Int J Comput Vis
DOI 10.1007/s11263-012-0521-z

Shape-Based Object Detection via Boundary Structure Segmentation

Alexander Toshev · Ben Taskar · Kostas Daniilidis

Received: 3 July 2011 / Accepted: 23 February 2012

Abstract We address the problem of object detection and segmentation using global holistic properties of object shape. Global shape representations are highly susceptible to clutter inevitably present in realistic images, and thus can be applied robustly only using a precise segmentation of the object. To this end, we propose a figure/ground segmentation method for extraction of image regions that resemble the global properties of a model boundary structure and are perceptually salient. Our shape representation, called the chordiogram, is based on geometric relationships of object boundary edges, while the perceptual saliency cues we use favor coherent regions distinct from the background. We formulate the segmentation problem as an integer quadratic program and use a semidefinite programming relaxation to solve it. The obtained solutions provide a segmentation of the object as well as a detection score used for object recognition. Our single-step approach achieves state-of-the-art performance on several object detection and segmentation benchmarks.

Keywords Shape representation · Shape matching · Object recognition and detection · Object segmentation

A. Toshev (✉)
Google Research, 1600 Amphitheatre Parkway, Mountain View, CA 94043, USA
e-mail: toshev@google.com

B. Taskar · K. Daniilidis
GRASP Lab, University of Pennsylvania, 3330 Walnut St, Philadelphia, PA 19104, USA

B. Taskar
e-mail: taskar@cis.upenn.edu

K. Daniilidis
e-mail: kostas@cis.upenn.edu

1 Introduction

A multitude of different object representations have been explored, ranging from texture and local features to region descriptors and object shape. Although local features based on image gradients and texture perform relatively well for some object classes, many classes are not modeled sufficiently by local descriptors. For objects primarily characterized by distinctive shape, local texture features typically provide weak descriptions. In this paper we focus on the problem of exploiting global shape properties for object detection. Moreover, we tightly couple these properties to object segmentation, which makes shape-based detection possible in cluttered scenes.

Shape is commonly defined in terms of the set of contours that describe the boundary of an object. In contrast to gradient- and texture-based representations, shape is more descriptive at a larger scale, ideally capturing the object of interest as a whole. This has been recognized by the Gestalt school of perception, which has established the principle of *holism* in visual perception (Palmer 1999; Koffka 1935). This principle suggests that an object should be perceived in its totality and not merely as an additive collection of individual parts. The essential goal of a holistic representation for object recognition is to capture not just the presence of object parts but also non-local relationships between these parts. In this work, our response to the mantra 'the whole is greater than the sum of its parts' is 'the whole is the sum of all the relationships between its parts', as we make precise below.

Some of the most notable holistic representations are based on global transforms, such as Fourier transform (Zhang and Lu 2003) or the Medial Axis Transform (Blum 1973). Unfortunately, such transforms assume a pre-segmented object shape as input. As a result, the above rep-

Published online: 22 March 2012

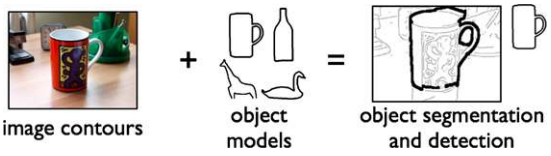


Fig. 1 Using BoSS to perform simultaneous shape-based object detection and segmentation in a cluttered scene

resentations cannot be used directly for object detection in realistic scenes which inevitably contain clutter.

To address the problems arising from clutter, a number of structural theories for object perception were introduced. According to this paradigm, an object can be decomposed and described as a configuration of atomic parts. Structuralism has inspired a number of approaches such as generalized cylinders (Marr 2010; Binford 1971), Recognition by Components Theory (Biederman 1987), and superquadrics (Pentland 1986). Although being well motivated from a perceptual point of view, the above approaches have not found wide applicability. First, the theories assume that one can extract the shape primitives in images, which is very difficult in realistic images. Second, even if one can obtain good primitive candidates from an image, the search for the correct shape is typically not straightforward and tractable (Grimson and Lozano-Perez 1987).

To alleviate the above problems, a number of approaches were proposed in recent years that use primitives which are simpler and easier to extract such as edgels (Huttenlocher et al. 1993), contour segments (Ferrari et al. 2008) or statistical descriptors of local or semi-local contours such as Shape Context (Belongie et al. 2002). The above local primitives are combined in a global configuration model. Depending on expressiveness of the model, inference can be intractable, such as graph matching where one captures all pairwise dependences among parts (Leordeanu et al. 2007), or tractable, such as, for example, dynamic programming (Ling and Jacobs 2007) in which case many of the dependences are left out. Another strategy is to capture all global dependences among parts in a less expressive model such as Thin-Plate Splines (Belongie et al. 2002) or Procrustes (Mcneill and Vijayakumar 2006). The above shape models present a step towards recognition in cluttered scenes but depart from the idea of holism.

In this work we advocate holistic shape-based recognition in realistic cluttered scenes. In particular, we propose a recognition method, called **Bo**undary **S**tructure **S**egmentation (BoSS).[1] This method relates the object detection, based on a novel holistic shape descriptor, to figure/ground segmentation and performs them *simultaneously* (see Fig. 1). While matching an input image with an object model, BoSS selects a foreground region with the following properties:

– *Similarity in Shape*: captured by a top-down process exploiting object-specific knowledge. Evidence from human perception indicates that familiarity with the target shape plays a large role in figure/ground assignment (Palmer 1999).
- *Perceptual Saliency*: captured by a bottom-up process based on general grouping principles, which apply to wide range of objects. In particular, the perceptual grouping component is based on configural cues of salient contours, color and texture coherence, and small perimeter prior.

Furthermore, the shape-based detection costs of matching several models to an image can be used to detect the corresponding object class as the one whose model has the smallest matching cost. In this way, object segmentation and detection are integrated in a unified framework. More precisely, the contributions of the approach are threefold:

Shape Representation We introduce a *global, boundary-based shape representation*, called *chordiogram*, which is defined as the distribution of all geometric relationships (relative location and normals) between pairs of boundary edges—called chords—whose normals relate to the segmentation interior. This representation captures the *boundary structure* of a segmentation as well as the position of the *interior* relative to the boundary. Moreover, the chordiogram is translation invariant and robust to shape deformations.

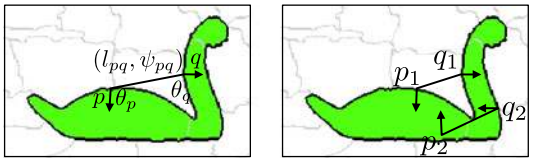
The chordiogram can be theoretically related to correspondence estimation techniques and thus to other common shape matching approaches. In particular, we show that the cost of chordiogram matching is a lower bound on the cost of the point correspondence estimation problem between two shapes. Furthermore, it is also equal to the cost of chord correspondence problem between two shapes. Thus the chordiogram provides approximate means to measure the cost of point correspondence estimation without the need of explicit inference.

Figure/Ground Segmentation We match the chordiogram while *simultaneously* extracting figure/ground segmentation. This is a key advantage of the representation, which relates the object boundary to its interior and thus to region segmentation. The perceptual grouping component of the segmentation model, which is defined in terms of configural cues of salient contours, color and texture coherence, and small perimeter prior, ensures that the detections constitute salient regions. More importantly, the joint matching and segmentation removes the irrelevant image contours during matching and allows us to obtain correct object detections and segmentation in highly cluttered images.

[1] A preliminary version of this work appeared in CVPR 2010 (Toshev et al. 2010).

Inference We pose BoSS in terms of selection of superpixels obtained via an initial over-segmentation. The selection problem is a hard combinatorial problem which has a concise formulation as an integer quadratic program consisting of two terms—a boundary structure matching term defined over superpixel boundaries, and a perceptual grouping term defined over superpixels. The terms are coupled via linear constraints relating the superpixels with their boundary. The resulting optimization problem is solved using a Semidefinite Programming relaxation and yields shape similarity and figure/ground segmentation *in a single step*.

We achieve state-of-the-art results on two challenging object detection tasks—94.3% detection rate at 0.3 fppi on ETHZ Shape Dataset (Ferrari et al. 2006) and 92.4% detection rate at 1.0 fppi on INRIA horses (Ferrari et al. 2007) as well as accurate object boundaries, evaluated on the former dataset.



(a) Chord features. (b) Normals of two chords.

Fig. 2 Chord features and orientation of the normals at boundary edges

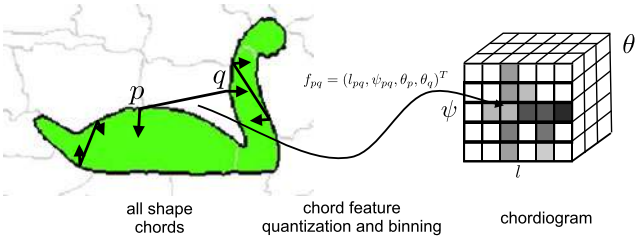


Fig. 3 For an input shape, all chord features are binned in the quantized chord feature space which is the resulting chordiogram

2 Chordiogram

We introduce a novel shape descriptor, called *chordiogram*. This descriptor adheres to the principle of holistic visual perception by describing each object contour in the context of the whole object. In other words, the contribution of an edge or a contour to the whole object representation depends on all other object contours. Furthermore, it captures both the boundary as well as the interior of the object. In addition, it is invariant to certain rigid transformations and robust to shape deformations. Most importantly, however, it can be applied in images with severe clutter, which allows for recognition in unsegmented images.

To define the chordiogram, consider the outline of a presegmented object as shown in Fig. 2(a) and denote by C a set of sampled boundary points of this outline (in the following we will include in C all the pixels lying on the outline). A pair of boundary edges p and q from C will be referred to as a chord. We can think of a chord as a way to express a dependency between edges p and q. We define features describing the geometry of the chord:

- *Length* l_{pq} and *orientation* ψ_{pq} of the vector $p \to q$.
- *Normalized normals* θ_p and θ_q to the boundary at p and q with respect to the chord orientation ψ_{pq}: $\theta_p = \theta'_p - \psi_{pq}$ and $\theta_q = \theta'_q - \psi_{pq}$, where θ'_p and θ'_q are the normals at p and q respectively.

Thus, the *chord features* can be written as a four-tuple:

$$f_{pq} = (l_{pq}, \psi_{pq}, \theta_p, \theta_q)^T. \tag{1}$$

We describe the shape of a segmented object by capturing the features of all chords. In this way we attempt to capture all dependencies among object boundary points and achieve a holistic description. More precisely, the chordiogram (denoted by ch) is defined as a K-dimensional histogram of all chord features, where the features are quantized into bins and the mth chordiogram element is given by:

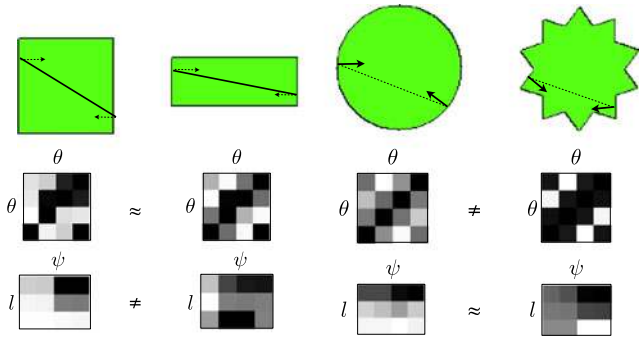
$$ch_m(C) = \#\big\{(p,q) \mid p,q \in C,\ f_{pq} \in bin(m)\big\}. \tag{2}$$

Note that the above definition can be applied not only to contours but also on any unordered point set C for which the points have normals associated with them.

The chordiogram construction process is visualized in Fig. 3. The length features l_{pq} are binned in b_l bins in log space, which allows for larger shape deformation between points lying further apart. The length h of the largest bin determines the scale of the descriptor—every chord whose two boundary points lie within distance h will be captured by the descriptor. To guarantee that the descriptor is global, we set h equal to the diameter of the object in case of presegmented object masks. The remaining three features are angles lying in $[0, 2\pi)$ and are binned uniformly—the chord orientation in b_r bins; the normal angles are binned in b_n angles. This binning strategy results in a $N = b_l \times b_r \times b_n^2$ dimensional shape descriptor at scale h.

The chord features are chosen such that they completely describe the geometry of a chord. When it comes to the chordiogram, the features capture different shape properties. The chord length and orientation capture global coarse shape properties (see Fig. 4(a)), while the fine information is captured by the normals (see Fig. 4(b)).

The chord features determine the invariance of the chordiogram to geometric transformations. Since we do not capture absolute location information, the resulting descriptor is translation invariant. However, the chord orientation feature prevents the descriptor from being rotation invariant.



(a) Global shape difference. (b) Local shape difference.

Fig. 4 For each pair of shapes (*upper row*), we show two chordiograms: one computed over the normal features θ only (*middle row*) and one over the chord length l and orientation ψ (*lower row*)

Similarly, the chord length feature prevents the chordiogram from being scale invariant. Removing those features would make the descriptor rotation and scale invariant, however, less descriptive. Hence, we have chose to keep these features and perform a search over scale and rotation.

To evaluate the dissimilarity between two shapes we can use any metric between the chordiograms extracted from the shapes. In the subsequent experiments we use L_1 distance between L_1-normalized chordiograms, which we will call *chordiogram distance*:

$$d(u,v) = \left\| u/\|u\|_1 - v/\|v\|_1 \right\|_1 \tag{3}$$

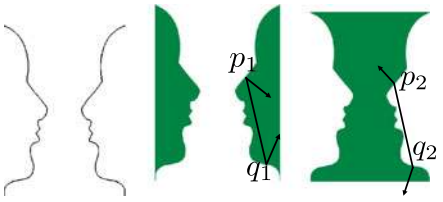
for two chordiograms u and v.

3 Properties and Analysis of the Chordiogram

In this section, we explore the properties of the chordiogram as a shape descriptor, motivate its holistic nature and present a theoretical analysis of the connection between chordiograms and point-set correspondence methods. In the next section, we show how chordiograms can be used in cluttered images via joint segmentation and detection.

