University of Crete

Department of Computer Science

Perceptually Relevant Mechanisms for the Description and Retrieval of Visual Information

Ph.D Thesis

Xenophon Zabulis

Heraklion, February 2002

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΕΠΙΣΤΗΜΗΣ ΥΠΟΛΟΓΙΣΤΩΝ

Αντιληπτικώς Σχετικοί Μηχανισμοί για την Περιγραφή και Ανάκληση της Οπτικής Πληροφορίας

Διατριβή που υποβλήθηκε από τον

Ξενοφώντα Ζαμπούλη

ως μερική απαίτηση για την απόκτηση του

ΔΙΔΑΚΤΟΡΙΚΟΥ ΔΙΠΛΩΜΑΤΟΣ

Ηράκλειο, Φεβρουάριος 2002

Σελίδα με υπογραφές

Acknowledgements

I would like to express my gratitude to my advisor, Professor Stelios Orphanoudakis, for trusting me to collaborate with him during the last several years. I would like to thank him for his encouragement, guidance, and advice, which I was privileged to enjoy, as well as the additional exposure to novel research, which he has offered to me by providing me with the opportunity to attend scientific conferences and seminars. I would also like to thank him for his contribution in the correction of this dissertation and the on-going help he still continues to generously offer.

Furthermore, I would like to sincerely thank the other members of the examination committee of this thesis, Professors Ioannis Kougioumoutzakis (Department of Philosophy & Social Studies, University of Crete), Adonis Moschovakis (Medical School, University of Crete), Rolf Pfeifer (Department of Information Technology, University of Zurich), Dimitris Plexousakis (Department of Computer Science, University of Crete), Siegfried Stiehl (Department of Computer Science, University of Hamburg), and Panos Trahanias (Department of Computer Science, University of Crete). I am deeply grateful for their corrections, comments, and insightful questions that facilitated a more careful consideration of certain scientific topics. I am greatly indebted to Professor Siegfried Stiehl for the extended time he has dedicated to this thesis and his significant contributions toward producing a final version of this manuscript. In addition, I would like to thank Professor Ioannis Kougioumoutzakis for the discussions we had on visual perception.

I would also like to thank the Institute of Computer Science (ICS) of the Foundation for Research and Technology - Hellas (FORTH), for providing me with generous financial support, modern research facilities, and a pleasant and inspiring working environment. As a student, I feel highly privileged and grateful to have received this support. Furthermore, I would like to thank the members of the Computer Vision and Robotics Laboratory (CVRL) of ICS and, in particular, Antonis Argyros, Panos Trahanias, and Eleutheria Tzova for their insightfull discussions and help. I would also like to thank Jon Sporring for his cooperation in a joint project, which inspired work that appears in this dissertation.

I also thank the Computer Science Department of the University of Crete for providing me with the opportunity to do my graduate studies, for giving me technical and financial support, and for making available facilities to do part of my work. Furthermore, I would like to thank the faculty of the department for the high quality of studies and for teaching me Computer Science. In addition, I would like to thank the departmental secretary Rena Kalaitzaki for the help she offered me on numerous occasions throughout my graduate studies.

Finally, I would like to thank the staff of ICS-FORTH and, in particular, the system administrators Panos Sikas and Vaggelis Karagiannis for the technical support that they provided, and the secretaries Liana Kefalaki and Maria Prevelianaki for their administrative help during the time I was affiliated with ICS-FORTH.

My sincere thanks also go to my friends for their support and encouragement, and for making my stay in Crete pleasant, exciting, and fruitful. To some, I owe extra thanks for their help in the preparation of the printed version of this manuscript.

Most importantly, I would like to thank my family, Chrysanthos, Dimitra, and Loukas for their support and encouragement through all the years of my education, and also for their contribution toward the achievement of every goal that I have ever pursued.

Αντιληπτικώς Σχετικοί Μηχανισμοί για την Περιγραφή και Ανάκληση της Οπτικής Πληροφορίας

Ξενοφών Ζαμπούλης

Διδακτορική Διατριβή

Τμήμα Επιστήμης Υπολογιστών
Πανεπιστήμιο Κρήτης

Περίληψη

Ο μεγάλος όγκος και ποικιλομορφία των ψηφιακών εικόνων που χρησιμοποιούνται σε διάφορα πεδία εφαρμογών έχουν αναδείξει την απαίτηση για τεχνικές αναζήτησης εικόνων με βάση το περιεχόμενό τους. Συγκεκριμένα, υπάρχει μια αυξανόμενη ανάγκη για την ανάπτυξη αυτόματων τεχνικών ανάλυσης και περιγραφής του περιεχομένου των εικόνων με σκοπό την αποδοτική ανάκλησή τους από μεγάλες συλλογές, με βάση το περιεχόμενό τους. Στη διατριβή αυτή παρουσιάζονται και εξετάζονται μηχανισμοί για την περιγραφή και ανάκληση οπτικής πληροφορίας με βάση ιδιότητες της αντίληψης, έχοντας ως κίνητρο τη βελτίωση της αντιστοίχισης μεταξύ των αποτελεσμάτων ανάκλησης εικόνων με βάση το περιεχόμενό τους και των προσδοκιών των τελικών χρηστών.

Οι προτεινόμενοι μηχανισμοί αφορούν την περιγραφή των πρωτογενών οπτικών γνωρισμάτων καθώς και των χωρικών τους διατάξεων και δίνουν έμφαση στην αναπαράσταση αυτής της πληροφορίας σε συνάρτηση με την κλίμακα παρατήρησης. Η αναπαράσταση αυτή χρησιμοποιείται στη συνέχεια για την εξαγωγή περιοχών της εικόνας οι οποίες παρουσιάζουν χαρακτηριστικές χωρικές διατάξεις γνωρισμάτων καθώς και για την αναγνώριση παραγόμενων από την κλίση της εικόνας κυρίαρχων δομικών στοιχείων. Τόσο για τα γνωρίσματα όσο και για τα στοιχεία αυτά είναι γνωστή η εξέχουσα περιγραφική τους σημασία ως συνιστωσών του οπτικού περιεχομένου. Η οργάνωση των κυρίαρχων δομικών στοιχείων σε αντιληπτικά σύνολα αποδίδει μια επιπρόσθετη συνιστώσα του οπτικού περιεχομένου. Τα γνωρίσματα τέτοιων αντιληπτικών συνόλων ενοποιούνται στη συνέχεια με πληροφορίες περί των χωρικών διατάξεων των πρωτογενών οπτικών γνωρισμάτων και χρησιμοποιούνται στην περιγραφή και ανάκληση της οπτικής πληροφορίας.

