Dissertation

submitted
to the
Combined Faculties for the Natural Sciences and for Mathematics
of the
Ruperto-Carola University of Heidelberg, Germany
for the degree of
Doctor of Natural Sciences

Put forward by
Master of Science. Pradeep Krishna Yarlagadda
Born in: India
Oral examination:

Beyond the Sum of Parts: Shape-based Object Detection and its Applications

Advisor: Prof. Dr. Björn Ommer

Abstract

The grand goal of Computer Vision is to generate an automatic description of an image based on its visual content. Such a description would lead to many exciting capabilities, for example, searching through the images based on their visual content rather than the textual tags attached to the images. Images and videos take an ever increasing share of the total information content in archives and on the internet. Hence, such automatic descriptions would provide powerful tools for organizing and indexing by means of the visual content. Category level object detection is an important step in generating such automatic image descriptions.

The major part of this thesis addresses the problems encountered in popular lines of approaches which utilize shape in various ways for object detection namely, i) Hough Voting, ii) Contour based Object Detection and iii) Chamfer Matching. The problems are tackled using the principles of emergence which states that the whole is more than the sum of its parts.

Hough Voting methods are popular because they efficiently handle the high complexity of multi-scale, category-level object detection in cluttered scenes. However, the primary weakness of this approach is that mutually dependent local observations independently vote for intrinsically global object properties such as object scale. All the votes are added up to obtain object hypotheses. The assumption is thus that object hypotheses are a sum of independent part votes. Popular representation schemes are, however, based on an overlapping sampling of semi-local image features with large spatial support (e.g. SIFT or geometric blur). Features are thus mutually dependent. The question arises as to how to incorporate the feature dependences into Hough Voting framework. In this thesis, the feature dependencies are modelled by an objective function that combines three intimately related problems: i) grouping of mutually dependent parts, ii) solving the correspondence problem conjointly for dependent parts, and iii) finding concerted object hypotheses using extended groups rather than based on local observations alone.

While Voting with dependent groups brings a significant improvement over standard Hough Voting, the interest points are still grouped in a query image during the detection stage. The grouping process can be made robust by grouping densely sampled interest points in training images yielding contours and evaluating the utility of contours over the full ensemble of training images. However, contour based object detection poses significant challenges for category-level object detection in cluttered scenes: Object form is an emergent property that cannot be perceived locally but becomes only available once the whole object has been detected and segregated from the background. To tackle this challenge, this thesis addresses the detection of objects and the assembling of their shape simultaneously, while avoiding fragile bottom-up grouping in query images altogether. Rather, the challenging problems of finding meaningful contours and discovering their spatially consistent placement are both shifted into the training stage. These challenges can be better handled using an ensemble of training samples rather than just a single query image. A dictionary of meaningful contours is then discovered using grouping based on co-activation patterns in all training images. Spatially consistent compositions of all contours are learned using maximum margin multiple instance learning. During recognition, objects are detected and their shape is explained simultaneously by optimizing a single cost function.

For finding the placement of an object template or its part in an edge map, Chamfer matching is a widely used technique because of its simplicity and speed. However, it treats objects as being a mere sum of the distance transformation of all their contour pixels, thus leading to spurious matches. This thesis takes account of the fact that boundary pixels are not all equally important by applying a discriminative approach to chamfer distance computation, thereby increasing its robustness. While this improves the behaviour in the foreground, chamfer matching is still prone to accidental responses in spurious background clutter. To estimate the accidentalness of a match, a small dictionary of simple background contours is utilized. These background elements are trained to focus at locations where, relative to the foreground, typically accidental matches occur. Finally, a max-margin classifier is employed to learn the co-placement of all background contours and the foreground template. Both the contributions bring significant improvements over state-of-the-art chamfer matching on standard benchmark datasets.

The final part of the thesis presents a case study where shape-based object representations provided semantic understanding of medieval manuscripts to art historians. To carry out the case study, a novel image dataset has been assembled from illuminations of 15th century manuscripts with ground-truth information about various objects of artistic interest such as crowns, swords. An approach has been developed for automatically extracting potential objects (for e.g. crowns) from the large image collection, then analysing the intra-class variability of objects by means of a low dimensional embedding. With the help of the resultant plot, the art historians were able to confirm different artistic workshops within the manuscript and could verify the variations of art within a particular school. Obtaining such insights manually is a tedious task and one has to go through and analyse all the object types from all the pages of the manuscript. In addition, a semi-supervised approach has been developed for analysing the variations within an artistic workshop, and extended further to understand the transitions across artistic styles by means of 1-d ordering of objects.

Zusammenfassung

Das große Ziel von Computer Vision ist, eine automatische Bildbeschreibung basierend auf dem visuellen Inhalt eines Bildes zu generieren. Eine solche Beschreibung eröffnet viele spannende Anwendungsmöglichkeiten, z.B. eine Bildsuche die direkt vom visuellen Inhalt eines Bildes ausgeht und somit nicht auf textliche Annotationen zurückgreifen muss. Da Bilder und Videos einen immer weiter anwachsenden Teil des gesamten Informationsgehalts in Archiven und dem Internet darstellen, wären solche automatische Bildbeschreibungen anhand des visuellen Inhalts mächtige Werkzeuge zur Organisation und Indexierung von Bildern. Objekterkennung auf Kategorieebene ist ein wichtiger Schritt um solche automatischen Bildbeschreibungen zu erstellen.

Der Hauptteil dieser Doktorarbeit beschäftigt sich mit den Problemen weitverbreiteter Ansätze, die auf verschiedene Weise die Form von Objekten nutzen : a) Hough Voting, b) konturbasierte Objekterkennung, und c) Chamfer Matching. Die Probleme werden mit Hilfe des Emergenzprinzips gelöst, das besagt, dass das Ganze mehr als die Summe seiner Teile ist.

Hough Voting Ansätze sind beliebt, da sie die Komplexität von multiskalen, kategoriebasierter Objekterkennung in verrauschten Bildern effizient handhaben. Ein wesentlicher Nachteil dieses Ansatzes ist, dass lokal gemachte voneinander abhängige Beobachtungen unabhängig voneinander für globale Objekteigenschaften abstimmen wie z.B. die Größe eines Objekts. Alle Votes werden aufaddiert um Objetkhypothesen zu erhalten. Daher wird angenommen, dass die Objekthypothesen die Summe von unabhängigen Votes einzelner Bildbestandteile sind. Gängige Darstellungsweisen führen allerdings ein Sampling von überlappenden semilokalen Bildeigenschaften durch die eine starke räumliche Unterstützung aufweisen (z.B. SIFT oder geometric blur). Die Merkmale sind daher voneinander abhängig. Daher stellt sich die Frage wie man Abhängigkeiten zwischen Merkmalen in das Hough Voting Framework integriert. In dieser Arbeit werden die Abhängigkeiten zwischen Merkmalen durch eine Zielfunktion beschrieben, die drei eng miteinander Verbundene Probleme verbindet: a) Gruppierung von voneinander abhängigen Bildteilen, b) gemeinsame Lösung des Korrespondenzproblems für abhängige Bildteile und c) das Finden von aufeinander abgestimmten Objekthypothesen mit Hilfe von erweiterten Gruppen statt ausschließlich mit Hilfe lokaler Beobachtungen.

Obwohl Voting mit abhängigen Gruppen eine signifikante Verbesserung gegenüber gewöhnlichem Hough Voting erzielt, werden die Interest Points immer noch während der Detektionsphase in einem Testbild gruppiert. Der Gruppierungsprozess kann robust gemacht werden, indem man dicht gesampelte Interest Points in Trainingsbildern gruppiert, um Konturen zu erhalten und den Nutzen dieser Konturen auf allen Trainingsbildern evaluiert. Jedoch bringt konturbasierte Objektdetektion signifikante Herausforderungen für die Objektdetektion auf Kategorieebene in verrauschten Szenen mit sich: Die Form eines Objekts ist eine entstehende Eigenschaft, die nicht lokal wahrgenommen werden kann sondern erst entsteht sobald das ganze Objekt detektiert und vom Hintergrund getrennt wurde. Um dieses Problem zu lösen befasst sich diese Arbeit gleichzeitig mit der Detektion von Objekten und der Konfiguration ihrer Form und vermeidet fehleranfällige bottom-up Gruppierung in Testbildern. Stattdessen werden die beiden schwierigen Probleme des Findens sinnvoller Konturen und deren räumlich konsistenter Platzierung in die Trainingsphase verschoben. Dieses Problem kann besser gehandhabt werden indem man eine Menge von

Trainingsbeispielen verwendet statt ein einzelnes Testbild. Ein Wörterbuch aus sinnvollen Konturen wird dann mit Hilfe von Gruppierungen basierend auf Koaktivierungsmustern aller Trainingsbilder erstellt. Räumlich konsistente Anordnungen aller Konturen werden mit Hilfe von Maximum-Margin Multiple Instance Learning gelernt. Während der Erkennung werden gleichzeitig Objekte detektiert und ihre Form durch die Optimierung einer einzigen Kostenfunktion erklärt.

Chamfer Matching ist Aufgrund seiner Einfachheit und Geschwindigkeit eine weitverbreitete Methode um die Platzierung eines Objekttemplates oder eines Teil des Templates in einem Kantenbild zu finden. Jedoch behandelt es Objekte als wären sie die bloße Summe der Distanztransformationen ihrer Konturpixel und führen so zu falschen Matches. Diese Arbeit berücksichtigt die Tatsache, dass nicht alle Konturpixel gleich wichtig sind und wendet einen diskriminativen Ansatz auf die Chamfer-Distanzberechnung an um so die Robustheit zu erhöhen. Obwohl damit das Verhalten im Vordergrund verbessert wird, ist Chamfer Matching immer noch anfällig für zufällige Matches in störendem Hintergrundrauschen. Um die Zufälligkeit eines Matches abzuschätzen wird ein kleines Wörterbuch einfacher Hintergrundkonturen verwendet. Diese Hintergrundelemente werden trainiert um sich auf Bereiche zu konzentrieren in denen relative zum Vordergrund typischerweise zufällige Matches auftreten. Schließlich wird ein Max-Margin Klassifikator verwendet um das gemeinsame Auftreten aller Hintergrundkonturen und des Vordergrundtemplates zu lernen. Beide Neuerungen bewirken eine signifikante Verbesserung gegenüber dem state-of-the-art Chamfer Matching auf den gebräuchlichen Benchmarkdatensätzen.

Den letzten Teil der Arbeit bildet eine Fallstudie , in der auf mittelalterliche Buchmalerei angewendete formbasierte Objektrepräsentation semantische Erkenntnisse für die Kunstgeschichte liefert. Um die Fallstudie durchzuführen wurde aus einem illustrierten Manuskript aus dem 15. Jahrhundert ein neuer Bilddatensatz zusammengestellt. Die annotierten Objekte in diesem Datensatz umfassen verschiedene Objekte von künstlerischem Interesse wie z.B. Kronen und Schwerter. Es wurde eine Methode entwickelt, um automatisch potentielle Objekte wie Kronen aus einer großen Bildsammlung zu extrahieren und dann die Variabilität innerhalb einer Objektklasse mit Hilfe eines niedrig dimensionalen Embeddings analysiert. Mit Hilfe der Ergebnisse konnten die Kunsthistoriker verschiedenen Werkstätten innerhalb des Manuskripts bestätigen und die Veränderungen der Formen innerhalb eines bestimmten Schulzusammenhangs verstehen. Solche Erkenntnisse per Hand zu generieren ist eine sehr zeitaufwendige Aufgabe, da man alle Objekttypen auf allen Seiten des Manusskripts durchgehen und vergleichen müsste. Darüber hinaus wurde ein semi-überwachter Ansatz für die Analyse der Variationen innerhalb einer Werkstatt entwickelt und weiter entwickelt um die übergänge zwischen Kunststilen bezüglich der 1-d Ordnung der Objekte zu verstehen.

Acknowledgements

Many people have contributed in various ways to the successful completion of this thesis. I thank all of them.

Thanks to my advisor Prof. Björn Ommer for creating a good research environment and more importantly for providing the motivation and inspiration to carry out good research work. I have learnt a lot from him such as how to write research articles, how to think about the big picture, paying attention to details, how to communicate effectively and such. I would like to thank Prof. Christoph Schnörr for helpful feedback regarding this thesis and for organizing interesting talks and seminars at HCI.

Thanks to all my colleagues at HCI (especially Hongwei, Antonio, Angela, Boris, Peter, Bernd, Joseph, Dorothea, Agnes) for interesting discussions at lunch and other venues over a wide range of topics from science, art, history, philosophy, politics, economics, etc. I am grateful for Tanja Kohl's immense help in all administrative matters. The network administrators Ole Hansen, Jürgen Moldenhauer and Markus Nullmeier have done a great job of making sure that all the computing goes smoothly for everyone at HCI.

Thanks to friends from IIT Roorkee and Brown University for the good times I had with them. Thanks to Jasmin Montabon for everything.

I am lucky to have a great set of parents, Mrs. Sasi Prabha Gummadi and Mr. Rama Rao Yarlagadda. I have come this far mainly because of their constant support and encouragement.

CONTENTS

CHAPTER 1

INTRODUCTION

The industrial revolution from the period 1750 to 1860 is considered to be one of the most important events in human history. The main characteristic of this period is the transition from manual labour to using machines in the manufacturing of goods. Using machines to manufacture parts or components of an object has two characteristics: i) high precision and ii) uniform industry standards for the parts. Having uniform industry standards has resulted in the concept of replaceable parts where parts manufactured by different machines are completely interchangeable in making up a product. Such innovations had a tremendous impact on increasing the productivity and resulted in an epoch of sustained economic growth in western world.

The advantage of using machines over manual labour is obvious in some of the tasks in today's world. An example is a lathe machine cutting a piece of wood with very low tolerance for dimensional errors. And then there are tasks such as playing a game of chess which require considerably more amount of intelligence than cutting a piece of wood. In this domain too, machines have been proven to match the capabilities of humans if not surpass them. Yet, there is a whole domain of tasks, such as interacting with the world using information from various senses including vision, where machines are far behind human capabilities. Automatically navigating through the streets falls under the category of difficult tasks for machines.

In terms of understanding and extracting useful information from visual sensors, rapid strides have been made since the inception of Computer Vision as a research area in 1950s. The grand goal of Computer Vision is to generate an automatic description of an image based on its visual content. Yet, the state-of-the-art vision systems are still far from passing Turing test [91] in terms of vision. A vision Turing test would be as follows. The Computer Vision system should try to generate automatic description of an image/scene given its visual content. A human observer has to tell if the description has been human generated or machine generated. If he cannot detect the difference, then the vision system is deemed to have passed the Turing test.

There is an ever increasing interest in the field of Computer Vision despite the enormous challenges that need to be tackled. Vast amount of research is being carried out to develop

vision based solutions [2] for problems in diverse fields such as agriculture, art and architecture, construction, service sector, surveillance, manufacturing and inspection, robotics, entertainment and media, environmental analysis, medicine, human computer interfaces and transportation.

An important building block for generating visual content based description of an image is Category level Object detection. The basic objective of Category level object detection is to identify and localize all instances of an object category in a variety of images. Challenges faced by such a system include large intra-class variability of objects, fine grained differences between various object classes, inter-class object similarity, running time complexity of the object search, large scale variation of the objects, object occlusion and illumination changes of the scene [71]. Different research trends have emerged to tackle the above challenges. Ponce et al. [77] provide an excellent overview of various trends in Object Detection.

A few examples illustrate the strengths and weaknesses of different trends. For handling occlusions, generative approaches are significantly better than discriminative models for object detection. To handle clutter, the most prominent approaches are currently part-based models using local or semi-local descriptors. Based on appearance patches [45, 60], SIFT [63], geometric blur [19], and other texton-like features [53] local image information is extracted and then combined in a spatial model. These models range from no spatial relationships like bag-of-features [33], conditionally independent parts in voting methods [60, 51, 68] and pictorial structures [40], over rigid, grid-like structures to joint models of all parts [45] like the constellation model. Hough Voting based approaches [60, 46, 74, 86, 73, 68, 51] efficiently handle the complexity issues in object detection.

Shape-based approaches are robust to intra-class object variations such as color and appearance, robust to moderate viewpoint variations. Shape-based models also provide an effective approach for accurately explaining meaningful object pixels in an image. There are various lines of work within Object Recognition which utilize shape in different ways. i) Hough Voting approaches such as [68, 73, 102] utilise the edge information in an image to sample interest points which vote for object properties such as the location of the object center, object's scale and aspect ratio. [87, 74] use the contours obtained from bottom-up grouping of edge pixels to vote for object properties. ii) Approaches such as [75, 64] formulate the detection problem as a many-to-one matching of contours from query images to a sparse set of model contours obtained from training images. iii) Chamfer matching is a widely used technique for detecting objects, especially in industrial inspection tasks because of its simplicity and speed.

The limitations of the state-of-the-art shape-based approaches in the above three lines of research work araise because of a fundamental issue. An object is treated as a mere sum of its constituent parts. In the case of Hough Voting approaches, each interest point is independently voting for object properties. Contour based approaches formulate the detection problem as an independent matching of constituent object contours. Chamfer matching obtains the matching costs for a template by summing over all the template pixels in the distance transform of the query image.

This thesis goes beyond the sum of the parts and treats an object as more than the sum of its constituents elements. This principle of emergence is first stated by the Greek philosopher Aristotle and has been extensively studied in the context of perceptual patterns by Gestalt school of psychologists, Wertheimer, Köhler and Koffka. The next sections describe how this philosophy is applied to solve the problems of shape-based object detection approaches.

1.1 Hough Voting

Hough Voting methods are popular because they efficiently handle the high complexity of multi-scale, category-level object detection in cluttered scenes. Hough transform was introduced in [16] for detecting simple geometrical entities such as lines and circles in images. Leibe et al. [60] used interest point descriptors like SIFT [63] in generalized Hough Voting to perform Category level Object Detection. This work has generated renewed interest in Hough Voting as evidenced by recent publications such as [46, 68, 51, 73].

However, the current voting approaches have a common weakness. Mutually dependent local observations independently vote for intrinsically global object properties such as object scale (c.f. Fig. 1.1). All the votes are added up to obtain object hypotheses. The assumption is thus that object hypotheses are a sum of independent part votes. Popular representation schemes are, however, based on an overlapping sampling of semi-local image features with large spatial support (e.g. SIFT or geometric blur). Features are thus mutually dependent. Moreover, matching individual interest points from query images to training images (referred to as correspondence problem) is highly unreliable as demonstrated in [102].

This thesis models the feature dependencies by deriving an objective function that combines three intimately related problems: i) grouping of mutually dependent parts, ii) solving the correspondence problem conjointly for dependent parts, and iii) finding concerted object hypotheses using extended groups rather than based on local observations alone. Under the new modelling paradigm, an object hypothesis is more than just a summation of individual part votes.

To detect objects in a novel image, a probabilistic edge map is computed in a first step using [66]. A uniform sampling of edge pixels yields points where local features are extracted on a single scale (we use geometric blur features [19]). Descriptors are then matched to the full ensemble of codebook features collected from all training images and the corresponding matching probabilities are stored. Because of independence assumption between voting elements, all the votes are simply added up in a Hough accumulator in standard Hough Voting. In contrast, this thesis groups the dependent voting elements and object hypotheses are jointly estimated for whole groups. The cost function which combines the different intimately related problems is optimized in an iterative routine. That way, all related points influence each others voting and correspondences and their voting influences their grouping, in turn. The grouping is initialized by a pairwise clustering of edge points. Measuring the co-occurrence of points in different levels of the hierarchical segmentation of the initial probabilistic edge map from [66] yields the necessary pairwise affinities.

1.2 Contour based Object Detection

Shape is a natural, highly prominent characteristic of objects that human vision utilizes everyday for detecting objects. But despite its expressiveness, shape poses significant challenges for category-level object detection in cluttered scenes: Object form is an emergent property that cannot be perceived locally but becomes only available once the whole object has been detected and segregated from the background. Fig. 1.2 demonstrates this phenomenon by some examples. Such emergent phenomena have been extensively studied by Gestalt school of psychologists such as Wertheimer, Köhler and Koffka. The majority of shape-based detection methods are based on spatially flexible matching algorithms and

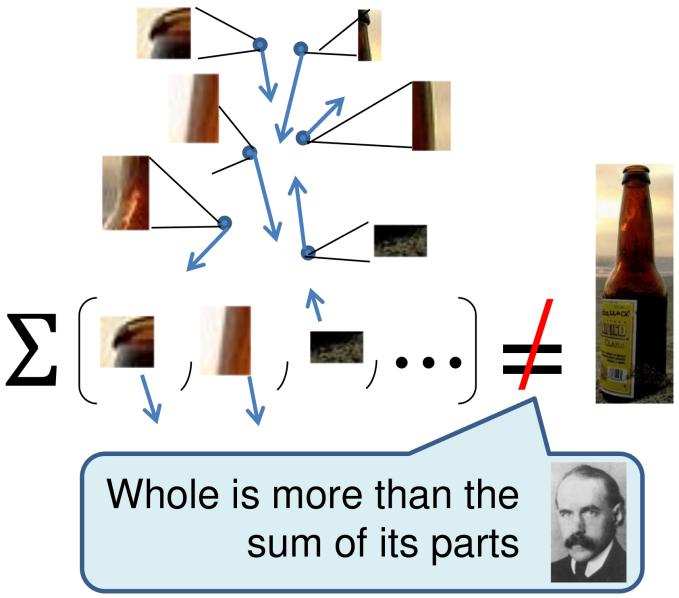
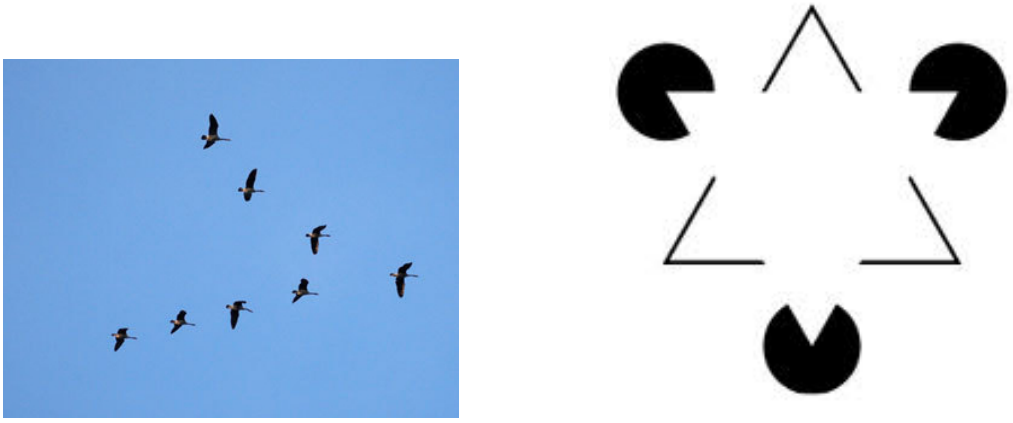


Figure 1.1: Limitation of Hough Voting approaches

deformable part configurations. For example, [78, 64] present a shape-based approach based on the partial matching of edge fragments. Jianbo Shi et al. [75] utilize a many-to-one matching of contours from query images to a sparse set of model contours. Both approaches require a bottom-up grouping of edge pixels in a query image which is a fragile process. Moreover, contours which are part of the object model in [75] are matched independently for detecting objects. This is against the principles of emergentism and it has limitations when dealing with articulated objects where the relative configuration of parts differs across different instances. [87, 74] measure the direct visual similarity between contours while building a dictionary of codebook contours. This creates problems when contours are corrupted by the bottom-up extraction process.

To tackle all of the above challenges, this thesis addresses the detection of objects and the assembling of their shape simultaneously, while avoiding fragile bottom-up grouping in query images altogether. Rather, the challenging problems of finding meaningful contours and discovering their spatially consistent placement are both shifted into the training stage. These challenges can be better handled using an ensemble of training samples rather than just a single query image. A dictionary of meaningful contours is then discovered using grouping based on co-activation patterns in all training images. The training images have only weak supervision information (the bounding box of the objects). Hence there are multiple placements for each model contour in a training image within the bounding box of each annotated object. Thus the most relevant co-activation pattern of all contours is treated as a hidden variable. The hidden variables are learnt along with the weights for the codebook co-activations in a max-margin multiple instance learning framework. During recognition, objects are detected and their shape is explained simultaneously by optimizing a single cost function. A joint placement for all codebook contours is sought after, rather than placing each contour independently.



a) b)

Figure 1.2: Figures demonstrate the emergence phenomenon in images.

1.3 Chamfer Matching

Chamfer matching was first introduced by Barrow et al. [17] to match two sets of contour fragments. Since then, Chamfer matching is a widely used technique for detecting objects because of its simplicity and speed. Thayananthan et al. [88] have compared shape context [18] and chamfer matching of templates for object detection in cluttered images. They report that chamfer matching is more robust in clutter than shape context. Nevertheless, false positives in cluttered background were found to be the major downside of chamfer matching. More recent research has made attempts to address this problem. Shotton et al. [87] proposed an improved matching scheme called oriented chamfer matching (OCM) that takes into account the orientation mismatch between pixels. In [62], an alternative approach (directional chamfer matching) for incorporating edge orientation has been proposed which solves the matching problem in an augmented space. Although [62] is the state-of-the-art in chamfer matching, it performs poorly in the presence of clutter as shown in Fig. 1.3.

In chamfer matching, the matching costs for a template are obtained by summing over all the template pixels in the distance transform of the query image. Thus, the objects are treated as being a mere sum of the distance transformation of all their contour pixels. However, Biederman [21], Attneave [14], and various experiments on illusionary contours demonstrate that object boundary pixels are not all equally important in object detection.

This thesis introduces discriminative distance transform to take account of the fact that boundary pixels are not all equally important. This discriminative approach to chamfer distance computation increases the robustness of the matching process. However, chamfer matching is still prone to accidental responses in spurious background clutter even with discriminative distance transform. Chamfer matching only matches the template contour and thus fails to discount the matching score by the accidentalness, i.e., the likelihood that this is a spurious match. To estimate the accidentalness of a match, a small dictionary of simple background contours is utilized. These background elements are trained to focus at locations where, relative to the foreground, typically accidental matches occur. Finally, a max-margin classifier is employed to jointly learn the relative importance of foreground template points as well as the co-placement of all background contours.

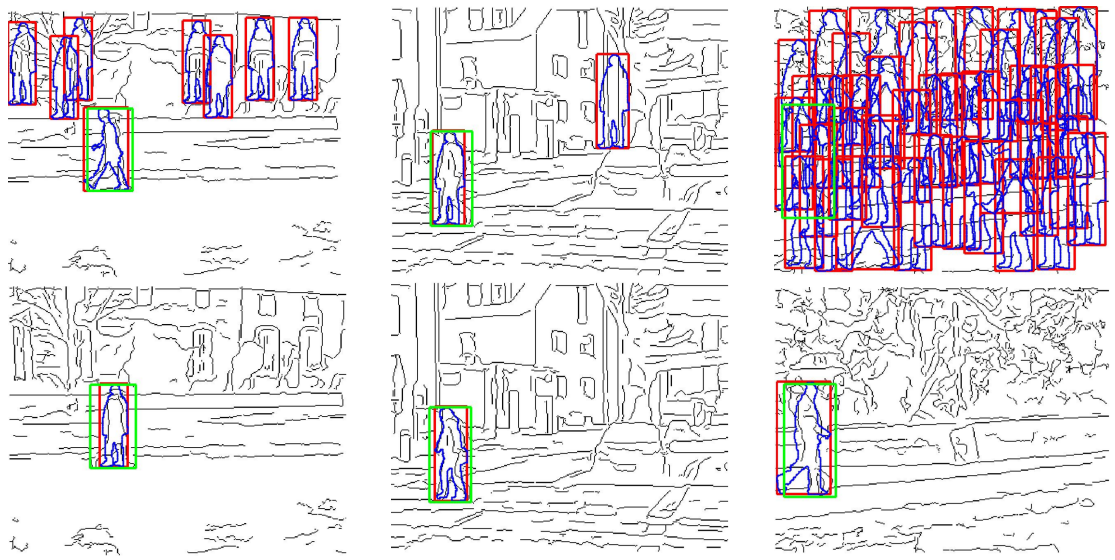


Figure 1.3: The top row shows the directional chamfer matching [62]. Bottom row shows the regularized chamfer matching proposed in this thesis. The ground-truth bounding box is shown in green and the top scoring object hypotheses are shown in red.

1.4 Semantic Understanding of Medieval Manuscripts

Recent digitization projects in the field of art history have lead to the creation of large amounts of visual data. To efficiently open up these resources for art historians, Computer Vision algorithms which advance beyond the analysis of individual pixels are required. Once such algorithms are in place, they provide a semantic understanding of visual data. More specifically, they help the users in tasks such as 1) searching through the image collections for different objects of interest like crowns and swords 2) identifying the sub-categories of an object type 3) identifying different artistic workshops to which the objects belong 4) understanding the variations of art within a particular school of design and 5) understanding the transition of art from one school of design to another. Manually performing the above tasks is a tedious process for humans which require a great deal of time and effort. Obviously no human user can view all of these images at the same time and, thus, relations between different images or the objects within are hard to discover. Revealing the structure that is inherent to a collection of images, i.e., the artistic variations of all instances of an object category such as medieval crowns, is consequently a very difficult task. The mere size of a dataset makes it difficult to see the greater whole. Computers on the other hand can easily handle thousands of images at the same time.

The final part of the thesis presents a case study where shape-based Computer Vision techniques provided semantic understanding of medieval manuscripts to art historians. Fig. 1.4 demonstrates the various objectives of the case study. To carry out the case study, a novel image dataset has been assembled from 15th century manuscript of medieval images with ground-truth information about various objects of artistic interest such as crowns, swords. An approach has been developed for automatically extracting objects (for e.g. crowns) from the large image collection, then analysing the intra-class variability of objects by means of a low dimensional embedding. With the help of the resulting plot, the art historians were able to confirm different artistic workshops within the manuscript

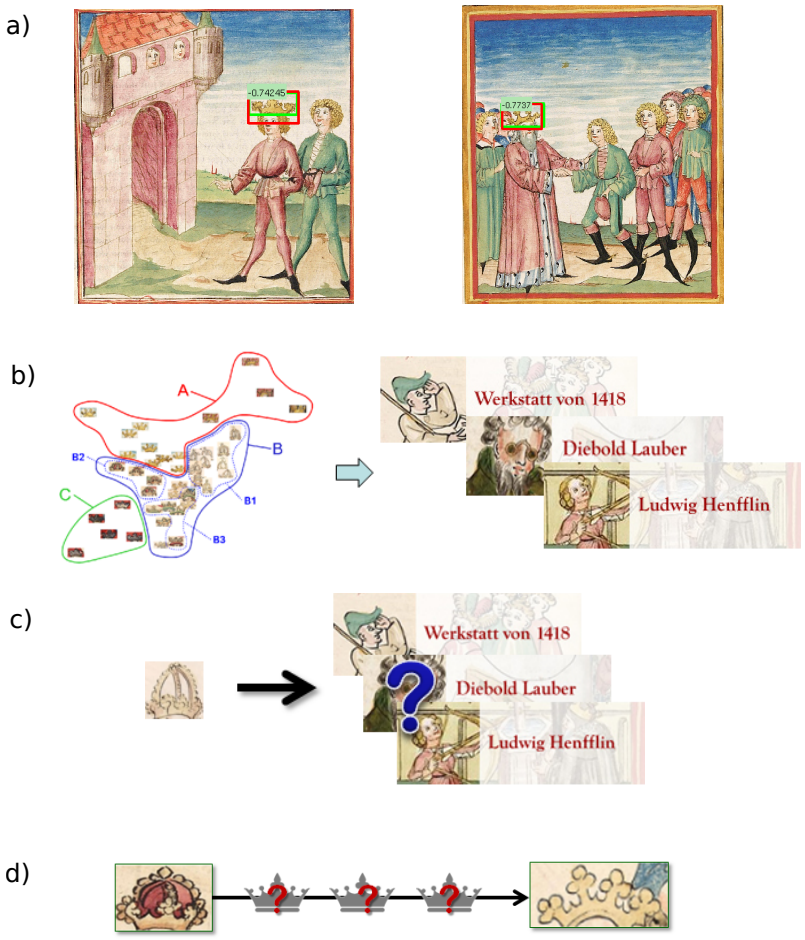


Figure 1.4: Various objectives of the Case Study on Medieval manuscripts. a) Object Detection b) Unsupervised discovery of different artistic workshops c) Supervised Classification of Objects according to different workshops d) Semi-supervised ordering and interpolation between exemplars

and understood the variations of art within a particular school of design. In addition, a semi-supervised approach has been developed for analysing the variations within an artistic workshop, and extended further to understand the transitions across artistic styles by means of 1-d ordering of objects.

1.5 Contributions

This section summarizes the contributions of this thesis.

- The fundamental limitation to Hough Voting methods, the assumption of independence between voting elements, is addressed. The mutual dependence between local features is modelled in a joint optimization framework.
- The advantage of modelling dependencies between voting elements over voting with independent elements is thoroughly demonstrated by various experimental evaluations.

- A novel approach for learning a codebook of meaningful contours from all training images is presented.
- Learning an object model from a codebook of meaningful contours is formulated as a max-margin multiple instance learning problem.
- The learnt object model is used to detect objects and assemble their shape simultaneously, while avoiding fragile bottom-up grouping in query images altogether.
- The primary weakness of chamfer matching, i.e. the false positives in background clutter is addressed by introducing discriminative distance transform and regularizing chamfer matching with generic background contours.
- A case study on Upper German medieval manuscripts is presented where Computer Vision techniques provided semantic understanding to art historians.
- A novel benchmarking dataset of Upper German manuscripts for addressing the pertinent research questions of art historians has been assembled.
- Objects such as crowns and swords which have a lot of meaning and significance to art historians have been detected.
- An unsupervised approach for discovering various artistic workshops in the manuscripts is presented.
- A semi-supervised approach for inducing 1-d ordering of objects based on the pairwise relationships is discussed.