3.1 Figure/Ground Organization

An important difference with most contour-based shape representations, is that the chordiogram captures the contour orientation relative to the object interior. Orienting the boundary normals with respect to the interior allows us to capture different interpretations of a contour, as shown in Fig. 5. This property will allow us to relate the descriptor to the segmentation of the image, as we will see in Sect. 4. In addition, it contributes to better discrimination, for example, between concave and convex structures (configurations $f_{p_1q_1}$ and $f_{p_2q_2}$ respectively in Fig. 2(b)), which otherwise would be indistinguishable.



(a) Contours. (b) Face. (c) Vase.

Fig. 5 Rubin's vase, whose contours are shown in (**a**), can have two different interpretations depending on the figure (see (**a**) and (**b**)). A purely contour-based shape descriptor would not be able to differentiate between these two interpretations. The chordiogram, however, is able to make this distinction through the orientation of the normals of its chords

(a) Centaur. (b) Horse.

Fig. 6 Two shapes which are perceptually different and have one identical part—torso. Since the chordiogram captures the parts in the context of the whole shape, the chordiogram distance between the two shapes is larger than the distance between the parts together (see text)

3.2 Gestaltism

The introduced descriptor is a *global* since it takes into account all possible chords—long chords as well as short chords. Thus we capture short-range as well as long-range geometric relations. To give some intuition about the holistic nature of the descriptor, consider the example of a horse and a centaur in Fig. 6, each of which can be thought of being composed of two parts—a head and a torso. Since the chordiogram captures not only the shape of the individual parts but also their relationship, the chordiogram distance between the two shapes:

$$d\left(ch^{horse}, ch^{centaur}\right) = 0.72$$

is larger than the distance between the isolated parts together:

$$d\left(ch^{horse}_{torso} + ch^{horse}_{head}, ch^{centaur}_{torso} + ch^{centaur}_{head}\right) = 0.46$$

In other words, each object part is captured in the context of the whole object, which we interpret is a holistic representation.

3.3 Shape Part Correspondence

A common paradigm in shape matching is to try to quantify the similarity between two shapes by establishing correspondences between points on the shapes. Correspondences between the points serve as an explanation of the match,

while the quality of the match is determined using a matching model (Yoshida and Sakoe 1982; Basri et al. 1998). The chordiogram, as defined in Sect. 2, does not capture any absolute boundary point information as part of the chord features, neither it captures any location relations among chords. As a result, it is not clear whether the chordiogram, as a histogram, can be used to establish correspondences among boundary points of two shapes.

In this section we relate the chordiogram to the graph matching problem, which is a widely used approach to the correspondence problem (Shapiro and Haralick 1979; Gold and Rangarajan 1996; Umeyama 1988), and obtain the following insights:

1. We provide a different interpretation of the chordiogram matching as bipartite matching among chords. We show that the chordiogram can be used to compute the cost of this bipartite matching efficiently without recovering any explicit correspondences.
2. We bound the chordiogram matching from above with the cost of a graph matching among points on the shape. This relates our descriptor to correspondence estimation.
3. Finally, we show how to estimate correspondences between shapes starting from the bipartite matching interpretation of our descriptor.

Next we set up the notation and tools needed for the subsequent analysis.

Graph matching Suppose that the two shapes, whose similarity needs to be assessed, are defined in terms of point sets:

$$P^s = \left\{p_1^s, \ldots, p_n^s\right\} \quad \text{for } s \in \{1, 2\}$$

For simplicity, we assume that both point sets have the same cardinality n. In this case, we can think of a shape as a complete graph, whose nodes are the above point set and the edges are the chords (see Sect. 2).

Chord distances Furthermore, a chord (i, j) from shape described with point set P^s, can be described by the bin into which it falls using a predefined binning scheme b. This can be written as a chordiogram $ch_{ij}^{b,s}$ built only on the point set $\{i, j\}$:

$$ch_{ij}^{b,s} = ch\big(\{i, j\}\big)$$

Using the definition from (2), the above chordiogram can be considered as a binary indicator vector which describes in which bin the chord falls into:

$$\left(ch_{ij}^{b,s}\right)_m = \begin{cases} 1 & \text{if } f_{ij}^s \in bin_b(m) \\ 0 & \text{otherwise} \end{cases}$$

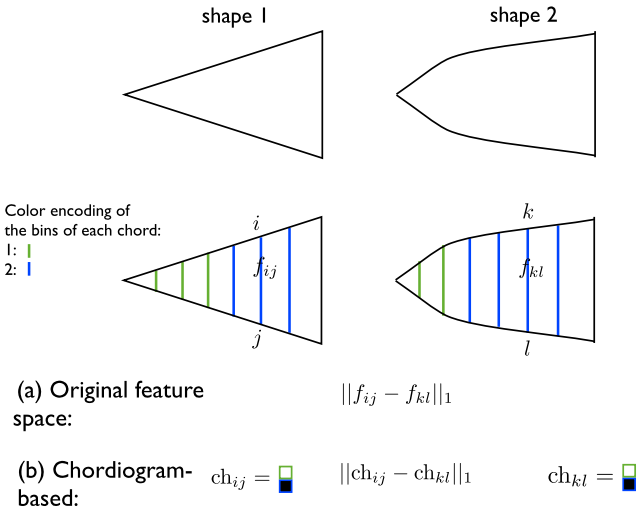


Fig. 7 *Top*: two similar shapes. *Middle*: for each of the two shapes, we show chords of different lengths for fixed orientation and normals. The colors of the chords correspond to the bins they fall in. *Bottom*: (**a**) One can use the feature vectors of the chords to compute a distance between them, or (**b**) a chordiogram for each chord can be defined and the distance between them can be used

Denote further by $ch^{b,s}$ the chordiogram for shape s using binning scheme b and $N = \binom{n}{2} = \|ch^1\| = \|ch^2\|$ the number of chords.

In the following exposition we will use a sequence of nested binning schemes, as defined in Indyk and Thaper (2003). Suppose that Δ is the diameter of the chord set of both shapes, where the diameter is defined in terms of the L_1 distance on the feature vector f_{ij} of a chord (i, j). Further, δ is the smallest L_1 distance among a pair of chords. We assume that each chord has a unique feature vector so that $\delta > 0$. Then the bth binning scheme is defined by partitioning each feature space using a grid of size $\delta 2^b$. The values of b are $\{-1, 0, 1, \ldots, \lceil \log_2(\Delta/\delta) \rceil\}$ such that they define together a fine to coarse hierarchical binning, where at the finest level each bin contains a single chord, while at the coarsest level all chords are contained in a single bin.

Using the above descriptors of a chord, we can define the following three *distances* $W_{ij;kl}$ *between chords* (i, j) *and* (k, l) from two different shapes, which characterize their dissimilarity:

– **Distance in feature space** (see Fig. 7(a)):

$$W_{ij;kl}^{orig} = \left\| f_{ij}^1 - f_{kl}^2 \right\|_1 \tag{4}$$

– **Chordiogram-based distance**: For a particular binning scheme b, one can declare two chords similar if they lie in the same bin, and dissimilar otherwise (see Fig. 7(b)). This can be expressed as follows:

$$W_{ij;kl}^{b} = \left\| ch_{ij}^{b,1} - ch_{kl}^{b,2} \right\|_1 \tag{5}$$

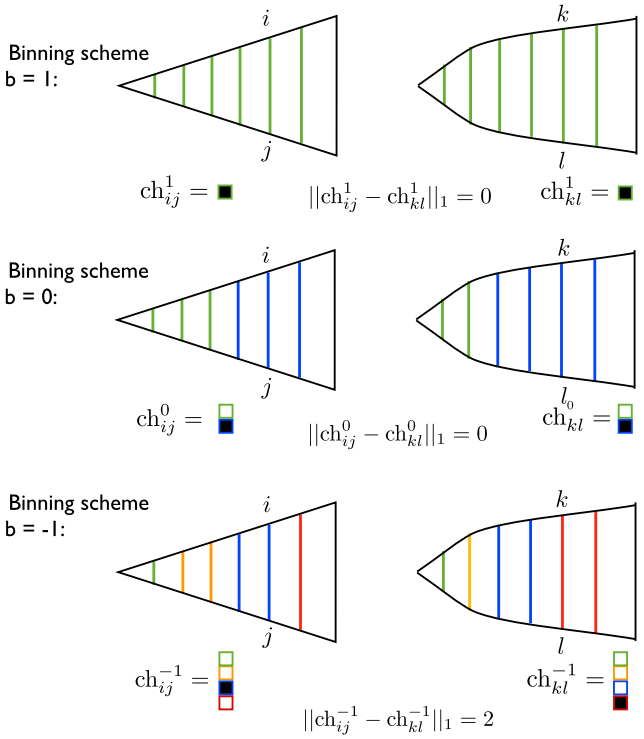


Fig. 8 For the two shapes from Fig. 7, we visualize chords and their bin membership for three different nested binning schemes. Note that for the two coarse binning schemes, the chords ij and kl fall in the same bin, while in the finer binning scheme they are assigned to different bins. Aggregating the distances over all binning schemes gives an approximation of chord distance in the original feature space (see text)

– **Multilevel chordiogram-based distance**: In addition to the above bin comparison distance, one can combine multiple binning schemes into a single distance (see Fig. 8):

$$W^{mbins}_{ij;kl} = \sum_{b=-1}^{B} \alpha_b \left\| ch^{b,1}_{ij} - ch^{b,2}_{kl} \right\|_1 \tag{6}$$

with positive weights α_b.

Graph Matching Formulation We would like to recover one-to-one correspondence between both graphs. For this purpose, we define a correspondence indicator variable

$$x_{ik} = \begin{cases} 1 & \text{if } p^1_i \text{ and } p^2_k \text{ are in correspondence} \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

Then, the *graph matching problem*, which evaluates the structural similarity between the graphs, can be formulated as follows:

$$\text{(GM):} \quad \min_x \sum_{ijkl} W_{ij;kl} x_{ik} x_{jl} \tag{8}$$

$$\text{subject to} \quad \sum_k x_{ik} = 1 \quad \text{for all } i \tag{9}$$

$$\sum_i x_{ik} = 1 \quad \text{for all } k \tag{10}$$

$$x_{ik} \in \{0, 1\} \quad \text{for all } i, k \tag{11}$$

where w can be any positive chord distance, such as the one defined in (4–6). The constraints (9–10) guarantee one-to-one correspondence, while the integral constraints (11) assure that the solution to the problem is a correspondence indicator variable, as defined in (7).

Graph Matching via Chord Matching Following Chekuri et al. (2005), we reformulate the above problem into an equivalent one, in which we introduce a new set of variables $X : X_{ijkl} = x_{ik} x_{jl}$. These variables can be thought of as correspondence variables between chords. Then problem (GM) from (8) can be formulated in terms of the chord correspondence variables. This new formulation has correspondence uniqueness and integrability constraints as GM. In addition, it has consistency constraints which guarantee that the obtained chord correspondences are consistent with a set of point correspondences (see Fig. 9):

$$\text{(GMC):} \quad \min_X \sum_{ijkl} W_{ij;kl} X_{ijkl} \tag{12}$$

$$\text{subject to} \quad \sum_{k,l} X_{ijkl} = 1 \quad \text{for all } i, j \tag{13}$$

$$\sum_{i,j} X_{ijkl} = 1 \quad \text{for all } k, l \tag{14}$$

$$\sum_l X_{ij_1kl} = \sum_l X_{ij_2kl} \quad \text{for all } i, k, j_1, j_2 \tag{15}$$

$$\sum_j X_{ijkl_1} = \sum_l X_{ijkl_2} \quad \text{for all } i, k, l_1, l_2 \tag{16}$$

$$X_{ijkl} \in \{0, 1\} \quad \text{for all } i, k \tag{17}$$

Constraints (13–14) stem directly from the definition of X and the constraints (9–11) on x. Further, the constraints (15–16) assure that corresponding chords agree on a unique correspondence between the points. This constraint can be derived from the following relationship between point and chord correspondences:

$$x_{ik} = x_{ik} \sum_l x_{jl} = \sum_l X_{ijkl} \quad \text{for all } j \tag{18}$$

Relaxation of Graph Matching To solve the integer program (GMC), one needs to resort to relaxations of the problem (see Fig. 10).

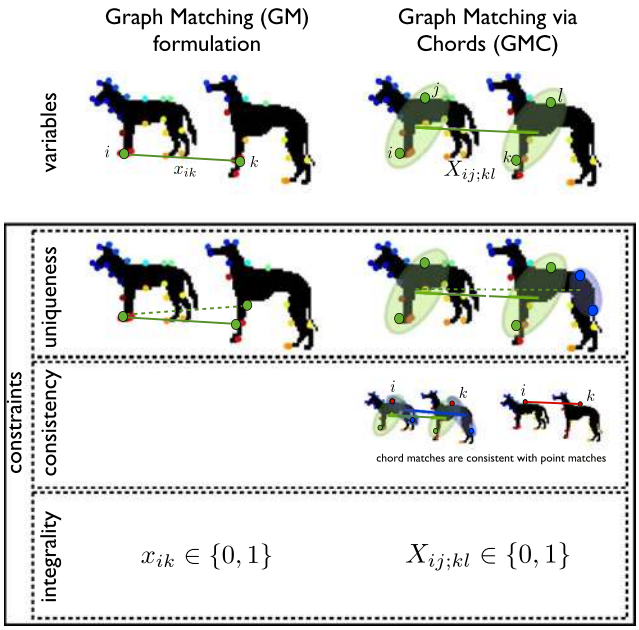


Fig. 9 Equivalent formulations of the graph matching problem. *Left*: The original graph matching formulation (GM) through point correspondence variables. *Right*: An equivalent formulation using chord correspondence variables (GMC)

The first tractable problem can be obtained by relaxing the integral constraints (17) to non-negativity constraints. As a result, one obtains the following exactly solvable linear program (Chekuri et al. 2005), which we call *point matching* (PM) indicating that it aims to recover point correspondences:

$$\text{(PM)}: \quad \min_X \quad W \cdot X \quad \text{subject to} \quad X \in \mathcal{P}_{\text{PM}} \tag{19}$$

where $W \cdot X = \sum_{ijkl} W_{ij;kl} X_{ijkl}$. The above constraint set $\mathcal{P}_{\text{PM}}$ is defined in terms of the following constraints:

$$\mathcal{P}_{\text{PM}} = \left\{ \begin{array}{l} \sum_{k,l} X_{ijkl} = 1 \text{ for all } i,\, j \\ \sum_{i,j} X_{ijkl} = 1 \text{ for all } k,\, l \\ \sum_{l} X_{ij_1kl} = \sum_{l} X_{ij_2kl} \text{ for all } i,k,\, j_1,\, j_2 \\ \sum_{j} X_{ijkl_1} = \sum_{l} X_{ijkl_2} \text{ for all } i,k,\, l_1,\, l_2 \\ X \geq 0 \end{array} \right\}$$