Αρχικά παρουσιάζεται μια μέθοδος αναπαράστασης πρωτογενών οπτικών γνωρισμάτων βασισμένη στη σύνοψη κλιμάκων του οπτικού περιεχομένου, η οποία είναι εμπνευσμένη από τη φυσιολογία της όρασης. Η προτεινόμενη αναπαράσταση χρησιμοποιεί την κανονικοποίηση, όσον αφορά την κλίμακα, της απόκρισης των συναρτήσεων ανίχνευσης γνωρισμάτων για τη σύνοψη των πληροφοριών περί οπτικών γνωρισμάτων από ένα εύρος κλιμάκων σε μια εικόνα. Η αναπαράσταση της σύνοψης κλιμάκων διευκολύνει την εισαγωγή μιας μεθόδου για την περιγραφή πρωτογενών οπτικών γνωρισμάτων στο εύρος κλιμάκων στο οποίο λαμβάνουν χώρα, την εξαγωγή άνω του ενός σημαίνοντος εύρους κλιμάκων από μια εικόνα καθώς και την ταξινόμηση των οπτικών γνωρισμάτων με βάση το εύρος κλιμάκων στο οποίο λαμβάνουν χώρα. Η επιπρόσθετη πληροφορία που παράγεται επιδεικνύεται ως χρήσιμη στην περιγραφή του περιεχομένου των εικόνων, όπως επίσης και σε ένα πλήθος διεργασιών επεξεργασίας εικόνων. Επιπρόσθετα, η αναπαράσταση σύνοψης κλιμάκων μπορεί να υπολογισθεί παράλληλα και επιδεικνύει υπολογιστικές και περιγραφικές ιδιότητες οι οποίες επεκτείνουν τον πρότυπο ορισμό της οπτικής πληροφορίας με βάση την κλίμακα παρατήρησης. Η μελέτη των πρωτογενών οπτικών γνωρισμάτων ολοκληρώνεται με την εξέταση του ρόλου τους στην ανάκληση εικόνων με βάση το περιεχόμενο .

Στη συνέχεια, η διερεύνηση για την σχετική με την αντίληψη περιγραφή και ανάκληση περιεχομένου εικόνων εστιάζεται στη δυνατότητα εξαγωγής και σύγκρισης, όσον αφορά στην οπτική τους ομοιότητα, περιοχών εικόνων οι οποίες επιδεικνύουν αναλλοίωτη διάταξη στο χώρο όσον αφορά με πρωτογενή οπτικά γνωρίσματα. Χρησιμοποιώντας τοπικούς περιγραφείς με μεταβλητό χωρικό εύρος δειγματοληψίας, εξάγεται μια πολυ-κλιματική αναπαράσταση των χωρικών διατάξεων των πρωτογενών οπτικών γνωρισμάτων. Η επέκταση της αναπαράστασης της σύνοψης κλιμάκων για τοπικούς περιγραφείς καθιστά εφικτή την, όσον αφορά στην κλίμακα, κανονικοποίηση τους. Αυτή η κανονικοποίηση μπορεί στη συνέχεια να χρησιμοποιηθεί για την αναλλοίωτη, όσον αφορά την κλίμακα, περιγραφή μεταβλητών χωρικών διατάξεων των πρωτογενών οπτικών γνωρισμάτων. Η ομαδοποίηση των, κανονικοποιημένων, όσον αφορά στην κλίμακα, τοπικών περιγραφέων διευκολύνει την εξαγωγή περιοχών από εικόνες που επιδεικνύουν αναλλοίωτη διάταξη πρωτογενών οπτικών γνωρισμάτων στο χώρο. Επίσης, η χρήση της αναπαράστασης της σύνοψης κλιμάκων για την αναπαράσταση χωρικών

διατάξεων των πρωτογενών οπτικών γνωρισμάτων καταλήγει σε μειωμένες απαιτήσεις όσον αφορά στη χωρητικότητα της μνήμης. Επιπρόσθετα, ορισμένα γνωρίσματα των τοπικών περιγραφέων προτείνονται για την εκλέπτυνση της περιγραφής των χωρικών διατάξεων των πρωτογενών οπτικών γνωρισμάτων και χρησιμοποιούνται ως κατηγορήματα στη διατύπωση οπτικών επερωτήσεων. Τέτοιου είδους γνωρίσματα αντιστοιχίζονται με οπτικές ιδιότητες των εικόνων προκειμένου να προσφερθούν οπτικές επερωτήσεις οι οποίες είναι κατανοητές από τους τελικούς χρήστες. Τελικά, η αποκτηθείσα αναπαράσταση των χωρικών διατάξεων των πρωτογενών οπτικών γνωρισμάτων χρησιμοποιείται για την αναδίφηση και ανάκληση παρόμοιων οπτικά εικόνων.

Προκειμένου να εμπλουτιστεί περαιτέρω η παραχθείσα περιγραφή του περιεχομένου εικόνων με γνωρίσματα που σχετίζονται με την αντίληψη, εξετάζεται η συνιστώσα που απορρέει από τη διαδικασία της αντιληπτικής οργάνωσης του περιεχομένου. Για το σκοπό αυτό παρουσιάζονται επίσης δύο προσεγγίσεις για την εξαγωγή και περιγραφή δύο κλάσεων αντιληπτικών συνόλων. Αυτές συνίστανται στην κλάση των αντιληπτικών συνόλων από γραμμές προοπτικής απεικόνισης και στην κλάση των περιγραμμάτων. Σχετικά με την πρώτη κλάση, αντιληπτικά σύνολα που συντίθενται από συγκλίνοντα ευθύγραμμα τμήματα εξάγονται από εικόνες, βάσει υποθέσεων περί της σύγκλισης τους προς ένα σημείο διαφυγής. Τέτοιου είδους υποθέσεις διατυπώνονται, αρχικά, με βάση την φωτεινή αντίθεση και το μέγεθος των ευθύγραμμων τμημάτων μιας εικόνας και στη συνέχεια εξετάζονται ως προς την εγκυρότητα τους, χρησιμοποιώντας υποστηρικτικές ή αντικρουόμενες αποδείξεις από την εικόνα. Τα ευθύγραμμα τμήματα για τα οποία οι υποθέσεις επαληθεύονται ομαδοποιούνται στο ίδιο σύνολο με γνωρίσματα τα οποία αφορούν την όψη του συνόλου. Ο ίδιος αλγόριθμος γενικεύεται για νοητά ευθύγραμμα τμήματα, αυτά δηλαδή τα οποία συντίθενται από την συγγραμική ύπαρξη τοπικών γνωρισμάτων της εικόνας, όπως γωνίες και στιγμές. Τα ανιχνευθέντα αντιληπτικά σύνολα και τα γνωρίσματά τους επιδεικνύονται ως χρήσιμα στην ανάκληση και ταξινόμηση εικόνων με βάση το περιεχόμενο τους. Σχετικά με τη δεύτερη κλάση αντιληπτικών συνόλων, παρουσιάζεται μια προσέγγιση για την περιγραφή και ανάκληση περιγραμμάτων, η οποία χρησιμοποιεί μια μέθοδο βασισμένη στην καμπυλότητα. Η μέθοδος αυτή ανιχνεύει αντιληπτικώς σχετικά και υπολογιστικώς σταθερά σημεία αναφοράς πάνω στα περιγράμματα. Συγκεκριμένα, τα ακρότατα της καμπυλότητας ακολουθούνται στο χώρο κλιμάκων του περιγράμματος και η κανονικοποιημένη, όσον αφορά στην κλίμακα, τιμή της καμπυλότητάς τους χρησιμοποιούνται στη διατύπωση ενός μέτρου για την διαπίστωση της σημασίας τους. Χρησιμοποιώντας τα πιο σημαίνοντα σημεία του περιγράμματος, επιτυγχάνεται η αποσύνθεση σε τμήματα των περιγραμμάτων, η οποία χρησιμοποιείται στην ευθυγράμμισή τους. Η δυνατότητα για ευθυγράμμιση περιγραμμάτων χρησιμοποιείται τελικά στην εύρεση παρόμοιων περιγραμμάτων, βάσει ενός μέτρου ομοιότητας το οποίο συλλέγει πληροφορίες περί της σχετικής μετατόπισης ευθυγραμμισμένων (αντίστοιχων) τμημάτων των περιγραμμάτων.