1.6 Organisation of the Thesis

This thesis is organized as follows:

Chapter 2 reviews the Probabilistic Hough Voting approaches. The limitations of standard Hough Voting are discussed and an approach is outlined to address these concerns.

Chapter 3 elaborates upon the outline for modelling feature dependencies in Hough Voting. Finding optimal correspondences between voting elements in training images and query image, grouping the voting elements in query image and finding optimal transformations of grouped entities are jointly modelled in a probabilistic optimization framework. The computational complexity and the convergence of iterative optimization are discussed. Useful variations such as many-to-one matching of each query point to multiple training points and learning the weights for each voting element in training images are also presented.

Chapter 4 provides the motivation for shifting from interest points to contours in object detection. Various shape-based approaches ranging from template matching to active shape models to shape hierarchies are reviewed. The challenges that need to be addressed by a shape-based detection are described and a solution is outlined. A brief review of useful machine learning algorithms is provided.

Chapter 5 begins by learning a codebook of meaningful contours from their co-activation pattern over an ensemble of training images. A discriminative approach for learning object models in a Max-Margin Multiple Instance Learning framework is described. In the detection stage, objects are detected from describing their shape by jointly placing all codebook contours.

Chapter 6 describes the shortcomings of state-of-the-art chamfer matching techniques like fast directional chamfer matching and provides two contributions to address the issues.

Chapter 7 provides a brief overview of work flow in art history and how Computer Vision could help simplify the work flow.

Chapter 8 describes the various vision techniques which helped the art historians gain semantic understanding of Upper German medieval manuscripts.

CHAPTER 2

PROBABILISTIC HOUGH VOTING

2.1 Overview

Hough transform based object detection techniques, first introduced in [16], are being widely used in the Computer Vision community. The basic strategy of such techniques is to let local interest points vote for parametrized object hypotheses, e.g. object locations, scales and aspect ratios. Thus, they avoid the computational complexity faced by other approaches such as sliding windows (e.g. [95, 34]) wherein a binary classifier is applied in rectangular sub-regions at all locations, scales and several aspect ratios for object detection. Generalizations of the Hough transform to arbitrary shapes, exemplar recognition [63], and category-level recognition [60, 46, 74, 86, 73, 68, 51] have successfully demonstrated the potential of voting based approaches, and their wide applicability.

2.2 Hough Voting with Independent Parts

Hough voting makes part-based object models with large numbers of parts feasible by letting all parts independently cast their votes for object hypotheses [60]. All these locally estimated object hypotheses are summed up in a Hough accumulator $\mathcal{H}^{\text{pnt}}(c, \mathbf{x}, \sigma)$ over scale space. Here, $\mathbf{x}$ and σ are the location and scale of an object hypothesis and c denotes its category. Moreover, a local part detected at location $\mathbf{x}_i^Q \in \mathbb{R}^2$ in a query image incorporates a feature vector $f_i^Q \in \mathbb{R}^N$ and a local estimate $\sigma_i^Q \in \mathbb{R}$ of object scale. The key assumption of Hough voting is that all parts are *independently* casting their votes for the object hypothesis,

$$\mathcal{H}^{\text{pnt}}(c, \mathbf{x}, \sigma) \propto \sum_i P\big(\mathbf{x}, \sigma | c, f_i^Q, \mathbf{x}_i^Q, \sigma_i^Q\big) P\big(c | f_i^Q, \mathbf{x}_i^Q, \sigma_i^Q\big) \tag{2.1}$$

Let f_j^T denote the j-th codebook vector or the j-th training sample, depending on whether vector quantization or a nearest neighbour approach is used. Without loss of generality we

can assume that the training object is centred at the origin so that the location $\mathbf{x}_j^T \in \mathbb{R}^2$ of f_j^T is the shift of the feature from the object center. Moreover, all training images are assumed to be scale normalized, i.e. they are rescaled so that objects are the same size. Summation over f_j^T and $\mathbf{x}_j^T$ then yields

$$\begin{aligned}\mathcal{H}^{\text{pnt}}(c, \mathbf{x}, \sigma) \propto \sum_{i,j} & P\big(\mathbf{x} - [\mathbf{x}_i^Q - \sigma_i^Q \mathbf{x}_j^T], \sigma - \sigma_i^Q\big) \\ & \times P(c|f_j^T)\ P(f_j^T|f_i^Q)\end{aligned} \tag{2.2}$$

The density in the first term $P\big(\mathbf{x} - [\mathbf{x}_i^Q - \sigma_i^Q \mathbf{x}_j^T], \sigma - \sigma_i^Q\big)$ can be approximated by Kernel density estimation using a kernel K with bandwidth $b(\sigma)$, scale dependent normalization $V_b(\sigma)$ and a distance function $d : \mathbb{R}^3 \times \mathbb{R}^3 \mapsto \mathbb{R}$ in scale space. Applying the balloon density estimator [30] yields

$$\begin{aligned}\mathcal{H}^{\text{pnt}}(c, \mathbf{x}, \sigma) \approx \sum_{i,j} & p\big(c|f_j^T\big) p\big(f_j^T|f_i^Q\big) \cdot \frac{1}{V_b(\sigma)} \\ & \times K\left(\frac{d\Big[(\mathbf{x}, \sigma)^\top; (\mathbf{x}_i^Q - \sigma_i^Q \mathbf{x}_j^T, \sigma_i)^\top\Big]}{b(\sigma)}\right)\end{aligned} \tag{2.3}$$

Candidate object hypotheses are then obtained by searching for local maxima of the objective function (2.3).

2.3 Review

Geometric Blur [20] and SIFT [63] are popular descriptors used for part representation in Hough Voting approaches. The parts are usually sampled from Globablized Probability Boundaries and UltraMetric Contour Maps [66] computed for training as well as query images. This section reviews the above mentioned useful ingredients in Hough Voting approaches [68, 73].

2.3.1 Geometric Blur Descriptor

The basic idea of Geometric Blur [20] is to convolve a spatial neighbourhood around an interest point with a spatially varying kernel. Such a descriptor summarizes the response of a signal under all affine transformations at a point. Typically, the signal is sparse such as the edge output of [66]. The output of the convolution is averaged to produce a robust signal. For an edge map I, the descriptor centred at location x is the following convolution

$$G_x(y) = \sum_{\times} I(x + y - \times)\eta(\times, \alpha(x - y) + \beta) \tag{2.4}$$

η denotes the Gaussian kernel whose standard deviation is a linear function of distance from the descriptor's center, x. Typically, Geometric Blur is computed at points $y_1, y_2, ..., y_k$

sampled from concentric circles around the interest point x and the values of α and β are chosen as 0.5 and 1 respectively [83]. All the sampled blur values are concatenated to form a feature descriptor f_j^T for an interest point. Normalized correlation between the feature descriptors f_j^T and f_i^Q is then used to compute the similarities between two interest points i and j.

2.3.2 SIFT Descriptor

Another popular feature used in earlier works of Hough Voting such as [60] is the SIFT descriptor. SIFT has been proposed by David Lowe in the seminal paper [63]. Given an image, interest points (refered to as keys in [63]) are obtained as a maxima and minima of a difference of Gaussian function applied in scale space [61]. An image is smoothed by convolving twice with a gaussian function $\eta(x)$ in the horizontal and vertical directions.

$$\eta(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{x^2}{2\sigma^2}) \tag{2.5}$$

The descriptor for a patch centred at each key point is computed as follows. First, the gradient magnitude M_{ij} and orientation R_{ij} are computed at each pixel P_{ij} of the image patch.

$$M_{ij} = \sqrt{(P_{ij} - P_{i+1,j})^2 + (P_{ij} - P_{i,j+1})^2} \tag{2.6}$$

$$R_{ij} = atan2(P_{ij} - P_{i+1,j}, P_{ij} - P_{i,j+1}) \tag{2.7}$$

The patch is assigned a canonical orientation based on a weighted histogram of orientation values R computed within the patch. Typically, the patch is divided into a regular 4 by 4 grid and orientation histograms are computed for each cell, with 8 bins. This yields a 128 dimensional feature vector describing the patch.

2.3.3 Globalized Probability Boundary

The basic idea of Globalized Probability Boundaries (referred to as gPb) [66] is to combine local information obtained from cues such as brightness, color and texture with global information obtained from spectral partitioning. Brightness, color and texture gradients are computed at multiple scales and are linearly combined into a single multi scale oriented signal mPb. Next, an affinity matrix W encoding the similarity between pixels is obtained from mPb using the intervening contour cue [48]. From W, global information is obtained by computing the eigenvectors v of the linear system

$$(D - W)v = \lambda Dv \tag{2.8}$$

D is defined as $D_{ii} = \sum_j W_{ij}$, $D_{ij} = 0 \ \forall \ i \neq j$. The eigen vectors v obtained from (2.8) are treated as images and are linearly combined to obtain the spectral signal sPb. Finally, weights are learnt for combining the local information from mPb with the global information from sPb to produce a state-of-the-art contour detector.

2.3.4 Ultrametric Contour Map

The gPb output from [66] is used as an input in [13] to obtain a hierarchy of regions. Ultrametric Contour Map [13] defines a duality between such a hierarchy of regions and closed, non-self intersecting weighted contours. Let $E(x, y, \theta)$ denote the gPb Output at an orientation θ. Define $E(x, y) = \max_\theta E(x, y, \theta)$. The regional minima of $E(x, y)$ are taken as seed locations for homogeneous segments and watershed transform is applied to yield the catchment basins P_0 of the minima. P_0 provide the regions at the finest level of the hierarchy.

Using the regions in P_0 as nodes in a graph, a hierarchy is constructed by means of a greedy merging algorithm. The hierarchy is represented by its dual, the Ultrametric Contour Map, a real valued image obtained by weighting each boundary between two regions by its scale of disappearance in the hierarchy.

2.4 Limitations of Hough Voting

This section describes the limitations of current voting approaches to object detection. Approaches such as [60, 46, 68, 51] generate hypotheses by summing over all local votes in a Hough accumulator. The underlying assumption is thus that *objects are a sum of their parts*. This assumption is against the fundamental conviction of Gestalt theory [96] that the whole object is more than the sum of its parts. And indeed, popular semi-local feature descriptors such as SIFT [63] or geometric blur [19] have a large spatial support so that different part descriptors in an image are overlapping and thus mutually dependent. To avoid missing critical image details, a recent trend has been to even increase sampling density which entails even more overlap. Hence, it is critical to model the dependence between the votes of these interest points rather than assuming conditional independence between their votes as in standard Hough voting approaches. Models with richer part dependences such as constellation models [45] or pictorial structures [43] have been proposed to address these issues, however these methods are limited by their complexity (number of parts and the number of parameters per part).

2.5 The Contribution

In our work [102], not only is dependence assumed between parts, but the dependence is used to tackle significant weaknesses of Hough voting that limit its performance: i) unreliability of matching the local interest points to training points based on their surrounding semi-local feature descriptors ii) intrinsically global object properties such as object scale [73] are estimated locally from each interest point. Since the parts are mutually dependent, the correspondence problem (matching query interest points to training interest points) is jointly solved for all dependent points. This improves the reliability of matching. Properties such as object scale are jointly estimated from all the constituents of the group. Therefore, the object properties are more reliable.

- The problem of grouping the dependent parts, finding the correspondences jointly and voting with the dependent points are integrated into a single objective function that is jointly optimized, since each subtask depends on the other two (c.f. Fig. 2.2).

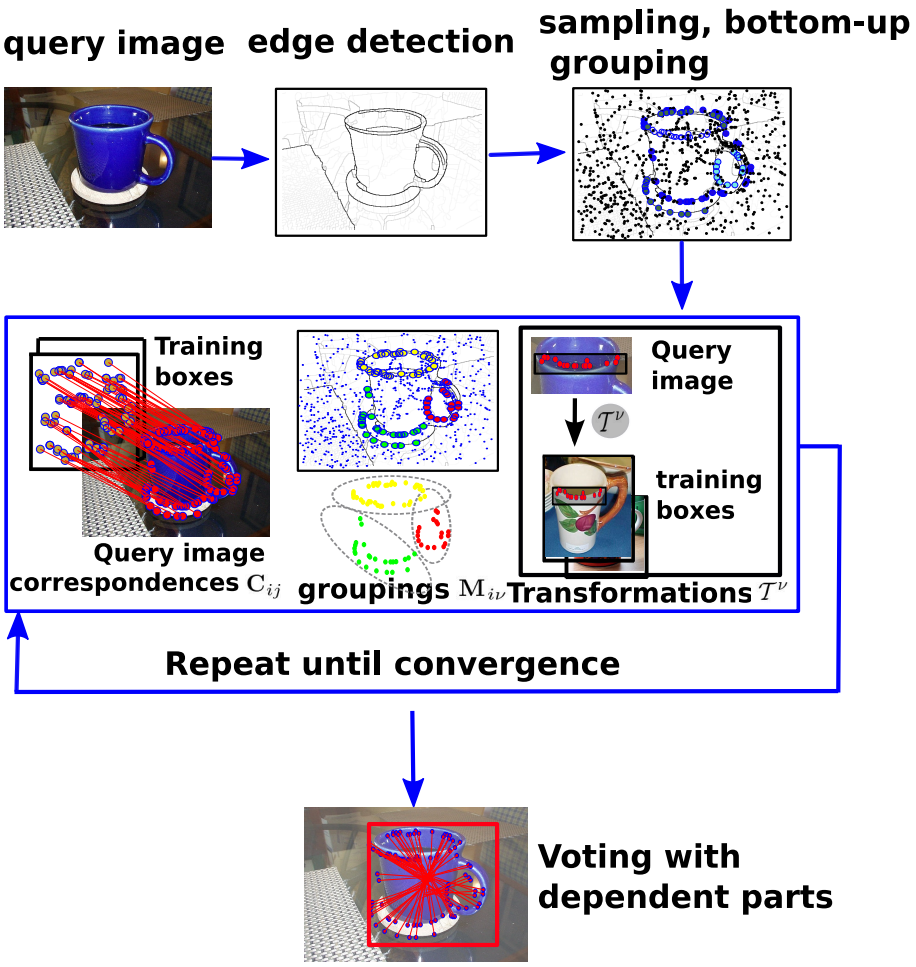


Figure 2.1: Processing pipeline of our voting approach

- The independence assumption between the votes of interest points is no longer made and rather the votes of points belonging to the same group are allowed to influence each other.

  The proposed voting paradigm is depicted in Fig. 2.1

2.6 Comparison of the Proposed Approach with Related Work

Methods such as [60, 46, 73, 68, 51]) let all parts independently cast their votes for the object hypothesis, thereby neglecting part dependence. In contrast to this, our approach models the dependencies between parts by establishing groups and letting all parts in a group jointly find a concerted object hypothesis. Without grouping, [19] transform a complete query image onto a training image. Therefore, this method is constrained to few distractions (e.g. little background clutter) and the presence of only one object in an image. In [46] Hough voting precedes the complex transformation of the complete object from [19] to limit the hypothesis space and reduce the influence of background clutter. However, the voting is limited by assuming independent part votes. The improvements over other voting methods are detailed below.

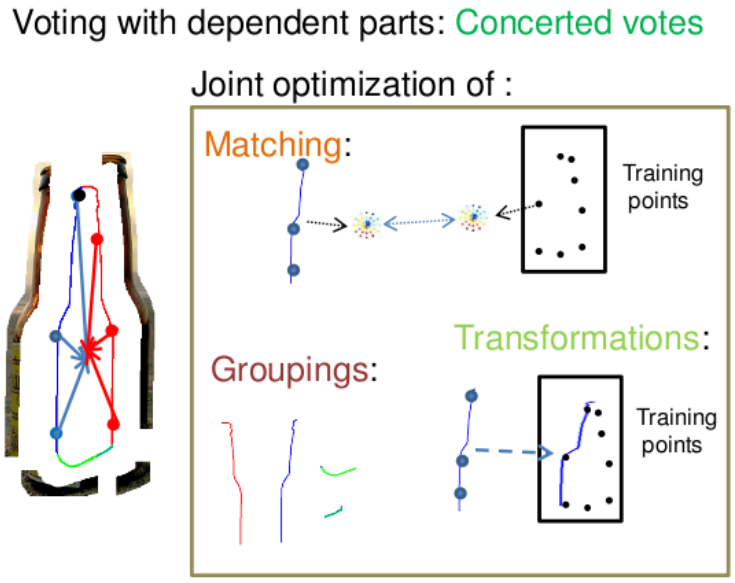


Figure 2.2: To incorporate dependencies into Hough Voting, three entities are jointly optimized i) matching of query interest points to training interest points ii) grouping the query interest points and iii) transformation matrices to be applied to the groups of query interest points to map them with training interest points

Joint Voting of Groups of Dependent Parts: Mutually dependent parts in a group assist each other in finding compatible correspondences and votes, rather than estimating these independently as in standard Hough voting. Thus groups yield votes with significantly less uncertainty than the individual part votes (c.f. Fig. 3.8). Intrinsically global parameters such as object scale are then obtained by global optimization rather than by local estimates (such as local scale estimation in [60, 27]). [73] could only model the uncertainty of each local part. However, the grouping of parts yields reliable estimates.

Joint Optimization of Grouping, Voting, and Correspondences: Identifying and grouping dependent parts, computing joint votes for complete groups, and solving the part correspondence problem are mutually dependent problems of object detection. The dependent problems are tackled jointly by iteratively optimizing a single objective function. Rather than letting each of these factors influence the others, [27] finds groups before using them to optimize correspondences in a model where parts are grouped with their k nearest neighbours. Estrada et al. [38] pursue the simpler problem of exemplar matching by only dealing with grouping and matching consecutively. Several extensions have been proposed to the standard Hough voting scheme, but the critical grouping of dependent parts has not been integrated into voting in any of those approaches. [74] extend the Implicit Shape Model by using curve fragments as parts that cast votes. Without incorporating a grouping stage into their voting, parts are still independently casting their votes. Amit et al. [9] propose a system limited to triplet groupings. In contrast to such rigid groupings, our approach combines flexible numbers of parts based on their vote consistency and geometrical distortion. In contrast to hierarchical grouping approaches, where later groupings build on earlier ones, our method does not require any greedy decisions that would prematurely commit to groupings in earlier stages but rather optimizes all groupings at the same time.

Linear Number of Consistency Constraints: Berg et al. [19] need a quadratic number of consistency constraints between all pairs of parts. In contrast, voting with dependent

groups imposes only a linear number of constraints between parts and the group they belong to. The details can be seen under Sect. 3.1.

Flexible Model vs. Rigid Template: Template-like descriptors such as HoG [34] or [58] have a rigid spatial layout that assumes objects to be box-shaped and non-articulated. Moreover, they require a computationally daunting search through hypothesis space although approximations such as branch-and-bound [57] have been proposed to deal with this issue. On the other end of the modelling spectrum are flexible parts-and-structure models [45, 43]. However, the modelling of part dependencies in [45] becomes prohibitive for anything but very small number of points and [43] restrict the dependencies to a single, manually selected reference part. In contrast to this, we incorporate dependencies in the powerful yet very efficient Hough voting framework. Moreover, we do not rely on pixel accurate labelling of foreground regions as in [60] but only utilize bounding box annotations. In contrast to [46, 19] who transform a query image onto training images using a complex, non-linear transformation we decompose the object and the background into groups and transform these onto the training samples using individual, linear transformations. That way, unrelated regions do not interfere in a single, complex transformation and regions of related parts can be described by simpler and thus more robust, linear models.

CHAPTER 3

HOUGH VOTING WITH GROUPS OF DEPENDENT PARTS

As motivated in Sect. 2.4 of chapter 2, it is necessary to model the dependency between Voting elements to address the limitations of Hough Voting approaches. Sect. 3.1 models the dependency by combining the feature matching term (3.1) commonly used in Hough Voting approaches with additional terms (3.3), (3.4) defined for groups of dependent elements in a query image. The cost function (3.19) guiding the feature matching, grouping the dependent parts and finding the transformations of different groups is derived in a probabilistic framework in Sect. 3.4.

3.1 Modelling the Dependency between the Voting Parts

Hough voting approaches to object detection let all local parts independently vote for a conjoint object hypothesis. Each local part from the query image is independently matched to parts from the training images based on the similarity of their feature descriptors. Let $\mathbf{C}_{ij} \in \{0, 1\}$ denote a matching of the i-th query part to the j-th training part, where $\mathbf{C}_{ij}$ can potentially capture many-to-many-matchings. Matching the feature vector of i-th part of a query image, f_i^Q, to the feature vector of the training part or training codebook vector f_j^T results in a distortion

$$\delta^1(i, j) = \left\| \mathbf{C}_{ij}(f_i^Q - f_j^T) \right\|_2 \tag{3.1}$$

Then the vote of the best matching part from the training images is utilized by the query part to hypothesize object location, scale and other object characteristics in the query image. Thus the voting process heavily relies on the accuracy of the feature descriptor matching between query and training parts. However, independently matching a query part to training part is not quite reliable as shown in Fig. 3.8.

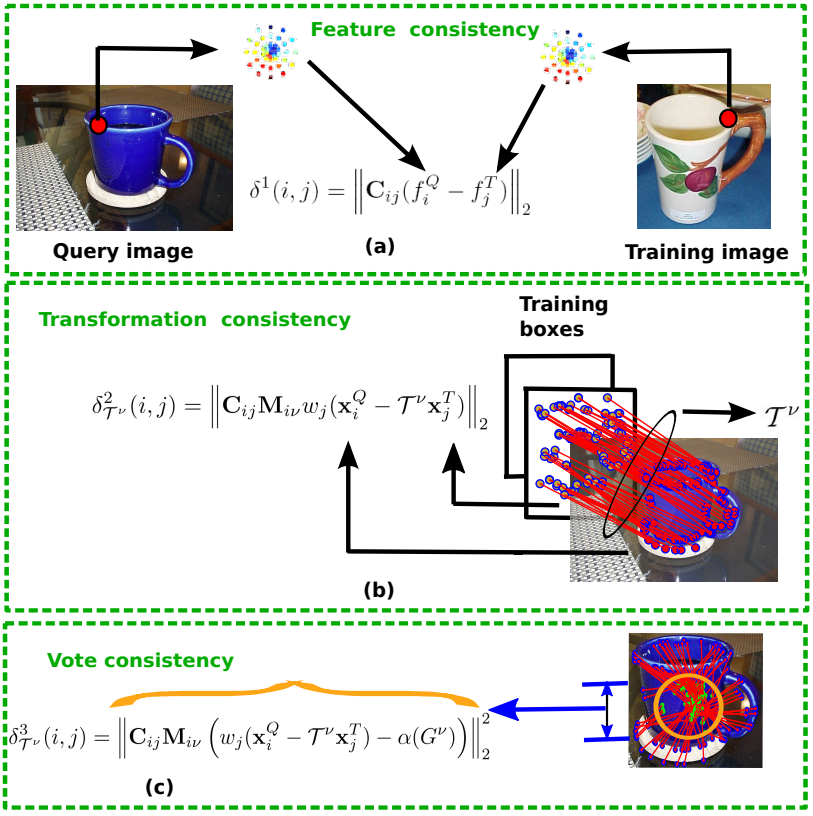


Figure 3.1: Panel a) shows the feature descriptor matching between training and the query part in equation (3.1). Panel b) shows the geometric distortion term resulting out of transforming a group of parts from query image to training parts as in equation (3.3). Panel c) shows the scatter of the votes as indicated by the orange circle depicting equation (3.4)

We improve the matching of query parts to training parts by placing additional constraints on what training parts can be matched to query parts. We note that there are mutual dependencies between query parts because of overlap due to large spatial support of their respective features and also since interest point detection has a bias towards related regions in background clutter [23]. The mutual dependencies between features lead to a bias in the voting process. A set of image descriptors which are all similar due to their large overlap have an unreasonably high influence. Our goal is now to utilize these dependencies which are hindering current voting approaches. Therefore, we let local parts that are mutually dependent form groups. Rather than letting all parts independently of one another, we force all parts of a group to jointly find their corresponding training parts under the constraint that there is minimal geometrical distortion of the query group. $\mathbf{M}_{i\nu}$ is an assignment matrix denoting the assignment of parts to groups, $\mathbf{M}_{i\nu} \in \{0,1\}$,$\sum_\nu \mathbf{M}_{i\nu} = 1$. Let part i belong to a group ν, be denoted as $\mathbf{M}_{i\nu} = 1$. All the parts which belong to group ν are matched to the training parts under the constraint that they undergo the same transformation $\mathcal{T}^\nu$ from query image to the training data, $\mathbf{x}_i^Q \stackrel{!}{=} \mathcal{T}^\nu \mathbf{x}_j^T$. Due to the relatedness of points in a group, transformations should be forced to be simple, eg. similarity transformations

$$\mathcal{T}^{\nu} = \begin{pmatrix} \sigma_x^{\nu}\cos(\theta) & -\sigma_y^{\nu}\sin(\theta) & t_x^{\nu} \\ \sigma_x^{\nu}\sin(\theta) & \sigma_y^{\nu}\cos(\theta) & t_y^{\nu} \\ 0 & 0 & 1 \end{pmatrix} \tag{3.2}$$

In effect, we are decomposing heterogeneous objects into groups of dependent parts so that piecewise linear transformations (one for each group) are sufficient rather than using a complex non-linear transformation for the whole scene as in [19, 46]. Let $G^{\nu} := \{i : \mathbf{M}_{i\nu} = 1\}$ denote all parts in a group ν and $|G^{\nu}| = \sum_i \mathbf{M}_{i\nu}$ denote the number of parts in the group. Let $|G|$ denote the number of groups in a query image. The geometric distortion caused by matching i-th part of a query image to the training part j is given by

$$\delta^2_{\mathcal{T}^{\nu}}(i,j) = \left\| \mathbf{C}_{ij}\mathbf{M}_{i\nu} w_j (\mathbf{x}_i^Q - \mathcal{T}^{\nu}\mathbf{x}_j^T) \right\|_2 \tag{3.3}$$

where w_j is the weight associated with the training part j. We will explain in later section how to obtain different weights for the training parts.

Another observation is that the best matching training parts for the query parts of a group could actually belong to different training images and as such do not have to produce consistent votes. Hence, we impose the constraint that all the matching training parts to a query group should produce consistent object hypotheses. We measure the scatter caused in the group votes from matching i-th part of a query image to the training part j by defining the following vote consistency term,

$$\delta^3_{\mathcal{T}^{\nu}}(i,j) = \left\| \mathbf{C}_{ij}\mathbf{M}_{i\nu} \left(w_j (\mathbf{x}_i^Q - \mathcal{T}^{\nu}\mathbf{x}_j^T) - \alpha(G^{\nu}) \right) \right\|_2^2 \tag{3.4}$$

$$\alpha(G^{\nu}) = \sum_{i \in G^{\nu}, j} \frac{\mathbf{C}_{ij}\mathbf{M}_{i\nu} w_j (\mathbf{x}_i^Q - \mathcal{T}^{\nu}\mathbf{x}_j^T)}{|G^{\nu}|} \tag{3.5}$$

$\alpha(G^{\nu})$ denotes the weighted average vote cast by all the query parts i which belong to group ν. In Fig. 3.1, the orange circle shows the geometric interpretation of equation (3.4).

3.2 Voting with Regions

The terms $\delta^1(i,j), \delta^2_{\mathcal{T}^{\nu}}(i,j), \delta^3_{\mathcal{T}^{\nu}}(i,j)$ defined in Sect. 3.1 can be generalized to a voting framework with regions as voting entities. Let h_i denote the feature vector of a region obtained by concatenation of histograms of colour, texture, etc. Similar to matching two interest points as defined in (3.1), the similarity between regions is measured by the χ^2 distance between their histograms.

$$\delta^1(i,j) = \chi^2(h_i, h_j) \tag{3.6}$$

Next, the transformation residual term, $\delta^2_{\mathcal{T}^{\nu}}(i,j)$, needs to be extended to regions. Let us consider a union of regions that are grouped together. The points $\mathbf{x}_i^Q$ in $\delta^2_{\mathcal{T}^{\nu}}(i,j)$ now

represent the points sampled from the external boundaries of the union. $\mathbf{x}_j^T$ represents the points sampled from the boundaries of regions in training images. The term $\mathbf{C}_{ij}$ in this case indicates the mapping of query regions to training regions.

As in the case for interest points, the term $\delta^3_{\mathcal{T}^\nu}(i,j)$ for the case of regions represents the consistency of votes cast by all the regions assigned to a group.

3.3 Finding Optimal Groups,Correspondences and Transformations

We seek optimal correspondences, group assignments and the transformations of the groups given the locations and the feature vectors of the query and the training parts, i.e. $P(C_i = j, M_i = \nu, \mathcal{T}^\nu | \{\mathbf{x}_i^Q\}, \{\mathbf{x}_j^T\}, \{f_i^Q\}, \{f_j^T\})$ is what we seek to optimize. $C_i = j$ denotes that query part i is matched to the training part j. $M_i = \nu$ denotes that query part i is assigned to group ν. Applying Bayes theorem, we obtain

$$\begin{aligned} &P(C_i, M_i, \mathcal{T}^\nu | \{\mathbf{x}_i^Q\}, \{\mathbf{x}_j^T\}, \{f_i^Q\}, \{f_j^T\}) = \\ &\frac{P(\{\mathbf{x}_i^Q\}, \{\mathbf{x}_j^T\}, \{f_i^Q\}, \{f_j^T\} | C_i, M_i, \mathcal{T}^\nu) P(C_i, M_i, \mathcal{T}^\nu)}{P(\{\mathbf{x}_i^Q\}, \{\mathbf{x}_j^T\}, \{f_i^Q\}, \{f_j^T\})} \end{aligned} \tag{3.7}$$

In Hough voting there is usually the assumption of independence between different features. We avoid this strong assumption and optimize all parts jointly. Looking at the ensemble of all query and training parts, their absolute locations and their feature descriptors can be assumed to be independent, considering all the information provided by all the other descriptors,

$$\begin{aligned} &P(\{\mathbf{x}_i^Q\}, \{\mathbf{x}_j^T\}, \{f_i^Q\}, \{f_j^T\} | C_i, M_i, \mathcal{T}^\nu) = \\ &P(\{\mathbf{x}_i^Q\}, \{\mathbf{x}_j^T\} | C_i, M_i, \mathcal{T}^\nu) P(\{f_i^Q\}, \{f_j^T\} | C_i, M_i, \mathcal{T}^\nu), \end{aligned} \tag{3.8}$$

$$\begin{aligned} P(\{\mathbf{x}_i^Q\}, \{\mathbf{x}_j^T\}, \{f_i^Q\}, \{f_j^T\}) = &P(\{\mathbf{x}_i^Q\}, \{\mathbf{x}_j^T\}) \times \\ &P(\{f_i^Q\}, \{f_j^T\}). \end{aligned} \tag{3.9}$$

Using a uniform prior on $C_i, M_i, \mathcal{T}^\nu$ and substituting equations (3.8) and (3.9) into (3.7), we obtain

$$\begin{aligned} &P(C_i, M_i, \mathcal{T}^\nu | \{\mathbf{x}_i^Q\}, \{\mathbf{x}_j^T\}, \{f_i^Q\}, \{f_j^T\}) \propto \\ &P(C_i, M_i, \mathcal{T}^\nu | \{\mathbf{x}_i^Q\}, \{\mathbf{x}_j^T\}) P(C_i, M_i, \mathcal{T}^\nu | \{f_i^Q\}, \{f_j^T\}) \end{aligned} \tag{3.10}$$

3.4 Deriving the Cost Function

Let us now define some probabilities which help in solving equation (3.10). $P(C_i = j|f_i^Q, \{f_j^T\})$ denotes the probability of matching query part i to training part j, conditioned upon the feature vector of query part f_i^Q and the feature vectors of all training parts $\{f_j^T\}$.

$$P(C_i = j|f_i^Q, \{f_j^T\}) = \frac{\exp(-\beta_1 \delta^1(i,j))}{\sum_j \exp(-\beta_1 \delta^1(i,j))} \tag{3.11}$$

$P(C_i = j, M_i = \nu, \mathcal{T}^\nu = T|\{f_i^Q\}, \{f_j^T\})$ denotes the joint probability of matching query part i to training part j, assigning the query part i to group ν and the transformation matrix $\mathcal{T}^\nu$ of the group being T, conditioned upon the training and the query feature vectors.

$$\begin{aligned} &P(C_i, M_i, \mathcal{T}^\nu|\{f_i^Q\}, \{f_j^T\}) = \\ &P(C_i|M_i, \mathcal{T}^\nu, \{f_i^Q\}, \{f_j^T\}) P(M_i, \mathcal{T}^\nu|\{f_i^Q\}, \{f_j^T\}) \end{aligned} \tag{3.12}$$

The transformation matrix $\mathcal{T}^\nu$ of the group ν and the group assignment of the query part i are independent of the feature vectors. Hence

$$\begin{aligned} &P(C_i, M_i, \mathcal{T}^\nu|\{f_i^Q\}, \{f_j^T\}) \\ &\propto P(C_i|M_i, \mathcal{T}^\nu, \{f_i^Q\}, \{f_j^T\}) = P(C_i|\{f_i^Q\}, \{f_j^T\}) \end{aligned} \tag{3.13}$$

Matching a query part i to training part j based on their feature vectors does not depend on any of the other query parts in the image. Hence

$$P(C_i|\{f_i^Q\}, \{f_j^T\}) = P(C_i|f_i^Q, \{f_j^T\}). \tag{3.14}$$

$P(C_i, M_i, \mathcal{T}^\nu|\mathbf{x}_i^Q, \{\mathbf{x}_j^T\}, \alpha(G^\nu))$ denotes the joint probability of matching query part i to training part j, assigning the query part i to group ν and the transformation matrix $\mathcal{T}^\nu$ of the group being T, conditioned upon the location of the query part, locations of all training parts and the average vote $\alpha(G^\nu)$ of the group. Equations (3.3) and (3.4) both capture the relationship between $C_i, M_i, \mathcal{T}^\nu$ and the locations of the query and the training parts, the average vote of the group. Hence, we define the conditional probability $P(C_i, M_i, \mathcal{T}^\nu|\mathbf{x}_i^Q, \{\mathbf{x}_j^T\}, \alpha(G^\nu))$ as a joint Gibbs distribution of the dissimilarities in (3.3) and (3.4),

$$\begin{aligned} &P(C_i, M_i, \mathcal{T}^\nu|\mathbf{x}_i^Q, \{\mathbf{x}_j^T\}, \alpha(G^\nu)) = \\ &\frac{\exp(-\beta_2 \delta^2_{\mathcal{T}^\nu}(i,j) - \beta_3 \delta^3_{\mathcal{T}^\nu}(i,j))}{\sum_j \sum_\nu \sum_{\mathcal{T}^\nu} \exp(-\beta_2 \delta^2_{\mathcal{T}^\nu}(i,j) - \beta_3 \delta^3_{\mathcal{T}^\nu}(i,j))} \end{aligned} \tag{3.15}$$

Conditioned on the average vote $\alpha(G^\nu)$), the unknowns $C_i, M_i, \mathcal{T}^\nu$ for a query part i are conditionally independent from other query parts,

$$\begin{aligned} &P(C_i, M_i, \mathcal{T}^\nu | \{\mathbf{x}_i^Q\}, \{\mathbf{x}_j^T\}, \alpha(G^\nu)) = \\ &P(C_i, M_i, \mathcal{T}^\nu | \mathbf{x}_i^Q, \{\mathbf{x}_j^T\}, \alpha(G^\nu)). \end{aligned} \tag{3.16}$$

Marginalizing over the possible values of $\alpha(G^\nu)$, yields the joint probability given only the locations of the query part and the training part.

$$\begin{aligned} &P(C_i, M_i, \mathcal{T}^\nu | \{\mathbf{x}_i^Q\}, \{\mathbf{x}_j^T\}) = \\ &\int_t P(C_i, M_i, \mathcal{T}^\nu | \{\mathbf{x}_i^Q\}, \{\mathbf{x}_j^T\}, \alpha(G^\nu) = t) P\left(\alpha(G^\nu) = t\right) \, dt \end{aligned} \tag{3.17}$$

We assume $P\left(\alpha(G^\nu)\right)$ has the form of a Dirac-delta distribution i.e. $P\left(\alpha(G^\nu) = t\right) = 1$ for $t = \alpha(G^\nu)^*$ and for all other t, $P\left(\alpha(G^\nu) = t\right) = 0$. Then equation (3.17) simplifies to

$$\begin{aligned} &P(C_i, M_i, \mathcal{T}^\nu | \{\mathbf{x}_i^Q\}, \{\mathbf{x}_j^T\}) = \\ &P(C_i, M_i, \mathcal{T}^\nu | \mathbf{x}_i^Q, \{\mathbf{x}_j^T\}, \alpha(G^\nu)^*) \end{aligned} \tag{3.18}$$

Subsequently, we will derive $\alpha(G^\nu)^*$.