A different relaxation would be to retain the integral constraints (17), but to remove the constraints (15–16) which guarantee that the chord correspondences translate into point correspondences. This corresponds to bipartite matching among the chords of the two shapes, which we will call *chord matching* (CM):

$$\text{(CM)}: \quad \min_X \quad W \cdot X \quad \text{subject to} \quad X \in \mathcal{P}_{\text{CM}} \tag{20}$$

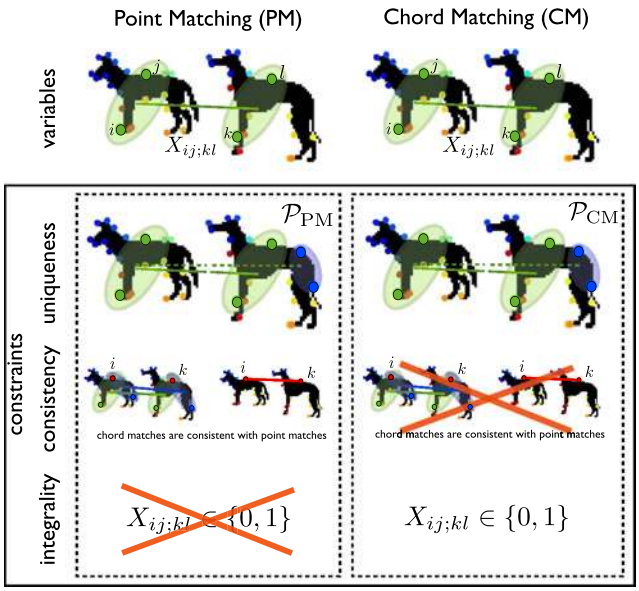


Fig. 10 Relaxation of GMC. *Left*: Point Matching (PM) is obtained by relaxing the integrability constraint. *Right*: Chord Matching (CM) is obtained by relaxing the consistency constraints

with constraints

$$\mathcal{P}_{\text{CM}} = \left\{ \begin{array}{l} \sum_{k,l} X_{ijkl} = 1 \text{ for all } i,\, j \\ \sum_{i,j} X_{ijkl} = 1 \text{ for all } k,\, l \\ X_{ijkl} \in \{0,1\} \text{ for all } i,\, j,\, k,\, l \end{array} \right\}$$

The latter program does not guarantee that the resulting chord correspondence can be directly translated to point correspondences. However, it is an integer program, which can be solved exactly using Max-Flow estimation algorithms.

Relations Between Graph Matching and Chordiogram Distance Using the above definition of graph matching and its relaxations, one can show that the chordiogram distance is closely related to the correspondence problem between two shapes. First, we show the relationship between the chordiogram and bipartite matching among chords:

Theorem 1 *Consider the chord matching problem* (*CM*) (*see* (20)) *with the multilevel chordiogram-based distance* (*see* (6)):

$$\min_X W^{mbins} \cdot X \quad \textit{subject to} \quad X \in \mathcal{P}_{\text{CM}}$$

The solution of this problem can be characterized as follows:

– *The minimum can be analytically computed using the chordiogram distance*:

$$\min_{X \in \mathcal{P}_{\text{CM}}} W^{mbins} \cdot X = \sum_{b=-1}^{B} \alpha_b \left\| ch^{b,1} - ch^{b,2} \right\|_1$$

for weights $\alpha_b = 2^b$.

– *All the minimizers can be described in terms of the chordiograms of the individual shapes with the following set*:

$$\mathcal{P}^*_{\mathrm{CM}} = \Bigg\{ X \in \mathcal{P}_{\mathrm{CM}} \Bigg| \sum_{\substack{(i,j)\in bin_b(m) \\ (k,l)\in bin_b(m)}} X_{ijkl} = \min\left\{ch^{b,1}_m, ch^{b,2}_m\right\} \quad \textit{for all bins } m \textit{ and schemes } b \Bigg\} \tag{21}$$

Furthermore, we can relate the chordiogram distance to point matching between shapes:

Theorem 2 *Suppose that* $X^*_{cm,orig}$ *is a minimizer of the chord matching problem* (*see* (20)) *using data terms* W^{orig} *based on the distance in the original feature space* (*see* (4)):

$$X^*_{cm,orig} \in \arg\min_X W^{orig} \cdot X \quad \textit{subject to} \quad X \in \mathcal{P}_{\mathrm{CM}}$$

Further, $X^*_{pm,mbins}$ *is a minimizer of the point matching problem* (*see* (19)) *using data terms* W^{mbins} *based on the multilevel chordiogram-based distance* (*see* (6)):

$$X^*_{pm,mbins} \in \arg\min_X W^{mbins} \cdot X \quad \textit{subject to} \quad X \in \mathcal{P}_{\mathrm{PM}}$$

Then, the following relationship holds:

$$\alpha W^{orig} \cdot X^*_{cm,orig} \leq \sum_{b=-1}^{B} \alpha_b \left\| ch^{b,1} - ch^{b,2} \right\|_1 \leq W^{mbins} \cdot X^*_{pm,mbins}$$

for a positive constant α.

The proof of both theorems is given in Appendix A. There are several insights we gain from the above theorems which relate our shape representation to matching points on the two shapes.

1. As shown in Theorem 1, the chordiogram distance is a minimizer of a bipartite matching among chords for a specific form of the chord distances. Thus, it quantifies the best possible correspondences among chords on two shapes without explicitly giving those correspondences. In addition, the chordiogram distance does not require any inference and thus it is more efficient.
2. As shown in the first inequality of Theorem 2, the chordiogram over several binning schemes is an upper bound of the bipartite matching for which the similarities are defined in the original chord feature space. This shows that by choosing several binning schemes for the chordiogram, we can obtain an approximation to the original distance in the chord feature space.
3. As shown in the second inequality of Theorem 2, the distance based on our shape descriptor is a lower bound of the linear programming approximation for establishing correspondences among points on two shapes.

Correspondence Recovery The above theorem is based on the fact that we can think of the chordiogram distance as a different relaxation of the original graph matching formulation. This allows for recovery of point correspondences—if we have $X \in \mathcal{P}_{\mathrm{PM}}$, then we can use (18) for an arbitrary j to estimate point correspondences. To obtain such an X, however, we will not solve (PM) directly, but rather use the solution for (CM) obtained from the chordiogram distance. More precisely, we will try to find $X \in \mathcal{P}_{\mathrm{PM}}$ closest to any minimizer of (CM):

$$\min_X \left\{ \left\| X - X^*_{cm} \right\|_2 \middle| X \in \mathcal{P}_{\mathrm{PM}}, X^*_{cm} \in \mathcal{P}^*_{\mathrm{CM}} \right\} \tag{22}$$

Note the above problem is an integer quadratic program, and thus NP-hard. To obtain an approximate solution, one can relax the above problem by replacing the integral constraints with nonnegativity constraints in the definition of $\mathcal{P}^*_{\mathrm{CM}}$:

$$\mathcal{P}^{**}_{\mathrm{CM}} = \left\{ \begin{array}{l} \sum_{k,l} X_{ijkl} = 1 \text{ for all } i, j \\ \sum_{i,j} X_{ijkl} = 1 \text{ for all } k, l \\ X_{ijkl} \geq 0 \text{ for all } i, j, k, l \\ \sum_{(i,j),(k,l)\in bin(m)} X_{ijkl} = \min\{ch^1_m, ch^2_m\} \end{array} \right\}$$

The above polytope $\mathcal{P}^{**}_{\mathrm{CM}}$ is a convex set and if we replace $\mathcal{P}^*_{\mathrm{CM}}$ with $\mathcal{P}^{**}_{\mathrm{CM}}$ in problem (22), then we obtain a convex program. The correspondence recovery procedure is summarized in Algorithm 1.

Algorithm 1 Correspondence estimation from chordiograms

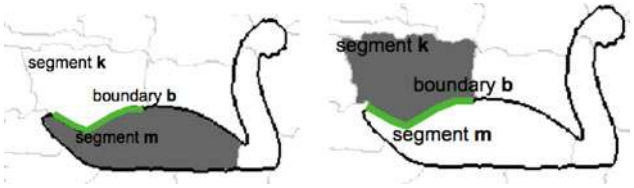
Require: Chordiograms ch^1, ch^2 of two shapes.

1: Define $\mathcal{P}^{**}_{\mathrm{CM}}$ using ch^1 and ch^2.
2: Solve program (22) and obtain minimizer $X^* \in \mathcal{P}_{\mathrm{PM}}$.
3: Recover correspondence indicator variables x from X^* using (18).
4: Obtain discrete indicators

$$\hat{x}_{ij} = \begin{cases} 1 & \text{iff } j = \arg\max_{j_1} \{x_{ij_1}\} \\ 0 & \text{otherwise} \end{cases}$$

Examples We show results of the correspondence recovery algorithm on selected pairs of shapes from MPEG 7 dataset (Latecki et al. 2000). From each shape, defined by the outline of the shape mask, we sample uniformly 30 points, which are to be put in correspondence. The chordiogram is

Fig. 11 Examples of recovered correspondence on pairs of shapes. Points, colored in the same color, are in correspondence



(a) Segment m is foreground (b) Segment k is foreground

Fig. 12 There are two cases in which boundary b can be an object boundary

computed using only the sample points. For the optimization problem in step 2 of the algorithm, we use the CVX optimization package (Grant and Boyd 2010). Results are shown in Fig. 11. As we can see, correct correspondences are recovered for most of the points for articulated as well as rigid objects. The main difficulties are in cases of strong articulation (see lizard in row 1, column 2, where animal's torso and tail are articulated differently), or lack of matching points (see elephant in row 1, column 3, where in the left object two legs are visible, while in the right object three legs are visible).

4 Boundary Structure Segmentation and Detection Model

The introduced chordiogram as a holistic and global representation can potentially suffer from all the irrelevant structure present in images, such as interior contours and background clutter. This is a major challenge in applying global object representations in realistic images, which include multiple objects and rich background structure.

To address this problem, we propose a chordiogram-based object detection model called **Bo**undary **S**tructure **S**egmentation (BoSS) model, which solves simultaneously for object *segmentation* and *detection*. First, we show how to relate region segmentation to chordiogram matching in Sects. 4.1 and 4.2. The bottom-up perceptual principles are described in Sect. 4.3. The BoSS model and inference are explained in full detail in Sects. 4.4 and 4.5.

4.1 Chordiogram Parameterization

In order to relate the chordiogram to image segmentation, we parameterize it in terms of variables that track selected segments and segment boundaries.

Oversegmentation As a starting point for our method, we assume that we have an over-segmentation of the input image. The property we require from the segments is that they do not cross object boundaries (most of the time). In this way, every object in the image is representable as a set of such segments and the object boundary as a set of segment boundary.

Segment Parametrization For each segment k obtained via the oversegmentation we introduce a segment indicator variable $s_k \in \{-1, 1\}$:

$$s_k = \begin{cases} 1 & \text{segment } k \text{ is foreground} \\ -1 & \text{otherwise} \end{cases} \tag{23}$$

We use N to denote the number of segments.

Segment Boundary Parameterization We denote by $\mathcal{B}$ the set of all boundary segments between pairs of neighboring segments, where the number of such boundary segments is $M = |\mathcal{B}|$. Note that a contour b is a boundary because exactly one of its neighboring segments k and m is foreground and the other is background (see Fig. 12). To differentiate between those two cases, for each contour b and its two neighboring segments k and m we include in $\mathcal{B}$ two boundaries: b^m and b^k. The first denotes the case when m is foreground and k is background; the second denotes the opposite case.

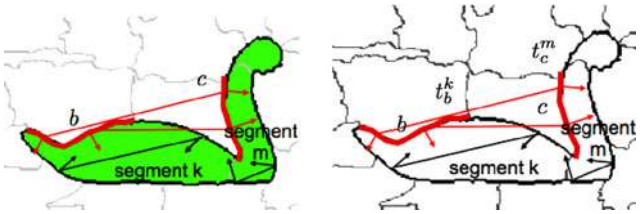
We introduce boundary indicator variables which indicate whether a segment boundary is an object boundary. This variable not only captures the state of the boundary but tracks which segment configuration causes this state. More precisely, for each boundary $b^k \in \mathcal{B}$ we introduce a boundary indicator variable $t_b^k \in \{0, 1\}$:

$$t_b^k = \begin{cases} 1 & \text{segment } k \text{ is foreground and} \\ & \text{segment } m \text{ is background} \\ 0 & \text{otherwise} \end{cases} \tag{24}$$

As a result, there are two variables associated with each boundary. If a segment boundary designates an object boundary, then exactly one of the variables has value 1. Otherwise both are 0. The relationship between the values of the boundary and segment variables is summarized in Table 1. This relationship can be expressed in terms of two constraints:

$$t_b^k - t_b^m = \frac{1}{2}(s_k - s_m) \tag{25}$$

$$t_b^k t_b^m = 0 \tag{26}$$



(a) Segmented object (b) Unknown object

Fig. 13 The chordiogram of an object can be decomposed in terms of chordiograms which relate pair of boundaries, as shown on the left. If the object is not segmented, the boundaries can be selected via the boundary indicator variables

Table 1 We present the relationship between boundary and segment indicator variables

Boundary		Segments	
t_b^k	t_b^m	s_k	s_m
1	0	1	−1
0	1	−1	1
0	0	1	1
0	0	−1	−1

Chordiogram Additivity To parameterize the chordiogram using the above variables, it will prove useful to provide an equivalent definition to (2). For a given segmented object, the chords connecting points on two boundaries b and c, caused by segments k and m being foreground respectively, can be described by a chordiogram $ch_{bc}^{km} \in \mathbb{R}^K$, $b^k, c^m \in \mathcal{B}$ (see Fig. 13(a)):

$$\left(ch_{bc}^{km}\right)_l = \#\left\{(p,q) \mid f_{pq} \in bin(l),\, p \in b^k, q \in c^m\right\} \tag{27}$$

The above quantity can be considered as boundary-pair chordiogram. Note that the boundary-pair chordiogram is a subset of the overall chordiogram. Then (2) can be expressed as a sum of all boundary-pair chordiograms for all pairs of boundaries. This has the following linear form:

$$ch = \sum_{b^k, c^m \in \mathcal{B}} ch_{bc}^{km} \tag{28}$$

The above decomposition will be referred to as *chordiogram additivity*—the descriptor can be expressed in an additive form in terms of relations between object parts. Note that this is not a contradiction to the holistic nature of the descriptor since the additive components are *not* object parts, but configurations between parts.

