in an unified model but is less repeatable under intra-class variability. On the other hand, groups of contour points [1], or curves [2, 7] encode local object parts at various scales in the image. In this case, the confidence measure for a reliable detection has to be aggregated spatially, but has the advantage of allowing more flexibility among object parts.

Our framework follows the latter approach; we approximate object contours using line-segments and groups of line segments are encoded by a descriptor. A multi-level description is used to capture local complexities in shape. Inter-shape distances between descriptors induce correspondences between contours.

In part-based representations, background parts often *hallucinate* [14] as objects and produce wrong correspondences. Therefore, a crucial step involves searching for object parts that form a *coherent whole*. The overall object is then, a group of contour segments that are spatially and structurally related to each other.

Our framework is summarized as follows. First, the input image is processed to extract and model local shape using the *Contour Segment Network* [2] model (Figure 1(a), (b)). We propose a multi-level shape description that handles local shapes of varied complexities and accumulate evidence across multiple layers of shape abstraction. See section 2. Next, to match contours between image and model, we define two metrics: (a) distance *between* the model and image contours (termed as inter-shape correspondence) to find likely candidates for a match. This is achieved by solving for one-to-one correspondences 2.2 and (b) spatial connectivities among contours *within* each image (termed as intra-shape contextual constraints) to find a single coherent whole that matches the model 2.3. To summarize, we exploit *inter-shape correspondences* and *intra-shape grouping cues* to identify a subset of connected line segments in the image that match the object model. We achieve this by means of a *relaxation labeling technique*. The results are demonstrated in 4, followed by our conclusion in 5.

Background and Related Work:
In our work, we frame object detection as contour segment matching between input and model image. The matches are established by integrating inter-shape correspondences and intra-shape grouping. Our method uses the shape description formulated in Contour Segment Network (CSN) [2] model. In their work, objects are detected by finding paths through the network resembling the model. The concept of object detection by contour matching has been previously applied in [14], in which shape contexts of points are extended to represent contour contexts. Continuity is included as curvature and distance measures on shape contexts in [11] and as centroid-boosting on k-point groups in [8]. Most of the above mentioned methods robustify shape matching by incorporating spatial constraints, that is solved by it-

erative approaches. In [14], contour sets are matched by 2-stage linear programming. [3] uses MCMC based labeling followed by contour based labeling. A matrix framework that combines the inter and intra-image cues is proposed in [17]. Cue combination is interpreted as structural graph matching and solved using EM and spectral techniques in [6]. Our cue integration method is similar in principle to the probabilistic Relaxation Labeling approach proposed in [5].

Our shape representation is based on line segments that are more robust and sparse as compared to point based approaches such as [14]. Other approaches that use CSN, like [2] and [8], use an empirically determined single level of shape abstraction. Moreover, the contour correspondences are built over individual line segment similarities. Our method differs from its predecessors in (1) We use a *multi-level* shape description to capture complexities of different object parts (2) We match groups of line segments across images for discriminative matches and use a novel mechanism to induce contour correspondences and (3) We follow a two-step intuitive approach for object detection: inter-shape correpondences are used to perform a dense search in the input image followed by a sparse, local search for the best matching candidate. We integrate these shape cues for contour segment matching in an iterative framework. Our algorithm is less vulnerable to background clutter, more adaptive to different complexities in shape and yields much higher detection rate than its predecessors. To the best of our knowledge, this is the first work that frames and solves the problem of object detection as a contour segments correspondence problem.

2 Structure Description and Matching

In our algorithm, an image is represented by contour segments terminating at points of high curvature (e.g., an *edge image*, see Figure 1(b)). We adopt the *Contour Segment Network* (CSN) [2] formulation to describe local shape. In CSN, contour segments are grouped based on spatial connectivity. An ordering of the segments is then enforced and a numerical descriptor encodes the structure as a vector with following attributes - (a) The relative distances between mid points of line segments (r_i, where $r_i = ||p_1 - p_i||, i = 2 \cdots k-1$), (b) Line segment orientations (θ_i, where $i = 1 \cdots k$) and (c) Length of the individual line segments, (l_i, where $i = 1 \cdots k$). The relative distances r_i and the segment lengths l_i are normalized by the distance between the farthest midpoints, making the descriptor scale-invariant.