Η ολοκληρωμένη χρήση πληροφορίας περί των χωρικών διευθετήσεων των πρωτογενών οπτικών γνωρισμάτων και των αντιληπτικών συνόλων στην ανάκληση οπτικής πληροφορίας με βάση το περιεχόμενο, καθίσταται δυνατή με την εισαγωγή μιας διεπιφάνειας χρήσης υψηλής σαφήνειας. Η διεπιφάνεια αυτή διευκολύνει την επιλογή

τόσο συγκεκριμένων συνιστωσών του οπτικού περιεχομένου όσο και των γνωρισμάτων τους και καθιστά δυνατή τη χρήση τους ως κατηγορημάτων σε οπτικές επερωτήσεις. Επιπρόσθετα, η κατ'επανάληψη παράθεση κριτηρίων ταξινόμησης προτείνεται ως μέθοδος για την ενοποίηση πληροφοριών σχετικών με το περιεχόμενο μιας εικόνας αλλά και για την διαδραστική ταξινόμηση συλλογών από εικόνες.

Perceptually Relevant Mechanisms for the Description and Retrieval of Visual Information

Xenophon Zabulis

Doctoral Dissertation

Department of Computer Science
University of Crete

Abstract

The large volume and variety of digital images, currently acquired and used in different application domains, has given rise to the requirement for content-based image management and retrieval techniques. In particular, there is an increasing need for the development of automated image content analysis and description techniques in order to retrieve images efficiently from large collections, based on their visual content. In this dissertation, mechanisms for the perceptually relevant description and retrieval of visual information are presented and discussed, motivated by the need to provide a better match between content-based image retrieval results and end user expectations.

The proposed mechanisms concern the description of primitive visual features and spatial arrangements of such features, and emphasize the representation of this information with respect to scale of observation. This scale dependent representation is subsequently used to extract image regions that exhibit a characteristic spatial arrangement of primitive features and identify gradient-derived dominant structural elements, which are both

known to be significant descriptive components of visual content. The organization of gradient-derived dominant structural elements into perceptual groups yields an additional component of visual content. Attributes of such perceptual groups are then integrated with information about the spatial arrangement of primitive visual features and used in the description and content-based retrieval of images.

Initially, the role of primitive visual features in the formation of image content is considered and a physiology-inspired method is presented for their representation, based on the scale-summarization of visual content. The proposed representation utilizes the scale-normalization of feature detection response functions to summarize visual feature information from a range of scales into a single image. This scale-summarized representation facilitates the introduction of a method for the description of primitive features at the range of image scales at which they occur, the extraction of more than one meaningful ranges of scales from an image, and the classification of primitive visual features with respect to the range of scales at which they occur. The additional information thus generated is demonstrated to be useful in the description of image content, as well as in a number of image processing tasks. Furthermore, the scale-summarized representation can be computed in parallel and exhibits computational and descriptive properties that extend the standard representation of visual information with respect to scale. The study of primitive visual features is concluded with a discussion about their role in content-based image retrieval.

Subsequently, the investigation for the perceptually relevant description and retrieval of image content is focused on the ability to extract and compare, with respect to their visual similarity, image regions that exhibit a constant spatial arrangement of primitive visual features. Using local descriptors with varying image sampling aperture, a multiscale representation of the spatial arrangements of primitive features is derived. The extension of the scale-summarized representation for local descriptors makes their scale-normalization possible. This scale-normalization can then be utilized for the constant description of scale-varying spatial arrangements of primitive features. The clustering of scale-normalized local descriptors facilitates the extraction of image regions that exhibit a constant spatial arrangement of primitive features, even when these features vary at scale. In addition, the scale-summarized representation of spatial arrangements of primitive features results in reduced memory capacity requirements. Furthermore, attributes of local descriptors are proposed for the refinement of the description of spatial arrangements of primitive features and are used as predicates in the formulation of visual queries. Such attributes are mapped onto visual properties of images to provide visual queries which are comprehensible by end users. Finally, the acquired representation of spatial arrangements of primitive features is utilized for the browsing and retrieval of visually similar images.

In order to further enrich the derived description of image content with perceptually relevant attributes, the component of visual content resulting from the process of perceptual organization is considered and two approaches are presented for the extraction and description of two classes of perceptual groups. These are the class of linearly perspective perceptual groups and the class of silhouette boundaries. Regarding the first

class, perceptual groups that consist of converging line segments are extracted from images, based on hypotheses about their convergence to a vanishing point. Such hypotheses are initially formulated, based on the contrast and the size of line segments in an image and subsequently tested, as to their validity, utilizing supporting or contradicting image evidence. Line segments for which the hypothesis is verified are assigned to the same group along with appearance-related attributes of that group. The same algorithm is generalized for subjective line segments, that is line segments that are composed of collinear local image features, such as corners or dots. The detected perceptual groups as well as their attributes are demonstrated to be of use in the content-based retrieval and classification of images. With respect to the second class of perceptual groups, an approach is presented for the description and retrieval of silhouette boundaries, which utilizes a curvature-based method to detect perceptually significant and computationally stable anchor points. In particular, curvature extrema are tracked along the scale-space of the contour and their scale-normalized curvature across scale is utilized to formulate a salience metric. Using the most salient contour points, a piecewise decomposition of contours is achieved and further utilized in their alignment. The ability to align contours is finally utilized in matching similar contours, based on a similarity metric which captures information about the relative displacement of aligned (corresponding) contour segments (pieces).