From equations (3.10), (3.11), (3.13), (3.15) and (3.18) , we obtain

$$\begin{aligned} &P(C_i, M_i, \mathcal{T}^\nu | \{\mathbf{x}_i^Q\}, \{\mathbf{x}_j^T\}, \{f_i^Q\}, \{f_j^T\}) \propto \\ &\frac{\exp(-\beta_1 \delta^1(i,j))}{\sum_j \exp(-\beta_1 \delta^1(i,j))} \times \\ &\frac{\exp(-\beta_2 \delta^2_{\mathcal{T}^\nu}(i,j) - \beta_3 \delta^3_{\mathcal{T}^\nu}(i,j))}{\sum_j \sum_\nu \sum_{\mathcal{T}^\nu} \exp(-\beta_2 \delta^2_{\mathcal{T}^\nu}(i,j) - \beta_3 \delta^3_{\mathcal{T}^\nu}(i,j))} \\ &\propto \exp(-\beta_1 \delta^1(i,j)) \exp(-\beta_2 \delta^2_{\mathcal{T}^\nu}(i,j) - \beta_3 \delta^3_{\mathcal{T}^\nu}(i,j)) \\ &= \exp(-\beta_1 \delta^1(i,j) - \beta_2 \delta^2_{\mathcal{T}^\nu}(i,j) - \beta_3 \delta^3_{\mathcal{T}^\nu}(i,j)) \end{aligned} \tag{3.19}$$

Maximizing $P(C_i, M_i, \mathcal{T}^\nu | \{\mathbf{x}_i^Q\}, \{\mathbf{x}_j^T\}, \{f_i^Q\}, \{f_j^T\})$ is equivalent to minimizing $\beta_1 \delta^1(i,j) + \beta_2 \delta^2_{\mathcal{T}^\nu}(i,j) + \beta_3 \delta^3_{\mathcal{T}^\nu}(i,j)$ over the parameters M_i, C_i, T^ν.

3.5 Optimization

Jointly solving for $M_i^*, C_i^*, T^{\nu *}$ from equation (3.19) by a brute-force search is intractable as illustrated by the following example. Consider a query image with 10^3 parts and assume there are 10 groups. Moreover, let the training set consist of 10^5 training parts. Then each query part has 10 possible group assignments, and 10^5 possible training part matches.

Since the transformation matrix of each group is a function of group assignments of query parts and the correspondences between query and training parts, transformation matrix of each group has to be estimated 10^5 times.

However, given two out of the three entities M_i^*, C_i^*, $T^{\nu *}$, it is possible to efficiently estimate the third entity. Hence, we adopt an iterative approach where we sequentially update each of the three entities,

$$M_i^* = \underset{\nu}{\operatorname{argmin}} \left(\beta_2 \delta^2_{\mathcal{T}^\nu}(i, C_i) + \beta_3 \delta^3_{\mathcal{T}^\nu}(i, C_i) \right), \tag{3.20}$$

$$T^{\nu *} = \underset{T^\nu}{\operatorname{argmin}} \left(\beta_2 \delta^2_{\mathcal{T}^\nu}(i, C_i) + \beta_3 \delta^3_{\mathcal{T}^\nu}(i, C_i) \right), \tag{3.21}$$

$$C_i^* = \underset{j}{\operatorname{argmin}} \left(\beta_1 \delta^1(i, j) + \beta_2 \delta^2_{\mathcal{T}^{M_i}}(i, j) + \beta_3 \delta^3_{\mathcal{T}^{M_i}}(i, j) \right). \tag{3.22}$$

Using the example introduced in the beginning of this optimization section, we illustrate the number of computations needed in the iterative approach. To estimate M_i^* in equation (3.20), we search over 10 possible group assignments. We estimate $T^{\nu *}$ in equation (3.21) once for every iteration (using Levenberg-Marquardt algorithm). To estimate C_i^* in equation (3.22), we search over 10^5 training parts which is feasible using approximate nearest neighbour search.

In the brute-force search, we need to estimate $T^{\nu *}$ 10^5 times for each group using Levenberg-Marquardt algorithm. In contrast, the Levenberg-Marquardt algorithm is only used once per group per each iteration, thus reducing complexity drastically.

We estimate $\alpha(G^\nu)^*$ for the Dirac-delta distribution of $P\left(\alpha(G^\nu)\right)$ in equation (3.17) in each iteration from the current estimates of M_i^*, C_i^*, and $T^{\nu *}$ using (3.5).

Equations (3.20),(3.21) and (3.22) are iterated until convergence of M_i^*, C_i^*, and $T^{\nu *}$. In each step of the iteration, we are solving for optimal parameters which increase the probability in equation (3.19) compared to the previous iteration. We have already seen in the illustrative example that the number of possible values for C_i and M_i are finite. Since $\mathcal{T}^\nu$ is a function of C_i and M_i, the number of possible values of $\mathcal{T}^\nu$ are also finite. Hence, the combined search space of the parameters is finite and the iterations are bound to converge to a local optimum. In practice, convergence is achieved within 5 iterations.

3.6 Avoiding Early Commitment

The term $\mathbf{C}_{ij}$ introduced earlier represents a many-to-one matching, i.e, each query part can be associated with one training part. During the voting phase, each query part would then cast a single vote for an object hypothesis. To avoid such early commitment by choosing the best possible match, we relax $\mathbf{C}_{ij}$ to include many-to-many matching. Each query part i now has k training parts associated to it. In this scenario, $\mathbf{C}_{ij} \in \{0, 1\}$ and $\sum_j \mathbf{C}_{ij} = k$. During the voting phase, each query part would then cast k votes for object hypotheses. For finding the correspondence $\mathbf{C}_{ij}$ for the i-th query part, we choose k best matching training parts according to equation (3.22). For finding $\mathbf{M}_{i\nu}$, we now have to search over all groups and the k training correspondences of query part i,

$$M_i^* = \underset{\nu}{\operatorname{argmin}} \min_{j:\mathbf{C}_{ij}=1} \left(\beta_2 \delta^2_{\mathcal{T}^\nu}(i, j) + \beta_3 \delta^3_{\mathcal{T}^\nu}(i, j) \right). \tag{3.23}$$

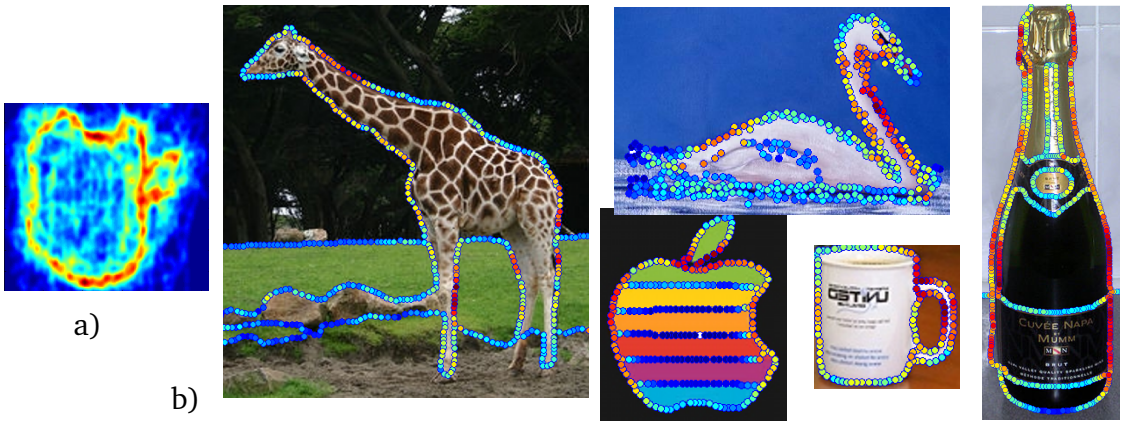


Figure 3.2: a) shows the probability density map (see Sect. 3.8 for details) generated for possible locations of relevant training parts of mugs in ETHZ images. Red indicates high probability and blue indicates low probability. Such probability density maps are used to obtain weights for points sampled on various objects in panel b). High weight is associated to characteristic parts on the outline of the objects (the neck and the rear of the swan, the sides of the bottle, tip and the sides of the apple, handle of the mug, the neck and the upper parts of legs of the giraffe) whereas low weight is associated to the interiors and exteriors of the object.

For finding $\mathcal{T}^\nu$, we use the best correspondence for each query part i in equation (3.21). Again, the search space for the parameters is finite in this case too. Also, in each iteration, the probability in equation (3.19) is increased compared to previous iteration. Hence, the iterations are again bound to converge to a local optimum.

3.7 Initialization of Groups and Correspondences

Good initialization of groups and correspondences is an important step for quick convergence to an optimal solution. Object detection in a query image starts by computing a probabilistic edge map [13] and uniformly sampling edge points. Next, we perform a bottom-up grouping on the probabilistic edges which serves as an initialization for the groups in the query image ν. Two edge points i, i' are considered to be connected on level s of the hierarchical ultra-metric contour map of [13], if they are on the boundary of the same region on this level. Let $1 = \mathbf{A}^s_{ii'} \in \{0, 1\}$ denote this case. Averaging over all levels, $\bar{\mathbf{A}}_{ii'} \propto \sum_s \mathbf{A}^s_{ii'}$, yields a similarity measure between points and pairwise clustering (using Ward's method) on this similarity matrix produces a grouping $\mathbf{M}_{i\nu}$ which we use to initialize the optimization of (3.19). We initialize $\mathbf{C}_{ij}$ by finding the best matching training part for each query part exclusively utilizing the geometric blur descriptors in equation (3.1).

Algorithm 3.1 Voting with groups of dependent parts: Joint optimization of groupings, correspondences, and transformations.

Input: • parts from query image: f_i^Q, $\mathbf{x}_i^Q$,
• UCM-connectivity [13] $\bar{\mathbf{A}}_{ii'}$
• parts from all training images: f_j^T, $\mathbf{x}_j^T$
Init: • pairwise clustering on $\bar{\mathbf{A}}_{ii'} \rightarrow \mathbf{M}_{i\nu}()$

do
  $C_i \leftarrow$ k best matching training parts from equation
  $\left(\beta_1\delta^1(i,j) + \beta_2\delta^2_{\mathcal{T}^\nu}(i,j) + \beta_3\delta^3_{\mathcal{T}^\nu}(i,j)\right)$
  $M_i \leftarrow \text{argmin}_{(\nu,\, j:\mathbf{C}_{ij}=1)} \left(\beta_2\delta^2_{\mathcal{T}^\nu}(i,j) + \beta_3\delta^3_{\mathcal{T}^\nu}(i,j)\right)$
  $\mathcal{T}^\nu \leftarrow \text{argmin}_{\mathcal{T}} \left(\beta_2\delta^2_{\mathcal{T}^\nu}(i,C_i) + \beta_3\delta^3_{\mathcal{T}^\nu}(i,C_i)\right)$

until convergence
$\mathcal{H}^{\text{grp}}(c,\mathbf{x},\sigma) \leftarrow$ Apply
$\left\{(\mathbf{x}^h,\sigma^h)^\top\right\}_h \leftarrow$ Local minima of $\mathcal{H}^{\text{grp}}$

3.8 Learning the Relevance of Training Parts

In the training set, there are repeatable parts and non-repeatable, spuriously occurring parts. Repeatable parts occur consistently in all the training images and they can be matched quite reliably across different images based on their descriptors (3.1). The training parts sampled from the handles of the mugs are an example for this type. Non-repeatable, spuriously occurring parts occur randomly in images and cannot be reliably matched to other parts based on their features. Parts sampled from the face of a mug or the background are an example for this type. Therefore, we associate a weight with each training part such that the stable training parts are assigned high weights and the spurious parts are assigned low weights.

In the training phase, all the images are resized so that the training object has same length across its diagonal. Training parts are then sampled from the resized version of the training images. We place a 2-d Gaussian at the location of each training part. The σ of the Gaussian is proportional to the Geometric blur matching distance between the training part and rest of the parts sampled from the other training images in its spatial vicinity. We sum up the Gaussian distributions to obtain a probability density for all possible locations of training parts which is shown in the first panel of Fig. 3.2. We utilize the probability density as a look-up table to obtain the weights of the training parts in the images.

Fig. 3.2 shows the weights of interest points from some of the training images(one from each object category of ETHZ shape database). Red indicates that the points have high weight and blue indicates the low weight associated with the interest points.

3.9 Hough Voting with Groups

After finding optimal groupings, group transformations, and correspondences, the votes from all groups have to be combined. In standard Hough voting, the votes of all parts are summed up, thus treating them as being independent, c.f. the discussions in [97, 7].

Figure 3.3: Left plot in panels (a) and (b) shows standard Hough voting which assumes mutual independence between features. Right plot in panels (a) and (b) shows the voting after joint optimization of correspondences, groups, and votes.

In our setting, all mutually dependent parts are combined in the same group. Evidently, all parts in a group are coupled by using the same transformation matrix $\mathcal{T}^\nu$ and the jointly optimized correspondences $\mathbf{C}_{ij}$. The joint optimization of correspondences and transformations forces these dependent parts to agree upon individual votes $(\mathbf{x}, \sigma)^\top$ that are mutually compatible,

$$(\mathbf{x}, \sigma)^\top = (\mathbf{x}_i^Q - \mathcal{T}^\nu \mathbf{x}_j^T C_i + t^\nu, \sigma^\nu)^\top, \tag{3.24}$$

where t^ν and σ^ν are the translation and scaling component of $\mathcal{T}^\nu$. The relevance or weight, $\mathcal{R}(G^\nu)$, of the vote from a group is given by

$$\mathcal{R}(G^\nu) = \sum_{i \in \nu} \left(\beta_1 \delta^1(i,j) + \beta_2 \delta^2_{\mathcal{T}^\nu}(i,j) + \beta_3 \delta^3_{\mathcal{T}^\nu}(i,j) \right). \tag{3.25}$$

The Hough accumulator for the voting of groups is obtained by summing over independent groups rather than over dependent parts as in standard Hough voting. Since groups are mutually independent, their summation is justified. Analogous to (2.2) we obtain

$$\begin{aligned} \mathcal{H}^{\text{grp}}(c, \mathbf{x}, \sigma) \quad &\propto \sum_\nu \frac{1}{|G^\nu| \mathcal{R}(G^\nu)} \\ \times \sum_{i \in G^\nu} \sum_j \mathbf{C}_{ij} \cdot P(\mathbf{x} - [\mathbf{x}_i^Q &- \mathcal{T}^\nu \mathbf{x}_j^T + t^\nu], \sigma - \sigma^\nu) \end{aligned}$$

where $P(\bullet)$ is obtained using the balloon density estimator [30] with Gaussian Kernel K, Kernel bandwidth b, and distance function in scale space $d : \mathbb{R}^3 \times \mathbb{R}^3 \mapsto \mathbb{R}$,

$$\begin{aligned} &P(\mathbf{x} - [\mathbf{x}_i^Q - \mathcal{T}^\nu \mathbf{x}_j^T + t^\nu], \sigma - \sigma^\nu) = \\ &K\left(\frac{d\left[(\mathbf{x}, \sigma)^\top; (\mathbf{x}_i^Q - \mathcal{T}^\nu \mathbf{x}_j^T + t^\nu, \sigma^\nu)^\top \right]}{b(\sigma)} \right) \end{aligned} \tag{3.26}$$

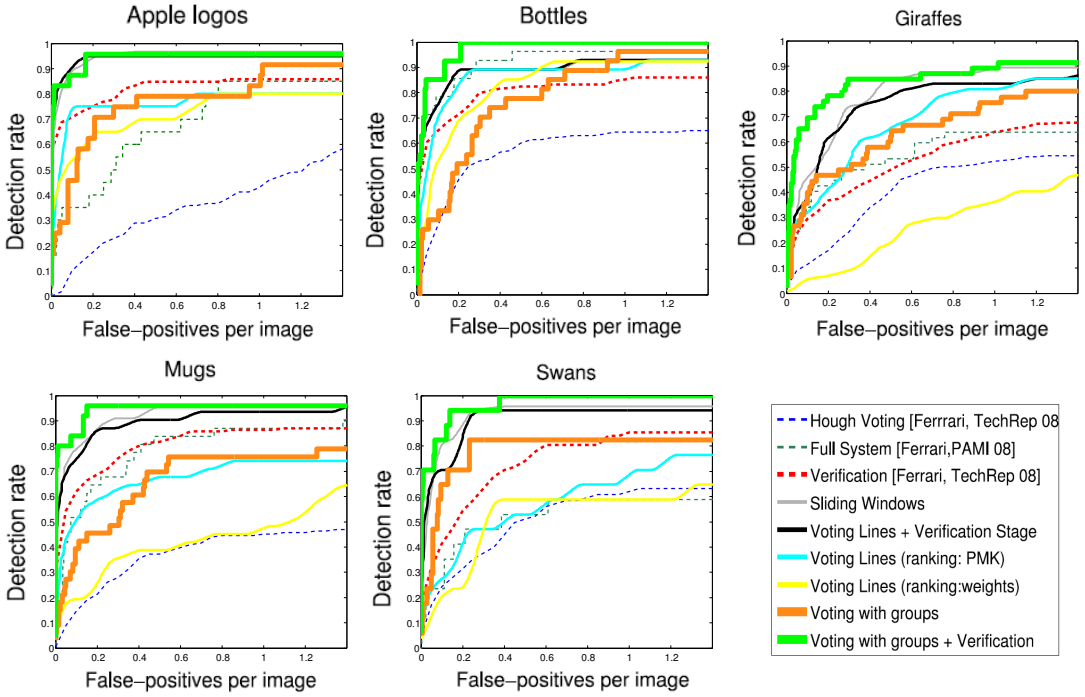


Figure 3.4: Detection performance measured in terms of fppi curves. On average our voting approach yields a 30% higher performance in terms of detection rate at 1 fppi compared to standard Hough voting and improves line voting [73] by 20%.

3.10 Experiments

We evaluate our approach on 3 standard benchmark datasets, i.e., the ETHZ shape dataset, Inria Horses dataset and Shape-15 dataset, which have been widely used for comparing voting-based methods [68, 73, 51, 102]. The datasets feature significant scale changes, intra-class variation, multiple-objects per image, intense background clutter, and out-of-plane rotations. We use the latest experimental protocol of Ferrari et al. [46]. We split the dataset into training and test images as suggested by [68]. The codebook is constructed from half the positive samples of a category. No negative training images are used and all remaining images are used for testing. For INRIA shape dataset, 50 horse images are used for training and the remaining 120 horse images plus 170 negative images are used for testing. For Weizmann and Shape-15 datasets, the images are already annotated as training and test images. In all experiments, the detection performance is measured using the PASCAL VOC criterion [46] (requiring the ratio of intersection and union of predicted and ground-truth bounding box to be greater than .5)

3.10.1 ETHZ Shape Dataset

Tab. 3.1 compares our approach with state-of-the-art voting methods on ETHZ. Voting with our groups of dependent parts outperforms all current voting based approaches. We achieve a gain of 30% over the Hough voting in [46], an improvement of 22% over [68], and 20% higher performance than [73]. Even compared with the local sliding window classification in [73] (PMK re-ranking) we obtain a slightly higher performance (4.4%). The PMK re-ranking is a separate classifier that performs verification of votes. Thus our

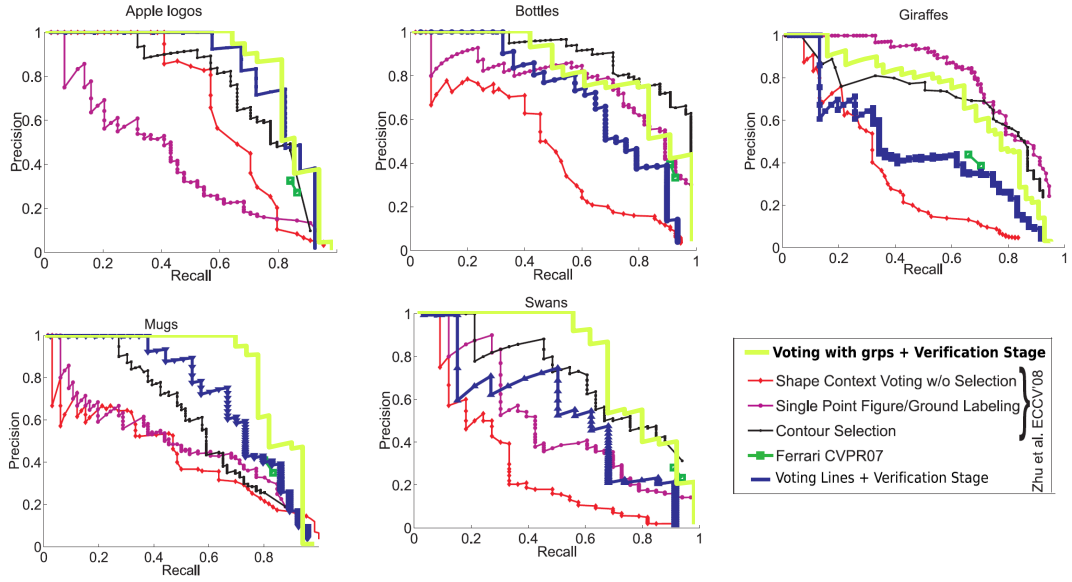


Figure 3.5: Comparing our voting+verification with the supervised approach [105]. [46] has shown that our training scenario is significantly harder and yields 13% lower recall at .4 fppi

voting method alone not only improves current Hough voting approaches, but also produces results beyond those of the verification stage of some of the methods. Since [80] report their detection rate only at 0.4 fppi, we are including their results in the text here instead of reporting them in Tab. 3.2. They achieve detection rates of $87.3, 87.3, 83.1, 85.8, 79.4$ for Applelogos, Bottles, Giraffes, Mugs and Swans respectively at 0.4 fppi. Compared to [102], we obtain a gain of 3%. This gain is attributed to two factors namely, the weights being utilized for the training parts and each part casting k votes for an object hypothesis.

The primary focus of this work is to improve Hough voting by modelling part dependence. Nevertheless, we also investigate the combined detector consisting of voting and a verification stage. We train an SVM model based on HOG feature vectors [34] and use it during the verification stage to search in the neighbourhood of hypotheses obtained from the voting stage. We obtain the negative examples for training the SVM model from the false positives generated by the voting stage in training images. Our results compare favourably with sliding window classification in [73]. We obtain a gain of 5.33% over sliding windows at 0.3 fppi. Compared to the best verification systems [68], we obtain a gain of 2.33% at 0.3 fppi. The voting with dependent regions paradigm described in Sect. 3.2 yielded a gain of 18% at 0.3 fppi over voting independently with each region. Thus, our grouping process is beneficial in a variety of voting scenarios. Although approaches such as [75, 64] yield state-of-the-art detection performance on ETHZ, they heavily rely upon long training contours for detecting the objects. For datasets such as INRIA Horses and Shape 15, this dependence on long training contours could lead to problems with such approaches.

We also compare the supervised methods of [105] which use one hand drawn object model against our detector (which only needs training images with bounding boxes). Without requiring the supervision information of [105], we are dealing with a significantly harder task. [46] showed a performance loss of 13% at 0.4 fppi. Nevertheless, we perform better on 3 out of 5 categories (actual values of [105] are unavailable). We obtain average precision of 0.89 for Apples, 0.85 for Bottles, 0.67 for Giraffes, 0.85 for Mugs, 0.81 for Swans. The

Approach	Applelogos	Bottles	Giraffes	Mugs	Swans	Average
$\mathcal{H}^{\text{grp}}$	**85.71**	**96.3**	**75.56**	**75.76**	**82.35**	**83.13**
PAS Voting [46]	43	64.4	52.2	45.1	62	53.3
M^2HT [68]	85	67	55	55	42.5	60.9
Voting lines [73]	80	92.4	36.2	47.5	58.8	63

Table 3.1: Performance comparison of $\mathcal{H}^{\text{grp}}$ **voting** with other state-of-the-art approaches on ETHZ Shape Classes at the detection rate of 1 fppi. The approach of [68] use positive as well as negative samples for training whereas we use only positive samples for training. Our voting yields a 30% higher performance than the Hough voting in [46], 22% gain over max-margin Hough voting [68], and 20% gain over line voting [73], thus significantly improving the state-of-the-art in voting.

Approach	Applelogos	Bottles	Giraffes	Mugs	Swans
$\mathcal{H}^{\text{grp}}$ Full sys	**95.83 /95.83**	**100 /100**	84.78/84.78	**96.44 /96.44**	94.12/ **100**
Full system [73]	95/95	89.3/89.3	70.5/75.4	87.3/90.3	94.1/94.1
Sliding Windows	95.8/ **96.6**	89.3/89.3	73.9/77.3	91/91.8	**94.8 /95.7**
Full system [46]	77.7/83.2	79.8/81.6	39.9/44.5	75.1/80	63.2/70.5
M^2HT [68]	95/95	92.9/96.4	**89.6 /89.6**	93.6/ **96.7**	88.2/88.2
[78]	93.3/93.3	97/97	79.2/81.9	84.6/86.3	92.6/92.6

Table 3.2: Performance comparison of $\mathcal{H}^{\text{grp}}$ **voting + verification** with other state-of-the-art approaches on ETHZ Shape Classes at 0.3/0.4 detection rates. $\mathcal{H}^{\text{grp}}$ Full sys yielded an average detection rate of 94.23/95.41 compared to the next best performance of 91.9/93.2 yielded by M^2HT [68]

mean average precision is 0.81.

Apart from initializing the grouping of interest points as in Sect. 3.7, we have also experimented with other possibilities for the grouping initializations. One such instance was to use the $\mathbf{A}^s_{ii'}$ matrix at a given level s for initializing the grouping. The results have proven to be robust with respect to initial grouping.

3.10.2 Multiple Training and Test Splits

We have also experimented with 5 splits of training and test images for each of the categories in ETHZ. For each split, we randomly chose the training and test images, however keeping the number of images the same as per the protocol in [46]. None of the 5 resultant splits is the exact replica of the split used to generate the results in Tab. 3.1. We computed the mean and standard deviation of detection rates for each category at 0.3/0.4 fppi. The results are summarized in Tab. 3.3. The low standard deviation indicates the stability of the approach over multiple splits.

Detection rate	Applelogos	Bottles	Giraffes	Mugs	Swans
at 0.3 fppi	97.22 ± 2.53	95.57 ± 4.06	80.95 ± 2.26	91.48 ± 2.86	97.64 ± 2.22
at 0.4 fppi	97.22 ± 2.53	97.05 ± 1.66	82.24 ± 1.84	92.89 ± 3.42	97.64 ± 2.22

Table 3.3: Performance of $\mathcal{H}^{\mathbf{grp}}$ **voting + verif** on ETHZ Shape Classes over 5 splits of random training and test images is reported in terms of mean detection rate and the corresponding standard deviation at 0.3 and 0.4 fppi

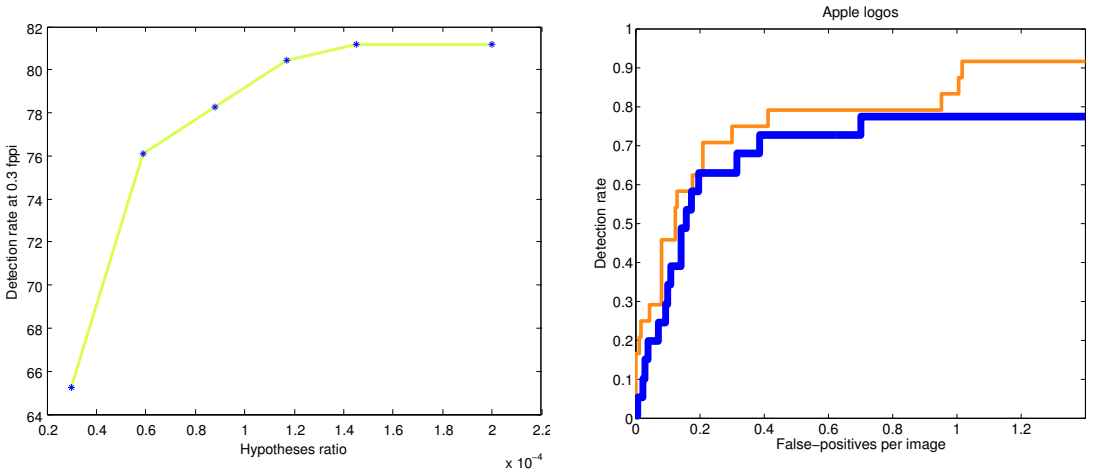


Figure 3.6: a) Detection rate at 0.3 fppi vs the hypotheses ratio for the category of Giraffes. See Sect. 3.10.6 for details. b) The orange curve shows the performance on Applelogos without occlusion and the blue curve shows the performance on synthetic occlusions. See Sect. 3.10.5 for details.

3.10.3 Reliability of Individual Votes vs Voting with Groups

Let us now compare the reliability of votes from individual parts with the reliability of object hypotheses produced by our groupings. Therefore, we map object query features (features from within the ground-truth bounding box) onto the positive training samples and we do the same for background query features. By comparing the matching costs we see how likely positive query features are mistaken to be background and vice versa. Then we are doing the same for groups, i.e. groupings (3.20) from the object and from the background are mapped onto positive training samples. Fig. 3.8 shows that groups have a significantly lower error rate $\mathcal{R}$ (30% vs. 77%) to be mapped onto wrong training samples. Thus group votes are significantly more reliable.

3.10.4 False Positives

Fig. 3.7 shows the false hypotheses obtained by standard Hough voting, the false hypotheses obtained by our voting with groups and the false hypotheses obtained from our full system under panels a),b) and c) respectively. We show 5 false positives for each method. Panel d) shows the edge image for the corresponding numbered figure from panel c). The index below each figure in panels a),b) and c) indicates the rank of the false hypothesis amongst all the hypotheses sorted according to their scores. A good detection system

Figure 3.7: This figure shows the top 5 false hypotheses when searching for *bottles* and *swans* in the test images of ETHZ shape dataset. Results are shown for a) standard Hough voting b) our Hough voting with groups c) our full system. d) shows the edge image for the corresponding numbered figure from c). The numbers below each figure indicate the occurrence of the false hypothesis amongst all the hypotheses sorted according to their score and we notice a significant shift in this ranking from a) to b) to c).