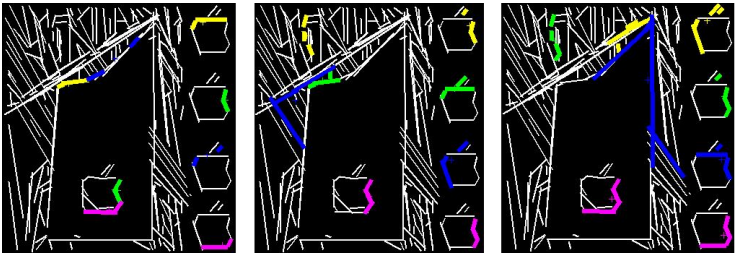


Figure 2: This figure shows the matching at different *structural supports*, from left to right, $k = 2, k = 3$ and $k = 4$. Each column has an input image and four instances of the model. Each instance shows a *structure* color-coded with the corresponding match in the image. Structural matches may arise both from the background and the object.

2.1 Multi-level Structure Description

We define a *structure* as a group of connected line segments encoded by a descriptor. *Structural support* is then, the number of line segments in a structure. An important consideration for local shape abstraction is the choice of the structural support that can best represent the complexities of an entire object shape. For example, a swan's beak or a bitten side of an apple can be illustrated accurately by a pair ($k = 2$ in CSN) of segments denoting a single point of high curvature. On the other hand, the curved side of an apple with continuous curvature variation is better represented by several contour segments that approximate the sampled curvature change. In general, there is no single value of k that can describe an entire shape effectively (as also noted in [2] and [8]).

A natural way to solve this scalability issue is to represent structures at multiple structural supports and filter out the irrelevant supports in a later process. Based on this idea, we generate and describe multi-support structures at each segment. Specifically, let $a, b \cdots$ and $\alpha, \beta, \cdots$ be segments in the input (D) and model (M) images. We generate structures $S_D^k = \cup_{1:k}\{a\}$ and $S_M^k = \cup_{1:k}\{\alpha\}$ at $k = 2, 3, 4$ and compute associated descriptors.

2.2 Inter-shape correspondences.

Our goal is to match contour segments between input and model images for which we require an inter-segment distance measure. The CSN apparatus provides a way to compare structures, i.e., groups of contour segments. Specifically, the distance between two k-structures in input and model is the Euclidean distance between their descriptors. In what follows, we describe a heuristic to compute inter-segment distances from the inter-structure distances.

An inter-structure distance $d^k_{S_D,S_M}$ is a distance between *two sets* of k *segments*. By slight twist of notation, we convert this distance to its segment-centric form, i.e., distance between k *sets* of *two segments* $= d^2_{(a,\alpha),(b,\beta),\cdots,k}$. I.e., each inter-structure distance $d^k_{S_D,S_M}$ is attributed to k ordered, pairwise, inter-segment distances with each $d_{a\alpha} = d_{S_D,S_M}$, where segment $a \in S_D, \alpha \in S_M$.

It has been observed in [2] that each segment in an image is typically connected to two to three other segments. Therefore, each segment can be a member of several structures. As each structure attributes a distance measure to its members, a pair of segments in two images may be associated by multiple inter-segment distances. These observations are illustrated in the Figure 3 in which three different $d_{a\alpha}$ values exist at $k = 3$.