The integrated use of information about spatial arrangements of primitive visual features and perceptual groups in content-based visual information retrieval is made possible by the introduction of a high specificity user interface, which facilitates the selection of individual visual content components and their attributes and their use as predicates in visual queries. In addition, the iterative introduction of image classification criteria is proposed as a method for integrating information about image content and interactively classifying image collections.

Contents

1 Introduction

The large volume and variety of digital images currently acquired and used in different application domains has given rise to the requirement for content-based image management and retrieval techniques. In particular, there is an increasing need for the development of automated image content analysis and description techniques in order to retrieve images efficiently from large collections, based on their visual content. Large collections of images can be found in many application domains such as journalism, advertising, entertainment, weather forecasting, map production, remote sensing, computer-aided design, architecture, vision-based robot navigation, medicine, etc. Thus, an important functionality of next generation image database and multimedia information systems will undoubtedly be the search and retrieval of images based on visual content.

In the first section of this chapter, representative application domains in which visual information is important are considered and characteristic components of the visual information retrieval task, such as image type, query goal, and context are identified and discussed. In the second section, existing image search engines are described as representative of the current state-of-the-art. The motivation for taking into account the physiology and psychology of biological vision in the task of content-based visual information browsing and retrieval is elaborated in the third section of this chapter. However, the task of visual information browsing and retrieval, based solely on pictorial information, has certain characteristics that impose limitations on the straightforward application of known perceptual mechanisms. These characteristics and corresponding limitations are presented and discussed in the fourth section of this chapter. In the fifth section, the visual information retrieval task is defined and put in the context of this dissertation, along with other terminology that is subsequently used. The last section of this chapter highlights the research contributions of this work.

1.1 Application domains

The work presented in this dissertation has been motivated by the existence of many application domains in which content-based access to visual information is often desirable and provides added value. In this section, some representative application domains that make use of visual information are presented and briefly discussed in order to demonstrate that the comprehen-

sion and management of visual information highly depends on observation task, image context, and the type of images used. More specifically, it is argued that some prior knowledge is required for the understanding of image content. This knowledge cannot be extracted from the image's visual features and it is usually referred to as image or application context. In the discussion that follows, terms such as "visual information" and "visual content description" are loosely used without being formally defined. The meaning of these terms is further clarified in the fifth section of this chapter.

Remote sensing has been an application domain in which digital images have been acquired and analyzed to achieve a variety of different goals. For example, remotely sensed images are used to construct maps in a variety of fields such as geography, meteorology, route planning and others. In most cases, the images used are acquired in a different way and represent different information, while in the same image different features may be of interest depending on the observation task. Thus, in meteorology, the occlusion of the ground by clouds may be of interest and, by analyzing a sequence of such images, one may obtain information on wind speed, while in geography the shape and location of ground features are of greater interest. The analysis and understanding of these images requires knowledge of the image type and the visual appearance of earth, sea, and sky elements. Thus, the diversity of image observation tasks, corresponding to a variety of features of interest, results in different contextual knowledge requirements and, most often, different image content analysis and description methods.

Surveillance is another application domain in which visual information is used not only to extract properties of the environment, but also to study patterns of behavior. The required functionalities of such systems include visual feature selection, extraction of interesting information, and recognition of prototypes or learned patterns. In each case, context-related knowledge is required for the identification of interesting patterns and behaviors.

In *art and design*, features such as color and form are intrinsically related to the visual impression generated and semantics perceived when looking at some composition. Depending on the image type and application, the required information may be contained in different visual features. For example, certain image features may contain the information needed to classify pieces of art with respect to the technique used in their composition. Both detailed and abstract image features may characterize the piece of art with respect to style, time period of creation, and artist. The visual information used in such applications is most often fused with other types of non-visual

context knowledge.

Medical imaging is an application domain in which understanding the method of image acquisition is crucial for content comprehension. As in the previous case of art, depending on the image type, different visual features are encountered or take on a particular significance, and different context knowledge is required in order to interpret the visual stimuli. Again, a different type of analysis is needed not only with respect to the image type, but also with respect to other types of non-visual information.

By considering different application domains, in which visual information is important, one quickly comes to the realization that the types of images resulting from the use of different sensors and acquisition methods may vary considerably. Furthermore, it is clear that the type of visual information targeted for extraction by workers in each application domain depends on the specific application requirements and the observation task. The comprehension of such visual information and the corresponding visual content, associated with each image or image type, are further determined by prior contextual knowledge. In each case, prior knowledge of the task and its context is required in order to trigger appropriate mechanisms (agents) of image content analysis and description.

In summary, the comprehension of visual content is dependent on image type, context, and image observation task. In addition, it is often the case that specific image features ought to be considered together with non-visual information or non-visual knowledge. Furthermore, visual cues that may be important in one application domain are not necessarily important in another. However, certain visual cues exist that are common to different application domains. For example, the description of color and form in an image are significant visual cues that find application in various domains such as art and design, remote sensing, and others. For these reasons, an approach is proposed in this dissertation, which is based on a collection of perceptually relevant visual content description and matching competences. Such competences can then be selectively activated and integrated with domain specific knowledge in order to satisfy specific application requirements.

1.2 Image search engines

In recent years, the rapid growth of the World Wide Web (WWW) has created a need for facilitating on-line access to visual information and user interaction with image repositories. Thus, a number of image databases are

available on the WWW and provide the capability of browsing image collections, based on pictorial content. Currently, most image search engines support the functionality of *query by example*, in order to allow the search for images that are similar to a given query image. The search is based solely on some similarity function of image features such as color, texture, or shape. The query response consists of images that provide a good match to the query image, based on an appropriate similarity metric. In most cases, the returned images exhibit a weak relationship or similarity to the query image. In other words, while users often aim at retrieving images containing particular objects or semantics, state-of-the-art generic image retrieval systems analyze image content based solely on low-level features. Some of the better known image search engines, currently available on the WWW, are considered below.

The *CANDID* system [50] was originally motivated by modern methods for searching databases containing free-text documents. The image comparison method uses signatures to represent the visual content of an image. In this system, a signature is typically represented by a histogram of the number of times that each "sub-string" of length N occurs in the document, where N is a predetermined value. Signatures represent features such as local textures, shapes, and colors. The general idea is that several features (local color, texture, and / or shape) are first computed at every pixel in the image, and then a probability density function that describes the distribution of these features, is estimated. This probability density function is the content signature for the given image. Given a query image, all database images are ranked with respect to their similarity to the query image.

Photobook [80] employs primitive content (e.g. color and texture) and model based retrieval (e.g. facial features). Retrieval is based on and may be assisted by interactive image segmentation and annotation with the help of a user interface agent. In addition, Photobook contains a set of interactive tools for browsing and searching images and image sequences. The novelty in this approach is that a direct search on image content is possible by use of "semantics-preserving" image compression. Such semantics are, however, manually defined. The system uses three descriptors, which support searching based on appearance, 2-D shape, and textual properties. The resulting descriptions may be combined to provide browsing and searching capabilities. Users can browse over large image databases quickly and efficiently by using both text annotation associated with the images and the descriptions of image content.