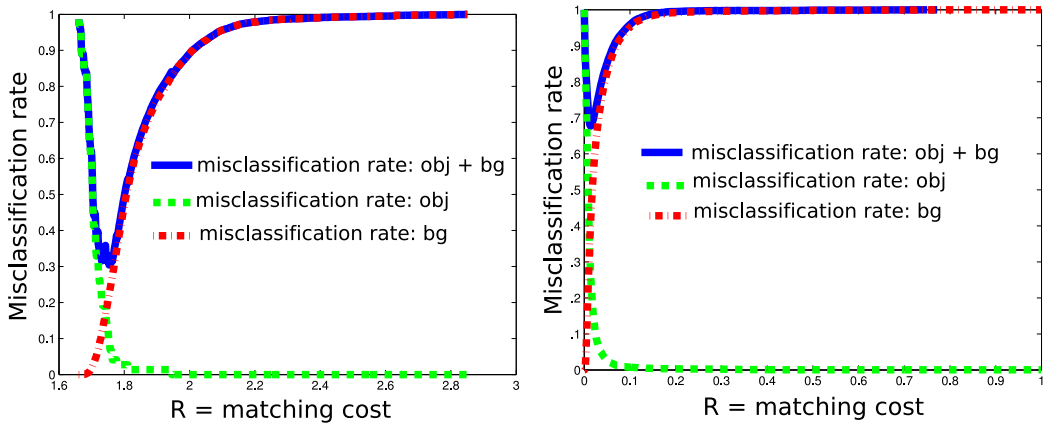


Figure 3.8: Reliability of groups,left plot vs. parts(singleton groups),right plot. The plots show the misclassification rate of groups and parts for different matching cost $\mathcal{R}$. The minimal error rate for parts is 77%, for groups 30% thereby underlining the increased reliability of groups.

would produce true positives with higher scores than the false positives. When the detection scores are sorted, false positives would thus get a higher index compared to true positives. For the standard Hough voting, the index for the false positives ranges from 2 to 9 for bottles and swans. For our voting with groups approach, the false positive index ranges from 8 to 20 for bottles and from 8 to 15 for swans, thus indicating the improvement in the performance compared to standard hough voting.

The false hypotheses in panel b) of Fig. 3.7 for bottles can be explained by long vertical lines (images 12, 16 and 20) and slanted long lines (image 8) which match well with the interest points from the sides of the training bottle images and hence casting strong votes for the presence of bottle. Image 11 is counted as a false positive because it fails the Pascal criterion of 50% overlap with the bounding box annotations provided along with the dataset. For the same reason, image 14 in panel c) is counted as a false positive. The rest of the false positives from panel c) can be explained by the presence of vertical lines (as can be seen from the edge maps in panel d) which populate the vertical orientation bins in the HoG feature vector employed in the verification stage.

The false positives in panel b) of Fig. 3.7 for swans can be explained by regions matching the bottom and the lower half of the swan's neck (images 13 and 15) and regions matching to the rear of the swan (image 8). 12 is counted as false positive because it has not been annotated in the dataset. The false positives in panel c) can be explained by the presence of edges in the orientation bins corresponding to the neck of the swan.

Fig. 3.9 and Fig. 3.10 show the top 5 hypotheses (including both true positives and false positives) obtained by standard Hough voting (panel b), the top 5 hypotheses obtained by our voting with groups (panel c) and the top 5 hypotheses obtained from our full system (panel d). The true positives are shown in green and the false positives are shown in red. The test image is shown in panel a). In Fig. 3.9, there are 6 true positives in the test image (only the mugs with their handles to the right side are considered as true positives). In the top 5 hypotheses, voting with groups has 3 true positives. All the top 5 hypotheses obtained from our full system are true positives. With standard hough voting, only 1 hypothesis out of the top 5 hypotheses is infact a true positive.The blue plot in each of the panels b), c) and d) is obtained from the sorted hypotheses scores for all the test images

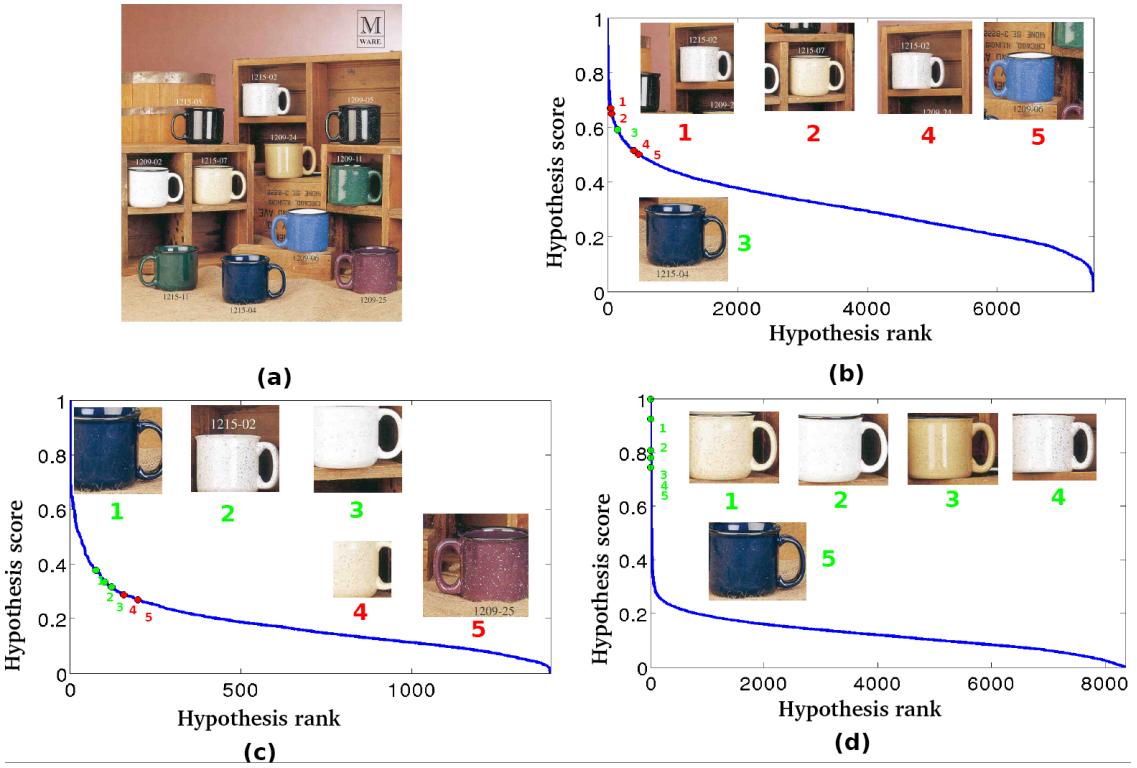


Figure 3.9: a) the input image b) top 5 hypotheses obtained from standard hough voting c) top 5 hypotheses obtained from our voting d) top 5 hypotheses obtained from our full system. The scores of the top 5 hypotheses are plotted along with the scores for all the hypotheses in all the test images. Hypotheses with numbers marked in green indicate true hypotheses and the hypotheses with numbers marked in red indicate false hypotheses.

of an object category. The scores obtained from each of the methods have been linearly scaled to map to $[0, 1]$ range (the y-axis of the plots).

The scores of the top 5 hypotheses are indicated on the blue plot with red and green markers. In panel b), there are two false positives which have higher score than the true positive. This is improved in voting with groups , where all the true positives have a higher score than the false positives (panel c). In Fig. 3.10, there is 1 true positive. The standard Hough voting does not have the true positive amongst the top 5 hypotheses, whereas voting with groups has the true positive as its best hypothesis. The left-most section of the blue curves, which is the most interesting section, has highest slope in d), less in c) and least in b). Note that the x-axis is rescaled. High slope indicates a sharp separation between the hypotheses perceived to be objects and the hypotheses perceived to be non-objects by the approach. In general, discriminative approaches are expected to produce better separation between positives and negatives and thus a higher slope than generative approaches. The blue curve from panels b) through d) confirms this thesis.

3.10.5 Occlusions

We have also tested the performance of our method in the case of occlusions. For this purpose, we have selected the category of Applelogos and created synthetic occlusions in

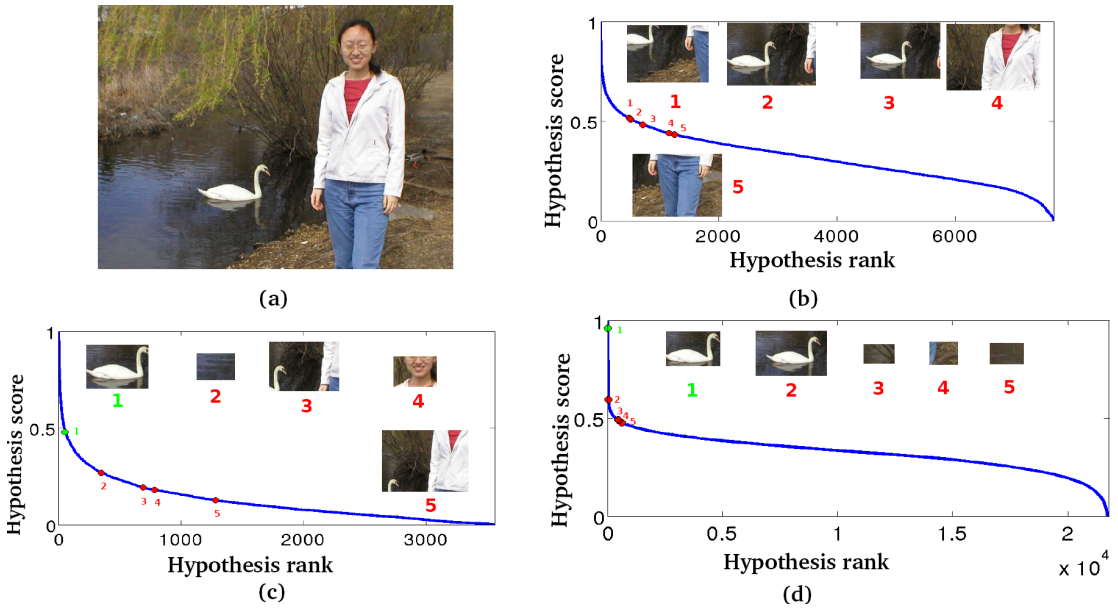


Figure 3.10: a) the input image b) top 5 hypotheses obtained from standard hough voting c) top 5 hypotheses obtained from our voting d) top 5 hypotheses obtained from our full system. The scores of the top 5 hypotheses are plotted along with the scores for all the hypotheses in all the test images. Hypotheses with numbers marked in green indicate true hypotheses and the hypotheses with numbers marked in red indicate false hypotheses.

test images by removing sampled interest points in a randomly selected patch covering about 25% of the object bounding box in test images. In Fig. 3.6, we report the detection rate vs fppi curve for the category of Applelogos, with and without occlusions. The result shows the robustness of the method in the presence of occlusions.

3.10.6 Computational Complexity

Sliding windows approach has to search over 10^4 hypotheses whereas our approach only needs on the order of 10 candidate hypotheses to achieve a performance better than that of sliding windows. Consequently, the gain in computational performance of our approach is between two and three orders of magnitude. Compared to preprocessing steps such as extraction of probabilistic edge maps and computation of geometric blur, our grouping, voting and correspondence optimization has insignificant running time (on the order of a few seconds).

Also, the common practice in most of the voting approaches is to obtain a short-list of object hypotheses and pass them to a verification stage. This two stage approach boosts the detection performance. Our voting stage greatly cuts down the number of hypotheses needed to be passed to a verification stage for achieving high detection rate. In order to demonstrate this utility, we have computed the detection accuracy achieved vs the ratio of number of hypotheses passed from voting to verification stage and the number of hypotheses evaluated in a sliding window approach. Fig. 3.6 shows the detection rate at 0.3 fppi for the category of Giraffes vs the hypotheses ratio. The curve saturates for a relatively small ratio of about 0.0001 indicating that our voting stage is indeed robust.

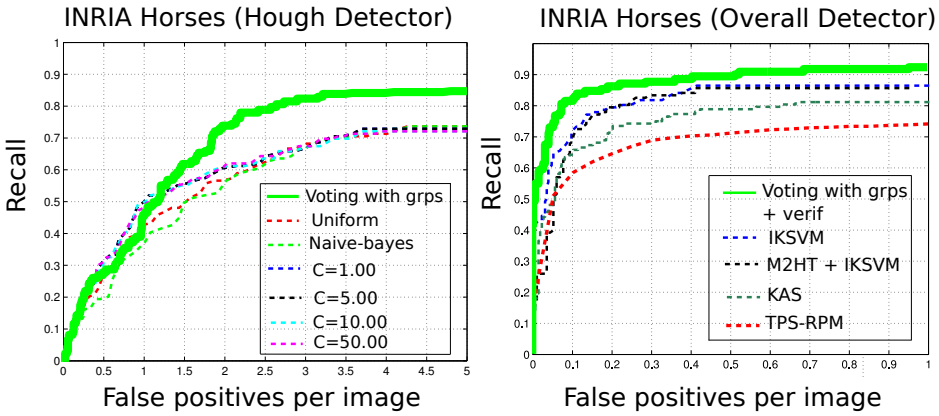


Figure 3.11: Detection plots on INRIA Horses dataset. Left plot compares the M^{2}HT detector for different parameters with our group voting. Voting with groups is superior to all. Right plot compares the overall detection results obtained from voting with groups plus verification with sliding windows (IKSVM) and state-of-the-art methods. At 1.0 FPPI we achieve a detection rate of 91% compared to the state of the art result of 86% (IKSVM) [68]

3.10.7 INRIA Horse Dataset

We have also evaluated the performance of voting with groups and the overall detector (voting + verification) on INRIA Horse dataset. The comparison is shown in Fig. 3.11. Voting with groups significantly outperforms the best voting methods so far (M^{2}HT detector), e.g., roughly 15% gain at 3 fppi. In terms of overall performance, we have a detection rate of 91% at 1 fppi compared to the state of the art results of 85.27% for M^{2}HT + IKSVM and 86% for sliding windows (IKSVM).

3.10.8 Shape-15 Dataset

To evaluate the presented approach on images with a larger number of categories, we utilize the Shape-15 dataset. This dataset, containing over 2000 images, has been assembled by combining five ETHZ categories with 10 categories from GRAZ-17 database. In order to compare with [80] and [47], we measure Equal Error Rates(EER) for missed detections and false positives per image and report the results in Tab. 3.4. We also report the EER obtained by sliding windows approach. Compared to the state-of-the-art, the proposed method achieves competitive performance thus underlining the value of grouping for Hough voting. For categories where this dataset combines large variations in pose (Giraffes, Bicyles, etc), voting with groups has an advantage compared to rigid sliding window paradigm.

3.11 Discussion

To address the primary shortcoming of voting methods, the assumption of part independence, we have introduced the grouping of mutually dependent parts into the voting procedure. Voting-based object detection has been formulated as an optimization problem that jointly optimizes groupings of dependent parts, correspondences between parts

Category	$\mathcal{H}^{\text{grp}}$ voting + verif	Method from [47]	Method from [80]
CarRear	**100**	90.6	89.9
Motorbike	**96**	82	82
Face	**100**	92.7	92.7
Carside	**93.9**	91	88.3
Cars Front	**94**	82.4	76.5
Person	58	**62.5**	58.3
Cup	**100**	85	80
Horse	72.5	**80.5**	79.3
Cow	93.5	**93.8**	89.2
Bicycle	**67.5**	63	61.1
Applelogos	**94**	81.8	79.5
Bottles	**100**	78.2	74.5
Giraffes	**81**	79	80.2
Mugs	**88.1**	81.8	80.3
Swans	**100**	72.7	69.7

Table 3.4: Comparison of Equal Error Rates(EER) for all the categories of Shape-15 dataset

and object models, and votes from groups to object hypotheses. This formulation tackles the fundamental problems of Hough voting methods, i) the unreliability of votes from local features by letting the dependent parts influence each others voting, and ii) the poor local estimates of global object properties such as object scale by jointly estimating the global properties at the group level. When addressing the correspondence problem, we have avoided making early decisions by maintaining a short-list of the potentially matching training points. Votes from local interest points have been reweighed by learning the relevance of each training point. Instead of each point independently casting a vote as in Hough voting, we utilized their dependencies to achieve concerted votes of groups. We have achieved a significant reduction in the number of object hypotheses (about three orders of magnitude) compared to the sliding window paradigm. Our model of part dependence in voting has demonstrated that it significantly improves the performance of probabilistic Hough voting in object detection. We have demonstrated the benefit of our approach on several benchmarks for voting based approaches.

CHAPTER 4

CONTOURS FOR OBJECT DETECTION

4.1 Overview

The first part of this thesis (chapters 2 and 3) dealt with object detection based on Hough Voting with interest points. The dependencies between interest points were incorporated into Voting framework thus giving a significant improvement over standard Hough Voting. However, a key thing to note is that the interest points were grouped in a query image during the detection stage. There is no grouping of interest points from the training images. However, shifting the grouping process from query image to training images carries an advantage. Whereas the grouping process in a query image depends on unreliable bottom-up information, the grouping process in training images can be made robust by evaluating the utility of groups over an ensemble of training images.

Groups of densely sampled interest points in training images yields contours. Thus, contour based object detection is the theme for this chapter and the next (chapters 4 and 5). This chapter reviews the various lines of research in contour based object detection and highlights the challenges faced by a contour based detection system. Tackling these challenges involves the usage of machine learning techniques. Hence, some useful machine learning techniques are also reviewed in this chapter.

4.2 Review of Contour based Object Detection

Contour based object detection has a long tradition in Computer Vision. Some of the recent work is reviewed in this section. There are several lines of work in contour based object detection such as template matching techniques [87, 62], Voting with contours [74, 87], partial shape matching [78, 75, 64], active shape models [32, 81], shape hierarchies [47] and parsing [8, 55, 90].

4.2.1 Template Matching

Template Matching is one of the most well established and widely used technique for object detection. Let $T = \{\mathbf{t}_i\}$ denote the set of pixels on the object template and $Q = \{\mathbf{q}_j\}$ denote the edge pixels in a query image. Because of its simplicity and efficiency, chamfer matching [17] has been used to localize the object template T in a query image Q.

Chamfer Matching

For a given location $\mathbf{x}$ of the template in the query image, chamfer matching finds the best $\mathbf{q}_j \in Q$ for each $\mathbf{t}_i \in T$ by minimizing the cost $|(\mathbf{t}_i + \mathbf{x}) - \mathbf{q}_j|$. Thus the chamfer distance for placing the template at location $\mathbf{x}$ is defined as

$$d_{CM}^{(T,Q)}(\mathbf{x}) = \frac{1}{|T|} \sum_{\mathbf{t}_i \in T} \min_{\mathbf{q}_j \in Q} |(\mathbf{t}_i + \mathbf{x}) - \mathbf{q}_j| \tag{4.1}$$

By incorporating the orientation information into (4.1), approaches such as [87] and [62] have attempted to make chamfer matching more robust.

Oriented Chamfer Matching

Let $\phi(\mathbf{t}_i)$ denote the edge orientation of the edge point $\mathbf{t}_i$. Oriented Chamfer Matching [87] augments (4.1) by defining an additional term d_{orient} based on the edge orientation on the template point $\mathbf{t}_i$ and its corresponding nearest query point $\mathbf{q}_j^*$ obtained by minimizing $|(\mathbf{t}_i + \mathbf{x}) - \mathbf{q}_j|$.

$$d_{orient}^{(T,Q)}(\mathbf{x}) = \frac{2}{\pi|T|} \sum_{\mathbf{t}_i \in T} |\phi(\mathbf{t}_i + \mathbf{x}) - \phi(\mathbf{q}_j^*)| \tag{4.2}$$

The oriented chamfer matching score is defined as

$$d_{OCM}^{(T,Q)}(\mathbf{x}) = (1 - \lambda) d_{CM}^{(T,Q)}(\mathbf{x}) + \lambda d_{orient}^{(T,Q)}(\mathbf{x}) \tag{4.3}$$

Directional Chamfer Matching

Instead of first finding the nearest point $\mathbf{q}_j^*$ based on (4.1) and then augmenting the cost with a directional term (4.2) based on $\mathbf{q}_j^*$, [62] demonstrate that it is beneficial to incorporate the directional term into (4.1) and directly compute the nearest point in an augmented space. Thus the directional chamfer distance for placing the template at location $\mathbf{x}$ is defined as

$$d_{DCM}^{(T,Q)}(\mathbf{x}) = \frac{1}{|T|} \sum_{\mathbf{t}_i \in T} \min_{\mathbf{q}_j \in Q} |(\mathbf{t}_i + \mathbf{x}) - \mathbf{q}_j| + \lambda |\phi(\mathbf{t}_i + \mathbf{x}) - \phi(\mathbf{q}_j)| \tag{4.4}$$

4.2.2 Voting with Contours

For approaches such as [74, 87], the basic idea is to automatically assemble a codebook of contours from training images. Based on the bottom-up information of each training image, a large set of contours are sampled. The relative location of object center is recorded for each contour. Such a large set of contours is distilled to a reasonable size codebook by clustering the contours based on their visual similarity. Next, the relative importance of each contour in detecting an object is learnt using Adaboost [50]. For detecting an object in a test image, each codebook contour is matched to the edge map of the test image yielding a set of possible placements for each contour. For each possible placement, the contour casts a weighted vote for the object properties such as its centre, scale and aspect ratio based on its relative importance learnt in the training stage. Finally, the local maxima in the Hough accumulator space are considered as object hypotheses.

The above approaches suffer from the following drawbacks. Clustering the contours based on visual similarity, e.g. based on the chamfer distance between contours c_i and c_j, has deficits. For instance, contours that are fractured or corrupted by noise can fall in different clusters although they are matched to similar locations in the training images. Another serious limitation is that each contour is independently matched to a query image. Such a matching does not take into account the spatial co-occurrences of various contours forming an object.

4.2.3 Partial Shape Matching

Approaches such as [78, 75, 64] perform the shape-based object detection task based on partial matching of edge fragments from training images to edge fragments in the query image. [75, 64] collect a sparse set of long contours obtained from within the object bounding boxes of training images. [78] on the other hand uses a hand drawn model for each object category. The bottom-up information is used in each query image to obtain a set of edge fragments. However, the edges obtained from bottom-up information in the query image fragment unpredictably.

To tackle this problem,[75] allows a many-to-one matching of long training contours to the contours from query images. The matching cost consists of the difference in the histogram bins of shape context descriptor and also the intersection of the histogram bin counts of the training and the query contours. The placement of model contours are treated as latent variables and weights for the histogram bins are learnt in a latent SVM framework [42].

[64] efficiently perform a partial matching of fragments based on integral image algorithm [94]. The partial edge matches are then grouped together to form an object hypothesis by formulating the grouping process as a maximum clique inference on weighted graph. The hypotheses are ranked based on their residual obtained from holistic transformation of model.

4.2.4 Active Shape Models

Techniques such as [31, 32] build flexible models of object shape, represented as point distributions, from a collection of training images. An object is basically represented as a set of labelled points and statistical distributions are maintained for the locations of

these points over all training images. To solve the correspondence problem for the points sampled from the object contours, criteria applicable for a pair of images as well as for all the images in the training set are used. After all the training objects are aligned, useful things such as principal components are computed where each object shape is represented as an ordered set of points in a high dimensional space. Such principal components can be used to synthesize new object shapes. Thus the active shape models follow Ulf Grenander's dictum 'pattern analysis equals pattern synthesis'.

In a query image, an object hypothesis is evaluated by successively projecting it onto the principal components and measuring the residual each time. The residual obtained after projecting a true hypothesis would be lower than the residual obtained by projecting a false hypothesis.

[81] applies a user defined set of shapes to detect closed contour cycles in a query image. These contour cycles are then abstracted and categorized by applying active shape models learnt from the user defined set of shapes.

4.2.5 Shape Hierarchies

[47] learns object representations as spatially flexible compositions of oriented edge fragments. The lower levels of the hierarchy consists of parts which are not category specific but rather quite generic. The higher levels of the hierarchy consists of category specific parts. Each part is represented by its position, scale and the principal axes of the ellipse encoding the variance of its position. Part hierarchy is learnt by identifying statistically significant compositions.

4.2.6 Parsing

[8] introduces connected segmentation trees (CST) for object detection. An object is represented by means of containment and neighbour relationships of its constituent regions denoted as nodes in a segmentation hierarchy. An object model is learnt as a common subgraph obtained by matching the CSTs of objects across several training images. Matching the model to a query image not only localizes the object but also yields its segmentation and the semantic explanation of the instance.

[90] demonstrate a Bayesian parsing framework on scenes consisting of faces and text. A parsing graph for a scene is output similar to parsing sentences in speech recognition. A parsing graph is constructed and reconfigured dynamically using a set of reversible Markov chain jumps.

[55] propose a contour based hierarchical object model that recursively decomposes an object into simple structures. Compositional rules are formulated to build the object models addressing the issues such as contour fragmentation and missing parts. An efficient inference algorithm based on coarse to fine search is used to rule out large portion of futile compositions and to parse complex objects in heavily cluttered scenes.

4.3 Review of Learning Algorithms

This section reviews the machine learning techniques which are going to be used in chapter 5 in formulating a novel state-of-the-art shape-based object detection approach.

4.3.1 Support Vector Machine

Let (x_i, y_i) be a tuplet denoting a feature vector x_i and its corresponding class label $y_i \in \{-1, 1\}$. Without loss of generality, we assume a two class classification scenario. Given a set of training tuplets consisting of feature vectors and their class labels, the objective of Support Vector Machines [93] is to find a hyperplane such that the margin between hyperplane and feature vectors is maximized. This is in contrast to a perceptron model where any arbitrary hyperplane between the classes is selected. Typically, the features are not linearly separable by means of a hyperplane. Hence, a function ϕ is applied to the feature vector x_i so as to map x_i to a much higher dimensional space where it is possible to linearly separate the feature vectors. Let $f(x) = sign(\mathbf{w}^T\phi(x) + b)$ denote the hyperplane separating the features belonging to different classes. Finding the parameters $\mathbf{w}^*, b^*$ corresponding to maximum margin is equivalent to solving the following optimization problem.

$$\begin{aligned}(\mathbf{w}^*, b^*) = &\underset{\mathbf{w},b}{\operatorname{argmin}} \frac{1}{2}||\mathbf{w}||^2, \\ &\text{s.t } \; y_i(\mathbf{w}^T\phi(x_i) + b) \geq 1 \;\; \forall i \in \{1, ..., N\}\end{aligned} \tag{4.5}$$

The optimization problem can be formulated using Lagrange multipliers as

$$L(\mathbf{w}, b, \alpha) = \frac{1}{2}||\mathbf{w}||^2 - \sum_{i=1}^{N} \alpha_i[y_i(\mathbf{w}^T\phi(x_i) + b) - 1] \tag{4.6}$$

Differentiating (4.6) with respect to $\mathbf{w}$ yields

$$\mathbf{w} - \sum_{i=1}^{N} \alpha_i y_i \phi(x_i) = 0 \tag{4.7}$$

Differentiating (4.6) with respect to b yields

$$\sum_{i=1}^{N} \alpha_i y_i = 0 \tag{4.8}$$

Using (4.7) and (4.8), (4.6) can be rewritten as

$$L(\alpha) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i\alpha_j y_i y_j \phi(x_i)\phi(x_j) - \sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i\alpha_j y_i y_j \phi(x_i)\phi(x_J) + \sum_{i=1}^{N} \alpha_i \tag{4.9}$$

Simplifying (4.9), the optimization problem in dual formulation is thus

$$\begin{aligned} L(\alpha) = & \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \phi(x_i) \phi(x_j) \\ & \text{s.t} \ \sum_{i=1}^{N} \alpha_i y_i = 0 \\ & \alpha_i \geq 0 \ \ \forall i \in \{1, ..., N\} \end{aligned} \tag{4.10}$$

In order to avoid over-fitting the hyperplane to the training data, slack variables are introduced in the optimization problem defined in (4.5). Instead of the feature points x_i strictly satisfying the constraint $y_i(\mathbf{w}^T \phi(x) + b) \geq 1$, they are allowed some slack ξ_i so that $y_i(\mathbf{w}^T \phi(x) + b) \geq 1 - \xi_i$. Subsequently, the optimization in (4.5) modifies to

$$\begin{aligned} (\mathbf{w}^*, b^*, \xi^*) = & \underset{\mathbf{w}, b}{\operatorname{argmin}} \frac{1}{2} ||\mathbf{w}||^2 + C \sum_i \xi_i, \\ & \text{s.t} \ \ y_i(\mathbf{w}^T \phi(x_i) + b) \geq 1 - \xi_i \ \ \forall i \in \{1, ..., N\} \end{aligned} \tag{4.11}$$

Using Lagrange multipliers and taking the derivates as in the case of (4.5), the dual formulation of the problem is obtained which is written below.

$$\begin{aligned} L(\alpha) = & \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \phi(x_i) \phi(x_j) \\ & \text{s.t} \ \sum_{i=1}^{N} \alpha_i y_i = 0 \\ & 0 \leq \alpha_i \leq C \ \ \forall i \in \{1, ..., N\} \end{aligned} \tag{4.12}$$

Because of the 'kernel trick', the dot product $\phi(x_i).\phi(x_j)$ in (4.10), (4.12) can be replaced by the kernel function $\mathcal{K}(x_i, x_j)$.

4.3.2 Multiple Instance Learning

Multiple Instance Learning [37] is an extension of the supervised classification task where the labels y_i for each individual feature vector x_i are not provided. Instead, the concept of a "bag" is introduced and the labels for the bags are known. Each bag B consists of multiple x_is. A bag B_I is labelled as positive if atleast one feature vector in that bag belongs to the positive class. On the other hand, a negative bag consists of features all of which belong to the negative class.

Multiple Instance Learning (MIL) arises naturally in a variety of applications ranging from classification of molecules in drug design to image indexing for content-based image retrieval. In the next chapter, we describe how MIL arises in the context of building object models based on spatial contour co-occurrences.

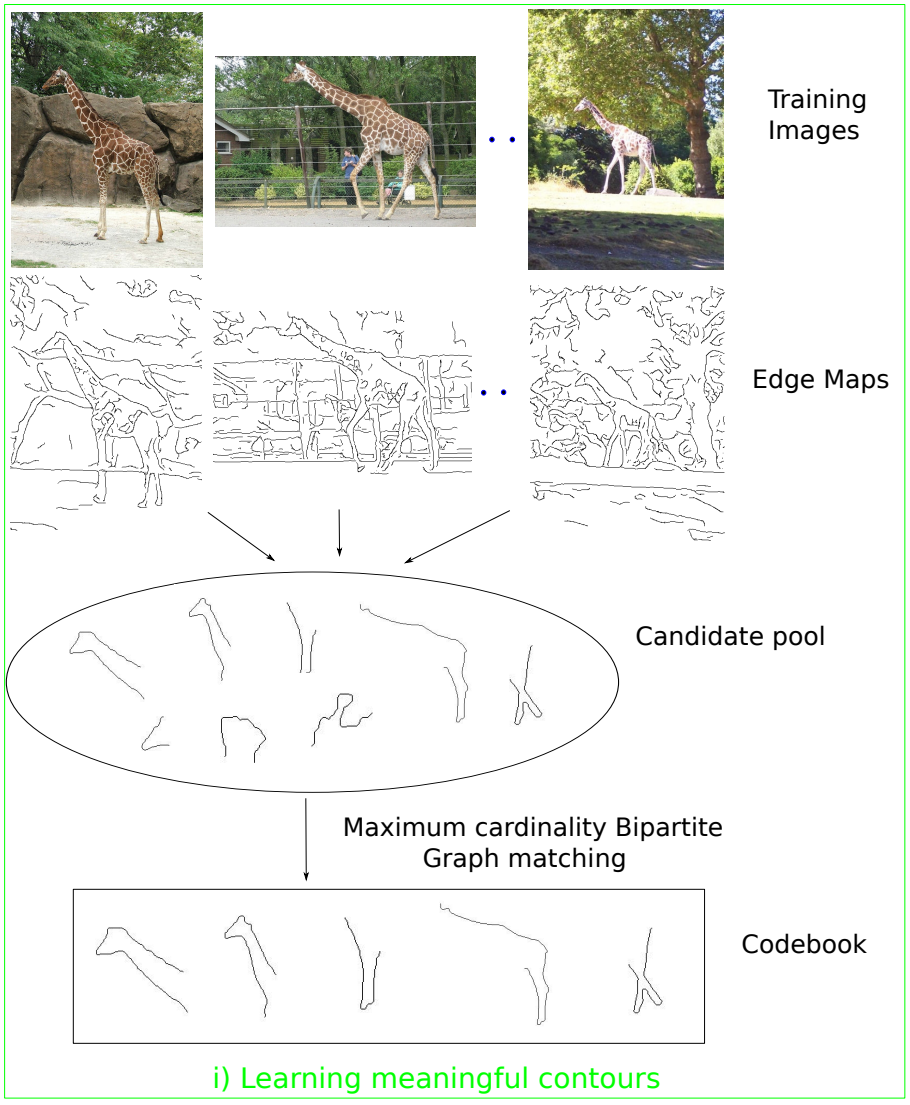


Figure 4.1: i) Obtaining meaningful contours from a collection of training images.