To choose a single, most suitable inter-segment distance, we observe the following. Given an appropriate k sized structure k-S in the input image, where all its member segments belong to the object shape, a line segment in k-S will be in perfect correspondence with a model segment. Then, out of all inter-structure distances attributed to this segment, the one induced by k-S will be the minimum. Hence, to find its distance from the model, we apply a minimum filter across all inter-structure distances and all k to compute the distance that is induced by the best matching structure. Note that, if the segment indeed belongs to the object, this distance will be smaller than if it does not. Formally, then:

$$d_{a\alpha} = min_k min_{i,j}(d(S^k_{i,D}, S^k_{j,M} | a \in_p S^k_{i,D}, \alpha \in_p S^k_{j,M})) \quad (1)$$

A graph based interpretation: The relations between image and model contour segments can be naturally encoded in a graphical framework. We express the input image as a graph and its contour segments as vertices V_D. Similarly, model image has vertices V_M. The inter-segment distances $d_{a\alpha}$ as computed in equation 1 are normalized and modified into

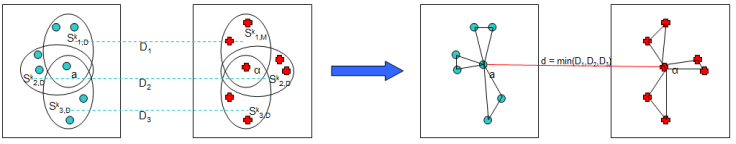


Figure 3: Conversion of inter-structure distance to inter-segment distance. The structures are marked $S^k_{i,G}$ with inter-structure distances D_i. The inter-segment distance is the minimum inter-structure distance. The structural grouping is translated to a clique to enforce intra-shape constraints.

probability scores by applying a Gaussian kernel. We generate a probabilitity matrix Q_{ic} as $|V_D|X|V_M|$ adjacency matrix (Equation 2), the elements of which express the initial likelihood of the segment matches. These probabilities depict the *inter-shape correspondences* which are refined by including *intra-shape constraints* as described next.

$$Q_{ic} := q(a,\alpha) = e^{-d^2_{a\alpha}} \tag{2}$$

2.3 Intra-shape Contextual Constraints.

Part-based representations model local shape, ignoring the spatial context of those parts. As a result, background parts often hallucinate as objects and create wrong matches. To mitigate erroneous detections, we include contextual constraints and search for a connected, coherent whole within the input.

We represent the contextual constraints of contour segments by the connectivities that underlie contour grouping in the formation of structures. We define *intra-shape* adjacency matrices in which two nodes in a graph are connected if their representative contour segments are members of a common structure.

$$\begin{aligned} D_{ab} &= 1, \text{ if } \exists S^k_D(a \in S^k_D, b \in S^k_D) & M_{\alpha\beta} &= 1, \text{ if } \exists S^k_M(\alpha \in S^k_M, \beta \in S^k_M) \\ &= 0, otherwise & &= 0, otherwise \end{aligned} \tag{3}$$

Specifically, the *intra-shape adjacency matrices* for input (D_{ab}) and model ($M_{\alpha\beta}$) are binary matrices where a pairwise relation of 1 implies that the two contour segments are connected and are within the spatial context of one another (Equation 3).

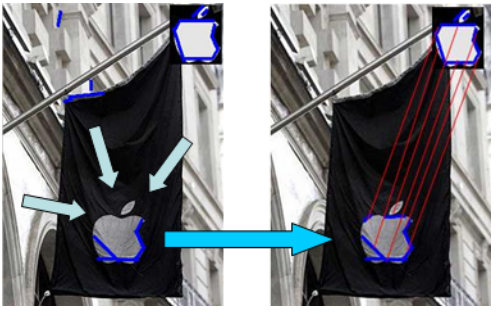


Figure 4: The first image shows the contour matches found using inter-shape correspondences only. These matches are enhanced by including contextual constraints that drive the matches towards a connected set of contour segments that match strongly with the model in the second image.

3 Combining Intra and Inter shape Cues.

The end goal of object detection is to find a *coherent whole*: a set of connected contour segments in the input image that matches best with the model segments. We achieve this by integrating *inter-shape correspondences* with *intra-shape groupings*.