Pic2seek [32] is an image retrieval system, which is implemented using photometric color and geometric invariant indices. The basic idea is to extract invariant features (independent of the imaging conditions) from each of the images in the database, which are subsequently matched with the invariant feature set derived from the query image. In this system, queries are primarily based on color, however some querying strategies based on edges, corners, and edge shapes are also supported.

The *Query By Image Content (QBIC) System* [73], supports browsing of an image collection based on different features of image content, such as color, texture, shape, location and their spatial layout. The adopted approach for assessing similarity is based on the computation of feature vectors and their comparison. The queries also include standard SQL and text / keyword predicates. In addition, QBIC has a rich user interface, which provides an end user with the ability to query for specific colored objects by selecting that color from a palette, to query for a particular texture from a set of selected texture patterns, to query for objects with a specific shape by drawing shapes on a "blackboard" etc. The graphical user interface provides the ability to the user to construct the queries, to view results, and to modify and resubmit queries.

In the *Virage* [39] image retrieval engine, the "internal properties" (primitives) of the image are computed from a predefined feature set, which contains the features of color, texture, and shape. The system also computes distance metrics between objects in feature space from their feature set. The similarity of images may be recomputed based on property weights assigned by the user, for properties supported by the system. The functions of image analysis, comparison and management are handled by the core module of the system.

The *Blobworld* [20] image retrieval system uses a unique image representation, based on image segmentation. To segment an image, the joint distribution of the color, texture, and location of each pixel in the image is modeled. After the image is segmented into regions, a description of each region's color, texture, and spatial characteristics is produced. While Blobworld is not oriented at object recognition, it considers the nature of images as combinations of objects. By finding image regions, which roughly correspond to objects, querying at object level is made possible.

More recently, an increased research effort in the areas of image databases and computer vision has resulted in the development of novel tools that support the content-based retrieval of images with additional functionalities,

such as learning, interaction, query space display, refinement of query specification and others. A recent state-of-the-art review of content-based image retrieval, including systems aspects such as database indexing, system architecture, and performance evaluation can be found in [102].

It should be pointed out that this dissertation does not attempt to introduce just another image retrieval engine. The emphasis is on developing perceptually relevant content-based image retrieval mechanisms, which can be used in image retrieval systems. Such mechanisms are expected to contribute toward an improved performance of similarity matching and content-based retrieval in image database systems, as well as over the WWW, by yielding query responses that are more compatible with human perception and better correspond to user expectations.

1.3 Interdisciplinary approach

Visual information retrieval concerns not only the objective structure of visual information, but also the subjective visual impression that is generated when observing an image. Image search engines dealing with image collections of great variety, such as those found on the WWW, often return images which are related to the query in a way that is not easily comprehended by humans. Besides the lack of knowledge concerning the intention of the query, such cases of failure partially stem from the fact that what constitutes image content is generally not well defined. Since vision is a natural process encountered in biological organisms, it is argued that advances in the fields of optics, ophthalmology, neurosciences, psychophysics and cognitive sciences could contribute to the field of content-based visual information retrieval. In particular, understanding visual perception mechanisms is expected to contribute towards the formulation of visual content description and matching methods that are more compatible with the way humans comprehend images than existing ones. Possible contributions of Computer Science in such an interdisciplinary approach are mainly related to the tasks of modeling and simulating the processes of visual perception, efficient management of visual information, extraction of meaningful information from sensory data, the testing of hypotheses using computer simulation, and performing experiments under controlled conditions.

The most fundamental argument for adopting an interdisciplinary approach to visual information management is based on the fact that, beyond image acquisition (studied by optics, ophthalmology, and neurosciences), the

analysis and comprehension of visual data takes place in the brain. Thus, visual perception is intrinsically determined by the way that the sensory stimulus is represented, transformed, and analyzed in the brain. This process employs mechanisms that, if understood, would afford insight to how visual impression is generated. Thus, a description of image content consists not only of the raw color or intensity data, but also of information derived from it.

Another reason that makes a biologically relevant approach of managing visual information interesting is that machine vision applications are inspired by tasks typically carried out by humans (and in some cases by other biological organisms, e.g. homing behavior), giving rise to the class of "biologically-inspired algorithms". The design and implementation of such algorithms however, requires knowledge of elements of biological vision. Due also to the fact that, in most cases, biological visual systems exhibit superior performance compared to artificial ones, mechanisms of visual information analysis encountered in nature cannot be overlooked. Furthermore, evidence for the existence of specialized brain modules, which perform specific tasks of visual information processing, supports the claim that understanding their structure and function can provide additional insight into possible functionalities of perceptual processes. At a higher perceptual level, not yet fully explored by the neurosciences, the contribution of Psychology is fundamental to the provision of a behavioral description of visual perception.

The concept that some visual tasks could be performed in a modular fashion influences the adopted approach to the problem of visual information retrieval. The ability to break down complex behaviors into simple, but perceptually relevant, mechanisms will facilitate the evaluation of such behaviors with respect to their usefulness and perceptual compatibility. A long term goal of this work is to integrate perceptually relevant visual content description and matching mechanisms on a single experimental platform that may be used to evaluate their comparative effectiveness and to conduct relevant psychophysical experiments. Such experiments could exploit the ability of a computer-based system to control the conditions of an experiment, in order to acquire information about the percepts derived from stimuli presented to observers. Besides the evaluation of visual content description and matching mechanisms, an additional goal of such experiments could be the testing of hypotheses concerning the function of perceptual processes that are related to vision. The visual content description and matching processes described in the following chapters have been hosted in a single system that facilitates

their selective activation. In addition, the system provides an environment and software tools for the storage and browsing of image collections, the formulation of visual queries using user-selected predicates, and the organized presentation of content-based retrieval results. This environment also facilitates the addition of new mechanisms that would extend such functionalities and support further research as described above.

In conclusion, it is apparent that there exists a need to define a new research direction in content-based image retrieval, which emphasizes visual content description and matching mechanisms that are compatible with human perception and may result in a qualitative enhancement of image retrieval methods. This new research direction ought to take into account the following observations:

- Depending on the application domain and retrieval task, as well as the type of images used, different visual properties of an image may be considered relevant.

- Depending on the application context and the requirements of an individual observer, the same visual stimuli may have different interpretations.

- Depending on the application, the criteria used in image browsing and the required accuracy of visual queries may vary. In the literature [102], the cases of *image-targeted* (the query for a specific image) and *category search* (the query for a class of images), as well as *associative image browsing* (interactive, multiple stage querying by examples, which are selected from previous query stages), indicate the diversity of required retrieval strategies.