4.3.3 Max-Margin Multiple Instance Learning

Let (B_I, Y_I) denote tuplets of feature bags and their corresponding labels. The instance labels y_i for each feature vector $x_i \in B_I$ and the bag label Y_I are related as follows.

$$\begin{aligned} \sum_{i \in I} \frac{y_i + 1}{2} \geq 1, \;\; \forall I \;\; \text{s.t} \;\; Y_I = 1 \\ y_i = -1, \forall I \;\; \text{s.t} \;\; Y_I = -1 \end{aligned} \tag{4.13}$$

The notion of margin between the classification hyperplane and features is extended to bags for the case of Max-Margin Multiple Instance Learning. For the case of positive bag, the most positive instance (denoted by the indicator variable $s(I)$ has to satisfy the following margin constraint,

$$(\mathbf{w}^T \phi(x_{s(I)}) + b) \geq 1 - \xi_I \tag{4.14}$$

For the case of negative bag, all the instances x_i inside the bag I have to satisfy the constraint,

$$-(\mathbf{w}^T \phi(x_i) - b) \geq 1 - \xi_I \tag{4.15}$$

The indicator variable $s(I)$ for each bag I is unknown and hence the optimization problem is formulated as

$$\begin{aligned} &\min_{\mathbf{w},b,\xi,s} \frac{1}{2}\|\mathbf{w}\|^2 + \rho \sum_{\forall I} \xi_I \\ &\text{s.t } \forall I \;\; Y_I = -1 \;\; \wedge - < \mathbf{w}, \phi(x_i) > -b \geq 1 - \xi_I, \forall i \in I \\ &\text{or } \; Y_I = 1 \; \wedge < \mathbf{w}, \phi(x_{s(h)}) > +b \geq 1 - \xi_I \\ &\text{and } \; \xi_h \geq 0 \end{aligned} \tag{4.16}$$

(4.16) is optimized by iteratively minimizing with respect to s keeping $\mathbf{w}, b, \xi$ fixed and minimizing with respect to $\mathbf{w}, b, \xi$ keeping s fixed. When s is fixed, (4.16) is reduced to a standard SVM problem. On the other hand, when $\mathbf{w}, b, \xi$ are fixed, s is found by choosing the most positive instance in a bag based on the SVM scores obtained from the current estimate of the parameters $\mathbf{w}, b, \xi$. The iterative process is repeated until the variable s converges, that is the most positive instance in the bag remains unchanged.

4.4 Challenges faced by Contour based Object Detection

Various contour-based models reviewed in Sect. 4.2 provide an effective approach for accurately explaining meaningful object pixels in an image. The fundamental challenge of contour representation is, however, that object form (i.e. the Gestalt) cannot be perceived locally. Unlike color or texture which can be captured by a small image region, the prototypical shape of an object like a giraffe cannot be understood based on local measurements. Shape is an emergent property that becomes apparent only after all the object boundary contours (or, in dual form, its regions) have been grouped. At the same time, invariance w.r.t. missing, occluded parts and intra-class variation require that incomplete Gestalt needs to be dealt with while inter-class similarity renders it futile to detect objects based on single contours, e.g., the leg of a giraffe might resemble the outline of a bottle.

This leads to a fundamental question: how can we represent shape, if it cannot be measured directly? Although there has been significant progress in edge detection and segmentation (e.g. [66, 28]), segmentation is an ill-posed problem and thus bottom-up contour extraction is intrinsically limited [24]. To avoid the shortcomings of purely image-driven contour extraction, we follow a model-based approach (e.g. [87]) where we search with model contours that have been learned during training. Given a set of training images, contours are extracted and verified against the other training images to make up for the unreliability of

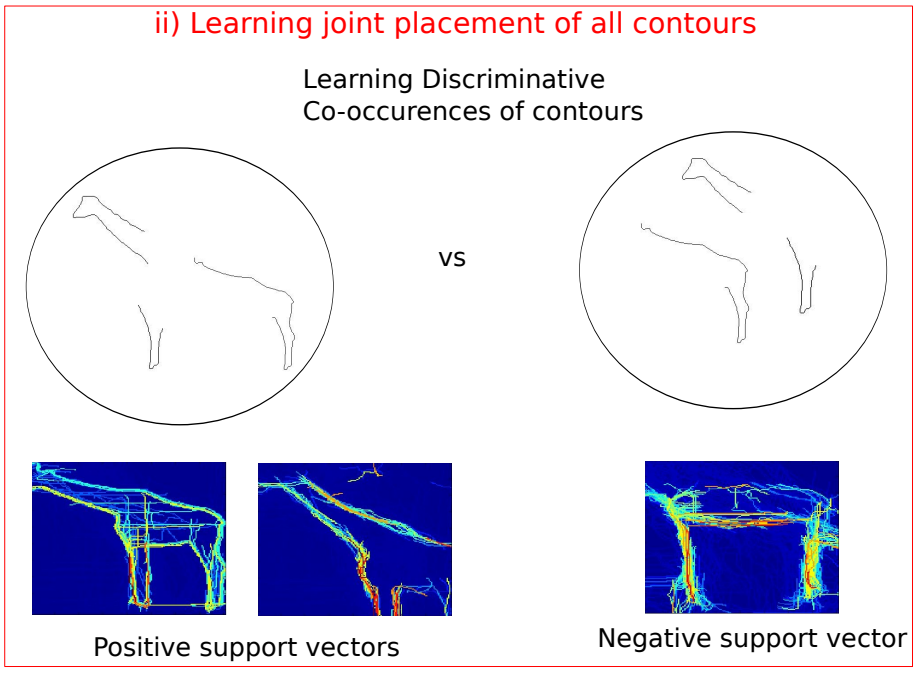


Figure 4.2: ii) learning discriminative contour co-occurrences.

the contour extraction process. This over-complete set of contours needs to be condensed into a feasible sized codebook. However, we do not follow the standard grouping based on visual similarity plus relative part location (e.g. [87]) as this fails when contours are corrupted by the extraction process. Rather we propose a clustering based on the activation pattern of contours where contours are grouped if they are activated similarly in a number of training images.

Although we now have a set of meaningful contours, matching them independently to novel query images (e.g. [87, 74]) still poses robustness issues due to the large intra-class variability. Therefore, we optimize the *joint placement* of all contours which maximally *discriminates* objects from non-objects.

But how can we learn meaningful co-placements of contours? During training these optimal *compositions* [22, 72] are not provided and the placement of individual contours is noisy. Therefore, we utilize *multiple instance learning* (MIL) and propose a number of candidate compositions of contours. Given positive and negative bounding boxes, MIL then selects a set of joint placements of codebook contours that are consistent among training images and optimally discriminate objects from non-objects. In addition each codebook contour receives a weight indicating how meaningful it is for discrimination.

Consequently, the difficult questions of selecting meaningful contours and finding consistent co-placements of these contours are shifted to the training phase. Here they can be addressed by optimization over an ensemble of training images rather then just a single query image.

To address the above challenges, i) we generate a dictionary of contours based on their co-activation patterns over an ensemble of training images (Fig. 4.1) ii) we learn the joint placement of all codebook contours that maximizes the discrimination between class and non-class structure using max-margin multiple instance learning (Fig. 4.2) and iii) we detect objects and assemble their shape at the same time by optimizing a single cost

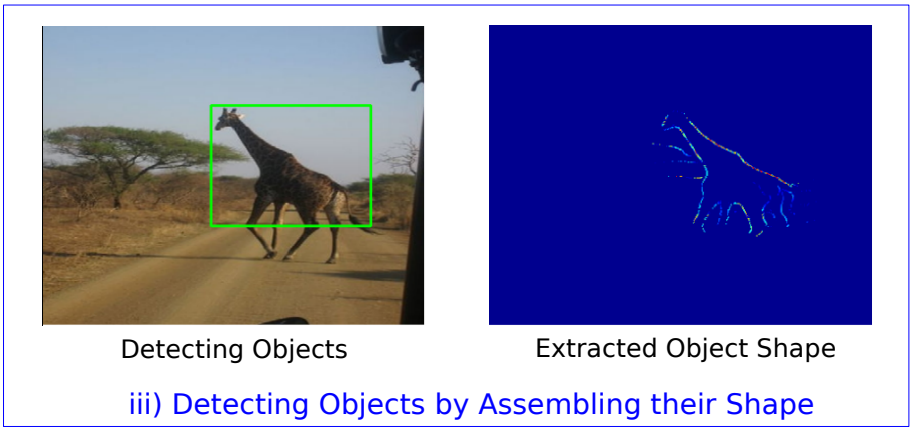


Figure 4.3: iii) using such co-occurrences in detecting an object and extracting its shape in a query image.

function that finds consistent joint placements of all dictionary contours. (Fig. 4.3)

CHAPTER 5

DETECTING OBJECTS BY ASSEMBLING THEIR SHAPE

5.1 Overview

This chapter addresses the three main challenges faced by a shape-based detection system, i) learning a dictionary of meaningful contours from an ensemble of training images Sect. 5.2 ii) learning a discriminative shape-based object model from the meaningful contours Sect. 5.3 and iii) detecting objects and simultaneously assembling their shape while avoiding bottom-up grouping in query images altogether Sect. 5.4.

5.2 Learning Meaningful Object Contours

To obtain a set of meaningful contours from the training images, we first compute the probabilistic edge maps for each image using [66]. We follow the standard procedure of normalizing the provided object bounding boxes so that they have the same scale and aspect ratio. Thereafter, we perform edge linking using the approach of [56]. In a next step, we extract a set of non-disjoint contours from each linked edge segment by first computing points of high curvature and considering the midpoints between them. Randomly selecting pairs of these points from an edge and taking the contour segments in between yields a set of candidate contours. Each contour has a shift vector $\mathbf{s}^{c_i}$ from its centroid to the center of the bounding box. Combining all the segments from all training images yields on the order of 10^4 contours. Many of these are redundant and the size of this set needs to be reduced to a compact, feasible sized subset of meaningful contours.

A common approach is to cluster contours based on their visual similarity, potentially also adding the relative location in the image [87]. However, such a clustering founded on visual similarity, e.g. based on the chamfer distance between contours c_i and c_j, has deficits. For instance, contours that are fractured or corrupted by noise can fall in different clusters although they are matched to similar locations in the training images.

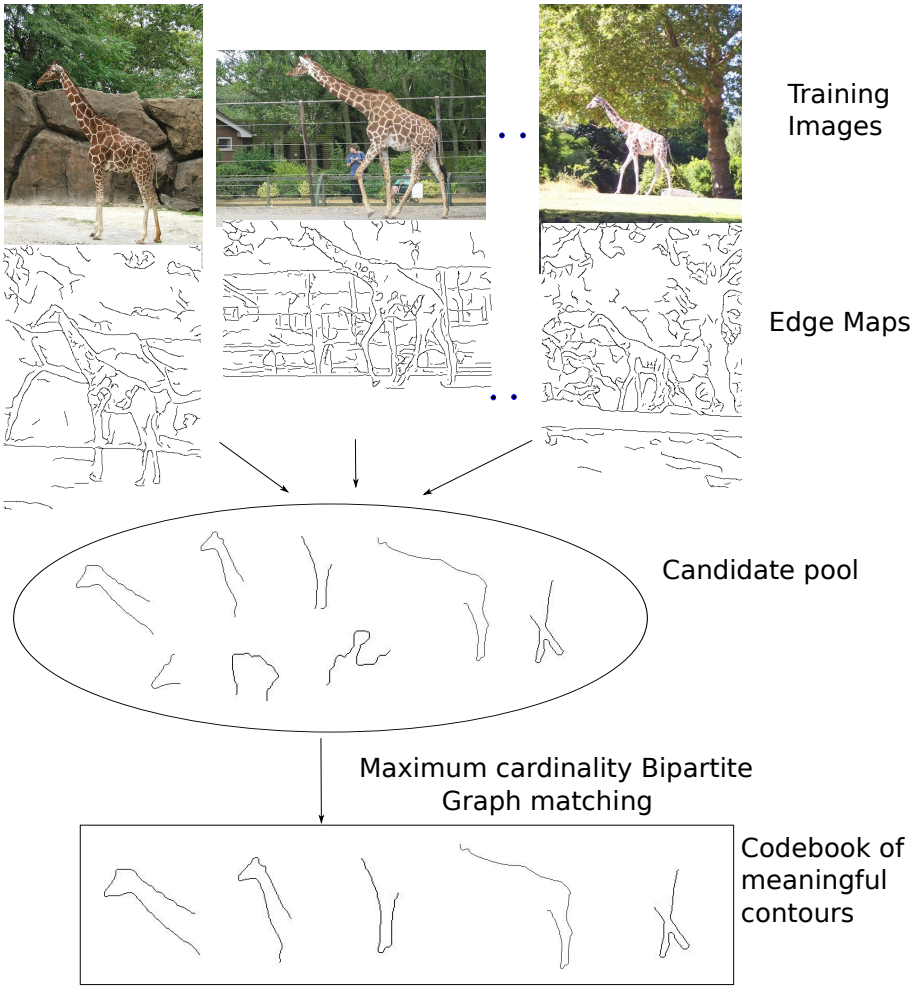


Figure 5.1: Figure shows the codebook generation process from edge maps of training images.

Therefore, we compute the pairwise dissimilarity matrix Δ_{ij} for all pairs c_i, c_j not by means of their visual similarity but based on where they match in an ensemble of training images. We use fast directional chamfer matching [62] for obtaining matching locations for each candidate contour in each training image. Let $A^i_{m,h}$ denote the m-th match of contour c_i in training bounding box h. $\mathcal{E}^h$ denotes the edge map of h. For the m-th match, we record the chamfer score $\gamma_m(c_i, \mathcal{E}^h)$ and the location of the match in the image $l_m(c_i, \mathcal{E}^h)$ (see Sect. 5.4.1),

$$A^i_{m,h} := \left(\gamma_m(c_i, \mathcal{E}^{t_h}), l_m(c_i, \mathcal{E}^{t_h}), \mathbf{s}^{c_i}\right)^\top \tag{5.1}$$

5.2.1 Clustering the Contours based on their Co-activation Patterns

We cluster the contours based on their activation patterns A^i over all the training images.

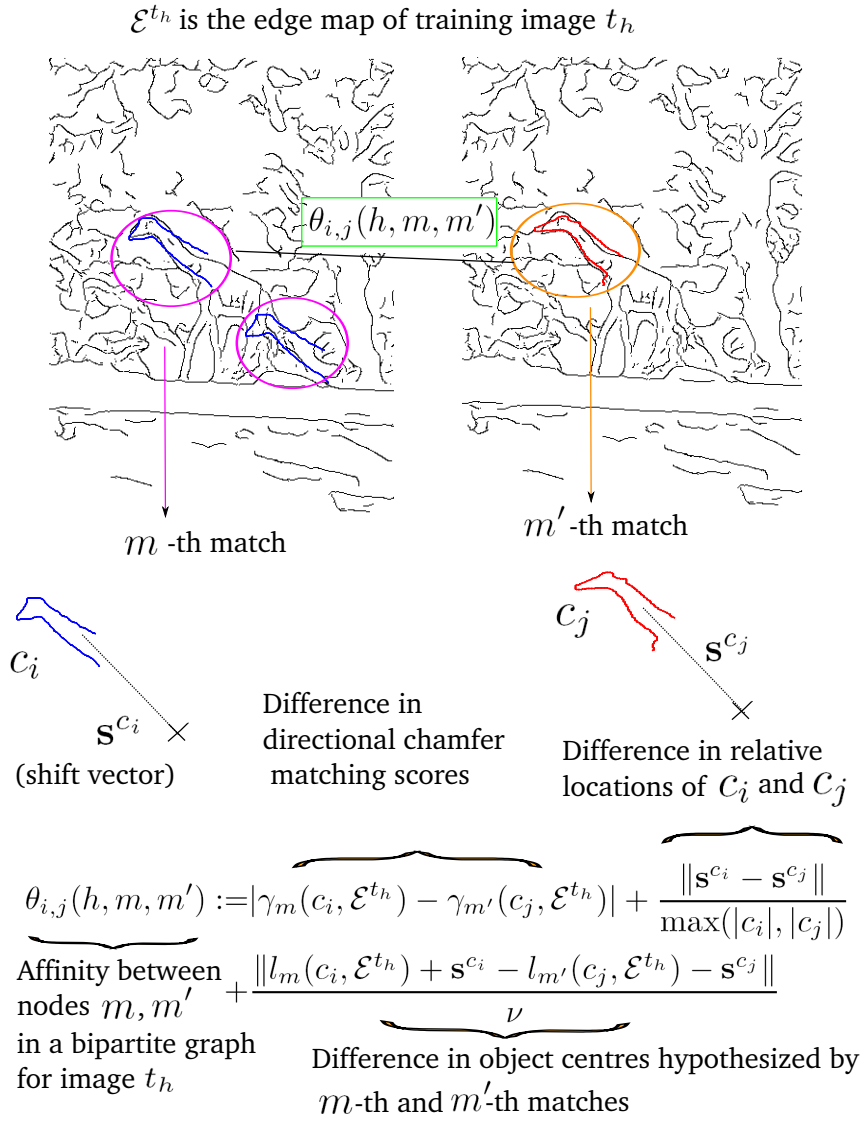


Figure 5.2: Bipartite matching to cluster contours.

$$A^i := \begin{bmatrix} A^i_{1,1} & A^i_{1,2} & \dots \\ A^i_{2,1} & \dots & \\ A^i_{3,1} & \dots & \\ \vdots & & \end{bmatrix} \tag{5.2}$$

We compute the dissimilarity matrix Δ_{ij} as $\Delta_{ij} := \sum_h \Theta\big(A^i_{\bullet,h}, A^j_{\bullet,h}\big)$. The dissimilarity Θ of both contours on training image h is obtained using *maximum cardinality bipartite matching*. For the bipartite matching, the elementary distance between the m-th match of c_i and the m'-th match of c_j is defined as

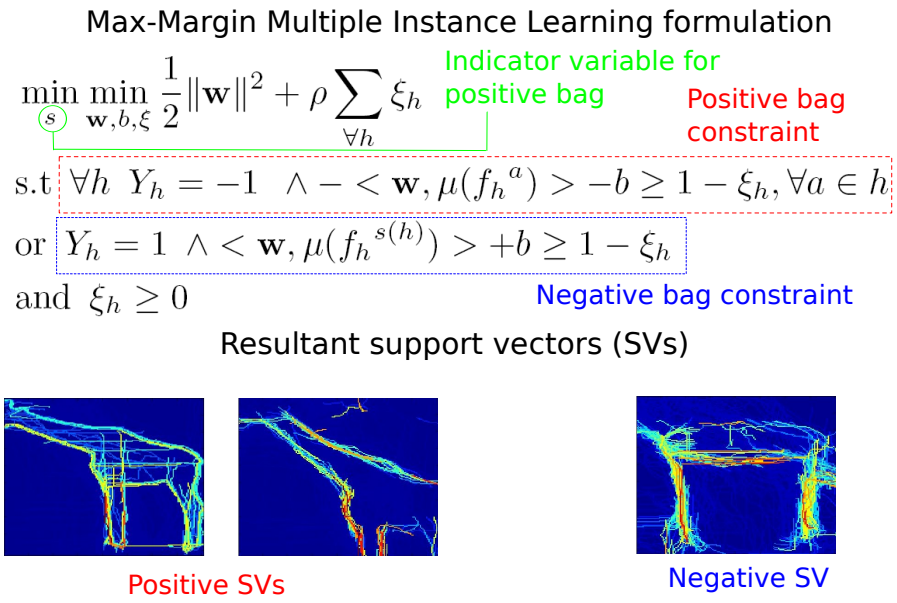


Figure 5.3: Learning the joint placement of all the contours in a Max-Margin Multiple Instance Learning framework

$$\theta_{i,j}(h,m,m') := |\gamma_m(c_i, \mathcal{E}^{t_h}) - \gamma_{m'}(c_j, \mathcal{E}^{t_h})| + \frac{\|\mathbf{s}^{c_i} - \mathbf{s}^{c_j}\|}{\max(|c_i|,|c_j|)} + \frac{\|l_m(c_i, \mathcal{E}^{t_h}) + \mathbf{s}^{c_i} - l_{m'}(c_j, \mathcal{E}^{t_h}) - \mathbf{s}^{c_j}\|}{\nu} \tag{5.3}$$

where ν is the average length of all object bounding box diagonals in the training data.

Given the pairwise dissimilarity matrix Δ_{ij} we perform pairwise clustering using *Ward's method* and obtain a codebook $\mathcal{C}$ that contains on the order of 10^2 contours. The representative for each cluster is the element that has maximal average affinity to all elements in this cluster.

5.3 Learning a Discriminative Model for Object Shape

Given the codebook $\mathcal{C}$, we need to learn how to jointly place all the contours so that the overall configuration optimally discriminates the shape of objects from non-objects. During the training stage, we are only provided ground-truth for the bounding box of objects, but obviously not for the placement of contours therein. As discussed before, relying on chamfer matching to yield an optimal match for each contour will result in spurious matches due to large intra-class variability and noise. Therefore, we consider multiple placements for each contour within the bounding box and learn to jointly place all contours. Therefore, candidate matches of contours are grouped and a MIL-based procedure [10] is used to find the group with best joint placement. Failing to learn the best joint placement and just selecting appropriate matches for all contours independently significantly degrades the performance—on average we observed a 10% drop on ETHZ shape dataset compared to the MIL-based procedure we propose in this work [103].

Let $\Gamma_i^h = (\gamma_1(c_i, \mathcal{E}^h), \gamma_2(c_i, \mathcal{E}^h), \dots)$ denote the matches for c_i in bounding box h. For the m-th match of c_i, we concatenate the chamfer score with the spatial consistency to form a 2-d feature vector ${f_h}^{i,m}$ that will be discussed in Sect. 5.4. The spatial consistency of a match measures how well the object hypothesis generated from m-th match of c_i agrees with the object bounding box h. Now we concatenate the 2-d feature representations of all contours to represent the joint placement of all parts. Let ${m_i}^a$ be some match for a contour $c_i \in \mathcal{C}$. Then we obtain a candidate configuration a for the placement of all parts represented by ${f_h}^a = ({f_h}^{1,m_1^a}, {f_h}^{2,m_2^a}, \dots, {f_h}^{|C|,m_{|C|}^a})$. We start with a contour $c_i \in \mathcal{C}$ and let each of its matches $\gamma_m(c_i, \mathcal{E}^h)$ predict an object hypothesis. Conditioned on this hypothesis, we obtain an object representation ${f_h}^a$ by choosing the spatial maximally consistent match for each of the other contours. By repeating this process for all codebook contours, we obtain a bag of candidate configurations $F_h = \{{f_h}^a\}$.

5.3.1 Max-Margin Multiple Instance Learning

However, not all the configurations in the bag F_h are meaningful. If for instance some contour c_i is providing a spurious match against background clutter within the bounding box then the resulting feature vectors are also affected. Therefore, we introduce an indicator variable $s(h) \in \{1, \dots, |F_h|\}$ which selects the most useful candidate configuration for describing the object bounding box. For negative bounding boxes which are obtained by randomly sampling boxes from regions not containing a positive box, all the configurations inside a bag are used as negative examples. Let $Y_h \in \{-1, 1\}$ denote the bag label and let μ be some non-linear function on the co-activation feature vectors ${f_h}^a$. Then we seek the weights $\mathbf{w}$ for each dimension of this transformed feature vector so that the most representative example (identified by $s(h)$) of a positive bag h with $Y_h = 1$ and all the examples of a negative bag h with $Y_h = -1$ have maximum margin separation. Therefore, for a positive bag, the following constraint has to be satisfied for the configuration identified by $s(h)$

$$< \mathbf{w}, \mu({f_h}^{s(h)}) > +b \geq 1 - \xi_h \tag{5.4}$$

And the following constraint has to be satisfied for all the configurations a in a negative bag.

$$- < \mathbf{w}, \mu({f_h}^a) > -b \geq 1 - \xi_h \tag{5.5}$$

Thus, we have the following max-margin multiple instance learning problem.

$$\begin{aligned} &\min_s \min_{\mathbf{w},b,\xi} \frac{1}{2}\|\mathbf{w}\|^2 + \rho \sum_{\forall h} \xi_h \\ &\text{s.t } \forall h \;\; Y_h = -1 \;\; \wedge - < \mathbf{w}, \mu({f_h}^a) > -b \geq 1 - \xi_h, \forall a \in h \\ &\text{or } \; Y_h = 1 \; \wedge < \mathbf{w}, \mu({f_h}^{s(h)}) > +b \geq 1 - \xi_h \\ &\text{and } \; \xi_h \geq 0 \end{aligned} \tag{5.6}$$

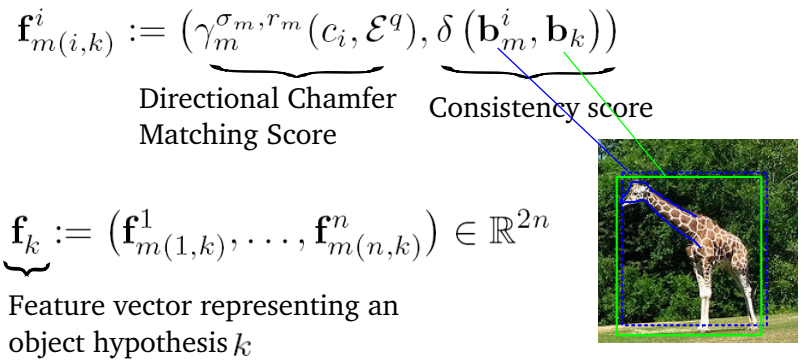


Figure 5.4: Consistency score.

Equation (5.6) is expressed in a compact form as

$$\min_s \min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 + \rho\Big(\sum_{Y_h=1} \max(0, 1 - < \mathbf{w}, \mu({f_h}^{s(h)}) > -b) + \sum_{Y_h=-1} \max(0, 1 + b + \max_a(< \mathbf{w}, \mu({f_h}^a) >))\Big) \tag{5.7}$$

Converting (5.7) into dual form and utilizing a kernel function $\mathcal{K}$(in our implementation, we use a second degree polynomial kernel) to compute the pairwise distances between original feature vectors $({f_h}^{a_1}, {f_h}^{a_2})$ eliminates the need to explicitly know the function μ. Therefore, equation (5.7) is optimized in its dual form by iteratively optimizing the indicator variables $s(h)$ and the usual SVM parameters i.e., the support vectors S_{h^a}, their co-efficients α_{h^a} and the offset b. For a positive bag, the dual variable $\alpha_{h^{s(h)}}$ has to satisfy $0 \leq \alpha_{h^{s(h)}} \leq \rho$. For a negative bag, the dual variable has to satisfy $0 \leq \sum_a \alpha_{h^a} \leq \rho$. Thus the effect of each configuration in a negative bag is limited to the box constraint ρ. The minimization starts by choosing $s(h)$ for each bag corresponding to the co-activation feature vector constructed from best match for each of the contours. After the optimization, we obtain the parameters α, S, b and use them in the cost function ψ,

$$\psi_{\alpha,S,b}(f) = \sum_{h:Y_h=1} \alpha_{h^{s(h)}} \mathcal{K}(f, S_{h^{s(h)}}) - \sum_{a,h:Y_h=-1} \alpha_{h^a} \mathcal{K}(f, S_{h^a}) + b. \tag{5.8}$$

In query images this cost function is applied to find a consistent joint placement f of all codebook contours and the score of ψ is used to rank and classify the resulting hypotheses.

5.4 Detecting Objects by Describing their Shape

To detect all instances of an object class in novel query images, their characteristic shape is to be extracted. To capture object shape, codebook contours need to be pieced together properly. Therefore, all these contours need to be jointly matched to a query image so that the grouping of all contours discriminates between objects from the class and all other structure. As a result, objects are segregated from background clutter which in turn improves classification and localization since distracting clutter is suppressed.

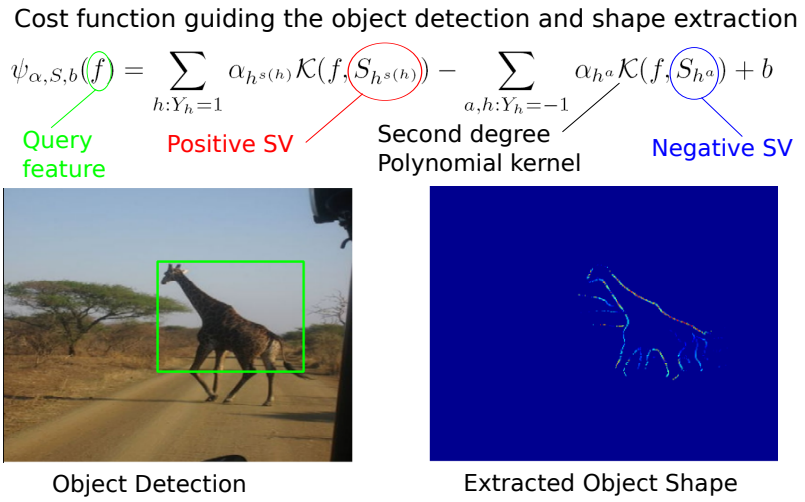


Figure 5.5: Detecting objects by jointly assembling all the contours.

5.4.1 Detecting Meaningful Contours

Let $\mathcal{E}^q$ be the edge map of the query image obtained by using [66]. $\mathcal{E}^{c_i}$ denotes the template edge map created from the codebook contour c_i. $\phi\left(\mathcal{E}_j^q\right)$ denotes the edge gradient orientation at the pixel $\mathcal{E}_j^q \in \mathbb{R}^2$ in the query image.

Given the dictionary $\mathcal{C} = \{c_1, \ldots, c_n\}$ of codebook contours for both objects and non-objects, each contour can be matched to a query image using fast directional chamfer matching [62].

As opposed to the training stage, object scale and aspect ratio are obviously unknown in a query image. Hence, each codebook contour has to be matched at different scales and aspect ratios to a query image. Applying directional chamfer matching [62] yields matches with scores $\gamma_m^{\sigma,r}(c_i, \mathcal{E}^q)$. The best match has for instance the matching score

$$\begin{aligned}\gamma_1^{\sigma,r}(c_i, \mathcal{E}^q) = |\mathcal{E}^{c_i}|^{-1} \sum_{\mathcal{E}_j^{c_i} \in \mathcal{E}^{c_i}} \min_{\mathcal{E}_k^q \in \mathcal{E}^q} \Bigg\{ \left\| \begin{bmatrix} \sigma r & 0 \\ 0 & \sigma \end{bmatrix} \mathcal{E}_j^{c_i} - \mathcal{E}_k^q \right\| \\ + \lambda \left| \phi\left(\begin{bmatrix} \sigma r & 0 \\ 0 & \sigma \end{bmatrix} \mathcal{E}_j^{c_i} \right) - \phi\left(\mathcal{E}_k^q\right) \right| \Bigg\}\end{aligned} \tag{5.9}$$

5.4.2 Representing Ensembles of Contours

Matching individual codebook contours to query images as done in [87, 74] is prone to yield spurious matches due to intra-class variations of contours. We cannot correctly detect objects by placing each contour individually. Rather, we need to represent an object hypothesis by jointly matching all contours from $\mathcal{C}$ and letting the model learned in Sect. 5.3 propose the right joint placement of contours. For each contour, we obtain multiple matches per scale and aspect ratio, yielding a set of scores $\Gamma = \left\{\gamma_1^{\sigma,r}(c_i, \mathcal{E}^q), \ldots, \gamma_k^{\sigma,r}(c_i, \mathcal{E}^q)\right\}$ and the corresponding coordinates of the matches $\mathcal{L} = \left\{l_1^{\sigma,r}(c_i, \mathcal{E}^q), \ldots, l_k^{\sigma,r}(c_i, \mathcal{E}^q)\right\}$. From this short-list of matches, we need to find the optimal match for each contour so that the

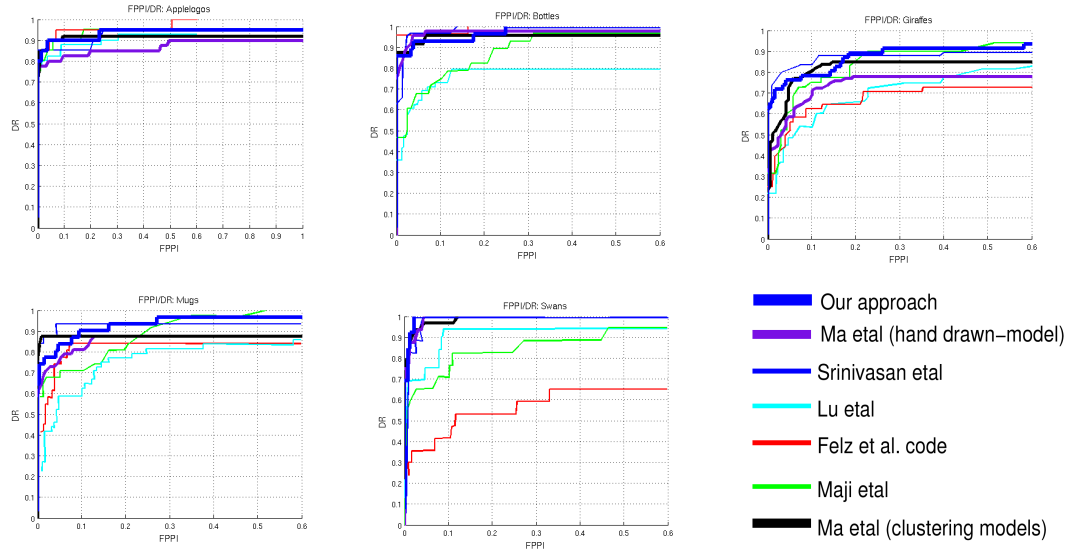


Figure 5.6: Comparison of our performance with other state-of-the-art approaches in terms of Detection Rate/FPPI Curves

overall configuration is maximally consistent with the joint placement of all contours from the training. As in Sect. 5.2 $\mathbf{s}^{c_i}$ denotes the shift vector of c_i. Then a candidate match $l_m^{\sigma,r}(c_i, \mathcal{E}^q)$ votes for an object bounding box

$$\mathbf{b}_m^i = \left((l_m^{\sigma,r}(c_i, \mathcal{E}^q))^\top - (\mathbf{s}^{c_i})^\top \begin{bmatrix} \sigma r & 0 \\ 0 & \sigma \end{bmatrix}, \sigma, r \right). \tag{5.10}$$

A short-list $\mathcal{B} = \{\mathbf{b}_1, \mathbf{b}_2, \dots\}$ of potential object hypotheses is created by collecting the hypotheses $\mathbf{b}_m^i$ of all contours c_i in a Hough accumulator [60] and performing the usual non-max suppression. Subsequently, we represent each candidate bounding box $\mathbf{b}_k \in \mathcal{B}$ using the co-activation pattern of all codebook contours. Therefore, we need to measure for each $\mathbf{b}_m^i$ its spatial consistency with an overall object hypothesis $\mathbf{b}_k$ using the standard Pascal VOC criterion [39]

$$\delta\left(\mathbf{b}_m^i, \mathbf{b}_k\right) := \frac{area(\mathbf{b}_m^i \cap \mathbf{b}_k)}{area(\mathbf{b}_m^i \cup \mathbf{b}_k)}. \tag{5.11}$$

Let $\hat{m}_{(i,k)}$ denote the m-th match of model contour c_i to the query image. For $\mathbf{b}_k$, the m-th match has the following directional chamfer and spatial consistency score

$$\mathbf{f}_{m(i,k)}^i := \left(\gamma_m^{\sigma,r}(c_i, \mathcal{E}^q), \delta\left(\mathbf{b}_m^i, \mathbf{b}_k\right)\right). \tag{5.12}$$

Thus the overall object hypothesis $\mathbf{b}_k$ can be represented by concatenating all the matching scores to obtain their co-activation pattern.

$$\mathbf{f}_k := \left(\mathbf{f}_{m(1,k)}^1, \dots, \mathbf{f}_{m(n,k)}^n\right) \in \mathbb{R}^{2n}. \tag{5.13}$$

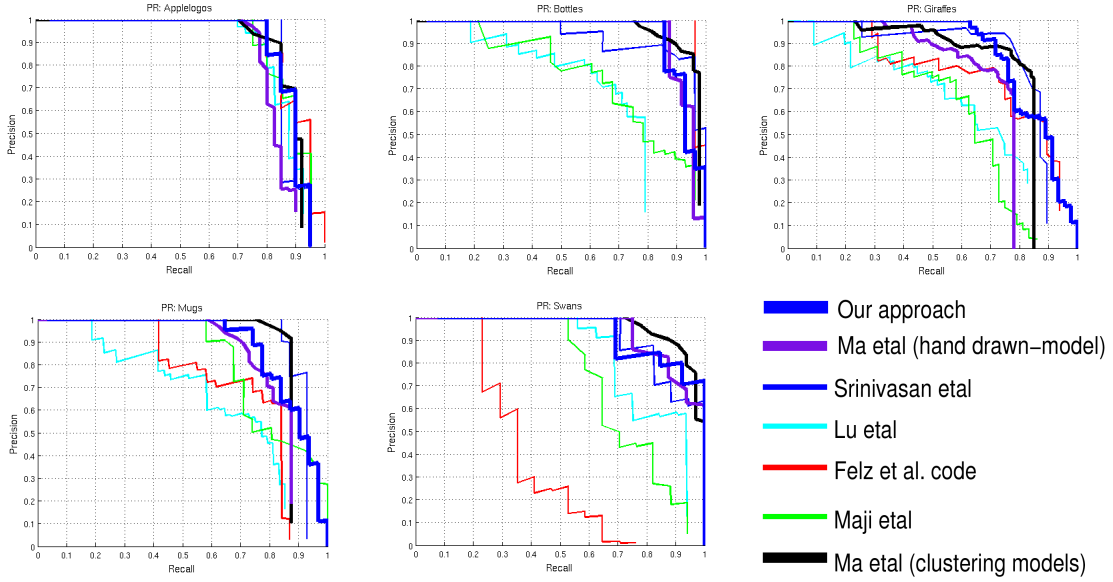


Figure 5.7: Comparison of our performance with other state-of-the-art approaches in terms of Precision Recall Curves

We cannot find the correct match $\hat{m}_{(i,k)}$ for each c_i independently. We thus need a joint optimization procedure to find a consistent match from the possible options for each contour. The hypothesis corresponding to the optimal placement of all the contours is then denoted by $\left(\mathbf{f}^1_{\hat{m}_{(1,k)}}, \ldots, \mathbf{f}^n_{\hat{m}_{(n,k)}}\right)$.