Formally, we seek an optimal match $V_D \rightarrow V_M$ such that a subset of input image segments $A = \{a, b, \cdots\}$ are assigned to model segments $\Lambda = \{\alpha, \beta, \cdots\}$. Moreover, the segments in set A should be connected via intra-shape adjacencies either directly, i.e., $D_{ab} = 1$ or indirectly, $D_{ac_1} = 1, D_{c_1,c_2} = 1, \cdots, D_{c_n,b} = 1$. The unmatched contour segments (and the background) in input image match dummy segments in V_M. These assignments can be easily expressed in a matrix framework. Let M be a binary matrix of size $|V_D|X|V_M|$ with each element defined as:

$$\begin{aligned} M := m_{a\alpha} &= 1, a \rightarrow \alpha \\ &= 0, otherwise \end{aligned}$$

This implies that a match between contour segment a and α, represented as an assignment $a \rightarrow \alpha$ is denoted by weight 1 in the assignment matrix. To induce a one-to-one correspondence between the nodes V_D and V_M we find the optimal assignment M.

$$M = argmax_{\hat{M}} \sum_a \sum_\alpha Q(a, \alpha) \hat{m}_{a\alpha} \tag{4}$$

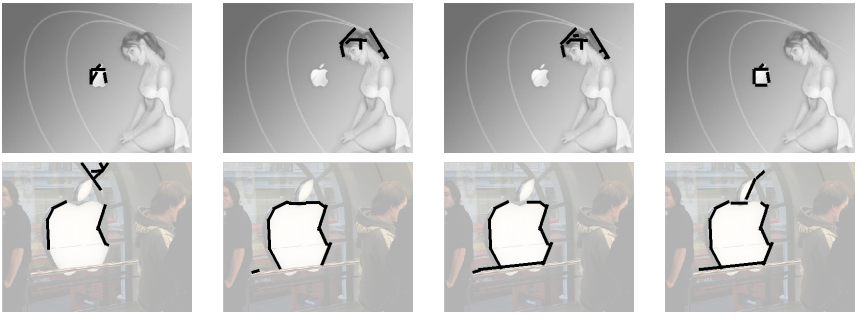
where $Q := Q(a, \alpha)$ is the probability matrix for this assignment. If this probability matrix is known, the assignment matrix M can be calculated by maximum weighted graph matching. The initial probability matrix is based on inter-shape correspondences $d_{a\alpha}$. To find a single connected entity, we need to softly bias the matches towards a group of connected contour segments. This can be achieved by integrating the inter-shape matches with intra-shape grouping in an iterative framework as described below.

Matching via Relaxation Labeling

To compute the optimal assigment M we see that the probabilities and assigments are interdependent functions. I.e., optimal assignment depends on the probabilities but the probabilities themselves are also influenced by a particular assignment. Thus the optimization problem needs to be solved iteratively. Our iterative approach is based on the relaxation labeling framework for contextual graph matching [5].

We break the problem into a two-step iterative approach. In the first step, we find an assignment M_t based on the probabilities Q_t between the nodes of input and model graph. The intra-shape cues are ignored at this stage and we formulate it as a bipartite graph matching which is solved using the polynomial-time Hungarian Algorithm. In the second step, we recalculate the probability matrix Q_{t+1} based on a *support function*. This is calculated as a function of matches M_t in the first step and the contextual cues expressed as intra-shape adjacency matrices D_{ab} and $M_{\alpha\beta}$. The iteration continues till a stable local point is reached and the corresponding M is the optimal assignment that identifies the object contours in the input image. The details of the algorithm are explained as follows.

The relaxation procedure starts by assigning to the data nodes the initial probabilities $Q_0 = Q_{ic}$ based on inter-contour distances as calculated in Equation 2. This induces a matching denoted by M_0, the initial labeling. The relaxation rule is then:



(a) (b) (c) (d)

Figure 5: Effects of different structural support, k. Contour segments detected by combining *inter-shape* and *intra-shape* cues for (a)$k = 2$(b)$k = 3$ (c)$k = 4$(d)$k = 2,3$ and 4, combined. In the top figure, the object contour is completely undetected for $k = 3$ and 4. In the bottom figure, even though all the k's approximately detect the same contours, higher support favors better detection. In general, the performance is best when information is combined over all the k's (column d).

$$Q_{a\alpha}^{t+1} = \frac{Q_{a\alpha}^{t} S_{a\alpha}^{t}}{\sum_{b}\sum_{\beta} Q_{b\beta}^{t} S_{b\beta}^{t}} \tag{5}$$

where Q^t is the probability matrix and the support function S^t weighs the probabilities according to the intra-shape contextual constraints.