Chordiogram Parameterization If we do not have a segmented object, we can select the object boundaries using the indicator variables (see Fig. 13(b)) and express the resulting image chordiogram as follows:

$$ch(t) = \sum_{b^k, c^m \in \mathcal{B}} ch_{bc}^{km} t_b^k t_c^m \tag{29}$$

The value of the lth bin can be expressed as a quadratic function:

$$ch(t)_l = \sum_{b^k, c^m \in \mathcal{B}} \left(ch_{bc}^{km}\right)_l t_b^k t_c^m = t^T Q_l t \tag{30}$$

for a matrix Q_l which contains the values of the boundary-pair chordiogram: $(Q_l)_{bk;cm} = (ch_{bc}^{km})_l$.

Note that in the above parameterization one needs to indicate not only the boundary but also its relationship to the neighboring segments. This information is already contained in the chordiogram, since as defined in Sect. 2, each chord captures the object interior via the orientation of the normals.

4.2 Shape Matching

After we have parameterized the chordiogram in terms of the boundary indicators (see (29)), we chose to compare it with the model ch^{model} using L_1 distance:

$$match(t, m) = \left\| ch^{model} - ch(t) \right\|_1 \tag{31}$$

The above shape matching cost evaluates the shape similarity between a model and a particular selection of segment boundaries. This motivates us to formulate the problem of shape matching as minimization of the above cost while taking into account the relation between boundaries and segments, as expressed in constraints in (25):

$$(SM): \quad \min_{t,s} \quad \left\| ch^{model} - ch(t) \right\|_1 \tag{32}$$

$$\text{s.t.} \quad t_b^k - t_b^m = \frac{1}{2}(s_k - s_m) \quad \text{for all } b^m, b^k \in \mathcal{B}$$

$$t_b^k t_b^m = 0, \quad t \in \{0,1\}^{2M}, \quad s \in \{-1,1\}^N.$$

Solving the above optimization problem produces:

- **Figure/ground segmentation:** The optimal values of the boundary and segment indicators encode the object interior and boundary.
- **Shape-based detection cost:** The minimum of the objective function quantifies the quality of the match based on shape similarity.

Solving the optimization problem for several object models, and selecting the best match, accomplishes joint shape-based detection and segmentation.

4.3 Perceptual Grouping

Our model can express grouping principles relating regions as well as boundaries.

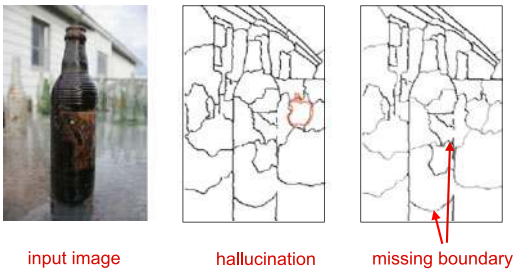


Fig. 14 *Left*: input image. *Middle*: if we use all segment boundaries, than non-existing objects can be easily hallucinated. *Right*: if we rely on an edge/contour detection, then we can miss correct boundaries, which the segmentation can potentially hallucinate

4.3.1 Region Grouping Principles

While matching the input image to a model, we would like to ensure that the resulting figure represents a *perceptually salient segmentation*, i.e. the resulting figure should be a coherent region or set of regions distinct from the background. This property can be expressed using the segment indicator variables, as introduced in Sect. 4.1, and a Min-Cut-type smoothness criterion. If we denote by $w_{e,g}$ the similarity between the appearance of segments k and m, then we can encourage region coherence by the standard graph cut score:

$$group_r(s) = -s^T W s = -1^T W 1 + 2 \sum_{\substack{k \in figure \\ m \in ground}} w_{k,m} \qquad (33)$$

for $s \in \{-1, 1\}^N$.

4.3.2 Boundary Grouping Principles

In many cases an edge/contour detector cannot detect all object boundaries since there is no evidence in the image (see Fig. 14, right). However, if we use segmentation we can hallucinate object boundaries and recover the missing ones (see Fig. 14, left). This comes with the danger that one can also hallucinate non-existing objects in the maze of segment boundaries.

To address this issue we propose to use all segment boundaries, while at the same time incurring a cost if we choose hallucinated ones. In this way we will be able to complete the bottom of the bottle in Fig. 14 by paying a small cost, while we will never detect the apple since the cost for hallucinating all boundaries will be prohibitively large.

For a boundary segment b, we denote by c_b the percent of the pixels of b *not covered by image edges* extracted using thresholded Probability of Boundary edge detector (Martin et al. 2004). Then the boundary cost is defined as

$$group_b(t) = c^T t = \sum_{b^k \in \mathcal{B}} c_b t_b^k \qquad (34)$$

for $t_b^k \in \{0, 1\}^{2M}$.

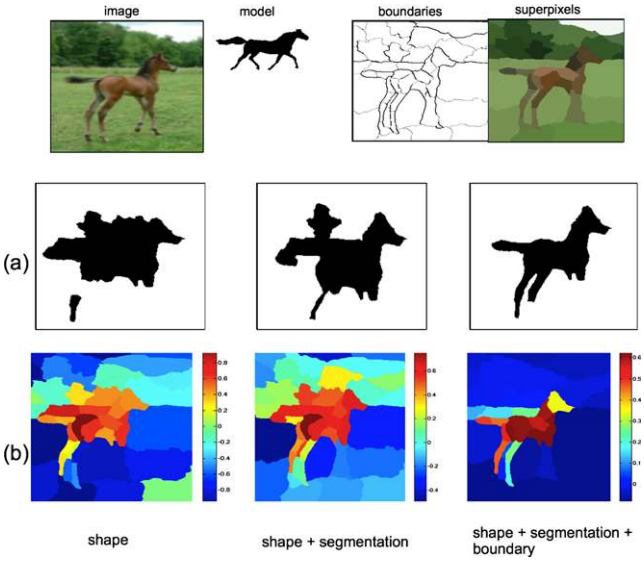


Fig. 15 For an input image and model, as shown in the *first row*, our algorithm computes an object segmentation displayed in (**a**) row. We present three solutions by using only the matching term from (31) in *first column*; the matching term together with the superpixel segmentation prior (see (33)) in *second column*; and the whole cost function consisting of the matching, segmentation and the boundary term in third column (see (35)). (**b**) We also show for the three cost combinations the relaxed values of the segmentation variable s, as explained in Sect. 4.5

4.4 BoSS Model

The BoSS model combines the costs from the previous sections. It solves for a shape match using cost (31) from Sect. 4.2, while at the same time applies grouping principles as formulated in costs (33) and (34) from Sect. 4.3:

$$\begin{aligned} \min_{t,s} \quad & match(t, m) + \delta group_r(s) + \gamma group_b(t) \\ \text{s.t.} \quad & t_b^k - t_b^m = \frac{1}{2}(s_k - s_m) \quad \text{for } b^k, b^m \in \mathcal{B} \\ & t_b^k t_b^m = 0, \; t \in \{0, 1\}^{2M}, \quad s \in \{-1, 1\}^N \end{aligned} \qquad (35)$$

where δ and γ are weights of the different terms. The difference from the problem (SM) in (32) lies in the addition of two grouping terms.

Term Contributions We examine the contribution of each term of the model on one concrete example presented in Fig. 15. The shown results were obtained using the optimization described in Sect. 4.5. By using only the matching term we are able to localize the object and obtain a rough mask, which however extends the back of the horse and ignores its legs (first column). The inclusion of the superpixel grouping bias helps to remove some of the erroneous superpixels above the object which have a different color than the horse (second column). Finally, if we add the boundary term, it serves as a sparsity regularization on t and re-

sults in a tighter segmentation (third column). Thus, the incorrect superpixels above the horse get removed, since they contain hallucinated boundaries not supported by edge response. Additionally, it recovers some of the legs, since they exhibit strong edge response along their boundary.

4.5 Inference

Both the Shape Matching problem formulated as an integer quadratic program (SM) in (32) and the BoSS program in (35) are in general NP-hard. This is not surprising since it is the problem of selecting from a set of exponentially many segments such that the resulting region has a desired shape and perceptual properties. To compute an approximate solution, we apply the Semi-definite Programming (SDP) relaxation (Goemans and Williamson 1995; Boyd and Vandenberghe 2004). Since the latter program is a superset of the former, we present an optimization scheme for the BoSS program only.

First, we re-write the objective as a linear function and a set of quadratic constraints. We introduce for the lth bin a variable β_l, which denotes the difference of the model and image chordiogram at this bin. Then the objective of the BoSS program can be expressed in terms of β and a quadratic constraint for each bin:

$$(BoSS): \quad \min_{t,s,\beta} \quad 1^T\beta - \delta s^T W s + \gamma c^T t \tag{36}$$

$$\text{s.t.} \quad t^T Q_l t - ch_l^{model} \leq \beta_l \tag{37}$$

$$ch_l^{model} - t^T Q_l t \leq \beta_l \tag{38}$$

$$t_b^k - t_b^m = \frac{1}{2}(s_k - s_m) \tag{39}$$

$$t_b^k t_b^m = 0 \tag{40}$$

$$t \in \{0,1\}^{2M}, \quad s \in \{-1,1\}^N \tag{41}$$

for all pairs of segment boundaries $b^k, b^m \in \mathcal{B}$. In the first two constraints (37) and (38) we use the chordiogram parameterization as defined in (30).

To apply the SDP relaxation, we introduce variables T and S, which bring both the quadratic terms (37) and (38) into linear form: $T = tt^T$; and the quadratic terms in (36) into linear form: $S = ss^T$. This allows us to state the relaxation as follows:

$$(\text{BoSS}_{sdp}): \quad \min_{t,s,\beta} \quad 1^T\beta - \delta tr\big(W^T S\big) + \gamma c^T t \tag{42}$$

$$\text{s.t.} \quad tr\big(Q_l^T T\big) - ch_l^{model} \leq \beta_l \tag{43}$$

$$ch_l^{model} - tr\big(Q_l^T T\big) \leq \beta_l \tag{44}$$

$$t_b^k - t_b^m = \frac{1}{2}(s_k - s_m) \tag{45}$$

$$T_{bk;bm} = 0 \tag{46}$$

$$t_b^k = T_{bk;bk} \quad \text{for } b^k \in \mathcal{B} \tag{47}$$

$$diag(S) = 1_n \tag{48}$$

$$\begin{pmatrix} T & t \\ t^T & 1 \end{pmatrix} \succeq 0 \tag{49}$$

$$\begin{pmatrix} S & s \\ s^T & 1 \end{pmatrix} \succeq 0 \tag{50}$$

The above problem was obtained from problem (36) in two steps. First, we relax the constraints $T = tt^T$ to $T \succeq tt^T$ and $S = ss^T$ to $S \succeq ss^T$ respectively, which by Schur complement are equivalent to (49) and (50) (Boyd and Vandenberghe 2004). Second, we weakly enforce the domain of the variables from the constraint (41). The $-1/1$-integer constraint on s is expressed as diagonal equality constraint on the relaxed S (see (48)), which can be interpreted as bounding the squared value of the elements of s by 1. The $0/1$-integer constraint (see (47)) is enforced by requiring that the diagonal and the first row of T have the same value. Since $T = tt^T$, this has the meaning that the elements of t are equal to their squared values, which is only true if they are 0 or 1. Finally, the boundary-region constraints, one of which is quadratic, naturally translate to linear constraints.

The above problem is a linear program with inequality constraints in the cone of positive semi-definite matrices. As such, it is convex and can be solved exactly with any standard optimization package which supports such problems.

Discretization Discrete solutions are obtained by thresholding s. Since s has N elements, there are at most N different discretizations, all of which are ranked using their distance to the model. If a threshold results in a set of several disconnected regions, we consider all possible subsets of this set. For each of the discretized segmentations, the matching cost function is evaluated. The algorithm outputs the top 5 ranked non-overlapping masks. Note that we are capable of detecting several instances of an object class since they result in several disconnected regions which are evaluated independently.

BoSS Algorithm The BoSS algorithms starts with an input image and a set of models. It solves the above optimization problem for each image-model pair at each scale. The best matching model gives the object segmentation as well as a detection cost—the chordiogram distance of the model to the obtained segmentation. The full details are presented in Algorithm 2.

5 Related Work

In the context of the proposed method, we review in this section relevant work.

Algorithm 2 BoSS algorithm.

Input: Model masks $m_1, \ldots, m_k$; image segmentation parametrized by t and s; scales $h_1, \ldots, h_p$.
for $i = 1 \ldots k$ **do**
  for $j = 1 \ldots p$ **do**
    $m_{i,j} \leftarrow$ rescale m_i to scale h_j:
    Compute $ch_{i,j}^{mod}$ of $m_{i,j}$ at scale h_j using (2).
    Solve relaxed BoSS problem (42) using $ch_{i,j}^{model}$.
    Discretize to obtain segmentation $s_{i,j}$.
    Compute $ch_{i,j}$ from $s_{i,j}$ at scale h_j using (2).
    Detection cost: $d_{i,j} \leftarrow \left\| \frac{ch_{i,j}}{\|ch_{i,j}\|} - \frac{ch_{i,j}^{model}}{\|ch_{i,j}^{model}\|} \right\|_1$.
  end for
end for
$(i^*, j^*) \leftarrow \arg\min_{i,j} d_{i,j}$.
Output: Segmentation s_{i^*,j^*} and detection cost d_{i^*,j^*}.