5.4.3 Modelling Shape by Jointly placing all Object Contours

To jointly find the optimal matches for all the codebook contours, we use the cost function ψ from equation (5.8). We utilize the second order polynomial kernel function $\mathcal{K}(f_{k_1}, f_{k_1}) = (1+ < f_{k_1}, f_{k_1} >)^2$. The optimal placement $m^*_{i,c}$ for each c_i can be computed using (5.8) conditioned on the placement of the other codebook contours. Thus, we employ a greedy algorithm to find the optimal placement of each c_i. We initialize the co-activation feature vector by best matches for each contour and then update the placement of contours one at a time. We visit the contours in a random schedule and update the contour placements. We reach rapid convergence for the cost function within 5 sweeps over all contours. Although techniques such as [70] could be potentially used for solving the joint placement problem, speed is an issue with such techniques. We found the sequential greedy approach to converge quickly and to produce competitive results which are described in the experimental section.

5.5 Experimental Evaluations

We report our experimental evaluations on the standard benchmark datasets for shape-based detection which have been widely used [68, 75, 64, 89, 78, 46], the ETHZ shape dataset and INRIA Horses dataset. These datasets feature significant intra-class variations, scale variations, different lighting conditions and articulations. To evaluate detection performance, we use the PASCAL criterion. Thus the detections are considered correct if

	Applelogos	Bottles	Giraffes	Mugs	Swans	**Mean**
Ours	95/95	**100/100**	**91.3/91.3**	**96.7/96.7**	**100/100**	**96.5/96.5**
[64]	92/92	97.9/97.9	85.4/85.4	87.5/87.5	**100/100**	92.6/92.6
[75]	95/95	**100/100**	87.2/89.6	93.6/93.6	**100/100**	95.2/95.6
[89]	**100/100**	96/97	86/91	90/91	98/100	94/96
[68]	95/95	92.9/96.4	89.6/89.6	93.6/96.7	88.2/88.2	91.9/93.2
[41]	95/95	**100/100**	72.9/72.9	83.9/83.9	58.8/64.7	82.1/83.3
[40]	95/95	96.3/100	84.7/84.7	**96.7/96.7**	94.1/94.1	93.3/94.1
[78]	93.3/93.3	97/97	79.2/81.9	84.6/86.3	92.6/92.6	89.3/90.5
[46]	77.7/83.2	79.8/81.6	39.9/44.5	75.1/80	63.2/70.5	67.1/72
[105]	80/80	92.9/92.9	68.1/68.1	64.5/74.2	82.4/82.4	77.6/79.5

Table 5.1: Comparison of detection rates for 0.3/0.4 fppi on ETHZ Shape Classes

the intersection of object hypothesis and the ground-truth over their corresponding union is greater than 50 %. Note that this is a stricter criterion than the 20 % overlap criterion used by [62] to report their performance on ETHZ shape classes. For performing our evaluations, we use the standard protocol described in [46], i.e., use the first half of images in each class for training, and test on the second half of this class as positive images plus all images in other classes as negative images. During the training stage, we only utilize the ground-truth bounding box annotations for the objects and build our shape model from this input.

We use the fast directional chamfer matching code provided by [24] (evaluates 1.05 million hypotheses per image in 0.42 seconds) to obtain matches for each contour. Our codebook contains on the order of 100 contours. Each test image needs a total processing time (matching all codebook contours and evaluating the model for all candidate hypotheses) on the order of seconds. During the training stage, the multiple instance learning converges within 10 iterations of alternating between indicator variables and dual variables (Sec.5.3). The whole training process is on the order of few hours.

During the testing stage, we search over 7 different scales and 3 different aspect ratios. We evaluate our approach in terms of detection rate over fppi(false positives per image) curves. The detection rates are reported in Tab. 5.1 at the usual threshold of 0.3/0.4 % fppi and we observe competitive performance compared to the state-of-the-art. The average detection rate is 96.5 % at 0.3 fppi thereby achieving a gain of 1.3 % over the best performing method so far. Our detection rates reach peak value before 0.3 fppi and hence the performance stays same at 0.3/0.4 fppi when comparing with other approaches. We achieve a mean average precision of 0.882 which is improving the performance of state-of-the-art methods summarized in Tab. 5.2). All in all, we observe a comprehensive gain over the current approaches in terms of various performance measures. Since we jointly explain each object hypothesis, we do not need a separate verification stage and we even outperform a two-stage detection system [68].

In Tab. 5.1 and Tab. 5.2, we have included the performance achieved by the latest code release of the popular sliding window based approach [40]. Thus, we are comparing ourselves not only with the state-of-the-art in shape-based methods but also against the currently best performing recognition system which utilizes many other cues besides shape. Compared to [40], we achieve a gain of 0.8 % in terms of mean average precision. Category-wise, we outperform on 4 categories. In terms of fppi/detection rate, there is an average gain of 3.2 % at 0.3 fppi.

Method	Ours	[75]	[64]	[68]	[41]	[40]
Mean Average Precision	**0.882**	0.872	0.877	0.771	0.712	0.874

Table 5.2: Comparison of Mean Average Precision (AP) on ETHZ Shape dataset

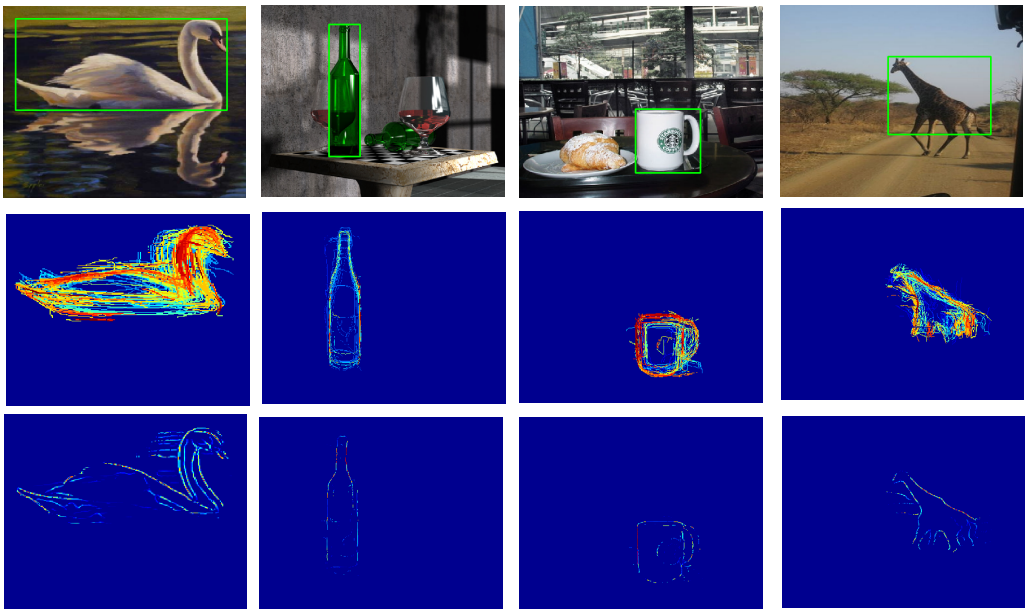


Figure 5.8: Detection results and the extracted shape.

For the INRIA Horses dataset, we compare our approach with the results reported by other current methods at 1 fppi in Tab. 5.3. We achieve a detection rate of 93.68 % compared to the current state-of-the-art performance of 92.4 % reported in [89].

5.6 Discussion

This work detects objects while, simultaneously, assembling their shape. Meaningful contours are obtained by clustering based on contour co-activation over the training images. The characteristic object shape is represented by learning consistent configurations of all model contours in a maximum margin MIL framework. Rather than placing each contour independently, a joint placement of all contours is sought that discriminates class from non-class structure. In a query image, detection and shape extraction are tackled jointly by optimizing a single cost function that yields optimal configurations of model contours and a classification. In the experimental validation the approach has shown competitive performance on widely used benchmark datasets for shape-based detection.

Method	Ours	[89]	[102]	[68]
Detection rate	**93.68**	92.4	87.3	85.3

Table 5.3: Comparison of detection rates for 1 fppi on INRIA Horses dataset

CHAPTER 6

DISCRIMINATIVE CHAMFER REGULARIZATION

For finding the placement of an object template or its part in an edge map, Chamfer matching is a widely used technique. Chapter 5 described a contour based approach where fast directional chamfer matching was used for the placement of a contour in an edge map. This chapter deals with a serious limitation of chamfer matching which yields spurious placements of a contour in background clutter.

The simplicity and speed of chamfer matching has benefited numerous application areas such as industrial inspection, Machine Vision, Robotic Perception. However, a serious limitation of chamfer matching is its susceptibility to background clutter. Although the inclusion of orientation information [87, 62] has improved the specificity, performance is still seriously affected by clutter. The primary reason for this is that the presence of individual model points in a query image is measured independently. A match with the object model is then represented by the sum of all the individual model point distance transformations. Consequently, i) all object pixels are treated as being independent and equally relevant, and ii) the model contour (the foreground) is prone to accidental matches with background clutter. As demonstrated by Biederman [22], Attneave [14], and various experiments on illusionary contours, object boundary pixels are not all equally important due to their statistical interdependence. Moreover, in dense background clutter the points on the model have a high likelihood to find good spurious matches [22, 14]. However, any arbitrary model would match to such a cluttered region, which consequently gives rise to matches with high accidentalness. Chamfer matching only matches the template contour and thus fails to discount the matching score by the accidentalness, i.e., the likelihood that this is a spurious match.

To improve the robustness of model matching, we learn the co-occurrence of model points (or rather their matches). To reduce the accidentalness of chamfer matching, we learn a flexible co-placement of generic background contours. Both these contributions are combined into a single discriminative learning algorithm. Our approach is built upon the publicly available, state-of-the-art directional chamfer matching approach [62] and we evaluate the proposed method on standard benchmark datasets for chamfer matching.

Figure 6.1: Pixel weights learned in a discriminative max-margin framework for various shape templates are visualized here. The pixels are weighted relative to the template and therefore are not comparable among different object classes. Red indicates high and blue low weight.

6.1 Max-Margin Chamfer Regularization

We base our study in this work [98] on the recently proposed improved fast directional chamfer approach [62]. The method by Liu et al. [62] achieves state-of-the-art performance in chamfer-based matching and it is publicly available, thus enabling our extension to be easily applicable. Let us now briefly review the fast directional chamfer matching [62] and introduce the required notation. Let $T = \{\mathbf{t}_i\}$ and $Q = \{\mathbf{q}_j\}$ be the sets of template and query edge map respectively. Let $\phi(\mathbf{t}_i)$ denote the edge orientation of the edge point $\mathbf{t}_i$.

For a given location $\mathbf{x}$ of the template in the query image, directional chamfer matching aims to find the best $\mathbf{q}_j \in Q$ for each $\mathbf{t}_i \in T$ by minimizing the cost $|(\mathbf{t}_i+\mathbf{x})-\mathbf{q}_j|+\lambda|\phi(\mathbf{t}_i+\mathbf{x})-\phi(\mathbf{q}_j)|$. λ denotes the weighting factor between location and orientation terms. Thus the directional chamfer distance for placing the template at location $\mathbf{x}$ is defined as

$$d_{DCM}^{(T,Q)}(\mathbf{x}) = \frac{1}{|T|}\sum_{\mathbf{t}_i \in T} \min_{\mathbf{q}_j \in Q} |(\mathbf{t}_i + \mathbf{x}) - \mathbf{q}_j| + \lambda|\phi(\mathbf{t}_i + \mathbf{x}) - \phi(\mathbf{q}_j)| \tag{6.1}$$

where λ denotes the weighting factor between location and orientation terms.

6.1.1 Learning the Relevance of Model Points

Not all the pixels on the shape template are equally important for detecting objects. Consider for instance the famous Kanizsa triangle. Provided only contour fragments around the corners, the whole triangle can be easily recognized. Similarly, Biederman [22] presents perceptual experiments with degraded contours that demonstrate the varying importance of different points on object contours. Another example is Attneave's cat [14], where for instance, points of high curvature are proposed as the most useful features for recognition. However, we do want to automatically learn, which parts of the model are important, rather than manually encoding a set of rules that define the importance of contour points.

In chamfer matching, matching costs for a template are obtained by summing over all the template pixels in the distance transform of the query image as in (6.1). Thus, all the pixels are implicitly considered to be equally important when computing the matching costs. To take into account the fact that not all pixels are equally important, we learn discriminative weights for the co-occurrence of individual template points, i.e., of their

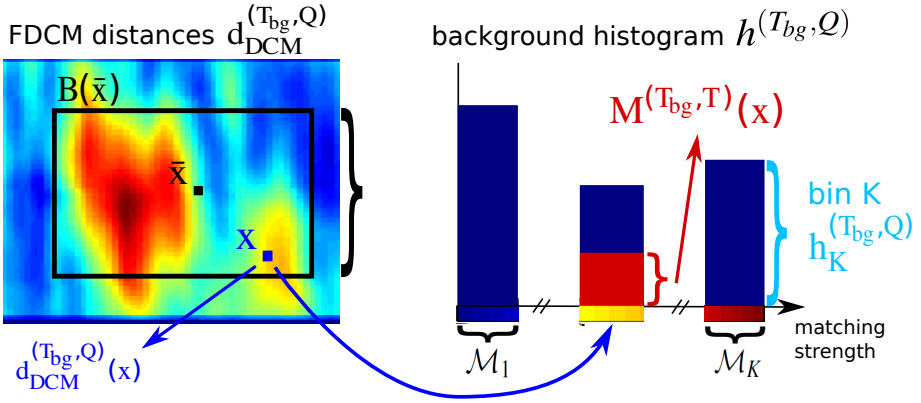


Figure 6.2: Construction of weighted background histograms from fast directional chamfer matching score maps, see (6.4). In the score map on the left, red indicates high matching score and blue indicates low matching score of the background template in the query image. For a bounding box region $B(\bar{\mathbf{x}})$ centred at $\bar{\mathbf{x}}$, each directional chamfer matching score $d_{DCM}^{(T_{bg},Q)}(\mathbf{x})$ is assigned to its corresponding histogram bin range $\mathcal{M}_k$ and casts a vote with weight $M^{(T_{bg},T)}(\mathbf{x})$ (see (6.3) and Fig. 6.3) to this bin.

matching costs $p_i^{(T,Q)}(\mathbf{x})$,

$$p_i^{(T,Q)}(\mathbf{x}) = \min_{\mathbf{q}_j \in Q} |(\mathbf{t}_i + \mathbf{x}) - \mathbf{q}_j| + \lambda |\phi(\mathbf{t}_i + \mathbf{x}) - \phi(\mathbf{q}_j)| \tag{6.2}$$

Adjacent template pixels are statistically dependent and, thus, we do average (6.2) over the direct neighbours of pixel i. The resulting $\bar{p}_i$ are then used to learn the importance of contour pixels. The discriminative learning algorithm that discovers the weights for the co-occurrences of pixels is described in Sect. 6.1.3. For visualization purpose, we learned the importance of each pixel using a linear SVM and display the resulting weights for various shape templates in Fig. 6.1.

6.1.2 Using Background Contours to Model Accidentalness

Chamfer matching is notoriously prone to spurious matches in background clutter. Although adding orientation information [62, 87] and learning the relevance of foreground pixels increase the specificity of the approach, they fail to eliminate false positives in intense clutter (for an example see Fig. 6.5). Consequently we need to measure the accidentalness of a match. We use a codebook of simple, generic contour segments, which obviously feature a low specificity and high accidentalness. To obtain the set of simple contour segments we collect differently oriented straight and curved lines (see Fig. 6.3 a)). These simple contours will be called background contours T_{bg} in the following. As a negative side effect these background contours will, however, also respond to the foreground object. To make up for the lack of specificity of individual contours we learn discriminative co-occurrence patterns of all of these background contours. These co-occurrence patterns identify matches to clutter and distinguish them from actual foreground matches. In contrast to [65], who manually combine tuples of normalizers consisting of one or two contours to form hand designed complex background templates, we propose to automatically learn flexible arrangements of all the background contours to improve detection accuracy.

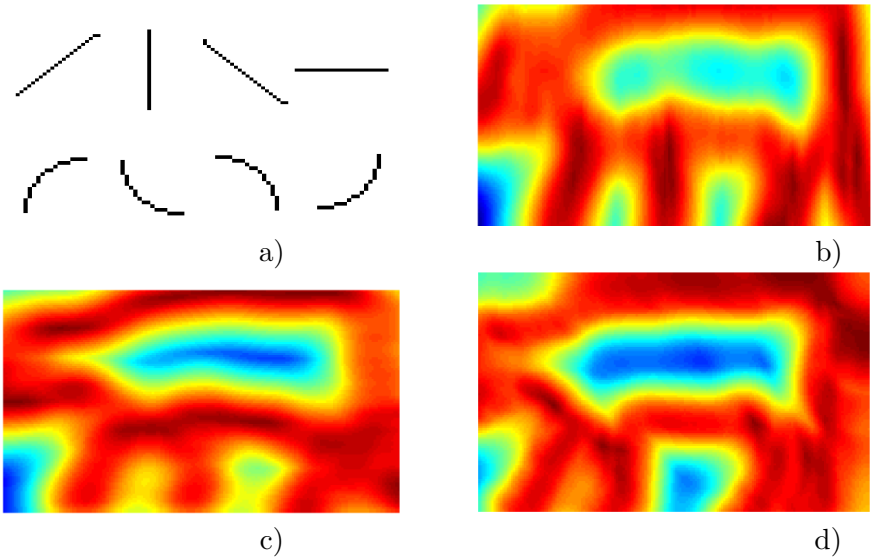


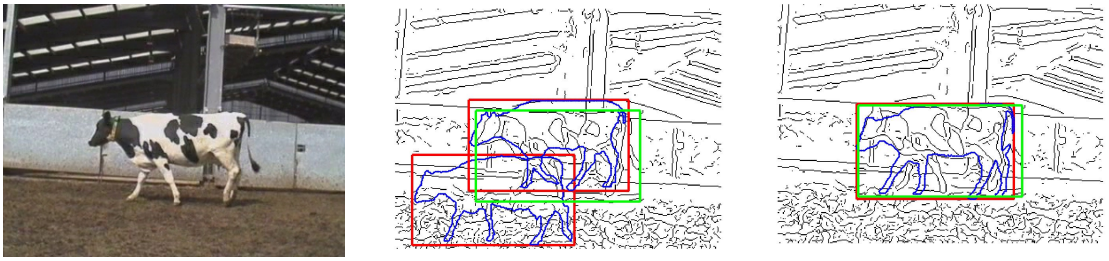
Figure 6.3: A set of simple background contours T_{bg} is shown in a). These background contours were placed relative to the cow shape mask shown in Fig. 6.1 to create masks described in (6.3). b)-d) show the resultant masks. b) shows the mask for the vertical line, c) shows the mask for the horizontal line and d) shows the mask for arc 3 in the second row of panel a). Red indicates high weight and blue indicates low weight.

False positives occurring in background clutter are caused by the edges in the query image at the locations where the foreground contour is placed. Consider a U-shaped template being matched to a query image. Clutter from the query image that is situated within the U does not interfere with the template. Only clutter that is close to the contour of the U will have an impact. Therefore, we need to check for spurious background contours in the neighbourhood of model contours, but not elsewhere. In contrast to this, [65] place background contours at a fixed single location, i.e., at the center of the model contour, thereby not measuring the susceptibility of the model contour to clutter. Rather than measuring the amount of clutter on the template contour where it actually matters, they check for clutter simply at the center of the object.

To measure where clutter typically interferes with the model contour we compute the directional chamfer matching score $d_{DCM}^{(T_{bg},T)}$ between each background contour and the object template. We consider placements of the background contour with better (lower) chamfer matching score to be more important since they occur on or close to the model contour. In order to weight these matching locations higher we create a mask

$$M^{(T_{bg},T)}(\mathbf{x}) = 1 - d_{DCM}^{(T_{bg},T)}(\mathbf{x}) \tag{6.3}$$

from the directional chamfer matching scores. Each combination of a foreground template and a background contour results in a different mask. Fig. 6.3 shows examples of these masks for different background contours. One can see that high weight is assigned where the background contour matches well to the foreground contour and low weight otherwise. Therefore matches of background contours inside the object are less important than those on the object boundary.



a) b) c)

Figure 6.4: Learning discriminative weights for the co-occurrences of $p_i^{(T,Q)}(\mathbf{x})$ improves the matching score of shape template as shown in the example here. The original image, the result obtained from directional chamfer matching and the result obtained from foreground reweighing are shown in panels a,b and c respectively. The ground-truth bounding box is shown in green and the top scoring object hypotheses are shown in red.

To describe the background matching costs for a hypothesis in a robust way we are building weighted histograms over chamfer matching scores $d_{DCM}^{(T_{bg},Q)}$ obtained from matching a background contour T_{bg} with the query image Q. Let $B(\bar{\mathbf{x}})$ be the bounding box region with center $\bar{\mathbf{x}}$ for a specific placement of the foreground template T in the query image Q (see Fig. 6.2). For each foreground hypothesis we build weighted histograms $h^{(T_{bg},Q)}$ over the directional chamfer matching scores $d_{DCM}^{(T_{bg},Q)}$ in the corresponding bounding box region. The weights introduced in (6.3) are used to weight the histogram votes. Therefore chamfer matching scores $d_{DCM}^{(T_{bg},Q)}$ are weighted according to their position relative to the foreground template. Each histogram consists of K bins where $\mathcal{M}_k$ is the range of the kth bin and $k = 1, ..., K$. We define a histogram bin $h_k^{(T_{bg},Q)}$ as

$$h_k^{(T_{bg},Q)} = \sum_{\substack{\mathbf{x}\in B(\bar{\mathbf{x}}) \\ d_{DCM}^{(T_{bg},Q)}(\mathbf{x})\in\mathcal{M}_k}} M^{(T_{bg},T)}(\mathbf{x}), \tag{6.4}$$

for each background contour T_{bg} on a certain position of the foreground template T in the query image Q (see Fig. 6.2).

6.1.3 Learning Chamfer Regularization

From above we know that we need to model the co-occurrence of all template points. Moreover, a codebook of simple generic contours needs to be matched close to the template contour where accidental matches typically occur. We combine these challenges in one discriminative approach.

The aim is to regularize directional chamfer matching by learning the characteristic co-occurrence of template pixels and the joint placement of background contours.

As training data this learning algorithm utilizes the object hypotheses obtained from running the directional chamfer matching code [62] on the training images.

A hypothesis j with an overlap greater than 80% with the ground-truth is labelled as positive $y_j = 1$. This ensures that only good hypotheses which are matching to the actual

object contours are selected as positive examples. For negative examples, we want to have the hypotheses where most of the object template matches in the background. Therefore, hypotheses with an overlap smaller than 40% with the ground-truth are labelled as negative $y_j = -1$. The learning algorithm is found to be robust to small variations in the cutoff values of 80% and 40% overlap criterion with the ground-truth. For each object hypothesis we build a feature vector $f_j = [\bar{p_1} \ ... \ \bar{p_L} \ h_1 \ ... \ h_G]$ consisting of the average pixel cost $\bar{p_i}$ and the corresponding background histograms h_i, where L is the number of template edge pixels and G is the number of background contours.

Let $\mathcal{K}(f_i, f_j)$ be a kernel that represents the similarity between feature vectors f_i, f_j. Subsequently, we use the radial basis kernel $\mathcal{K}(f_i, f_j) = \exp(-\gamma \|f_i - f_j\|^2)$. It is common practice in the field of kernel machines, to interpret the kernel $\mathcal{K}(f_i, f_j)$ as a dot product of transformed features $\psi(f_i), \psi(f_j)$. Here ψ represents the mapping of the feature vector into a higher dimensional space. Due to the seminal 'kernel trick' [25] it is sufficient to define the kernel $\mathcal{K}$ without explicitly representing the mapping ψ. We then seek weights w to be applied on $\psi(f_i)$ so that the margin between positive and negative hypotheses in the transformed space is maximized. To model the joint co-occurrences of foreground and background contours we need to utilize a non-linear kernel that captures the relationship between foreground and background pairs, triples, quadruples and so on. From the polynomial kernel $\mathcal{K}(f_i, f_j) =< f_i, f_j >^2$ of degree 2 one can easily determine, that the mapping function ψ comprises all possible second order terms. It is straightforward, that a polynomial kernel of degree d comprises all possible combinations between feature dimensions up to degree d. Since the Taylor expansion of the RBF kernel is a infinite set of features corresponding to polynomial terms it comprises an infinite amount of feature combinations. We need to optimize the following max-margin classification problem to learn the weights w.

$$\begin{aligned} &\min_{w,b,\xi} \frac{1}{2}\|w\|_2^2 + C\sum_{j=1}^{N}\xi_j \\ &\text{subject to :} \qquad y_j(w^T\psi(f_j) + b) \geq 1 - \xi_j \quad \wedge \quad \xi_j \geq 0, \quad \forall j \end{aligned} \tag{6.5}$$

where N is the number of training samples, b is the offset, C is the penalty and ξ_j are slack variables allowing for margin violations. Commonly (6.5) is converted into its dual form and solved for the dual SVM parameters, the support vectors S_i, their coefficients α_i and the offset b.

After training the combined model of foreground relevance and background accidentalness from (6.5) let us now utilize this model to improve upon the directional chamfer matching cost function (6.1). This improved, regularized chamfer distance $d^{(T,Q)}_{RDCM}(\mathbf{x})$ again measures the distortion cost of object hypotheses f_j. f_j denotes the feature vector of j-th object hypothesis obtained by the placement of object template T at location $\mathbf{x}$ in the query image Q. Since a non-linear radial basis kernel is employed, the regularized chamfer distance is obtained using the dual SVM parameters, obtained by solving the SVM optimization problem from (6.5) in its dual form,

$$d^{(T,Q)}_{RDCM}(\mathbf{x}) = 1 - \Big(\sum_i \alpha_i \mathcal{K}(f_j, S_i) + b\Big). \tag{6.6}$$

As in standard chamfer matching, candidate hypotheses are obtained by applying non-maximum suppression onto the regularized distances d_{RDCM}.

	Pedestrians	Cows	Giraffes	Mugs
DCM	3.0	88.1	27.0	10.1
Foreground Regularization	6.8	89.2	36.3	27.3
Regularized Chamfer Matching	**11.2**	**91.9**	**43.0**	**27.3**

Table 6.1: Comparison of **average precision** (in %) for three datasets namely, TUD Pedestrians, Cows and the ETHZ giraffes and mugs. We compare the basis of our approach (DCM) with the extension from Sec. 6.1.1 and our final learning of regularized chamfer matching.

6.2 Experimental Evaluations

We now evaluate the discriminative chamfer regularization on several datasets which are commonly used for evaluation of chamfer matching. In particular, we compare with the directional chamfer matching (DCM) [62], which our model is built upon and with normalized oriented chamfer matching (NOCM) [65], which is a state-of-the-art extension to chamfer matching.

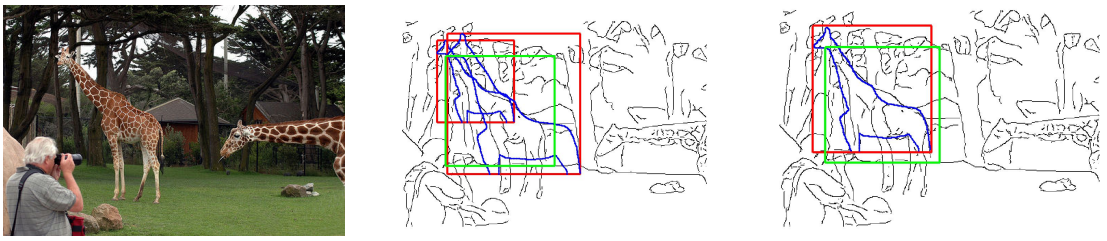
To obtain the edge maps used in the following we are utilizing the probabilistic boundary detector suggested in [69]. Furthermore we are using the support vector machine implementation of [29]. To perform directional chamfer matching, we are using the publicly available code of [62]. We use the same parameters from the downloaded version of the code for all the datasets. We used the same set of background shapes, as shown in Fig. 6.3, for all the datasets. The sizes of the background contours were adjusted relative to the size of foreground templates for each dataset. To measure the performance of our detection system we are using standard PASCAL overlap criterion.

In the first part of our experimental evaluation we are analysing the individual contributions of the suggested foreground and background regularization and compare their performance to that of DCM on which we build our approach. In the second part we compare the performance of our combined object detector to state-of-the-art chamfer matching approaches NOCM and HDT.

6.2.1 Evaluating Foreground and Background Regularization

Subsequently, we evaluate the gain achieved by the proposed foreground and background regularization on context of category-level object detection in three standard datasets and we compare our results with the DCM baseline on which we build our approach.

The first dataset we are using is the TUD pedestrian dataset. As suggested in [65] we are using the larger training set, consisting of 400 side-view pedestrians, to build our detector and test our approach on the provided test-set consisting of 250 test images. We use five masks of the training images as shape templates. The second dataset is the Cow dataset from the PASCAL Object Recognition Database Collection [59] which consists of 111 images in which cows appear with quite different articulation. We are following the protocol used in [65] to divide the dataset into training and testing sets. We use five masks of the training images as our shape templates. Finally, we evaluate on two challenging categories from the ETHZ shape dataset [44], giraffes and mugs. One hand-drawn template for each category is provided along with the dataset.



a) b) c)

Figure 6.5: Learning co-occurrences of foreground shape template alone is not enough as shown in the example here. The original image, the result obtained from foreground reweighing and the result obtained from the combined foreground and background regularization are shown in panels a,b and c respectively. The spurious hypothesis resulting in panel b is suppressed by means of the combined regularization learned in (6.5)

Our approach is efficient, as looking up the background contour costs form the integral image has negligible running time compared to computing the distance transformation of directional chamfer matching [62]. In Tab. 6.1 we are presenting our results for the DCM baseline, the performance of our foreground regularization method and of our combined detector. These experiments show that foreground regularization alone is already improving the average precision on all of these object categories. Additionally applying the background regularization is suppressing even more false positives in cluttered background.