Calculating the support function: To qualify as an object contour, a segment in an input should not only have a high inter-shape correspondence with the model but should also lie within the context of an object. In other words, a segment can qualify as a match if it is connected to segments that also match to the object. The role of the relaxation based iteration is to bias the matches towards the strongest candidate that fulfils the inter- and intra-shape constraints. The support function is used to induce this bias in the probability matrix.

We employ an indicator function $I_{b\to\beta}^{a\to\alpha}$ such that

$$\begin{aligned} I_{b\to\beta}^{a\to\alpha} = D_{ab} M_{\alpha\beta} m_{b\beta} &= 1, if (a,b) \in E_D, (\alpha,\beta) \in E_M, a \to \alpha, b \to \beta \\ &= 0, otherwise \end{aligned}$$

The above states that the indicator value is unity if (1) b is connected to a and (2) is matched to node β which is connected to α i.e., the label assignment of a. When either of these two conditions are not met, the quantity is zero. Note that $M_{\alpha\beta}$ refers to the intra-shape adjacency matrix for the model graph whereas $m_{b\beta}$ is an element of the assignment matrix M.

We define support function at a node as the joint probability of the nodes that are connected to it. Assuming independence, this joint probability is the product of the individual probabilities. Thus, the support function distribution can be considered as a log-normal distribution which is normalized by finding the n^{th} root of the product (the geometric mean), where $n = \sum_b \sum_\beta I_{b\to\beta}^{a\to\alpha}$, the number of nodes connected to α.

$$S_{a\alpha}^{t} = \left\{ \prod_{b}\prod_{\beta} Q(b,\beta) I_{b\to\beta}^{a\to\alpha} \right\}^{1/n} \tag{6}$$

	Apple Logo	Bottle	Swan	Giraffe
$k=2$	68.9	59.6	83.3	75.8
$k=3$	84.4	53.2	88.9	80.4
$k=4$	88.9	85.1	86.1	68.9
Multi-level	**93.3**	**91.5**	**94.0**	**89.6**
in [2]	57.0	90.0	75.0	63.0

Table 1: Detection rates at 0.30 *FPPI*.

4 Experimental Results

We evaluate our detection algorithm on the ETHZ shape dataset, that contains object categories at various scales, illumination and with large cluttered backgrounds. The method is tested on 4 shape classes namely, *Apple logo* (45 images), *bottles* (47 images), *giraffe* (87 images) and *swans* (36 images). The object model, which is also in the dataset, is a single line drawing of each shape in the test category.

We choose 3 levels of structural support at $k = 2, 3$ and 4. During relaxation labeling, we iterate between maximum weighted matching and updation of probability scores till the solution remains unchanged over two consecutive iterations. This signals a stable solution. The contour segments that match to the model at this stage are labeled as object contours. To localize the object, we compute the extremeties of the largest set of connected, matched contours and frame them with a bounding box. Any other set is counted as a false positive, except in images with multiple instances of the same object, in which they are ignored. The output of this system is a *bounding box* and the *detected contours*.

Detection rate/False Positives Per Image (DR/FPPI) is used for quantitative evaluation. A detection is considered correct if the ground-truth overlaps the detected bounding box over 50% of the region. However, if the bounding box exceeds the ground-truth by 20%, the detection is incorrect.

We compared our results with previous work [2] as shown in Table 1. The detection rates at 0.30 FPPI averaged over each dataset is shown in the table. The results illustrate that our algorithm performs remarkably well on all the four object classes, with an average detection rate of 92.17%. The Apple logo category is particularly interesting because this dataset has maximum clutter and variations in images. Our method outperformes the preceding work by a huge margin of 36%. To show our performance in localizing the actual contours of the shape, we show a few examples in Figure 6. We highlight the two important aspects of our algorithm, that lead to this improved performance.