Given the above observations, it is clear that a generic, with respect to application domain, approach to the content-based management of visual information requires the capability of adapting to the diverse description and matching requirements of each domain. This capability can be provided by visual content description mechanisms, specialized with respect to application requirements and context-related knowledge. The approach presented in this dissertation emphasizes the use of perceptually relevant image descriptors, obtained from an analysis of different visual cues encountered in images.

1.4 Limitations

In this section, certain limitations on the straightforward application of known visual perception mechanisms to the task of "image browsing and retrieval by content" are presented and discussed. This discussion is motivated by the need to show that, in applying knowledge about the perception of the environment, one is often limited by the nature of two-dimensional images and by the lack of information about the functionality of top-down perceptual mechanisms for image understanding. Such limitations confine the scope of content-based image retrieval to the visual information extracted from two-dimensional static images. Some of the most compelling factors imposing limitations on the automated and perceptually relevant description of images are:

- *Image segmentation.* Segmenting an image into regions that are meaningful with respect to a particular application is critical in image understanding. However, segmenting an image into regions that correspond to distinct physical objects, using solely two-dimensional visual information, is difficult or even impossible to achieve. This is due primarily to the projectively metameric nature of image content and to the lack of three-dimensional models for every possible identifiable physical object. In addition, absence of motion, stereo, and information about the illumination of a scene (also mentioned below) restrict the ability to detect solid surfaces in images.

- *Motion and binocular vision* are sources of rich visual information. Visual cues provided by motion and stereo facilitate the extraction of object boundaries, as well as the estimation of scene structure. On a semantic level, certain types of motion may constitute intense attentional attractors, dominating an observer's attention. Similarly, stereoptic images can be directly used to estimate scene structure, thus contributing to the identification of distinct physical objects and scene understanding. In static images, such visual cues are absent.

- *Illumination.* Knowledge of scene illumination plays an important role in the correct estimation of an object's reflectance spectrum. Human visual perception approximately normalizes perceived spectra with respect to global scene illumination, a phenomenon known as "color constancy". However, in the general case of image acquisition, the illumination of a scene is not homogeneous and it is typically unknown.

- *Object recognition.* The ability to identify specific objects in images would support the retrieval of semantically similar images. Images containing the same or "similar" objects, or even a contextually relevant object, may be considered as semantically related. In addition to models of identifiable objects, thesauri associating contextually relevant objects would be required for such a task. Even in this hypothetical case, object semantics may vary depending on the image observation task and context, as well as on the expectation of finding a particular object in a certain visual scene. For example, a tree trunk, which has been cut and is lying on the ground, may be characterized as a "place to sit" when taking a walk in the forest, while it could not be matched with any chair, stool or sofa model, used for the same purpose [34].

- *Context.* As already discussed, the context of a query by image content and the type of images used have a strong effect on how the content of these images is perceived, described, and matched, with respect to visual similarity. Contextual information and knowledge of the world are essential in deriving an appropriate image representation and may influence the role and significance of specific objects in such interpretations. Furthermore, the type of images in a content-based search and retrieval task may play an important role in determining which preprocessing methods are to be used for feature extraction.

- *Time.* Biological visual systems employ several physiological adaptation behaviors over time, such as intensity or chromatic adaptation [69], as well as motion adaptation [35]. Perceptual adaptation phenomena are also observed in the interpretation of static visual stimuli and mostly refer to changes in perception that reduce sensory discrepancies that have been caused by stimulus transformations [109]. Furthermore, given enough observation time, certain image features or details may be emphasized in the viewer's perception, depending on his / her cognitive background and observation task. In this dissertation, a contextually uncommitted analysis of visual content is attempted, taking only into account the early stages of visual perception.

- *Feedback.* Image feature extraction in biological vision systems may be adjusted depending on viewpoint, illumination, query target, learning, adaptation and other factors. Feedback connections exist in the visual cortex, however their functionality has not yet been clearly understood.

Certain image preprocessing methodologies may use feedback to improve feature extraction, but a generic framework for this is yet to be formulated.

The above limitations correspond to cases in which the information that is available to the observer of the physical world is different than that encoded in a single static image. Such limitations force us to focus the scope of visual queries; for example, given the lack of knowledge about the illumination of some scene, one can expect that a color-based visual query will yield color-metameric [1] results.

1.5 The visual information retrieval task

"Information retrieval deals with the representation, storage, organization of, and access to information items. The representation and organization of the information items should provide the user with easy access to the information in which he is interested." [112].

In the domain of visual information retrieval, characterization of a user's information requirements is not a simple problem. The notion of similarity is rather broad, even for textual queries, and a generic method for precisely retrieving the requested documents is yet to be formulated. Often, the underlying reason is the imprecise specification of information requirements. A typical example is the one of Internet search engines, when one searches solely on the basis of certain key words. For example, a search for the word "tree" targets a very different content for a biologist and a computer scientist. Furthermore, similarity may be defined not only at the lexical and syntactical level, but also at more abstract levels. Since words have synonyms and point to notions, similarity may also be defined on a semantic level. Finally, words may be used to form innuendos and may be overloaded with more than one meaning. Similarity assessment in these cases cannot be elaborated solely on the basis of lexical and syntactical knowledge. In order for one to be able to reason at that level, contextual knowledge is required. On the other hand, the field of textual information retrieval has introduced several heuristics concerning the capture of a document's context (or essence). Such are the detection of frequently repeated words in a document, the extraction

[1] *Color metamers:* A set of reflectance spectra which differ, but yield the same or similar tristimulus values under at least one set of viewing conditions.

of characteristic keywords, the dominance of words being contained in titles, and others.

As in the case of textual information, visual information may consist of primitive as well as composite components. Usually, spatial distributions of motion, color, light, and image structural elements (e.g. edgels, corners, etc.) are characterized as primitive components (or features) of visual information. This intuitive characterization correlates with neurophysiological findings regarding the existence of separate specialized brain modules. However, as in the textual example, primitive or composite visual information components may be related to semantic information. Typical examples of how context contributes to visual information understanding are the perception of surface shape from shading, color perception, physical surface discrimination, egomotion estimation, and others. Taking into account syntactical similarity, but neglecting context, can lead to counter intuitive results, such as those sometimes yielded by search engines. For example, an orange can be confused with the sun, due to shape and color similarity.

Depending on the application domain and the corresponding type of images, primitive token queries can sometimes yield satisfactory results. However, if the query is targeted at higher level aspects of visual information such as visual impression, object recognition, visual similarity etc., more information is required. Furthermore, depending on the visual task, different primitive tokens may be of interest. For example, in the task of driving, the driver and the passengers may receive identical visual stimuli, but they typically tend to observe different aspects of them. Thus, the topics of consciousness, context, observation task, subjectivity of the observer, and attentional selectivity are considered as important factors in the visual comprehension process.

In the remaining of this section, some definitions of the terminology used throughout this dissertation are given:

Visual information will henceforth refer to "the bits of knowledge" [16] that contribute to the interpretation of the image with respect to the observation task. Part of the set of these bits of knowledge may not reside in the image.