For the TUD Pedestrian dataset the images in the testing set are given at a very high resolution which yields very low average precision for the directional chamfer matching which is around 3%. The low baseline can be attributed to the high resolution of the test images, since it is known that chamfer matching is sensitive to all the fine details in the edge map. Our suggested foreground regularization more than doubled the average precision to the baseline. Adding the background regularization brought a further gain of 4.5%.
For the Cow dataset directional chamfer matching yields very good performance around 88% average precision. Nevertheless, our combined detector could still improve performance about 4% by exploiting the advantages of foreground and background regularization. In Fig. 6.4 one can see how foreground reweighing is improving the alignment with the ground-truth and that it also suppresses false positives.
The background normalization becomes particularly useful in cases of challenging objects appearing in images with a lot of clutter like the ETHZ giraffes. Performance improves by 16% in terms of average precision using our combined detector. 7% out of this gain could be attributed to background regularization. The example in Fig. 6.5 shows that foreground regularization is not always able to suppress false positives in cluttered background and how background regularization can handle such cases.
For rather simple objects like ETHZ mugs we observed that explaining the foreground more accurately is more important than suppressing false detections in cluttered background. We observed 17.3% improvement in average precision by learning the co-occurrence of template pixels while our combined detector is giving results in the same range.

All in all our combined detector using foreground and background regularization is achieving significant gain on all of the four categories compared to directional chamfer matching. Additional detection results comparing the regularized chamfer matching to directional chamfer matching are provided in Fig. 6.6.

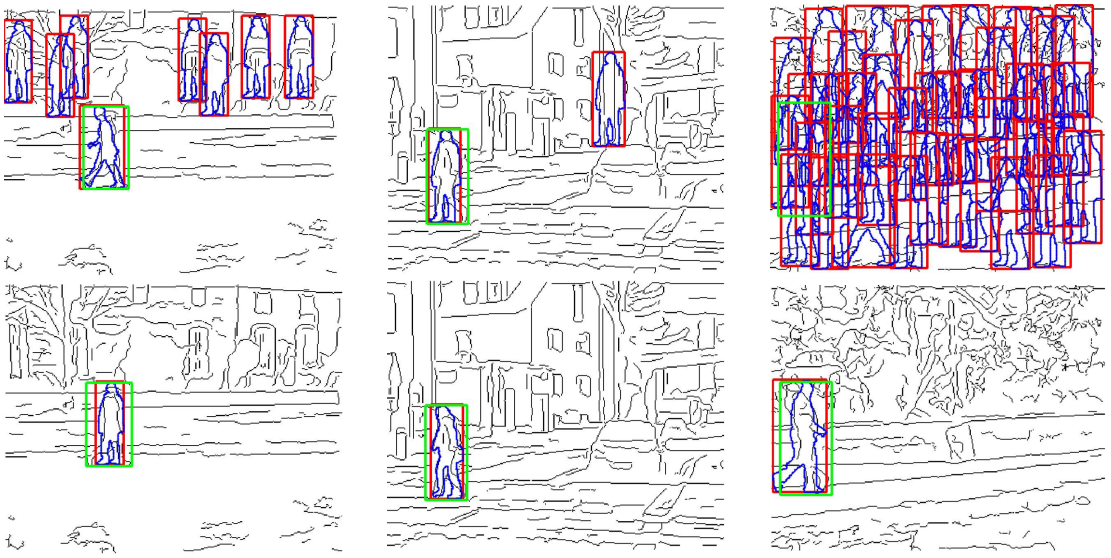


Figure 6.6: The top row shows detection results using the directional chamfer matching method. Bottom row shows the improved detections applying our regularized chamfer matching. The groundtruth bounding box is shown in green and the top scoring object hypotheses are shown in red.

6.2.2 Comparison with Chamfer Matching Methods

Furthermore we are comparing our method with two other state-of-the-art approaches on three datasets. The first method we are comparing our approach to is the normalized oriented chamfer matching by Ma et al. [65] (NOCD) since they also incorporate background into chamfer matching. We also compare our approach with the work of Zhu et al. [104] who utilize a novel probabilistic model called hierarchical deformable template model (HDT). [104] use one example learning in their evaluation whereas we utilize 5 templates for the TUD Cows and TUD Pedestrians.

[65] have reported results on two datasets: the TUD Pedestrian dataset [11] and the Cow dataset [59]. [104] have evaluated their method on the Cow dataset. Both approaches are reporting their results in terms of detection rate at 10% precision. In the previous section we are reporting in terms of average precision, since it is taking into account the area under the precision recall curve instead of just one point and therefore is a much more robust measure. However, to compare ourselves to [65, 104], we are reporting results in terms of detection rate at 10% precision.

Tab. 6.2 shows the results for the Cow dataset and the TUD Pedestrian dataset. We observed that to make the DCM baseline comparable to the OCM baseline the edge maps in the test images need to be downscaled. Hence, we report our final detection performance on the downscaled version of the test images. The results indicate that chamfer regularization is significantly improving performance on the Cow dataset compared to HDT and NOCM. For TUD Pedestrians we gain 10% in detection rate compared to NOCM, when running the directional chamfer matching on downscaled test images. All in all our results confirm that the regularized chamfer matching method is significantly improving over state-of-the-art chamfer matching techniques.

	Cows	Peds
Chamfer Matching	73.9	4.4
NOCM [65]	91.0	70.0
HDT [104]	88.2	-
Regularized Chamfer Matching	**98.3**	**80.0**

Table 6.2: Comparison in terms of **detection rate** (in %) at 10% precision on the Cow dataset and the TUD Pedestrian dataset with standard chamfer matching, NOCM and HDT.

6.3 Discussion

This work has addressed two issues that limit the performance of the established and widely used chamfer matching technique, its susceptibility to clutter due to accidental matches and the fact that all model points are treated as being independent and equally relevant. By learning the co-occurrence of model points we have modelled the varying relevance of different foreground pixels and increased the specificity of the model. By allowing a codebook of simple, generic contours to be flexibly placed along the model contour where spurious matches are most likely, accidental matches can be discovered. Learning the joint placement of all of these generic background contours does then suppress accidental matches to clutter. Both extensions are integrated in a single discriminative learning approach and the method is based upon a publicly available, state-of-the-art chamfer method thus demonstrating its simple and wide applicability. The approach has been shown to successfully improve current chamfer matching approaches on standard datasets.

CHAPTER 7

CASE-STUDY ON MEDIEVAL MANUSCRIPTS

The final part of this thesis (chapters 7 and 8) makes use of shape-based object representations to provide a semantic understanding of image collections. To demonstrate this capability, a Case-Study has been carried out on Upper German manuscript of medieval images. Sect. 7.1 highlights the goal of the Case-Study and describes the benchmark dataset which has been assembled for the same purpose. Also, the current way of accessing the datasets in art history and the corresponding limitations are described in Sect. 7.3. Finally, some useful concepts are reviewed towards the end of this chapter which form the basis for the approaches presented in Chapter 8.

7.1 Goals of the Case-Study

The large amounts of visual data that recent digitization projects are providing to the field of cultural heritage call for methods from scientific computing to efficiently open up these resources. This interdisciplinary cooperation requires algorithms which advance beyond a mere analysis of individual pixels onto a stage where the semantics of images can be analysed. Thus the goals of the Case-Study are 1) searching through the image collections for different objects of interest such as crowns, swords 2) identifying the subcategories of an object type 3) identifying different artistic workshops to which the objects belong 4) understanding the variations of art within a particular school of design and 5) understanding the transition of art from one school of design to another.

Manually performing the above tasks is a tedious process for humans which require a great deal of time and effort. Obviously no human user can view all of these images at the same time and, thus, relations between different images or the objects within are hard to discover. Revealing the structure that is inherent to a collection of images, i.e., the artistic variations of all instances of an object category such as medieval crowns, is consequently a very difficult task. The mere size of a dataset makes it difficult to see the greater whole. Computers on the other hand can easily handle thousands of images at the same time

Figure 7.1: Manuscript pages.

To demonstrate the capability of algorithms in performing the above tasks, we have assembled a novel image dataset that is highly significant for the humanities due to its unusual completeness of late medieval workshop production. From the Computer Vision point of view, this dataset is the first of its kind to enable benchmarking of object retrieval in pre-modern tinted drawings. The next section describes the details of the dataset that we have assembled.

7.2 Benchmark Dataset

We have assembled a novel, annotated benchmark image dataset [100, 99] for cultural heritage from a corpus of 27 late medieval paper manuscripts held by Heidelberg University Library [76]. Produced between 1417 and 1477 in three important Upper German workshops, this corpus is rare in its magnitude and, in addition, offers an exceptional homogeneity concerning its date of origin, its provenance and its technical execution. More than 2,000 half- or full-page tinted drawings illustrate religious and devotional texts, chronicles and courtly epics. We start from object categories which have a high semantic validity since they belong to the realm of medieval symbols of power [84]. So we can ensure right from the start that our approach has the highest possible connectivity to research in the humanities, e.g to art history and history with a focus on ritual practices [85] or on material culture.

Breakthroughs entailed by a novel benchmark dataset: Our motivation for introducing a novel benchmark dataset is spurred by the influence the Berkeley Segmentation Dataset (BSDS) [49] has had on the development and evaluation of segmentation algorithms. Before BSDS, measuring segmentation performance was mostly subjective and algorithms were difficult to compare. The new BSDS dataset with its ground-truth annotation has, for the first time, provided an objective performance measure for segmentation. This has stimulated algorithm development which lead to previously unexpected breakthroughs in segmentation performance. The F-measure, which is a suitable metric for comparing the performance of segmentation algorithms, has only seen a slight increase in the years before BSDS. Early segmentation algorithms such as Roberts (1965) [79] and Canny (1986) [26]

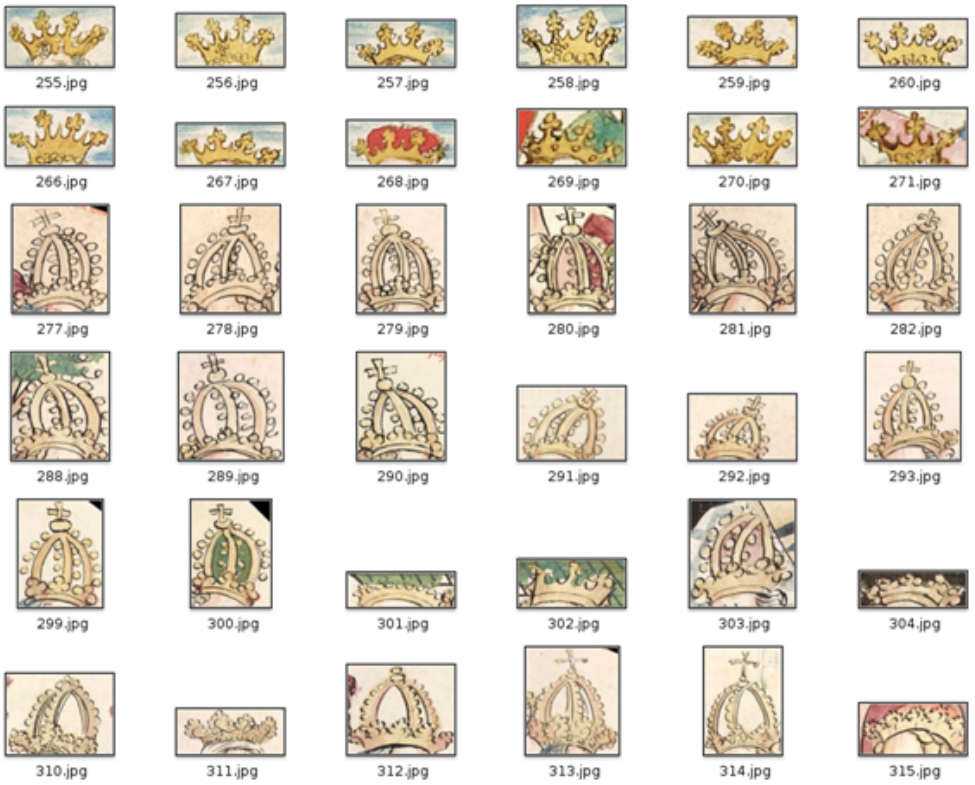


Figure 7.2: Examples of crowns from the dataset.

achieved F-measures of 0.47 and 0.53, respectively. In the short time since the introduction of BSDS in 2001, contributions such as [12] have increased the performance to 0.7 while human performance stands at 0.79.

Annotating the data: In order to generate ground-truth localizations for objects in the images, we developed an interactive annotation system. Using the expertise of an art historian we have gathered ground-truth annotations. Cubic splines are used to fit a bounding region to the principal curvature of an object. This helps excluding more background from the bounding boxes compared to rectangular bounding boxes.

7.3 Current Indexing of Art History Databases and its Limitations

Image databases in the field of cultural heritage are normally made accessible via textual annotations referring to the representational content of the images [15]. Therefore, searching for objects depends on either the controlled vocabularies of the search systems or the textual content of free descriptions. In both cases only that can be found what has been considered in the process of manual indexing; and it can only be found in the specific form in which it has been verbalized. The inevitability of textual descriptions generates numerous problems, for example concerning the scope and detail of the taxonomies, their compatibility beyond linguistic [54], professional or cultural boundaries, their focus on specific aspects of the content according to specific scientific interests or not least the qualification and training of the cataloguer.

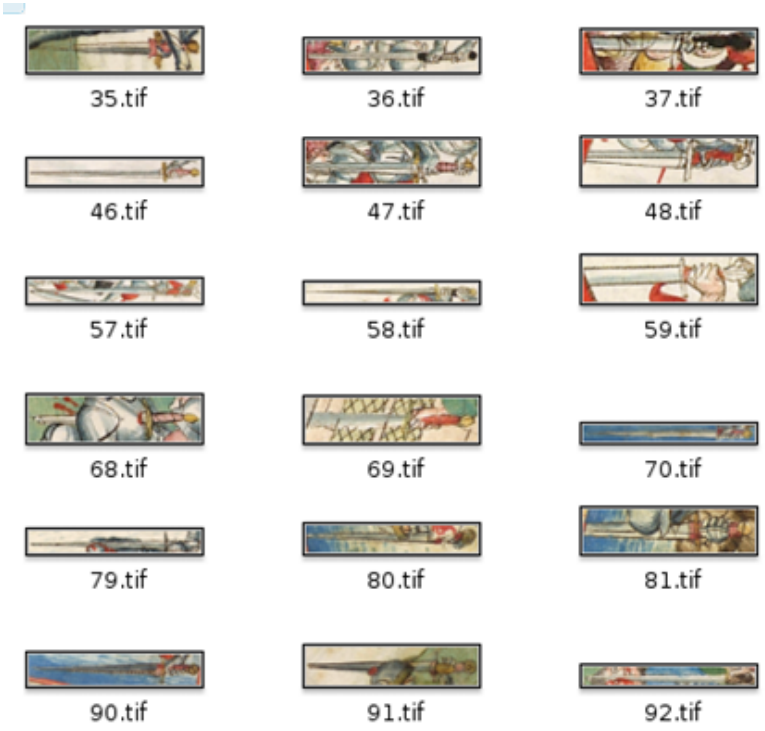


Figure 7.3: Examples of swords from the dataset.

One of the most sophisticated search systems is ICONCLASS [92]. Yet, despite its high level of differentiation it has severe limits in a global perspective because it was developed only to cover Western art and iconography. Therefore its ability to index for instance transcultural image resources such as the database of the Cluster of Excellence 'Asia and Europe in a Global Context' at the University of Heidelberg [4] is limited. Furthermore, object definition schemes are featuring a very limited differentiation. Consider the object category 'crown'. The hierarchy of objects ends with this general notion and does not offer varying types of crowns. To focus the object retrieval on subtypes is, in contrast, possible in the case of REALonline, the most important image database in the field of medieval and early modern material culture [5]. Here, the controlled vocabulary contains a few compounds like 'Bügelkrone' or 'Kronhut'. But whereas the main division 'Kleidung-Amtstracht' is searchable in German and in English, these subdivisions are available only in German, thus raising difficulties of translation. Problems such as the lack of detail and connectivity are even greater in the case of heterogeneous databases generated by the input from different contexts such as HeidICON [1], Prometheus [6] or ARTstore [3]. In such cases, the cataloguing of the image content is almost arbitrary due to the uncontrolled textual descriptions. Finally, a basic problem of all these databases is the fact that, due to the serious efforts of manual indexing in terms of cost and time, the fast-growing number of digital images will simply remain undiscovered because of lacking annotations.

To make image databases accessible in a quicker, more reliable, detailed and differentiated way, the images need to be searched based on the visual content, rather than accompanying textual annotations, for the objects in question.

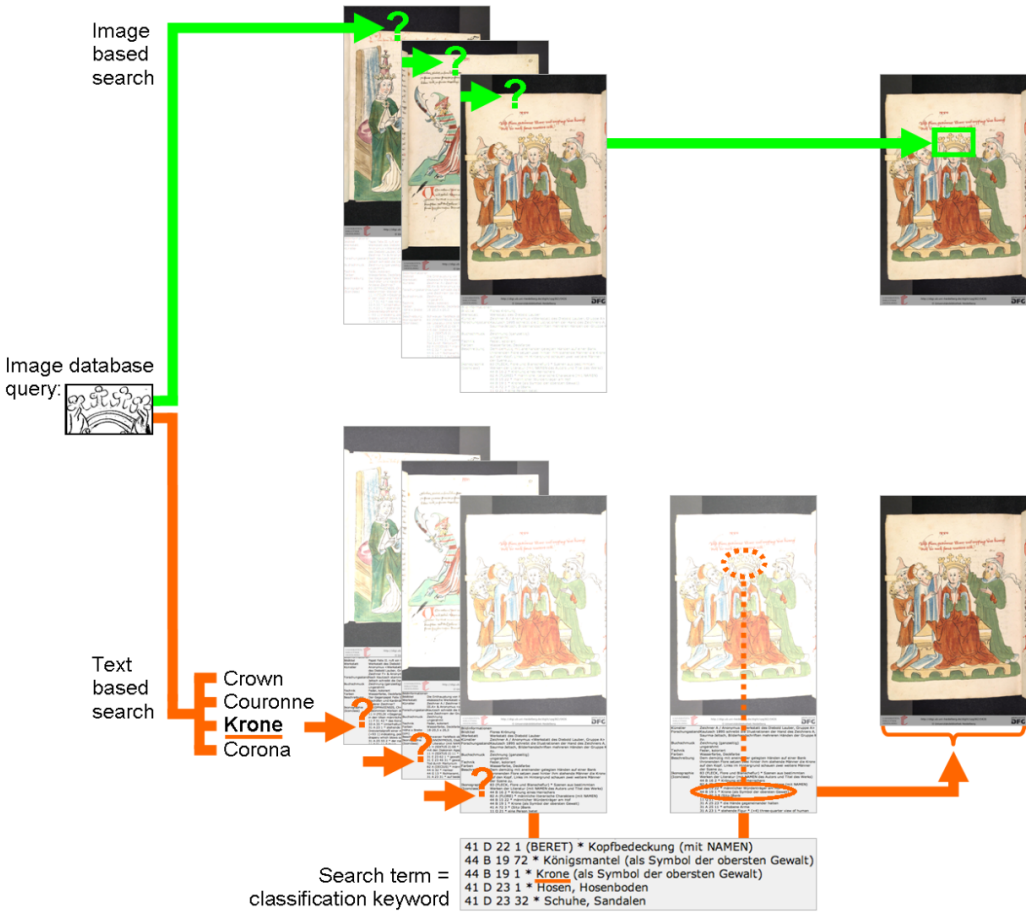


Figure 7.4: Text based vs image based retrieval.

7.4 Review

This section provides a review of two useful concepts, Histogram of Oriented Gradients (HoG) and Multi-dimensional Scaling (MDS), which are going to be used in Chapter 8 for providing a semantic understanding of image collections. HoG is chosen to represent objects such as crowns and swords in the database. MDS is a statistical technique useful for projecting the high dimensional HoG reprentation of objects into a low dimensional space, such as 2-d space which provides a simple visual summary of all object instances.

7.4.1 Histogram of Oriented Gradients

One of the most popular object representations in Computer Vision is Histogram of Oriented Gradients (HoG) [34]. Under this representation, an object is described by the distribution of intensity gradients in its corresponding image patch. The image patch is divided into small equally sized connected cells. A separate histogram based on gradient directions per pixel is computed for each cell. The cell histograms are normalized over larger spatially connected blocks so as to make the histograms invariant to local illumination and contrast changes. The final HoG descriptor of the object is formed by concatenating the cell histograms from all the normalized blocks.

7.4.2 Multi-dimensional Scaling

Multi-dimensional Scaling [52] is one of the popular techniques to map data points from a high dimensional space to low dimensional manifold. Let $\mathbf{f}_1, \mathbf{f}_2, ..., \mathbf{f}_N$ denote the data points in a high dimensional space $\mathcal{R}^p$. Let d_{ij} denote the distance between instances i and j. The objective of MDS is to seek embeddings $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N \in \mathcal{R}^k$ of the original data points so that the distances between the points are preserved in the $\mathcal{R}^k$ space. More specifically, the following stress function is minimized.

$$\mathbf{x}^* = argmin_{\mathbf{x}} \sum_{i \neq j} (\|\mathbf{x}_i - \mathbf{x}_j\| - d_{ij})^2. \tag{7.1}$$

x denotes the configuration $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N$ of all the N points in $\mathcal{R}^k$ space. The least squares cost function in (7.1) is optimized by gradient descent approach. Least squares falls under the metric scaling methods because the actual distances between the data points are approximated. Non metric scaling on the other hand seeks to minimize stress function of the form

$$\mathbf{x}^* = argmin_{\mathbf{x}} \frac{\sum_{i \neq j} (\|\mathbf{x}_i - \mathbf{x}_j\| - \theta(d_{ij}))^2}{\sum_{i \neq j} \|\mathbf{x}_i - \mathbf{x}_j\|^2}. \tag{7.2}$$

θ denotes an arbitrary increasing function. With θ fixed, x is optimized in (7.2) using gradient descent approach. With x fixed in (7.2), the best monotonic approximation $\theta(d_{ij})$ to $\|\mathbf{x}_i - \mathbf{x}_j\|$ is found by isotonic regression. These two steps are iterated until convergence.

CHAPTER 8

SEMANTIC UNDERSTANDING OF IMAGE COLLECTIONS

This chapter describes the various components of a semantic understanding system for medieval manuscripts, i) object detection (Sect. 8.1) ii) low dimensional embedding of objects (Sect. 8.2) leading to workshop identification iii) classification of objects into various workshops (Sect. 8.3) iv) A semi-supervised approach (Sect. 8.4) for analysing the intra-class variability of objects and for inducing one dimensional ordering of crowns between a pair of selected crowns.

8.1 Object Detection

8.1.1 Object Analysis

The most basic component for object analysis and object recognition is choosing an appropriate mathematical representation for objects which lays the foundation for recognition and further analysis. We utilize a shape-based representation of objects since shape is an important cue in medieval manuscripts.

Extracting artistic drawings to represent shape: We have discovered from experiments that the images when represented in HSV color space, particularly the saturation component, provide a good starting point for object boundary extraction. Object boundaries are essentially ridges in an image with few pixels thickness. To detect such ridges, we apply a filter which smooths the image along the direction orthogonal to the ridge and sharpens the image along the direction of the ridge,called the ridge detection filter [56]. It is defined by the following formula.

$$G(x, y, \sigma_x, \sigma_y) = \frac{1}{\pi * \sigma_x^2} * \left(1 - \frac{x^2}{2 * \sigma_x^2}\right) * \exp\left(-\frac{x^2}{{\sigma_x}^2} - \frac{y^2}{{\sigma_y}^2}\right) \tag{8.1}$$

Figure 8.1: Input image and ridge output.

Coordinates x,y denote image location, σ_x, σ_y determine the support of the ridge filter along the x and y directions. (8.1) defines the ridge filter assuming that the ridge is oriented along the x-axis. This formula is easily extended for detecting ridges at an orientation θ.

At each point in the image, optimization over the parameters σ_x, σ_y and θ yields the maximal filter response. Fig. 8.1 shows an input image and the result of applying the ridge filter to the input.

Shape representation: Ridges are represented using orientation histograms. We compute these Histograms of Oriented Gradients (HoG) [35] on a dense grid of uniformly spaced cells in the image. We combine histograms from 4 different scales and 9 orientations into a 765 dimensional feature vector.

Detection algorithm:

Objects are detected by classifying image regions as object or background using a support vector machine with intersection kernel [67]. This detection algorithm scans the image on multiple scales and orientations. Image regions are represented using the shape representation from subsection 3.2 and a color histogram. The necessary codebook of representative colors is obtained by first quantizing training image using minimum variance quantization into a set of 100 prototypical clusters per image. The bias towards large, homogeneous regions is resolved by clustering all these prototypes into an overall set of 30 prototypical colors. We count an object hypothesis as correct if $\frac{A_h \cap A_g}{A_h \cup A_g} >= 0.4$

where A_h and A_g is the area of the predicted and the ground-truth bounding box, respectively. The precision-recall curve in part a) of Fig. 8.3 shows the detection performance achieved by the presented approach.

The precision recall curves in Fig. 8.3 show scope for improvement as the curves are far from reaching the saturation stage. A closer look at the detection results revealed a lot of false positives in the images which were not sufficiently represented during the training stage of the SVM. To deal with this issue, we have incorporated a bootstrap training procedure to focus on difficult negative samples as is motivated by [36, 42]. Training starts as before by learning an SVM model on all positive training samples and an equally sized, random set of negative samples, i.e. bounding boxes drawn from the background. In the next round, negative samples which are either incorrectly classified by the model or fall inside the margin (defined by the SVM classifier) are added to the training set. Also,

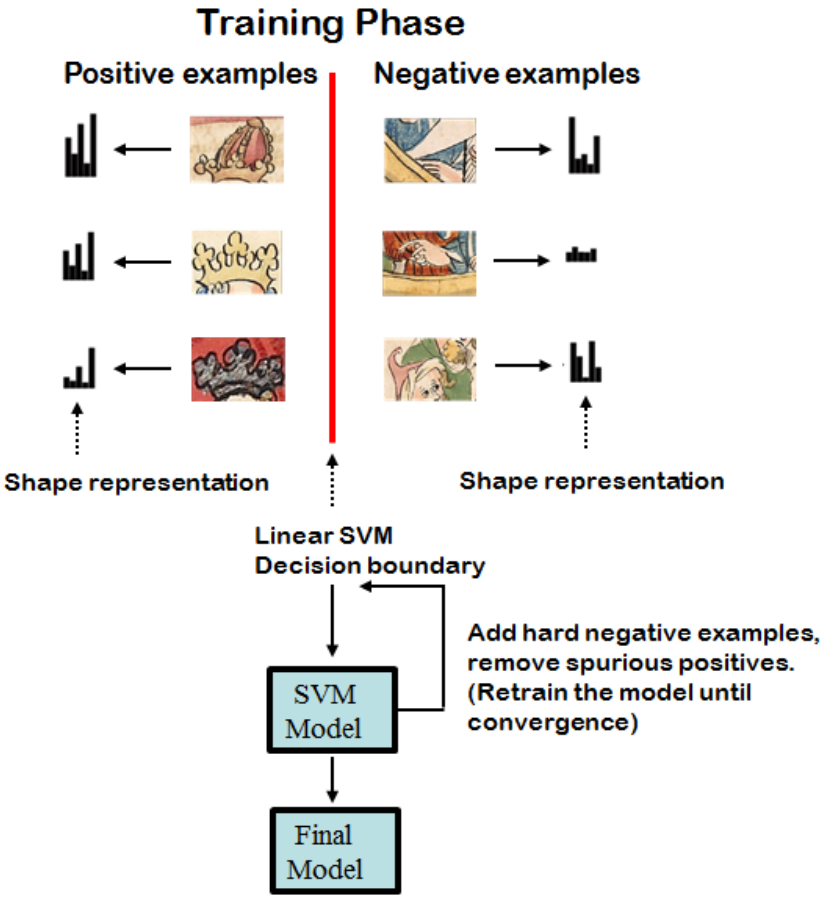


Figure 8.2: Training procedure with bootstrapping the hard examples.

positive samples which are classified correctly and fall outside the margin are removed from the training set. This process is repeated iteratively until there are no new hard negative samples that can be added to the training set. This iterative training procedure resulted in a significant improvement in the detection performance and the resulting PR curves are presented in Fig. 8.4 along with two examples of detections in test images.

8.2 MDS Analysis on Object Hypotheses in Image Collections

We capture the relationship between various object instances in the database in a single plot by embedding high dimensional HoG feature vectors into a low dimensional space. Such a plot makes it convenient for researchers from cultural heritage to discover relationships without having to study thousands of images. In a first step pairwise clustering based on HoG descriptors is employed to discover the hierarchical substructure of crowns. Then we compute the pairwise distances for samples in the vicinity of the cluster prototypes. Thereafter, a distance preserving low-dimensional embedding is computed to project the 765 dimensional feature vectors onto a 2-d subspace that is visualized in Fig. 8.6. The embedding of the crowns in the two dimensional space is given by locations $\mathbf{x}_i \in \mathbb{R}^2$ which are computed jointly using

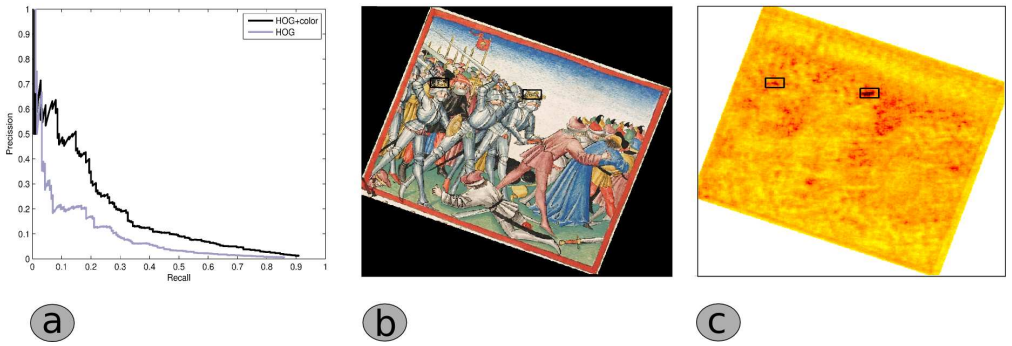


Figure 8.3: a) Precision recall curve for crowns obtained from HoG and HoG plus color features. b) Crowns detected in a test image. c) Response of our object detector at each image location.

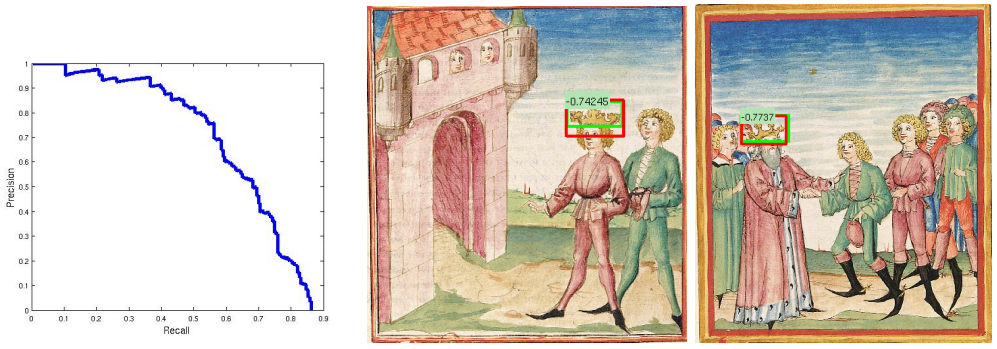


Figure 8.4: a) Precision recall curve for crowns obtained by using a bootstrapping training procedure. b) and c) Crowns detected in test images along with the SVM scores.

$$\mathbf{x}^* = argmin_{\mathbf{x}} \frac{\sum_{i \neq j} (\|\mathbf{x}_i - \mathbf{x}_j\| - d_{ij})^2}{\sum_{i \neq j} \|\mathbf{x}_i - \mathbf{x}_j\|^2}. \tag{8.2}$$

d_{ij} denote the distances between crowns i and j in the original 765 feature dimensional space.

This procedure has extracted relationships, variations and substructure of an object category out of hundreds of images and makes these directly apparent.

The plot displays two central findings of our recognition system and thus reveal the potential of the approach: i) the high type-variability within a category and ii) the different principles of artistic design. In particular, our clusters for the category 'crown' show that to the simple crown circlet (A) varied elements like arches (B1), lined arches (B2), torus-shaped brims (B3), hats, or helmets are added. Thus, objects provide advanced semantic information concerning e.g. social hierarchies, which is not displayed by the common taxonomies. Since an automated image-based search does not suffer from the desiderata of annotation taxonomies, it becomes a crucial instrument to assist with the detailed differentiation of such subtypes, combining data from large numbers of images and organizing the compositional complexity of objects into a hierarchy of formal variants. Moreover, the clustering and visualization in a MDS-plot (Fig. 8.6) features different principles of artistic design, which are characteristic for different workshops engaged with the illustrations.