First, to emphasize the importance of variable structural support, we display the results obtained at different k (see Figure 5). It is interesting to note that the detection of Apple logo improves significantly at higher structural support, whereas the swan and giraffe categories are better detected at $k = 3$. One reason that might explain this anomaly is that natural shapes of animal images are more deformable than the Apple which is a brand logo and hence is mostly consistent across images. Smaller, simpler structures are more likely to match correctly in natural shapes that longer, more complex structures. Most of the previous works on CSN have considered a single spatial support in their shape representations ($k = 2$ in [2] and and $k = 3$ in [8]). By considering different structural supports k, we obtain the best possible match across three levels of representation. Due to this, the results from multiple structural supports clearly supercedes the individual supports for all the four classes.

Second, contextual constraints help robustify the detection and minimize the detection of false contour segments. As noted in [14], background contours often hallucinate as ob-

Figure 6: Contour detection results.

jects and create wrong matches. Inclusion of contextual constraints leads to detection of a single (or a small number of) connected candidate(s) that match best with the model. After including contextual constraints, the number of correctly detected contours that belong to the object increased by 34% over the entire dataset. Thus, even without integrating multiple levels of structural support we obtain better detection rates than [2] (see first three rows of table 1).

5 Conclusions

Line segment matching has been used before for stereo analysis and image registration [4, 9]. In this paper we proposed a novel framework to use line-segment matching as a method for object detection and localization in a cluttered image. Our approach is simple and intuitive; a line drawing is used an object model and contour segments in an input image that share similar structure and context as model contours are deemed as the detected object. We show that contour line segments are able to represent local structures of deformable shapes efficiently when used in a multi-level framework. Inter and intra shape cues are exploited to delineate a single connected set of contours that matches closely to the model. Our method outperforms previous work in object detection and can be also be used to localize the object

contour in an image.

References

[1] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 24 (4):509–522, April 2002.

[2] V. Ferrari, T. Tuytelaars, and L.J. Van Gool. Object detection by contour segment networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages III: 14–28, 2006.

[3] A. Gupta, J.B. Shi, and L.S. Davis. A "shape aware" model for semi-supervised learning of objects and its context. In *Proceedings of Advances in Neural Information Processing (NIPS)*, 2008.

[4] G. Karimian, A.A. Raie, and K. Faez. A new efficient stereo line segment matching algorithm based on more effective usage of the photometric, geometric and structural information. *Transactions Institute Elec. Info. and Comm. Eng.*, E89-D(7):2012–2020, July 2006.

[5] A. Kostin, J.V. Kittler, and W.J. Christmas. Object recognition by symmetrised graph matching using relaxation labelling with an inhibitory mechanism. *Pattern Recognition Letters (PRL)*, 26(3):381–393, February 2005.

[6] B. Luo and E.R. Hancock. Structural graph matching using the em algorithm and singular value decomposition. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 23(10):1120–1136, October 2001.

[7] A. Opelt, A. Pinz, and A. Zisserman. A boundary-fragment-model for object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages II: 575–588, 2006.

[8] S. Ravishankar, A. Jain, and A. Mittal. Multi-stage contour based detection of deformable objects. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages I: 483–496, 2008.

[9] C. Schmid and A. Zisserman. Automatic line matching across views. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 666–671, 1997.

[10] J.P. Tarel and D.B. Cooper. A new complex basis for implicit polynomial curves and its simple exploitation for pose estimation and invariant recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 111–117, 1998.

[11] A. Thayananthan, B. Stenger, P.H.S. Torr, and R. Cipolla. Shape context and chamfer matching in cluttered scenes. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages I: 127–133, 2003.

[12] A. Toshev, J.B. Shi, and K. Daniilidis. Image matching via saliency region correspondences. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2007.

[13] J. Xie, P.A. Heng, and M. Shah. Shape matching and modeling using skeletal context. *Pattern Recognition (PR)*, 41(5):1773–1784, May 2008.

[14] Q.H. Zhu, L.M. Wang, Y. Wu, and J.B. Shi. Contour context selection for object detection: A set-to-set contour matching approach. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages II: 774–787, 2008.