Visual content will henceforth refer to the visual information that resides in the image, while *visual content component* will refer to a part or subset of this information.

Visual information management will henceforth refer to the processes of representing, describing, and retrieving visual information.

Representation will henceforth refer to data structures that reside in computer memory and are used to symbolize some entity. This entity will mainly correspond to visual content or a visual content component.

Description will henceforth refer to information extracted from a representation, with respect to some purpose.

Finally, the term *perceptually relevant* will henceforth be used to characterize computational processes that correspond (are equivalent or compatible) to a known perceptual process. The term will be mainly used to discriminate visual content description computational processes that are based on or inspired from the functionality of the human perceptual system from others that are based on the raw data of the two-dimensional image matrix.

1.6 Research contributions

In the remaining chapters of this dissertation, a taxonomy of visual content components is used for their study and the presentation of the proposed methods for content-based image retrieval. This taxonomy provides the ability to refine and, thus, increase the specificity of visual queries. In the proposed taxonomy, the class of primitive visual features is first considered and two broad classes of visual content components, derived from them, are examined. Primitive visual features are considered as a basic component of visual content, since the spatial organization of such features gives rise to other, more complex, visual content components. The two components of visual content that can be "composed" from primitive visual features and that are examined in this work are: *(a)* image regions that are determined by a spatially constant arrangement of primitive features, and *(b)* perceptually organized spatial arrangements of edge or boundary-related primitive features.

More analytically, the content of the remaining chapters is the following:

Second chapter In the second chapter of this dissertation, the need to describe and represent primitive visual features is addressed. Based on the observation that primitive features occur at different scales, it is required that: *(a)* the image "scale-space" is taken into account and *(b)* primitive features are differentiated with respect to the scale or range of scales at which they occur, in order to enhance their descriptive power and to facilitate the refinement of associated visual queries.

The standard Scale-Space image representation is utilized for the extraction of features at multiple scales. Single scale-selection is subsequently used

to estimate the scale at which a particular primitive feature occurs. Two drawbacks of this method are that: *(a)* it exhibits several computational difficulties, *(b)* more than one scale may be meaningful.

In order to overcome these difficulties, the Scale-Summarized Representation (SSR) is introduced as a method to represent and classify primitive features with respect to scale. This representation can be applied to a broad range of features and exhibits reduced memory capacity requirements and algorithmic simplicity. Furthermore, the detection of more than one meaningful image scales, at a single image point, is possible. In addition, the SSR of visual features is useful in image processing tasks, which are based on estimates of the size of local structure.

At the end of the chapter, a discussion on the utilization of knowledge about the observed environment in the refinement of visual queries is presented.

Third chapter In the third chapter, a solution to the problem of extracting image regions that are characterized by an approximately constant arrangement of primitive features is presented. A review of the relevant literature indicates that such regions constitute an important component of image content. Furthermore, it is desirable that the description of arrangements of primitive features is comprehensible in perceptual terms, so that it can be better appreciated by end users.

The proposed representation of spatial arrangements of primitive features is based on local descriptors. The scale dependence of feature arrangements is also considered and a multiscale description is formulated for their representation. The representation is based on local descriptors and is instantiated utilizing local histograms. This representation is memory consuming and the SSR is utilized for the reduction of memory requirements.

The SSR is also used for the scale-normalization of local descriptors, so that spatial arrangements of primitive features that exhibit variation with respect to scale are uniformly represented. This scale-normalization can be used to obtain clusters of similar local descriptors, corresponding to image regions extracted from the constant or scale-varying expression of feature arrangements in the image.

Finally, attributes of the local descriptors are used to describe feature arrangements in a human comprehensible way by mapping these attributes onto image properties. The resulting description is utilized in the task of

visual query formulation and content-based visual information retrieval.

Fourth chapter In the fourth chapter, the component of visual content that is derived from the perceptual organization of edge or boundary-related features is considered and its role in content-based visual information retrieval is investigated.

In this context, laws of perceptual organization are reviewed and existing methods for the perceptual organization of linear segments are found to suffer from lack of consideration for the perspective nature of visual content. Thus, a grouping method is formulated, which perceptually organizes parallel linear segments by taking into account the perspective nature of visual content.

In addition, image contours, which are typically derived from the perceptual organization of edge features, are identified as a characteristic component of visual content. A method for the perceptually relevant selection of anchor points in contour representations is proposed and is utilized in the formulation of a technique for the description and similarity matching of contours.

Fifth chapter In the fifth chapter, problems originating from missing information that is essential to performing content-based image retrieval in a generic way are discussed. Furthermore, methods are proposed that overcome certain difficulties associated with missing information about the target of visual information retrieval and provide for the interactive refinement of visual queries. Descriptions associated with the various visual content components are used to obtain an integrated description of visual content.

In addition, a user interface is introduced that captures the user preference for specific visual content components, in order to reduce the ambiguity caused by missing information about the target of visual information retrieval.

Finally, the classification of image collections using successive application of classifiers is proposed as a method to interactively refine visual queries.

Sixth chapter In the sixth and last chapter, conclusions are drawn and possible objectives of future work are presented and discussed.

In summary, the remainder of this dissertation is organized as follows: Chapter 2 deals with the visual content of primitive image features and Chapter 3 presents methods for the extraction of visual information contained in

spatial arrangements of primitive image features. In Chapter 4, the perceptual organization of features, which in turn yields visual entities of higher information order, is considered. Chapter 5 addresses issues concerning the content-based browsing and retrieval of images. Finally, in Chapter 6 an overview of contributions made by this dissertation is presented, followed by a discussion of open issues and the objectives of future work.

2 Primitive Visual Features

In this chapter, primitive visual features are considered as basic elements of visual content. Initially, the perception of primitive visual content is studied and several of its properties are reviewed. The motivation for this review is the identification of perceived primitive visual features, based on an understanding of the physiological mechanisms of visual perception. In this context, the scale at which individual features are observed is considered as a significant component of visual content, since it is an attribute of several primitive features.

In the second section of this chapter, components of primitive visual content are identified and a framework is introduced for the perceptually relevant representation of this information. In this context, elements of feature extraction and multiscale representation are reviewed and applied to primitive content description tasks. Inspired from the physiology of early stimulus representation in primates, the Scale Summarized Representation (SSR), which is introduced in the second section can be utilized for the classification and processing of visual content with respect to scale. In addition, the SSR framework supports the integration of visual content from different scales in a description optimized with respect to memory requirements. The ability to execute this feature extraction process in parallel, combined with its reduced memory requirements results in a computationally plausible platform for the real-time and compact content representation of visual content.

In the last section of this chapter, the use of primitive feature elements in the description and management of visual content is demonstrated and discussed.