Holistic Representations Some of the first attempts to define holistic representations are based on global transforms of the input object shape. Examples include Fourier coefficients of a contour distance function (Zhang and Lu 2003) and Zernicke moments applied on the object mask (Zhang and Lu 2003). Another class of holistic shape representations was initiated by the development of the Medial Axis Transform by Blum, which is defined as the set of centers of maximally inscribed circles in a closed shape (Blum 1973). This set can be thought of as a skeleton of the shape, which is computed globally, and reveals geometrical as well as topological shape properties. Depending how those properties are captured, the medial axis has led to the development of Shocks, Shock graphs (Kimia et al. 1995; Siddiqi et al. 1999; Sebastian et al. 2004; Trinh and Kimia 2011) as well as M-reps in medical imaging (Pizer et al. 1999). To deal with the instability of the medial axis to small boundary protrusions a more robust transform based on the Poisson equation has been proposed (Gorelick and Basri 2009).

More recently, Zhu et al. (2008) proposed a holistic shape matching approach which selects relevant object contours while matching Shape Contexts (Belongie et al. 2002). In a follow-up work, the above matching has been combined with discriminative learning to leverage salient object contours (Srinivasan et al. 2010).

The presented BoSS model does not try to establish a point correspondence between the model and the object shape. In many cases, however, an explicit correspondence estimation between the two shapes lies in the core of a shape matching technique. Spectral graph matching in conjunction with geometric features of edgels and pairs of edgels has been used by Leordeanu et al. (2007). A parametric statistical framework, which models the shape deformation of the point set is the Active Shape Model (Cootes 1995).

Simpler models which do not capture all pairwise relationships between shape parts depart from the idea of holism but allow for tractable inference. This is commonly done by treating a shape as a linearly ordered point set instead of unorganized point set as the chordiogram assumes. Lu et al. (2009) explore particle filtering to search for a set of object contours. Felzenszwalb and Schwartz (2007) propose a hierarchical representation by decomposing a contour into a tree of subcontours and using dynamic programming to perform matching. A globally optimal shape matching and segmentation based on the Minimum Ratio Cycle algorithm was introduced by Schoenemann and Cremers (2007). Dynamic programming has been also applied in a multi-stage framework to search for a chain of object contours (Ravishankar et al. 2008). A similar approach to shape-based recognition is to search for a chain of image contours which best matches to a model in a contour network extracted from the image (Ferrari et al. 2006).

The chordiogram uses edgels as atomic shape parts. A different approach is to use contour segments as parts. For example, a descriptor of groups of adjacent contour segments was introduced in conjunction with an SVM classifier for the purpose of recognition (Ferrari et al. 2008). Boundary fragments scored using a classifier and geometrically related to an object center have been explored as well (Opelt et al. 2006; Shotton et al. 2005). The simple fragment configuration model allows for efficient inference using a voting scheme.

Statistical Representations The presented descriptor in this work captures relationships among edgels in a statistical fashion. Similarly, geometric hashing has been used to describe purely geometric properties (Lamdan et al. 1990) as well as topological properties at a global scale (Carlsson 1999). A widely used descriptor, called Shape Context (Belongie et al. 2002) captures a semi-local distribution of edges. Its descriptive power has been extended to more deformed and articulated shapes (Ling and Jacobs 2007).

Histograms of geometric properties of sets of points have been used to match 3D models (Osada et al. 2002). These histograms can be interpreted as distributions of shape functions, where each function represents a property of a small set of points.

Recognition and Segmentation Close interplay between segmentation and recognition has been studied by Yu and Shi (2003) who guide segmentation using part detections and do not use global shape descriptors. Segment shape descriptors based on the Poisson equation have been used for detection and segmentation (Gorelick and Basri 2009). Leibe et al. (2008) combine recognition and segmentation in a probabilistic framework. Recently, Gu et al. (2009) use global shape features on image segments. However, segmentation is a preprocessing step, decoupled from the subsequent matching.

Object dependent segmentation has been addressed in prior work (Borenstein et al. 2004; Levin and Weiss 2006). Both methods combine bottom-up segmentation with top-down matching, using templates of object parts as a way to match shape. An explicit reasoning about figure/ground organization has been proposed by Ren et al. (2005) who use shapemes for local shape matching. Although these approaches have segmentation and boundary priors they employ only local shape descriptors.

6 Experiments

In this section we evaluate both the chordiogram on its own as well as BoSS on several established benchmarks. The parameter of the model and its implementation details are described in Sect. 6.1. In Sect. 6.2 we evaluate the performance of the chordiogram on the task of recognition of presegmented objects. In Sect. 6.3 we present recognition and segmentation results of our chordiogram-based method BoSS on two datasets of real cluttered images.

6.1 Implementation Details

We use the chordiogram on presegmented objects with parameters $b_l = 4, b_r = 8, b_n = 8$, resulting in a 2048-dimensional descriptor. The number of bins was selected such that on the one hand it results in a discriminative descriptor and on the other hand the dimensionality of the descriptor is not too large. When we use the chordiogram in the BoSS model, we use $b_l = 3, b_r = 4, b_n = 4$, resulting in a 196-dimensional descriptor. A lower dimensional descriptor is used for computational reasons—in the BoSS inference in (36) we introduce a variable for each chordiogram bin and thus a larger descriptor would result in a harder optimization.

To obtain superpixels we oversegment the image using NCuts (Cour et al. 2005) with $n = 45$ segments. The number of segments was chosen such that the resulting segmentation covers most of the object boundaries. The grouping cues used to define the affinity matrix W^{pixels} are color and intervening contours (Yu and Shi 2003) based on Probability of Boundary edge detector (Martin et al. 2004).

To define the segmentation term (33) in our model we can use any affinity matrix. We choose to use the same grouping cues as for segmentation above. For each pair of superpixels k and m we average the pixel affinities to obtain an affinity matrix over the superpixels: $W_{km}^{superpixels} = \frac{1}{a_k a_m} \sum_{p \in k, q \in m} \widehat{W}_{pq}^{pixels}$, where a_k and a_m are the size of the superpixels k and m respectively. Above, $\widehat{W}^{pixels}$ is obtained from the top n eigenvectors E of W^{pixels}: $\widehat{W}^{pixels} = E \Lambda E^T \approx W^{pixels}$, where Λ are the corresponding eigenvalues. This low-rank approximation represents a smoothed version of the original matrix and reduces the noise in the original affinities. Finally, the weights of the term in (35) were chosen to be $\delta = 0.01$ and $\gamma = 0.6$ on five images from ETHZ dataset and held constant for all experiments.

For the optimization we use SeDuMi (Sturm 1999) which is based on the Primal-Dual Interior Point Method. To compute the number of variables in the SDP, one can assume that each superpixel has at most C neighboring superpixels. Hence we obtain $M = Cn$ boundary variables. Thus, if we denote by D the dimensionality of the chordiogram, then the total variable number in the relaxed problem is bounded by $n^2 + C^2 n^2 + D \in O(n^2)$. In our experiments, we have $n = 45$ and the value of C is less than 5 which results in less than 200 boundary segment variables. The empirical running time of the optimization is around 30–45 seconds on a 3.50 GHz processor. Note that for other applications the number of needed superpixels n to segment an object might be larger than 45 which will increase the running time of the algorithm.

6.2 Chordiogram Evaluation

To evaluate the performance of the chordiogram for the task of object recognition, we perform experiments on the MPEG-7 CE-Shape 1 part B dataset (Latecki et al. 2000). This dataset is used for evaluation of shape-based classification and retrieval. It consists of 1400 binary object masks representing 70 different classes, each class having 20 examples. The recognition rate reported for this dataset is the Bullseye score: each shape is matched to all shapes and the percentage of the 20 possible correct matches among the top 40 matches is recorded; the score is the average percentage over all shapes.

To compute a distance between two binary object masks using the chordiogram, we first scale-normalize the masks. Since the chordiogram is not rotation invariant, we rotate each mask b_r times using b_r rotations of angle $\{0, \frac{2\pi}{b_r}, \ldots, (b_r - 1)\frac{2\pi}{b_r}\}$ around the object mask center of mass, compute the chordiogram and normalize it by setting its L_1 norm to 1. Thus, we obtain b_r descriptors $\{ch_i^{(1)}, \ldots, ch_i^{(b_r)}\}$ for the ith object. The distance between two objects i and j is defined as the smallest distance in L_1 sense among all rotated chordiograms:

$$d(i,j) = \min_{\theta_i, \theta_j} \left\{ \left\| ch_i^{(\theta_i)} - ch_j^{(\theta_j)} \right\|_1 | \theta_i, \theta_j \in \{1, \ldots, b_r\} \right\}$$

The bullseye score of the chordiogram in comparison to other shape matching approaches is presented in Table 2. Using the above setup, we achieve a score of 80.85%. We outperform most of the approaches with exception of Shape Trees by Felzenszwalb and Schwartz (2007), Hierarchical Procrustes by Mcneill and Vijayakumar (2006) and Inner Distance Shape Context by Ling and Jacobs (2007). The

Table 2 Bullseye score of the chordiogram and other shape matching methods on the MPEG dataset

Method	Bulleye score
Mokhtarian et al. (1997)	75.44%
Latecki and Lakamper (2000)	76.45%
Belongie et al. (2002)	76.51%
Sebastian et al. (2003)	78.16%
Tu and Yuille (2004)	80.03%
chordiogram	**80.85%**
Ling and Jacobs (2007)	85.40%
Mcneill and Vijayakumar (2006)	86.35%
Felzenszwalb and Schwartz (2007)	87.70%

main reason is that the latter methods are based on metrics which are computed along the shape contour, while our approach uses Euclidean distances to capture shape. As a result these methods deal better with non-rigid deformations and articulations than the chordiogram.

However, the use of rigid metrics to capture relationships between contours allows for a parameterization of the chordiogram in terms of image segmentation and thus deals with image clutter, as we will see in the next section. An additional advantage of the chordiogram is that its distance is simply a L_1 norm computation, while the above approaches require an inference of some sort.

6.3 BoSS Evaluation

In this section we turn to the evaluation of our complete model BoSS on two datasets consisting of real images.

6.3.1 ETHZ Shape Dataset

The ETHZ Shape Dataset (Ferrari et al. 2010) consists of 255 images of 5 different object classes—Applelogos (40 images), Bottles (48 images), Mugs (48 images), Giraffes (87 images) and Swans (32 images). The dataset is designed in such a way that the selected object classes do not have distinctive appearance and the only representation, which can be used to detect object class instances, is their shape. As a result, this dataset has been widely used for evaluation of shape-based detection methods. Some of the challenges in this dataset are highly cluttered images—in the background as well as internal spurious contours; wide variation of object scale; multiple instances of an object in the same image. However, the depicted objects are fully included in the images and are not occluded. Also, the used objects vary in shape but are not articulated (detection of the giraffe's legs is not part of the task).

We apply the BoSS model using hand-drawn object outlines as shape models, one model per class. These models were supplied with the dataset. We use 7 different scales, such that the scale of each model, defined as the diameter of its bounding box, ranges from 100 to 300 pixels. We use non-maximum suppression—for every two hypotheses, whose bounding boxes overlap by more than 50%, we retain the one with the higher score and discard the other one.

Detection Results In order to compute precision, recall and detection rates, traditionally two detection criteria were established. According to the 20% overlap detection criterion we declare a detection if the intersection of the hypothesis and ground truth bounding boxes overlap more than 20% with the each of them. A stricter criterion is the Pascal criterion which declares a detection if the intersection of the hypothesis and groundtruth bounding boxes is at least 50% of their union.

The results of BoSS under both criteria are presented and compared to other methods in Table 3 and Fig. 16. Under the 20% overlap criterion we achieve state-of-the-art performance of 91.2%/93.0% detection rate at 0.3/0.4 false positives per image (fppi). Under the stricter Pascal criterion we achieve 86.1%/88.6% detection rate at 0.3/0.4 fppi without any learning. With learning, which we call reranking (see below), we achieve state-of-the-art detection rates of 94.3%/96.0%.

For Applelogos, Swans and Bottles, the results for both criteria are almost the same, which shows that we achieve good localization of the objects. For Giraffes and Mugs results are slightly lower due to imperfect segmentation (some segments leak into the background or missed parts)—the detections which are correct under the weaker 20% overlap criterion, are not counted as correct under the Pascal criterion.

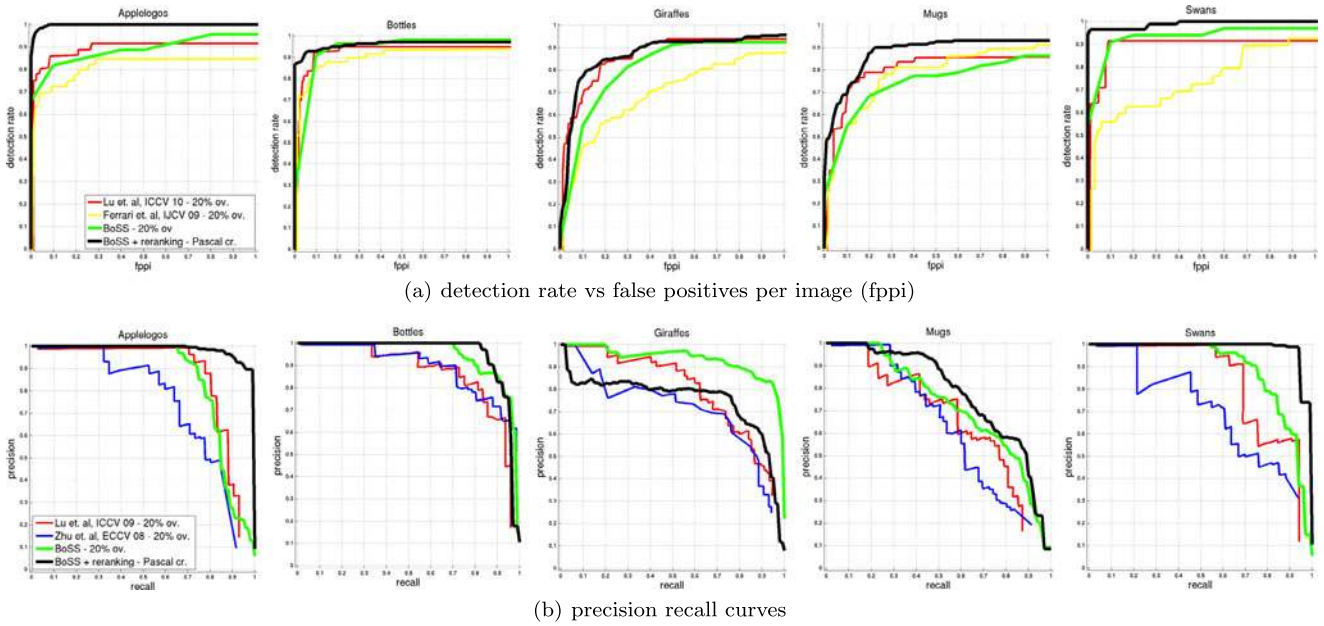
In Fig. 17 we show examples of typical detections in the datasets described above. Our method is capable of detecting objects of various scales in highly cluttered images, even when the object is small and most of the image contours and segments are not part of the object. Note that the translation invariance of the chordiogram allows us to find the object without having to search exhaustively for location. Additionally, the segmentation gives us a pixel-level object localization which is much more precise compared to the bounding-box localization used by other methods.