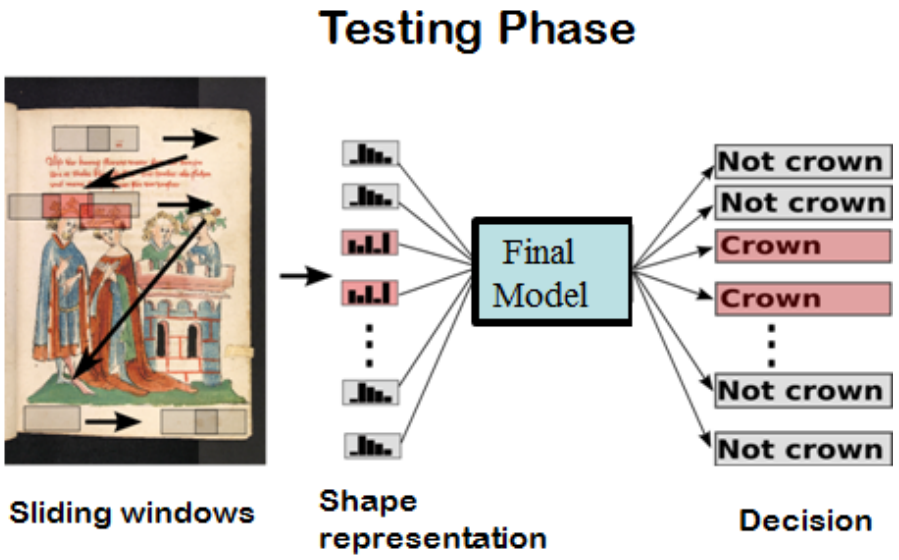


Figure 8.5: Detecting an object using SVM Model.

Workshops pred.:	A	B	C
A	0.9836	0.0163	0
B	0.0365	0.9634	0
C	0.0083	0.0083	0.9833

Table 8.1: confusion Matrix

Group (B) indicates the concise and accurate style, mainly based on definite contours, of the Hagenau workshop of Diebold Lauber [82], group (A) the more delicate and sketchy style of the Swabian workshop of Ludwig Henfflin, and group (C) the particular summary style of the so-called 'Alsatian Workshop of 1418'. This detection of specific drawing styles is a highly relevant starting point to differentiate large-scale datasets by workshops, single teams within a workshop, or even by individual draftsmen.

8.3 Workshop Classification

Based on this visualization, art historians have provided us with ground-truth information so that we can conduct a quantitative evaluation: they have labelled all crowns in the dataset with the workshop that they come from based on formal criteria [82]. There are 137 crowns in our dataset that belong to group A (the workshop of Ludwig Henfflin), 106 crowns belong to group B (the workshop of Diebold Lauber) and 23 crowns belong to group C (the Alsatian workshop). We then incorporate a discriminative approach for predicting the workshop that a crown belongs to. This multi-class classification problem is tackled using the features from before and incorporating SVM in a one-versus-all manner. For evaluation, we apply 10-fold cross-validation: In each round, 50 % of the crowns from each group have been used for training and the remaining 50 % of the crowns are used for testing by holding back their labels. The classification results of the crowns according to the workshops are presented in Tab. 8.1 in the form of a confusion matrix.

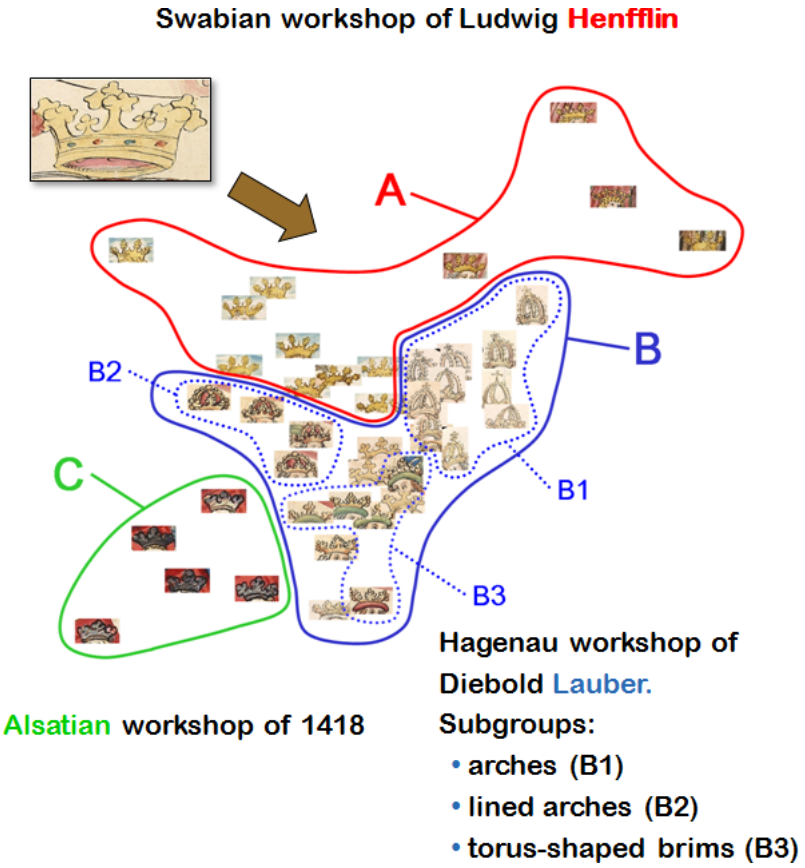


Figure 8.6: Visualization of Intra-Category variability and substructure of crowns. Group A shows the Swabian workshop of Ludwig Henfflin. Group B shows the Hagenau workshop of Diebold Lauber with the subgroups of crowns with arches (B1), crowns with lined arches (B2) and crowns with torus-shaped brims (B3). Group C shows the Alsatian workshop of 1418.

8.4 Semi-Supervised Analysis of Intra-Category Object Variability

Fig. 8.6 has helped the historians in visualizing the characteristics of different artistic workshops. However, the completely unsupervised mapping, defined by (8.2), from the high dimensional feature space to the 2-d space cannot preserve all the pairwise relationships between the crowns. This is an inherent limitation of any projection from higher dimensional feature space into a lower dimensional space that can be visualized. This limitation is particularly problematic for art historians when trying to infer the object relationship between crowns which belong to the same workshop, since these distances are more affected by the mapping from (8.2).

However, consider the following simple case. An arbitrary crown C has distances d_1, d_2 and d_3 from three crowns C_{R_1}, C_{R_2} and C_{R_3}. Given the distance triplet (d_1,d_2,d_3) we can assign 2-d locations to these four crowns such that the distances between C and C_{R_1}, C_{R_2} and C_{R_3} are preserved. In fact, if we fix the crowns C_{R_1}, C_{R_2} and C_{R_3} as landmark crowns with respect to which we obtain the distance triplets, we can find a 2-d configuration of crowns such that all the distance triplets are preserved. This simple but important insight leads us to a semi-supervised approach [101] where the user can choose the landmark

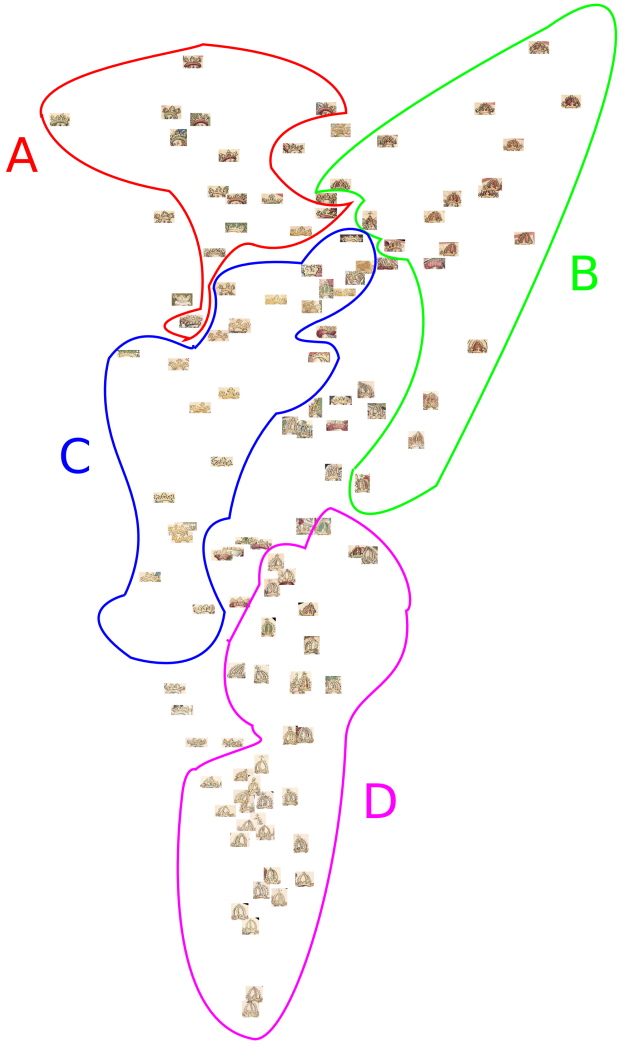


Figure 8.7: Configuration of crowns belonging to the Hagenau workshop in a probability simplex where the three landmark crowns are all chosen from the Hagenau workshop.

crowns, all the other crowns are projected into 2-d space preserving the distance triplets.

We start by obtaining three landmark crowns provided as input by the user. In a first experiment, one crown from each workshop was provided as landmark. Next, we compute the distance triplets for the rest of the crowns in the database. Then we choose the location of the landmark crowns at the three corners of an equilateral triangle in 2-d space (which we refer to as 'probability simplex') such that the side of the triangle is greater than the maximum of the distance triplet values. Next, we find a mapping for each of the crowns into the interior of probability simplex such that the distance of the crown from the three corners of the equilateral triangle is proportional to its pre-computed distance triplet (d_1,d_2,d_3). Fig. 8.7 shows the organization of crowns from the Hagenau workshop in a probability simplex.

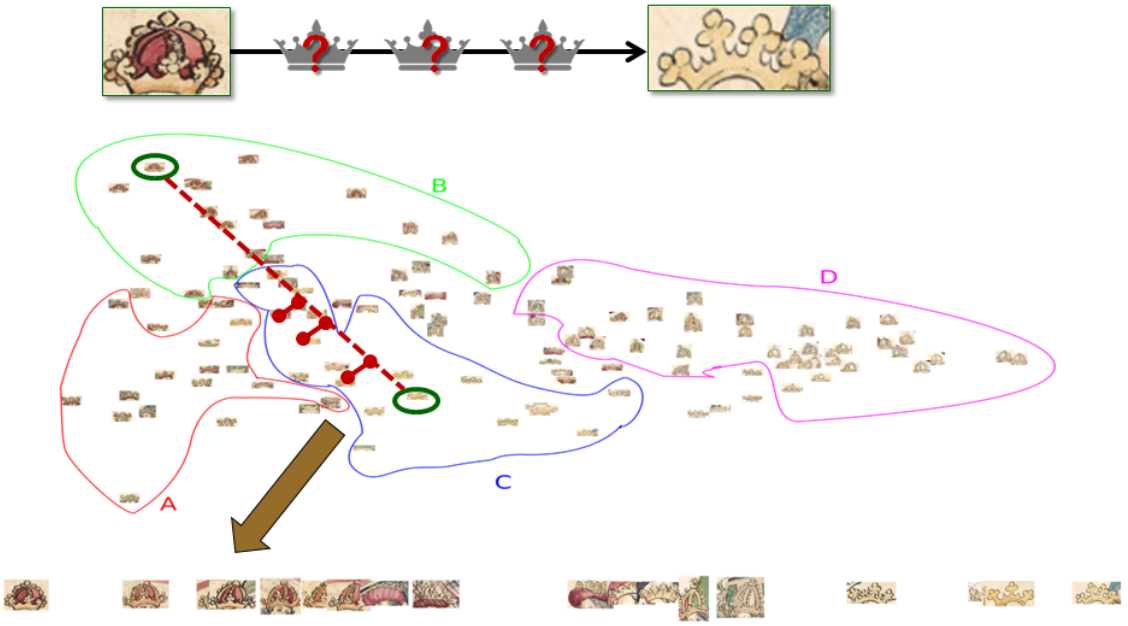


Figure 8.8: One dimensional ordering of crowns between two pairs of user chosen crowns

8.5 Inducing 1-d Ordering based on Pairwise Object Relationships

Given any two objects from a dataset the question arises how all the other objects in the database relate to these two exemplars. In particular, (i) can we find instances that help to *interpolate* between the selected reference exemplars, and (ii) can we *order* all those instances? Such an ordering is valuable for art history as it is directly visualizing relationships between the exemplars, it is illustrating smooth transitions in artistic style, and it could even reveal relationships between artists.

Given two crowns from the probability simplex, we compute the geodesic between the crowns (in this case, a straight line joining the two crowns in the probability simplex). Next, we project the rest of the crowns onto this geodesic and measure the distances between the projections onto the geodesics and the instances themselves. We retain the crowns with small distances. Then, we generate a one dimensional ordering of the crowns by showing the user selected crowns at the two ends of the geodesic and the retained crowns at the projected locations onto the geodesics.

Fig. 8.8 shows two examples where two pairs of crowns from Hagenau workshop were provided as user input. Notice that a smooth transition can be observed in the one-dimensional ordering of crowns in both the examples.

8.6 Discussion

The present case study on the Upper German manuscripts of Heidelberg University Library demonstrates the deep insight into medieval object representation and its artistic context provided by the proposed image analysis algorithms. The object detection algorithm has been successful in detecting objects in highly cluttered scenes despite large intra-class

variations. An unsupervised and a semi-supervised algorithm have been proposed based on a top-down object model. The approach decomposes the large intra-class variability of categories and visualizes the inherent structure of all objects in a dataset within a single 2D projection. The automatic analysis reveals subtypes based on their difference in artistic design and successfully classifies objects by the artistic workshop that has drawn them. Finally, an approach for ordering instances of an object category has been presented which also provides an illustration of the transitions in artistic style that are inherent to the image collection.

CHAPTER 9

CONCLUSIONS

This thesis identified the common shortcoming of object detection techniques which use shape in various ways such as 1) Hough Voting approaches which utilize interest points sampled from the edge maps of the images 2) Contour based approaches which detect objects by utilizing an ensemble of contours from training images and 3) template based approaches such as Chamfer Matchinig which search for an object or its part there-of in a query image by means of distance transform. The state-of-the-art approaches in all the above lines of work treat the object as an independent summation of its constituent parts.

In Hough Voting approaches, each object part independently votes for global object properties such as its scale, aspect ratio, etc. The independence assumption between the voting elements leads to weak and spurious votes as evidenced from the experimental results. To address the shortcoming, the dependencies between voting elements have been systematically modeled in a probabilistic framework where the grouping of interest points, finding the correspondences and finding the transformations of grouped entities are jointly solved. Voting with dependent entities lead to concerted votes and reduction of spurious object hypotheses as clearly evidenced from experiments carried out on a number of state-of-the-art databases for voting approaches.

Contour based approaches face the fundamental challenge that object form is an emergent property that cannot be obtained by independently placing contours to assemble object shape. This thesis addressed the problem of detecting objects from contours by jointly assembling the object shape from a codebook of meaningful contours. Large number of bottom-up contours are sampled from within the bounding boxes of objects in training images and are reduced to a dictionary of meaningful contours by a novel way of measuring affinity between two contours. The introduced affinity measure is needed instead of measuring direct visual similarity between contours because contours might break differently over an ensemble of training images during the bottom-up process. The codebook learnt from the novel affinity measure has been found to be robust to such breaking down of bottom-up contours. Finally, assembling the object shape from the dictionary of codebook contours has been formulated as a max-margin multiple instance learning problem. In this formulation, the positive bag contains the object hypotheses formed from multiple possible placements for each codebook contour within an object bounding box. Such

a formulation is required since the training images only contain the bounding boxes of objects and there is no further information regarding the correct placement of contours. The benefit of the proposed approach has been demonstrated by experimental comparisons with state-of-the-art contour based approaches. Another noteworthy aspect of the proposed approach is that, unlike other contour based approaches, there is no need for bottom-up grouping in query images which is usually unreliable.

Template based approaches such as Chamfer Matching are widely used in Computer Vision and especially in Machine Vision and Industrial Inspection applications because of their speed. A significant drawback of Chamfer Matching is the false matches of the template in the background clutter. The false positives can be attributed to two factors. Firstly, each point on the template is treated as equally important in computing the distance transform score in chamfer matching. Secondly, Chamfer Matching fails to take into account the accidentalness of a match in the background. The first issue has been addressed by introducing discriminative distance transform (DDT) where the relative importance of different points on a template has been learnt in a Max-Margin framework. The experimental results have shown that DDT brings a noticeable gain in performance. However, spurious responses in background clutter were still found to be an issue which has been addressed by the second contribution. A small dictionary of generic background contours has been utilized and the co-occurence values of the background contours relative to the placement of foreground have been used to discriminate between clutter and true placements of the object template. This has indeed lead to the reduction of spurious matches in background clutter as evidenced from experimental results.

The final part of the thesis presented a case study where it has been demonstrated that it is possible to obtain semantic understanding of image collections using a simple combination of shape-based representation for objects, standard statistical and Computer Vision techniques. For the purpose of case study, a novel benchmark dataset of upper German manuscripts from 15-th century has been assembled with ground-truth information about various objects of artistic interest such as crowns, swords. Such objects have been extracted using an approach presented in this thesis which relies upon shape-based representation of images. By finding a low dimensional embedding of objects in 2-d space using Multi-dimensional Scaling, intra-category variability of objects has been analysed. The resultant 2-d plot has lead the art historians to confirm their notion of different artistic workshops within the manuscripts. Further, a semi-supervised approach has been presented for not only anlysing the variations within an artistic workshop but also to understand the transitions across artistic styles by means of 1-d ordering of objects.

List of Publications

This dissertation has led to the following scientific publications:

- Yarlagadda, P., and Ommer, B. From Meaningful Contours to Discriminative Object Shape. In European Conference on Computer Vision (2012).
- Yarlagadda, P., Eigenstetter, A., and Ommer, B. Learning Discriminative Chamfer Regularization. In British Machine Vision Conference (2012).
- Yarlagadda, P., Monroy, A., Carque, B., and Ommer, B. Recognition and Analysis of Objects in Medieval Images. In Asian Conference on Computer Vision (e-heritage) (2010).
- Yarlagadda, P., Monroy, A., Carque, B., and Ommer, B. Top-down Analysis of Low-level Object Relatedness Leading to Semantic Understanding of Medieval Image Collections. In Computer Vision and Image Analysis of art, SPIE (2011).
- Yarlagadda, P., Monroy, A., and Ommer, B. Voting by Grouping Dependent Parts. In European Conference on Computer Vision (2010).
- Yarlagadda, P., Monroy, A., Carque, B., and Ommer, B. Towards a Computer-based Understanding of Medieval Images. In Scientific Computing and Cultural Heritage (2009).

BIBLIOGRAPHY

[1] http://heidicon.ub.uni-heidelberg.de.

[2] http://homepages.inf.ed.ac.uk/rbf/cvonline/applic.htm.

[3] http://www.artstor.org.

[4] http://www.asia-europe.uni-heidelberg.de/research/heidelberg-research-architecture/hra-databases-1/transcultural-image-database/the-image-database.

[5] http://www.imareal.oeaw.ac.at/realonline/.

[6] http://www.prometheus-bildarchiv.de.

[7] A. Lehmann, B. L., and van Gool, L. Prism principled implicit shape model. In *BMVC* (2008).

[8] Ahuja, N., and Todorovic, S. Connected segmentation tree: A joint representation of region layout and hierarchy. In *CVPR* (2008).

[9] Amit, Y., and Geman, D. A computational model for visual selection. In *Neural Computation* (1999).

[10] Andrews, S., Tsochantaridis, I., and Hofmann, T. Support vector machines for multiple-instance learning. In *NIPS* (2003).

[11] Andriluka, M., Roth, S., and Schiele, B. People-tracking-by-detection and people-detection-by-tracking. *CVPR* (2008).

[12] Arbelaez, P., Fowlkes, C., Maire, M., and Malik, J. Using contours to detect and localize junctions in natural images. In *Intl. Conf. on Comp. Vision and Pat. Rec.* (2008).

[13] Arbelaez, P., Maire, M., Fowlkes, C., and Malik, J. From contours to regions: An empirical evaluation. In *CVPR* (2009).

[14] Attneave, F. Some informational aspects of visual perception. *Psychological review 61*, 3 (1954), 183–193.

[15] Baca, M., Harpring, P., Lanzi, E., McRae, L., and Whiteside, A. *Cataloging Cultural Objects. A Guide to Describing Cultural Works and Their Images.* 2006.

[16] Ballard, D. Generalizing the hough transform to detect arbitrary shapes. *Pat.Rec. 13* (1981).

[17] Barrow, H. G., Tenenbaum, J. M., Bolles, R. C., and Wolf, H. C. Parametric correspondence and chamfer matching: Two new techniques for image matching. *Int. Joint Conf. Artificial Intelligence* (1977), 659–663.

[18] Belongie, S., Malik, J., and Puzicha, J. Shape matching and object recognition using shape contexts. *PAMI 24*, 4 (2002), 509–522.

[19] Berg, A. C., Berg, T. L., and Malik, J. Shape matching and object recognition using low distortion correspondence. In *CVPR* (2005), pp. 26–33.

[20] Berg, A. C., and Malik, J. Geometric blur for template matching. In *CVPR* (2001), pp. 607–614.

[21] Biederman, I. Recognition-by-components: A theory of human image understanding. *Psychological review 4*, 2 (1987), 115–147.

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

[23] Boiman, O., Shechtman, E., and Irani, M. In defense of nearest-neighbor based image classification. In *CVPR* (2008).

[24] Borenstein, E., and Ullman, S. Combined top-down/bottom-up segmentation. *PAMI 30*, 12 (2008), 2109–2125.

[25] Boser, B. E., Guyon, I. M., and Vapnik, V. N. A training algorithm for optimal margin classifiers. *5th Annual ACM Workshop on COLT* (1992), 144–152.

[26] Canny, J. A computational approach to edge detection. *IEEE Trans.Pat. Analysis and Machine Intelligence* (1986), 679–714.

[27] Carneiro, G., and Lowe, D. Sparse flexible models of local features. In *ECCV* (2006).

[28] Carreira, J., and Sminchisescu, C. Constrained Parametric Min-Cuts for Automatic Object Segmentation. In *CVPR* (2010).

[29] Chang, C.-C., and Lin, C.-J. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology 2*, 3 (2011), 27:1–27:27. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[30] Comaniciu, D., Ramesh, V., and Meer, P. The variable bandwidth mean shift and data-driven scale selection. In *ICCV* (2001), pp. 438–445.

[31] Cootes, T., and Taylor, C. Active shape models. In *BMVC* (1992).

[32] Cristinacce, D., and Cootes, T. Boosted regression active shape models. In *BMVC* (2007).

[33] Csurka, G., Dance, C. R., Fan, L., Willamowski, J., and Bray, C. Visual categorization with bags of keypoints. In *ECCV* (2004), *Workshop Stat. Learn. in Comp. Vis.*

[34] Dalal, N., and Triggs, B. Histograms of oriented gradients for human detection. In *CVPR* (2005), pp. 886–893.

[35] Dalal, N., and Triggs, B. Histograms of oriented gradients for human detection. In *Intl. Conf. on Comp. Vision and Pat. Rec.* (2005).

[36] Davison, A., and Hinkley, D. *Bootstrap Methods and their Application. Cambridge: Cambridge Series in Statistical and Probabilistic Mathematics.* 1997.

[37] Dietterich, T. G., and Lathrop, R. H. Solving the multiple-instance problem with axis-parallel rectangles. *Artificial Intelligence 89* (1997), 31–71.

[38] Estrada, F. J., Fua, P., Lepetit, V., and Susstrunk, S. Appearance-based keypoint clustering. In *CVPR* (2009).

[39] Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A. The PASCAL Visual Object Classes Challenge 2008 (VOC2008) Results. http://www.pascal-network.org/challenges/VOC/voc2008/workshop.

[40] Felzenszwalb, P., Girshick, R., McAllester, D., and Ramanan, D. Object detection with discriminatively trained part-based models. *PAMI 32*, 9 (2010), 1627–1645.

[41] Felzenszwalb, P., McAllester, D., and Ramanan, D. A discriminatively trained, multiscale, deformable part model. In *CVPR* (2008).

[42] Felzenszwalb, P. F., Girshick, R. B., Mcallester, D., and Ramanan, D. Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2010).

[43] Felzenszwalb, P. F., and Huttenlocher, D. P. Pictorial structures for object recognition. *IJCV 61*, 1 (2005).

[44] Ferarri, V., Tuytelaars, T., and Gool, L. V. Object detection by contour segment networks. *ECCV* (2006).

[45] Fergus, R., Perona, P., and Zisserman, A. Object class recognition by unsupervised scale-invariant learning. In *CVPR* (2003), pp. 264–271.

[46] Ferrari, V., Jurie, F., and Schmid, C. From images to shape models for object detection. *IJCV* (2009).

[47] Fidler, S., and Leonardis, A. Towards scalable representations of object categories: Learning a hierarchy of parts. In *CVPR* (2007).

[48] Fowlkes, C., Martin, D., and Malik, J. Learning affinity functions for image segmentation: Combining patch-based and gradient-based approaches. In *CVPR* (2003).

[49] Fowlkes, C., Tal, D., Martin, D., and Malik, J. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Intl Conf. on Comp. Vision* (2001).

[50] Freund, Y., and Schapire, R. A decision-theoretic generalization of on-line learning and an application to boosting, 1995.

[51] Gall, J., and Lempitsky, V. Class-specific hough forests for object detection. In *CVPR* (2009).

[52] Hastie.T, Tibshirani.R, and Friedman.J. *The elements of Statistical Learning.* Springer, 2008.

[53] Julesz, B. Textons, the elements of texture perception and their interactions. *Nature 29*, 290 (1981), 91–97.

[54] Kerscher, G. Thesaurus-Verwendung und internationalisierung in Bilddatenbanken. *Kunstchronik 57* (2008), 606–608.

[55] Kokkinos, I., and Yuille, A. L. Hop: Hierarchical object parsing. In *CVPR* (2009).

[56] Kovesi, P. D. MATLAB and Octave functions for computer vision and image processing.

[57] Lampert, C. H., Blaschko, M. B., and Hofmann, T. Beyond sliding windows: Object localization by efficient subwindow search. In *CVPR* (2008).

[58] Lazebnik, S., Schmid, C., and Ponce, J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR* (2006), pp. 2169–2178.

[59] Leibe, B., Leonardis, A., and Schiele, B. Combined object categorization and segmentation with an implicit shape model. In *ECCV* (2004), *Workshop Stat. Learn. in Comp. Vis.*

[60] Leibe, B., Leonardis, A., and Schiele, B. Robust object detection with interleaved categorization and segmentation. *IJCV 77*, 1-3 (2008), 259–289.

[61] Lindeberg, T. Feature detection with automatic scale selection. *IJCV 30*, 2 (1998), 77–116.

[62] Liu, M., Tuzel, O., A.Veeraraghavan, and Chellappa, R. Fast directional chamfer matching. In *CVPR* (2010).

[63] Lowe, D. Object recognition from local scale-invariant features. In *ICCV* (1999).

[64] Ma, T., and Latecki, L. From partial shape matching through local deformation to robust global shape similarity for object detection. In *CVPR* (2011).

[65] Ma, T., Yang, X., and L.Latecki. Boosting chamfer matching by learning chamfer distance normalization. In *ECCV* (2010).

[66] Maire, M., Arbelaez, P., Fowlkes, C., and Malik, J. Using contours to detect and localize junctions in natural images. In *CVPR* (2008).

[67] Maji, S., Berg, A., and Malik, J. Classification using intersection kernel support vector machines is efficient. In *Intl. Conf. on Comp. Vision and Pat. Rec.* (2008).

[68] Maji, S., and Malik, J. Object detection using a max-margin hough transform. In *CVPR* (2009).

[69] Martin, D., Fowlkes, C., and Malik, C. Learning to detect natural image boundaries using local brightness, color and texture cues. *PAMI 26*, 5 (2004), 530–549.

[70] Narasimhan, M., and Bilmes, J. A submodular-supermodular procedure with applications to discriminative structure learning. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence* (2005), pp. 401–412.

[71] Ommer, B. *Seeing the Objects Behind the Parts: Learning Compositional Models for Visual Recognition.* VDM Verlag, 2008.

[72] Ommer, B., and Buhmann, J. Learning the compositional nature of visual object categories for recognition. *PAMI 32*, 3 (2010), 501–516.

[73] Ommer, B., and Malik, J. Multi-scale object detection by clustering lines. ICCV:.

[74] Opelt, A., Pinz, A., and Zisserman, A. Incremental learning of object detectors using a visual shape alphabet. In *CVPR* (2006), pp. 3–10.

[75] P. Srinivasan, Q. Z., and Shi, J. Many-to-one contour matching for describing and discriminating object shape. In *CVPR* (2010).

[76] Pietzsch, E., Effinger, M., and Spyra, U. Digitalisierung und Erschließung spätmittelalterlicher Bilderhandschriften aus der Bibliotheca Palatina. H. Thaller,editor. Digitale Bausteine für die geisteswissenschaftliche Forschung.

[77] Ponce, J., Hebert, M., Schmid, C., and Zisserman, A. *Toward Category-Level Object Recognition*. Springer, 2006.

[78] Riemenschneider, H., Donoser, M., and Bischof, H. Using partial edge contour matches for efficient object category localization. In *ECCV* (2010).

[79] Roberts, L. Machine perception of three-dimensional solids. *Optical and electro-optical information processing* (1965), 159–197.

[80] S. Fidler, M. B., and Leonardis, A. A coarse-to-fine taxonomy of constellations for fast multi-class object detection. In *ECCV* (2010).

[81] Sala, P., and Dickinson, S. Contour grouping and abstraction using simple part models. In *ECCV* (2010).

[82] Saurma-Jeltsch, L. E. *Spätformen mittelalterlicher Buchherstellung. Bilderhandschriften aus der Werkstatt Diebold Laubers in Hagenau*. 2001. 2 vols.

[83] schlecht, J., and Ommer, B. Contour-based object detection. In *BMVC* (2011).

[84] Schramm, P. E. *Herrschaftszeichen und Staatssymbolik*. 1954. 3 vols.

[85] Schwedler, G., Meyer, C., and Zimmermann, K., Eds. *Rituale und die Ordnung der Welt*. 2008.

[86] Shotton, J., Blake, A., and Cipolla, R. Contour-based learning for object detection. In *ICCV* (2005).

[87] Shotton, J., Blake, A., and Cipolla, R. Multi-scale categorical object rrecognition using contour fragments. *PAMI 30*, 7 (2007), 1270–1281.

[88] Thayananthan, A., Stenger, B., Torr, P., and Cipolla, R. Shape context and chamfer matching in cluttered scenes. *CVPR* (2003).

[89] Toshev, A., Taskar, B., and Daniilidis, K. Object detection via boundary structure segmentation. In *CVPR* (2010), pp. 950–957.

[90] Tu, Z., Chen, X., Yuille, A., and Zhu, S. Image parsing: Unifying segmentation, detection, and recognition. vol. 2.

[91] Turing, A. Computing machinery and intelligence. *Mind 59* (1950), 433–460.

[92] van Straten, R. *Iconography, Indexing, ICONCLASS. A Handbook*. 1994.

[93] Vapnik, V. N. *Statistical Learning Theory*. Wiley, New York, NY, 1998.

[94] Viola, P. A., and Jones, M. J. Rapid object detection using a boosted cascade of simple features. In *CVPR* (2001), pp. 511–518.

[95] Viola, P. A., and Jones, M. J. Robust real-time face detection. *IJCV 57*, 2 (2004), 137–154.

[96] Wertheimer, M. Untersuchungen zur Lehre von der Gestalt I. Prinzipielle Bemerkungen. *Psychologische Forschung 1* (1922), 47–58. [Abridged English translation in W.D. Ellis, editor. *A Source Book of Gestalt Psychology*. New York, NY: Harcourt, Brace, 1938].

[97] Williams, C., and Allan, M. On a connection between object localization with a generative template of features and pose-space prediction methods. Tech. rep., University of Edinburg, Edinburg, 2006.

[98] Yarlagadda, P., Eigenstetter, A., and Ommer, B. Learning discriminative chamfer regularization. In *British Machine Vision Conference* (2012).

[99] Yarlagadda, P., Monroy, A., Carque, B., and Ommer, B. Towards a computer-based understanding of medieval images. In *Scientific Computing and Cultural Heritage* (2009).

[100] Yarlagadda, P., Monroy, A., Carque, B., and Ommer, B. Recognition and analysis of objects in medieval images. In *Asian Conference on Computer Vision (e-heritage)* (2010).

[101] Yarlagadda, P., Monroy, A., Carque, B., and Ommer, B. Top-down analysis of low-level object relatedness leading to semantic understanding of medieval image collections. In *Computer Vision and Image Analysis of art, SPIE* (2011).

[102] Yarlagadda, P., Monroy, A., and Ommer, B. Voting by grouping dependent parts. In *European Conference on Computer Vision* (2010).

[103] Yarlagadda, P., and Ommer, B. From meaningful contours to discriminative object shape. In *European Conference on Computer Vision* (2012).

[104] Zhu, L., Chen, Y., and Yuille, A. Learning a hierarchical deformable template for rapid deformable object parsing. *PAMI 32*, 6 (2010), 1029–1043.

[105] Zhu, Q. H., Wang, L. M., Wu, Y., and Shi, J. B. Contour context selection for object detection: A set-to-set contour matching approach. In *ECCV* (2008).