Our approach is robust against local shape variations as well as global transformations. As shown in Fig. 18(a), using a single mug model BoSS obtains detections of objects whose shape deviates from the model in various ways: aspect ration, global shape, shape of parts, etc. In addition, it tolerates global transformations as minor rotations and foreshortening (see Fig. 18(b)).

The major sources for incorrect detections are accidental alignments with background contours, which we call hallucinations, and partially incorrect boundaries (see Fig. 19). The former cause shows the limitation of shape—one can

Table 3 Detection rates at 0.3/0.4 false positives per image, using the 20% overlap and Pascal criteria. († use only hand-drawn models; * use strongly labeled training data with bounding boxes, while we use hand-drawn models and weakly labeled data (no bounding boxes) in the reranking; $^{\sharp}$ considers in the experiments only at most one object per image and does not detect multiple objects per image; $^{\circ}$ uses a slightly weaker detection criterion than Pascal)

	Algorithm	Apple logos	Bottles	Giraffes	Mugs	Swans	Average
20% over.	BoSS†	86.4%/88.6%	**96.4**%/**98.2**%	**97.8**%/**97.8**%	**84.8**%/86.4%	93.4% /93.4%	**91.2**%/93.0%
	(Lu et al. 2009)$^{\dagger\sharp}$	**92.5**%/**92.5**%	95.8%/95.8%	86.2%/92.0%	83.3%/92.0%	**93.8**%/**93.8**%	90.3%/**93.2**%
	(Fritz and Schiele 2008)*	–/89.9%	–/76.8%	–/90.5%	–/82.7%	–/84.0%	–/84.8%
	(Ferrari et al. 2010)†	84.1%/86.4%	90.9%/92.7%	65.6%/70.3%	80.3%/83.4%	90.9%/93.9%	82.4%/85.3%
Pascal crit.	BoSS†	86.4%/88.6%	96.4%/96.4%	81.3%/86.8%	72.7%/77.3%	93.9%/93.9%	86.1%/88.6%
	BoSS$^{*}_{rerank}$	100%/100%	96.3%/97.1%	86.1% /91.7%	90.1%/91.5%	98.8%/100%	94.3%/96.0%
	(Maji and Malik 2009)*	95.0%/95.0%	92.9%/96.4%	89.6%/89.6%	93.6%/96.7%	88.2%/88.2%	91.9%/93.2%
	(Srinivasan et al. 2010)*	95.0%/95.0%	100%/100%	87.2% /89.6%	93.6%/93.6%	100%/100%	95.2%/95.6%
	(Gu et al. 2009)*	90.6%/–	94.8%/–	79.8%/–	83.2%/–	86.8%/–	87.1%/–
	(Ravishankar et al. 2008)$^{\dagger\circ}$	95.5%/97.7%	90.9%/92.7%	91.2%/93.4%	93.7%/95.3%	93.9%/96.9%	93.0%/95.2%



(a) detection rate vs false positives per image (fppi)

(b) precision recall curves

Fig. 16 Results on ETHZ Shape dataset. Results using BoSS are shown using 20% overlap as well as after reranking using the stricter Pascal criterion. Both consistently outperform other approaches, evaluated using the weaker 20% overlap criterion

sometimes find a constellation of contours which resemble the model outline. Some of those cases can be ruled out by using perceptual grouping principle. However, in other cases the lack of an appearance model is limiting.

Reranking In order to compare with approaches on the ETHZ Shape Dataset which use supervision, we use weakly labeled data to rerank the detections obtained from BoSS. We use only the labels of the training images to train a classifier but not the bounding boxes. This classifier can be used to rerank new hypotheses obtained from BoSS.

More precisely, we use half of the dataset as training and the other half as test (we use 5 random splits). We use BoSS to mine for positive and negative examples. The top detection in a training image using a model which represents the label of that image is considered a positive example; all other detections are negative examples. The chordiograms of these examples are used as features to train one-vs-all SVM (Joachims 1999) for each class. During test time, each detection is scored using the output of the SVM corresponding to the model used to obtain this detection.

Note that this is a different setup of supervision which requires less labeling—while we need one hand-drawn model

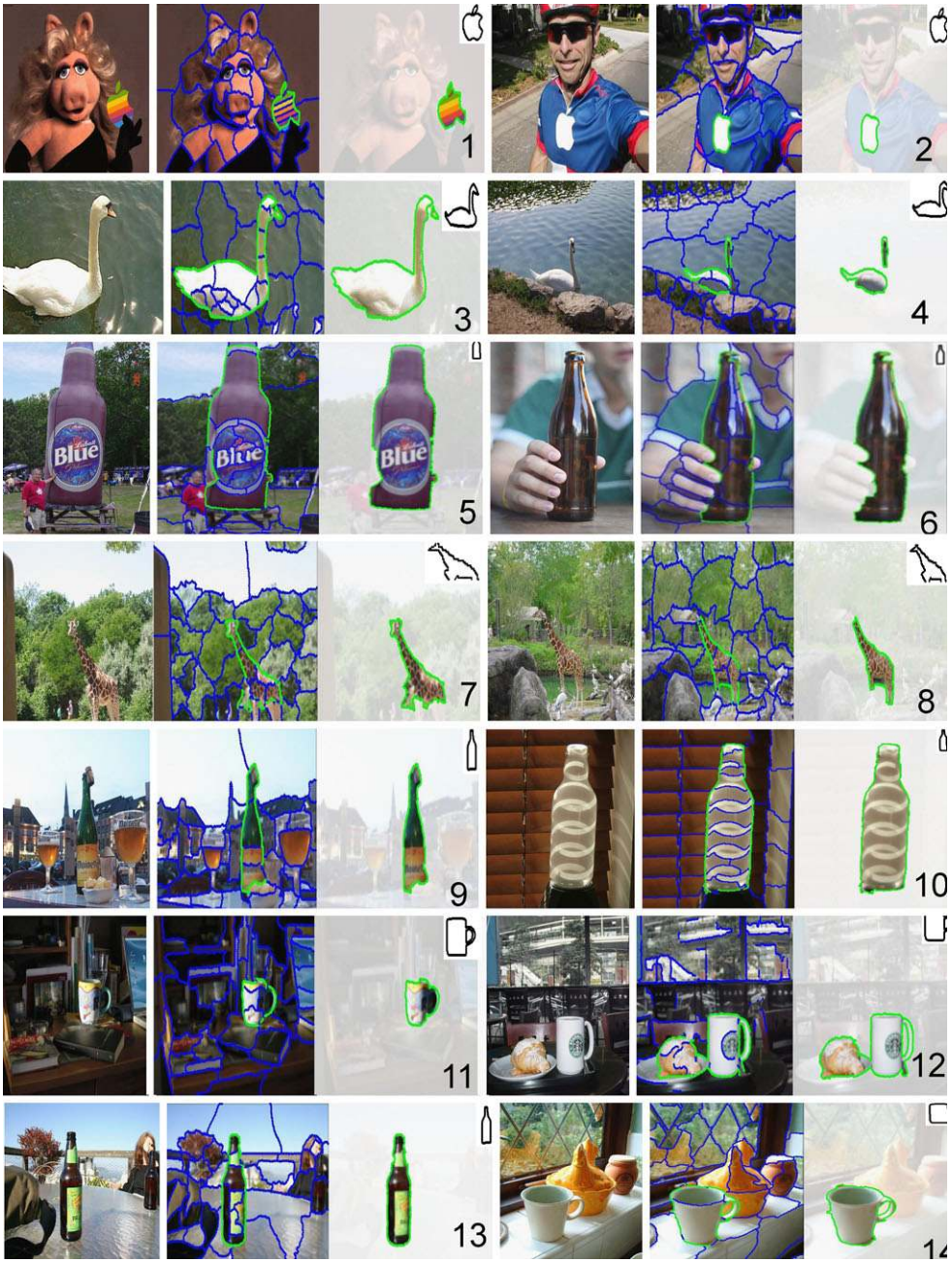


Fig. 17 Example detection on ETHZ Shape Dataset. For each example, we show on the left side the input image and in the middle and the right side the segmentation for the best matching model. In particular, we show in the middle the selected segment boundaries in *green*. On the right the selected object mask and the best matching model are displayed (Color figure online)

per class to obtain detections via BoSS, we do not use the bounding boxes but only the labels of the training images to score them. We argue that the effort to obtain a model is constant while segmenting images by hand is much more time consuming. Although the hand-drawn models are the driving force for object detection, the weakly labeled data is used to learn a discriminative chordiogram-based model which takes into account the shape deformations present in the dataset and not captured in the hand-drawn model. The majority of the approaches in Table 3, which use learning, use bounding-boxes as labeling but no hand-drawn models.

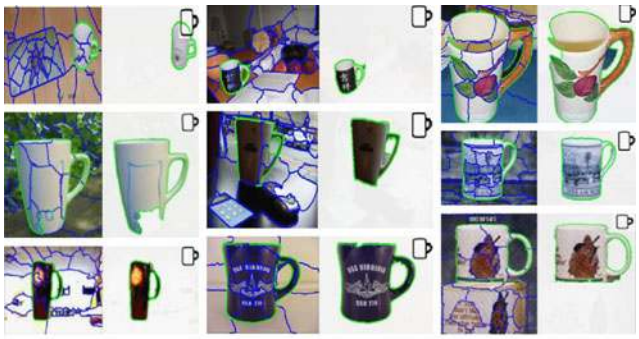
The results are shown in Table 3. The weak supervision leads to 94.3%/96.0% detection rate under Pascal criterion, which is an improvement of approx. 5% over BoSS. It is attributed to the discriminatively learned weights of the chordiogram's bins. This corresponds to discriminatively learning object shape variations and builds on the power of BoSS to deal with clutter.

Segmentation In addition to the detection results, we evaluate the quality of the detected object boundaries and object masks. For evaluation of the former we follow the test settings of Ferrari et al. (2010).[2] We report recall and precision

[2] A detected boundary point is considered a true positive if it lies within t pixels of a ground truth boundary point, where t is set to 4% of the

Table 4 Precision/recall of the detected object boundaries and pixel classification error of the detected object masks for the ETHZ Shape Dataset. We present results using only the shape matching cost (see (32)) as well as the full cost—BoSS—which consists of shape matching as well as perceptual grouping terms (see (35))

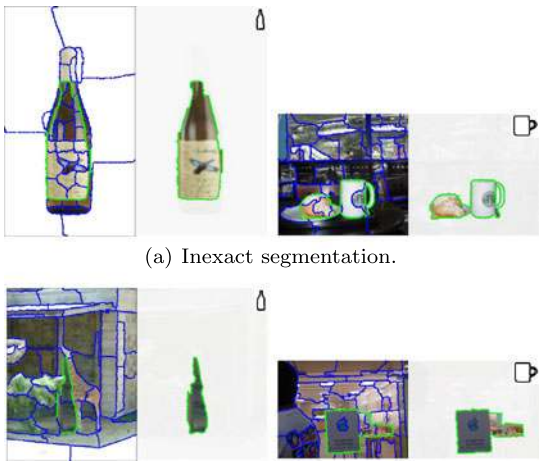
	Boundary precision/recall			Pixel error	
	SM	BoSS	Ferrari et al. (2010)	SM	BoSS
Applelogos	91.9%/97.1%	91.8%/97.5%	91.6%/93.9%	2.0%	1.6%
Bottles	89.4%/91.1%	90.3%/92.5%	83.4%/84.5%	2.8%	2.7%
Giraffes	75.4%/81.3%	76.8%/82.4%	68.5%/77.3%	6.2%	5.9%
Mugs	77.7%/89.1%	86.5%/90.5%	84.4%/77.6%	5.5%	3.6%
Swans	81.0%/86.8%	85.8%/87.6%	77.7%/77.2%	6.7%	4.9%



(a) Shape variations.

(b) Perspective and rotation distortions.

Fig. 18 Example detections on the ETHZ Shape Dataset which show the robustness of the chordiogram and BoSS to shape variations. For each example, we show on the left side the selected segment boundaries in green, and on the right the selected object mask. We use the same model to obtain those detections. Note, however, that the detected mugs may have different aspect ratio, largely varying shape of the body (rectangle or cone), and shape and size of the handle



(a) Inexact segmentation.

(b) Hallucinations.

Fig. 19 Examples of missdetections

of the detected boundaries in correctly detected images in Table 4. We achieve higher recall at higher precision compared to Ferrari et al. (2010).[3] This is mainly result of the fact that BoSS attempts to recover a closed contour and in this way the complete object boundary. These statistics show that the combination of shape matching and figure/ground organization results in precise boundaries (>87% for all classes except Giraffes). The slightly lower results for Giraffes is due to the legs which are not fully captured in the provided class models. We also provide object mask evaluation as percentage of the image pixels classified incorrectly by the detected mask (see Table 4). For all classes we achieve less than 6% error, and especially classes with small shape variation such as Bottles and Applelogos we have precise masks (< 3% error).

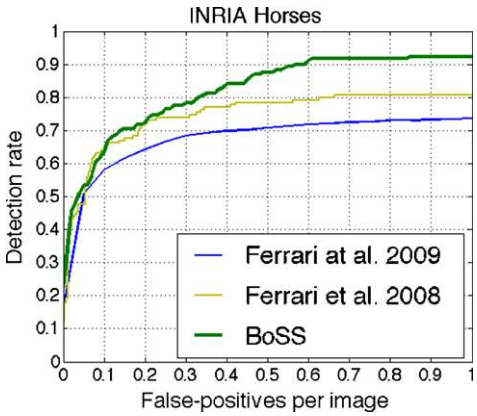
To analyze the contribution of the perceptual terms, we apply BoSS on the ETHZ Shape Dataset without the perceptual terms (see program SM in (32)) and compare the resulting segmentations and object boundaries to the one obtained using the full BoSS model. The results are compared in Table 4. Although SM performs pretty comparable to the full model, its boundary and pixel precisions are slightly below the ones obtain via BoSS—on average SM has 4.6% pixel error, while BoSS reduces it to 3.7%. Perceptual grouping tends to correct shape-based segmentation in cases where the shape match is not very good, but the bottom-up grouping is based on a strong signal.

6.3.2 *INRIA Horses Dataset*

The INRIA Horses Dataset has 340 images, half of which contain horses and the other half have background objects. This dataset presents challenges not only in terms of clutter and scale variation, but also in articulation, since the horses

diagonal of the ground truth mask. Based on this definition, one computes recall and precision.

[3]It should be noted that we use hand-drawn models while Ferrari et al. (2010) uses the models learned from the labeled data.



(a)

Method	Det. rate
BoSS	92.4%
Maji and Malik (2009)	85.3%
Ferrari et al. (2008)	80.8%
Ferrari et al. (2010)	73.8%

(b)

Fig. 20 (**a**) Detection rate vs false positives per image (fppi) for our and other approaches on INRIA Horse dataset. (**b**) Detection rates at 1 fppi

are in different poses. Also, some of the objects are partially occluded.

We use 6 horse models representing different poses for the INRIA Horse Dataset (see Fig. 21). In these experiments we used 10 scales such that the scale of the model, defined as the diameter of its bounding box, ranges from 55 to 450 pixels. Similarly to the previous dataset, we use non-maximum suppression—for every two hypotheses, whose bounding boxes overlap by more than 50%, we retain the one with the higher score and discard the other one.

Detection Results On INRIA Horses dataset, we achieve state of the art detection rate of 92.4% at 1.0 fppi (see Fig. 20). Examples of detections of horses in different poses, scales and in cluttered images are shown in Fig. 21.

6.3.3 BoSS vs. Multiple Segmentation-Based Approaches

Most of the applications of segmentation in computer vision serve as coarsening of the input space. In the case of general object recognition, one often computes texture-based descriptors for each segment (Shotton et al. 2009), groups of segments (Malisiewicz and Efros 2008) or bag-of word descriptors of segments (Russell et al. 2006). In such approaches, a pre-segmentation is considered useful if a segment or groups of segments overlap sufficiently well with

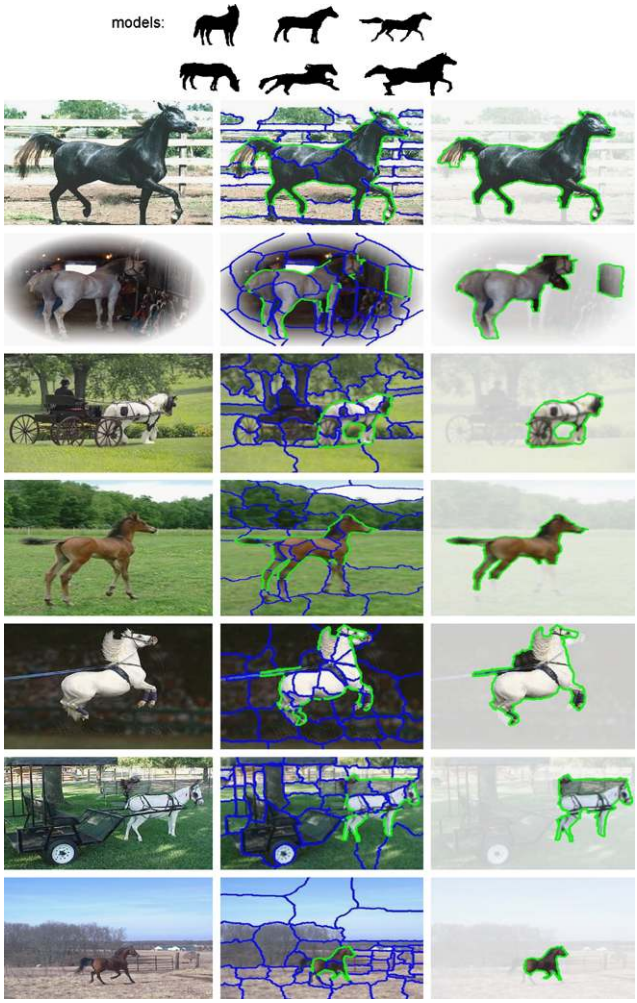
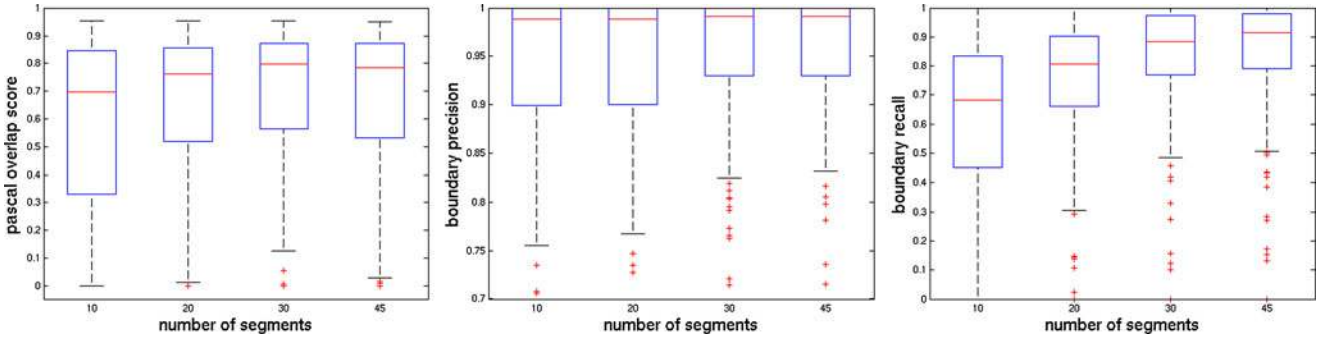


Fig. 21 Examples of detections on the INRIA Horses Dataset. For each image we show the selected superpixel boundaries on the left and the detected object segmentation on the right. *Bottom right*: 6 models used in the experiments

the object of interest. Therefore, using small groups of segments or multiple segmentations is often enough to capture an object.

In the case of shape-based object detection, it is important to capture the correct object boundaries in a segment selection. Therefore, even if the overlap of a segment or a group of segments with an object of interest is large, these segments may not capture the shape of the object at all.

To see the importance of being able to select all possible groups of segments, we compare the BoSS model to chordiogram-based detection over segments computed via multiple segmentations. More precisely, we use three different segmentations per image with 10, 20, and 30 segments. For each segmentation, we compute groups of connected segments of up to 5 segments. This results on average in 5337 groups of segments per image. We consider each group of segments as a hypothesis for an object segmentation. To evaluate how likely a hypothesis is an object of a particu-



(a) Pascal overlap error. (b) Boundary precision. (c) Boundary recall.

Fig. 22 We present three different measures for the quality of the segmentation. For each measure, we use all images from the ETHZ Shape Dataset and pre-segmentations with 10, 20, 30, and 45 segments. We display for each measure and pre-segmentation, the median in *red*, the 25% and 75% quantile as *blue boxes*, and the range of the values as *black lines* (Color figure online)

Table 5 Detection rates of Group of Segments (GoS) and BoSS at 0.3 and 0.4 fppi for the five classes of the ETHZ Shape Dataset

	Applelogos	Bottles	Giraffes
GoS	38.6%/43.2%	85.5%/87.3%	46.2%/52.8%
BoSS	86.4%/88.6%	96.4%/96.4%	81.3%/86.8%

	Mugs	Swans	Average
GoS	50%/50%	78.8%/78.8%	59.1%/62.4%
BoSS	72.7%/77.3%	93.9%/93.9%	86.1%/88.6%

lar class, we compute the chordiogram distance between the hypothesis and the object model.

The detection rates for the five classes of the ETHZ Shape Dataset are presented in Table 5. We can see that using only groups of segments, the detection rate drops, the main reason being that a selection of up to 5 segments is not sufficient to capture all object boundaries. This is apparently drastic for Applelogos and Mugs, which are large and occupy most of the image. Of course, one can increase the size of the groups, however their number groups exponentially with their size. Therefore, it would become less feasible to compute the chordiogram for all groups of larger sizes.

6.3.4 Number of Input Superpixels

As justified above, being able to select any possible combination of segments as a figure segmentation is of paramount importance when it comes to shape-based object detection. Using more segments could potentially result in a better object segmentation since one should be able to model finer details of an object shape. However, having more segments comes at a higher cost since the optimization problem in Sect. 4.5 will be carried over a larger number of variables.

To evaluate the importance of the number of segments in the final object segmentation, we run BoSS with a pre-segmentation on the ETHZ Shape Dataset with 10, 20, 30 and 45 superpixels. For every level of input pre-segmentation, we evaluate the obtained object segmentation using the ground truth model and scale for each image. We use the Pascal overlap score. To better evaluate the quality of the boundaries of the segmentation, we also compute boundary precision/recall, as used in the evaluation of the segmentation in Sect. 6.3.

The results for those three measures over the whole dataset for the four setups are summarized in Fig. 22. We can see that the Pascal overlap scores improve with increasing number of segments. Moreover, the values become closer to the median, which indicates that with increasing number of segments the quality of the segmentation improves for more images. Similar behavior can be observed for boundary precision/recall. The biggest improvement is in the recall—as we have more segments, we obtain larger portions of the object boundaries better. Also, we can see that there is a clear improvement from 10 to 20 and from 20 to 30 segments. However, the observed improvement beyond 30 segments is small. This means that using 30 segments for this dataset is sufficient to capture most of the objects. Hence, we use 45 segments in the preceding experiments.

7 Conclusion

In this paper we introduce a novel shape descriptor, called chordiogram, and a shape-based segmentation and recognition approach, called Boundary Structure Segmentation (BoSS).

The chordiogram is a global descriptor, which is motivated by the idea of holism introduced by the Gestalt school of perception. As such, the descriptor capture the object

shape as a whole. Moreover, the chordiogram can be parameterized in terms of image segments. As such it can be related to perceptual grouping principles in the image, such as consistency in region appearance and small hallucinations of object boundaries. This allows us to combine the chordiogram with perceptual grouping in the unified approach (BoSS). We perform simultaneous shape matching and segmentation and as a result, enable holistic shape-based object detection in cluttered scenes.

The approach is analyzed both theoretically and empirically. We show that the chordiogram can be viewed as an approximation of graph matching techniques for shape matching. Furthermore, we show very good performance of the descriptor for the task of shape retrieval. We evaluated BoSS for both object recognition and precise object localization on two datasets of realistic images and achieves state of the art results on both tasks.

Acknowledgements This work has been partially supported by the National Science Foundation grants NSF-OIA-1028009 and NSF-DGE-0966142 and by the Army Research Laboratory accomplished under the Robotics CTA Cooperative Agreement Number W911NF-10-2-0016. Ben Taskar was also supported by the DARPA Computer Science Study Group Award. Parts of the results presented in this paper have previously appeared in Toshev et al. (2010).

Appendix A: Proofs of Theorems 1 and 2

Theorem 1 *Consider the chord matching problem* (*CM*) (*see* (20)) *with the multilevel chordiogram-based distance* (*see* (6)):

$$\min_X \quad W^{mbins} \cdot X \quad \textit{subject to} \quad X \in \mathcal{P}_{\mathrm{CM}}$$

The solution of this problem can be characterized as follows:

- *The minimum can be analytically computed using the chordiogram distance*:

$$\min_{X \in \mathcal{P}_{\mathrm{CM}}} W^{mbins} \cdot X = \sum_{b=-1}^{B} \alpha_b \left\| ch^{b,1} - ch^{b,2} \right\|_1$$

  for weights $\alpha_b = 2^b$.
- *All the minimizers can be described in terms of the chordiograms of the individual shapes with the following set*:

$$\mathcal{P}^*_{\mathrm{CM}} = \left\{ X \in \mathcal{P}_{\mathrm{CM}} \,\middle|\, \sum_{\substack{(i,j) \in bin_b(m) \\ (k,l) \in bin_b(m)}} X_{ijkl} = \min\left\{ ch^{b,1}_m, ch^{b,2}_m \right\} \right.$$
$$\left. \textit{for all bins } m \textit{ and schemes } b \right\} \tag{51}$$

Proof First we will show that the chordiogram matching lower bounds the problem (CM) for all $X \in \mathcal{P}_{\mathrm{CM}}$. In a second step, we will show that for $X^* \in \mathcal{P}^*_{\mathrm{CM}}$ the bound turns into an equality.

Lower Bound for (CM) Suppose that $X \in \mathcal{P}_{\mathrm{CM}}$. Then, one can show that

$$\sum_{b=-1}^{B} \alpha_b \left\| ch^{b,1} - ch^{b,2} \right\| \tag{52}$$

$$= \sum_{b=-1}^{B} \alpha_b \left\| \sum_{i,j} ch^{b,1}_{ij} - \sum_{k,l} ch^{b,2}_{kl} \right\|_1$$

(def. of chordiogram)

$$= \sum_{b=-1}^{B} \alpha_b \left\| \sum_{i,j} \left(\sum_{k,l} X_{ijkl} \right) ch^{b,1}_{ij} - \sum_{k,l} \left(\sum_{i,j} X_{ijkl} \right) ch^{b,2}_{kl} \right\|_1 \tag{53}$$

$$= \sum_{b=-1}^{B} \alpha_b \left\| \sum_{i,j,k,l} \left(ch^{b,1}_{ij} - ch^{b,2}_{kl} \right) X_{ijkl} \right\|_1$$

$$\leq \sum_{i,j,k,l} \sum_{b=-1}^{B} \alpha_b \left\| ch^{b,1}_{ij} - ch^{b,2}_{kl} \right\|_1 X_{ijkl} \tag{54}$$

$$= \sum_{i,j,k,l} W^{mbins}_{ij;kl} X_{ijkl} \quad \text{(by (6))}$$

$$= W^{mbins} \cdot X$$

Line (53) is derived using the correspondence uniqueness, while line (54) uses the positivity of the variables.

Minimizers for (CM) As a second step, we will show that for each $X^* \in \mathcal{P}^*_{\mathrm{CM}}$ the above inequality turns into an equality.

Consider for a moment a concrete bin m using finest binning scheme $b = -1$. We can use the bin indices of the chords to define a matching between them. More precisely, we put chords in correspondence if they lie in the same bin. After this procedure there will remain chords which are not in any correspondence. The correspondence assignment for such chords is deferred for a coarser binning scheme.

Now we turn to the description of the correspondence assignment for a particular binning scheme b. For the sake of brevity we will skip the binning scheme index b. Suppose that X gives a chord mapping for which d_m denotes the number of chords from shape 1 from bin m mapped to chords from shape 2 which are also in bin m; a_m chords from shape 1 from bin m mapped to chords not in bin m;

and c_m chords from shape 1 not in bin m mapped to chords from shape 2 in bin m. From the definition of d_m we have

$$d_m = \sum_{\substack{(i,j)\in bin(m)\\(k,l)\in bin(m)}} X_{ijkl} \tag{55}$$

Since ch^1_m counts all the chords lying in bin m from shape 1, which can be mapped either to chords in bin m or not in bin m from the other shape, then $ch^1_m = a_m + d_m$. Similarly, $ch^2_m = d_m + c_m$. Therefore, $|ch^1_m - ch^2_m| = |a_m - c_m|$.

Also, since the $\sum_{ijkl} |(ch^1_{ij})_m - (ch^2_{kl})_m|_1 X_{ijkl} = a_m + c_m$. Thus, we can express the gap in the above inequality derivation for a single binning scheme as:

$$W^b \cdot X - \left\|ch^1 - ch^2\right\|_1 = \sum_m \left(a_m + c_m - |a_m - c_m|\right)$$

X is a minimizer for (CM) exactly when the above gap equals zero, i.e. $a_m + c_m - |a_m - c_m| = 0$ for all m. This is equivalent to $\min\{a_m, c_m\} = 0$, which holds iff $d_m = \min\{ch^1_m, ch^2_m\}$. The latter identity together with (55) gives the desired characterization.

Now, suppose that $d^b_m = \min\{ch^{1,b}_m, ch^{2,b}_m\}$ holds for all binning schemes from the definition of multiple-bin distance between chords from (6). This means that all gaps disappear:

$$W^b \cdot X - \left\|ch^{b,1} - ch^{b,2}\right\|_1 = 0 \quad \text{for all } b \in \{-1, 0, \ldots, B\}$$

with $B = \lceil \log(\Delta/\delta) \rceil$ as defined in Sect. 3.3. Combining the above inequalities together with weights α_b gives the equality relationship in the theorem. □

Theorem 2 *Suppose that $X^*_{cm,orig}$ is a minimizer of the chord matching problem (see (20)) using data terms W^{orig} based on the distance in the original feature space (see (4)):*

$$X^*_{cm,orig} \in \arg\min_X W^{orig} \cdot X \quad \textit{subject to} \quad X \in \mathcal{P}_{\mathrm{CM}}$$

Further, $X^*_{pm,mbins}$ is a minimizer of the point matching problem (see (19)) using data terms W^{mbins} based on the multilevel chordiogram-based distance (see (6)):

$$X^*_{pm,mbins} \in \arg\min_X W^{mbins} \cdot X \quad \textit{subject to} \quad X \in \mathcal{P}_{\mathrm{PM}}$$

Then, the following relationship holds:

$$\begin{aligned}\alpha W^{orig} \cdot X^*_{cm,orig} &\le \sum_{b=-1}^{B} \alpha_b \left\|ch^{b,1} - ch^{b,2}\right\|_1 \\ &\le W^{mbins} \cdot X^*_{pm,mbins}\end{aligned}$$

for a positive constant α.

Proof We show both inequalities separately.

First Inequality The left inequality is result of a direct application of Lemma 1 from Indyk and Thaper (2003). Note that the point sets, which are considered in Indyk and Thaper (2003), correspond to the chords sets in our setting. Then there is a constant α such that the chordiogram distance is lower bounded by the weighted bipartite matching among the chords, where the weights are defined in terms of the L_1 distance in the chord feature space:

$$\alpha\left(W^{orig} \cdot X^*_{cm,orig}\right) \le \sum_{b=-1}^{B} \alpha_b \left\|ch^{b,1} - ch^{b,2}\right\|_1$$

Second Inequality From the previous theorem, we have that the middle term is the minimum of the (CM) problem with data terms W^{mbins}. It is known that the minimum of the (CM) problem interpreted as a bipartite matching is smaller that the minimum of the (PM) problem interpreted as linear programming relaxation of the graph matching. This gives us the second inequality. □

References

Basri, R., Costa, L., Geiger, D., & Jacobs, D. (1998). Determining the similarity of deformable shapes. *Vision Research*, *38*(15–16), 2365–2385.

Belongie, S., Malik, J., & Puzicha, J. (2002). Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(4), 509–522.

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review*, *94*(2), 115–147.

Binford, T. O. (1971). Visual perception by computer. In *IEEE conference on systems and control*.

Blum, H. (1973). Biological shape and visual science. *Journal of Theoretical Biology*, *38*(2), 205–287.

Borenstein, E., Sharon, E., & Ullman, S. (2004). Combining top-down and bottom-up segmentation. In *IEEE computer society conference on computer vision and pattern recognition*.

Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge: Cambridge University Press.

Carlsson, S. (1999). Order structure, correspondence and shape based categories. In *International workshop on shape, contour and grouping*.

Chekuri, C., Khanna, S., Naor, J., & Zosin, L. (2005). A linear programming formulation and approximation algorithms for the metric labeling problem. *SIAM Journal on Discrete Mathematics*, *18*(3), 608–625.

Cootes, T. (1995). Active shape models-their training and application. *Computer Vision and Image Understanding*, *61*(1), 38–59.

Cour, T., Benezit, F., & Shi, J. (2005). Spectral segmentation with multiscale graph decomposition. In *IEEE computer society conference on computer vision and pattern recognition*.

Felzenszwalb, P., & Schwartz, J. (2007). Hierarchical matching of deformable shapes. In *IEEE computer society conference on computer vision and pattern recognition*.

Ferrari, V., Tuytelaars, T., & Gool, L. V. (2006). Object detection by contour segment networks. In *European conference on computer vision*.

Ferrari, V., Jurie, F., & Schmid, C. (2007). Accurate object detection with deformable shape models learnt from images. In *IEEE computer society conference on computer vision and pattern recognition*.

Ferrari, V., Fevrier, L., Jurie, F., & Schmid, C. (2008). Groups of adjacent contour segments for object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *30*(1), 36–51.

Ferrari, V., Jurie, F., & Schmid, C. (2010). From images to shape models for object detection. *International Journal of Computer Vision*, *87*(3), 284–303.

Fritz, M., & Schiele, B. (2008). Decomposition, discovery and detection of visual categories using topic models. In *IEEE computer society conference on computer vision and pattern recognition*.

Goemans, M. X., & Williamson, D. (1995). Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM*, *42*(6), 1115–1145.

Gold, S., & Rangarajan, A. (1996). A graduated assignment algorithm for graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *18*(4), 377–388.

Gorelick, L., & Basri, R. (2009). Shape based detection and top-down delineation using image segments. *International Journal of Computer Vision*, *83*(3), 211–232.

Grant, M., & Boyd, S. (2010). CVX: Matlab software for disciplined convex programming, version 1.21. http://cvxr.com/cvx.

Grimson, W., & Lozano-Perez, T. (1987). Localizing overlapping parts by searching the interpretation tree. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *9*(4), 469–482.

Gu, C., Lim, J., Arbelaez, P., & Malik, J. (2009). Recognition using regions. In *IEEE computer society conference on computer vision and pattern recognition*.

Huttenlocher, D., Klanderman, D., & Rucklige, A. (1993). Comparing images using the Hausdorff distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *15*(9), 850–863.

Indyk, P., & Thaper, N. (2003). Fast image retrieval via embeddings. In *3rd international workshop on statistical and computational theories of vision*.

Joachims, T. (1999). Making large-scale svm learning practical. In *Advances in kernel methods—support vector learning*.

Kimia, B., Tannenbaum, A., & Zucker, S. (1995). Shapes, shocks, and deformations I: the components of two-dimensional shape and the reaction-diffusion space. *International Journal of Computer Vision*, *15*(3), 189–224.

Koffka, K. (1935). *Principles of gestalt psychology*. London: Lund Humphries.

Lamdan, Y., Schwartz, J., & Wolfson, H. (1990). Affine invariant model-based object recognition. *IEEE Transactions on Robotics and Automation*, *6*(5), 578–589.

Latecki, L., & Lakamper, R. (2000). Shape similarity measure based on correspondence of visual parts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(10), 1185–1190.

Latecki, L., Lakamper, R., & Eckhardt, U. (2000). Shape descriptors for non-rigid shapes with a single closed contour. In *IEEE computer society conference on computer vision and pattern recognition*.

Leibe, B., Leonardis, A., & Schiele, B. (2008). Robust object detection with interleaved categorization and segmentation. *International Journal of Computer Vision*, *77*(1), 259–289.

Leordeanu, M., Hebert, M., & Sukthankar, R. (2007). Beyond local appearance: Category recognition from pairwise interactions of simple features. In *IEEE computer society conference on computer vision and pattern recognition*.

Levin, A., & Weiss, Y. (2006). Learning to combine bottom-up and top-down segmentation. In *European conference on computer vision*.

Ling, H., & Jacobs, D. (2007). Shape classification using the inner-distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(2), 286–299.

Lu, C., Latecki, L. J., Adluru, N., Yang, X., & Ling, H. (2009). Shape guided contour grouping with particle filters. In *International conference on computer vision*.

Maji, S., & Malik, J. (2009). Object detection using a max-margin hough transform. In *IEEE computer society conference on computer vision and pattern recognition*.

Malisiewicz, T., & Efros, A. A. (2008). Recognition by association via learning per-exemplar distances. In *IEEE computer society conference on computer vision and pattern recognition*.

Marr, D. (2010). *Vision: A computational investigation into the human representation and processing of visual information*. New York: Henry Holt.

Martin, D., Fowlkes, C., & Malik, J. (2004). Learning to detect natural image boundaries using local brightness, color, and texture cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *26*(5), 530–549.

Mcneill, G., & Vijayakumar, S. (2006). Hierarchical Procrustes matching for shape retrieval. In *IEEE computer society conference on computer vision and pattern recognition*.

Mokhtarian, F., Abbasi, S., & Kittler, J. (1997). Efficient and robust retrieval by shape content through curvature scale space. *Image Databases and Multi-Media Search*, 51–58.

Opelt, A., Pinz, A., & Zisserman, A. (2006). A boundary-fragment-model for object detection. In *European conference on computer vision*.

Osada, R., Funkhouser, T., Chazelle, B., & Dobkin, D. (2002). Shape distributions. *ACM Transactions on Graphics*, *21*(4), 807–832.

Palmer, S. E. (1999). *Vision science: Photons to phenomenology*. Cambridge: The MIT Press.

Pentland, A. (1986). Perceptual organization and the representation of natural form. *Artificial Intelligence*, *28*(3), 293–331.

Pizer, S., Fritsch, D., Yushkevich, P., Johnson, V., & Chaney, E. (1999). Segmentation, registration, and measurement of shape variation via image object shape. *IEEE Transactions on Medical Imaging*, *18*(10), 851–865.

Ravishankar, S., Jain, A., & Mittal, A. (2008). Multi-stage contour based detection of deformable objects. In *European conference on computer vision*.

Ren, X., Fowlkes, C., & Malik, J. (2005). Cue integration in figure/ground labeling. In *Neural information processing systems*.

Russell, B., Efros, A. A., Sivic, J., Freeman, B., & Zisserman, A. (2006). Using multiple segmentations to discover objects and their extent in image collections. In *IEEE computer society conference on computer vision and pattern recognition*.

Schoenemann, T., & Cremers, D. (2007). Globally optimal image segmentation with an elastic shape prior. In *International conference on computer vision*.

Sebastian, T., Klein, P., & Kimia, B. (2003). On aligning curves. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *25*(1), 116–125.

Sebastian, T., Klein, P., & Kimia, B. (2004). Recognition of shapes by editing their shock graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *26*(5), 550–571.

Shapiro, L., & Haralick, R. (1979). *Structural descriptions and inexact matching* (Technical report CS79011-R). Computer Science, Virginia Tech.

Shotton, J., Blake, A., & Chipolla, R. (2005). Contour-based learning for object detection. In *International conference on computer vision*.

Shotton, J., Winn, J., Rother, C., & Criminisi, A. (2009). Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *International Journal of Computer Vision*, *81*(1), 2–23.

Siddiqi, K., Shokoufandeh, A., Dickinson, S., & Zucker, S. (1999). Shock graphs and shape matching. *International Journal of Computer Vision*, *35*(1), 13–32.

Srinivasan, P., Zhu, Q., & Shi, J. (2010). Many-to-one contour matching for describing and discriminating object shape. In *IEEE computer society conference on computer vision and pattern recognition*.

Sturm, J. F. (1999). Using sedumi 1.02, a matlab toolbox for optimization over symmetric cones. *Optimization Methods & Software*, *11*(12), 625–653.

Toshev, A., Taskar, B., & Daniilidis, K. (2010). Object detection via boundary structure segmentation. In *IEEE computer society conference on computer vision and pattern recognition*.

Trinh, N. H., & Kimia, B. B. (2011). Skeleton search: Category-specific object recognition and segmentation using a skeletal shape model. *International Journal of Computer Vision*, *94*(2), 215–240.

Tu, Z., & Yuille, A. (2004). Shape matching and recognition–using generative models and informative features. In *Seventh European conference on computer vision*.

Umeyama, S. (1988). An eigendecomposition approach to weighted graph matching problems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *10*(5), 695–703.

Yoshida, K., & Sakoe, H. (1982). Online handwritten character recognition for a personal computer system. *IEEE Transactions on Consumer Electronics*, *3*, 202–209.

Yu, S. X., & Shi, J. (2003). Multiclass spectral clustering. In *International conference on computer vision*.

Zhang, D., & Lu, G. (2003). Evaluation of mpeg-7 shape descriptors against other shape descriptors. *Multimedia Systems*, *9*(1), 15–30.

Zhu, Q., Wang, L., Wu, Y., & Shi, J. (2008). Contour context selection for object detection: A set-to-set contour matching approach. In *European conference on computer vision*.