2.1 Review of the physiology of early vision

The review of the physiology of early vision in this section, serves the goal of understanding the type of information extracted at early visual stages. The discussion targets the objective identification of perceived primitive visual features, based on an understanding of the physiological mechanisms of visual perception. Thus, the physiological mechanisms of light transduction, as well as early stimulus representation and processing, are reviewed.

2.1.1 Light acquisition

Images are acquired by visual systems from the transduction of light by photoreceptors, which are spatially distributed on a light-sensitive area. Their physiology deeply impacts the type of the acquired visual content. In artificial light acquisition systems, such as digital photography, the receptors are CCD elements, while in conventional photography molecules chemically react after their exposure to light. In natural visual systems, photoreceptors perform the first step in the formation of visual content, which is the conversion of light into an electrical signal.

Despite similarities between light acquisition methods, differences also exist and are significant in understanding the types of visual content acquired in each case. In addition, in biological systems, some processing of the visual signal occurs even in the first layers of receptor cells. This processing is observed to influence the representation and perception of the visual stimulus. In the remainder of this subsection, such perceptual processes are reviewed.

Photoreceptors adapt to perceived illumination by altering the gain of transduction, thus varying the range of light intensities over which they can respond. In particular, biological visual systems deal with the problem of operating over an enormous intensity range by *adapting* to light intensity. The whole range is typically not covered at one time, since the full range of intensities is not encountered often at a given illumination. However, adaptation as well as aftereffects are not considered here in any detail, as this study is only concerned with acquired, static visual information.

Furthermore, the properties of the image acquisition process do not remain constant across the retina, in the human visual system. The information content of ganglion cells consists of the responses of three types of cones, whose output is mainly determined by their absorption spectra. These spectra are commonly referred to as short (S), medium (M), and long (L), taking the maximum absorption value at a wavelength of $440nm$, $530nm$, and $560nm$ respectively. However, the number of each type and their spatial distribution over the retina are not uniform. The ratio of L to M to S cones is approximately $10 : 5 : 1$ [26]. Moreover, the receptors in the center of the fovea are almost exclusively M and L cones, whereas the proportion of S cones increases farther from the center. This image acquisition property of the human visual system is also not considered in this work, since elements of digital images are typically acquired through the same type of photoreceptors (CCD elements).

The set of stimuli acquired by retinal photoreceptors *converges* to optic nerve fibers that project the output from ganglion cells to later visual stages. Approximately 10^9 receptors converge to 10^6 nerve fibers, demonstrating the severe compression of receptor output [25]. Speculations exist regarding the development of the visual system and how it may be related to the way in which visual information is compressed.

2.1.2 Light receptor organization and interaction

The spatial organization of photoreceptors on the retina plays an important role in the formation of visual content. The location of a single cell in the receptor grid implicitly attributes the receptor response with its coordinates in that grid (or retinal image). This order, or spatial arrangement, is preserved during the signal projection onto cortical areas performing further stimulus analysis. In this subsection, the topics of resolution and retinal receptor arrangement are reviewed, along with their impact on the perceived visual content.

The spatial density of light receptors is directly related to the spatial frequency content. The sampling theorem [74] gives the maximal frequency, which can be represented, as $f = 1/2d$, where d is the sampling interval of the receptor grid. In addition, it is known that the retinal receptor density is not constant throughout the receptor grid. Retinal receptors exhibit a log-polar distribution [95], resulting in an increased resolution in the central area of the visual field and reduced resolution at its periphery. Thus, the central region of a retinal image is characterized by superior detail[2]. However, the spatial arrangement of photoreceptors are not considered further in this dissertation, since digital images are typically of constant resolution.

Another property of the perceived visual content originates from the variety of ganglion cell types. Their spatial organization and their projection to the Lateral Geniculate Nucleus (LGN) is the topic of the remainder of this subsection. The visual signals, acquired by retinal photoreceptors, simultaneously and independently lead to "ON" and "OFF" ganglion cells of varying receptive field area. With respect to receptive field area and spatial distribution density, two types of ganglion cells, M and P, project selectively to the magno and parvo cells in the LGN [61]. This selective projection induces a discrimination of the visual stimuli with respect to the size of the ganglion

[2]That is also why observers adjust their eyes or gaze towards the target of attention.

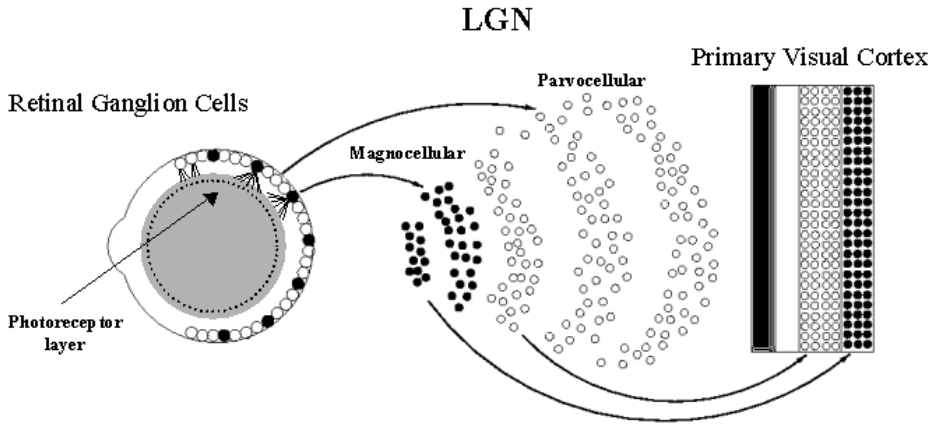

Figure 1: Projection of retinal stimulus to the LGN and Primary Visual Cortex. (Adapted from [61].)

receptive field, from which they originate, which is represented in six layers in the LGN. This independent transmission of primitive visual content is maintained, after non-linear projection, in the Primary Visual Cortex. The described circuitry is illustrated in Figure 1. The neural encoding of the light stimulus, projecting from the retina to the LGN, exhibits a feature-selective behavior based on the receptive field size [100]. In particular, samples provided by M ganglion cells, with large(r) receptive fields, project to magno cells in LGN, which are mainly color-blind and selective to transient stimuli. In addition, they exhibit high-contrast sensitivity and a fast neural response. In contrast, samples projected from P cells, with small(er) receptive fields, to parvo cells in LGN are mainly color sensitive, exhibit low contrast sensitivity, and yield a slower neural response.

The representation of image structure in different layers defines a representation, which can be modeled as a series of responses of cells with increasing receptive field centered at each grid point. The center-surround receptive field type of such cells implies the relatively strong response of cells exhibiting a receptive field size that matches with the size of local image structure, as illustrated in Figure 2. In the left figure, the small receptive field of the center-surround cell implies a weak response. In contrast, the large receptive field shall elicit a stronger response: the excitatory center and the inhibitory surround are both stimulated by stimuli of opposite polarity, thus yielding an

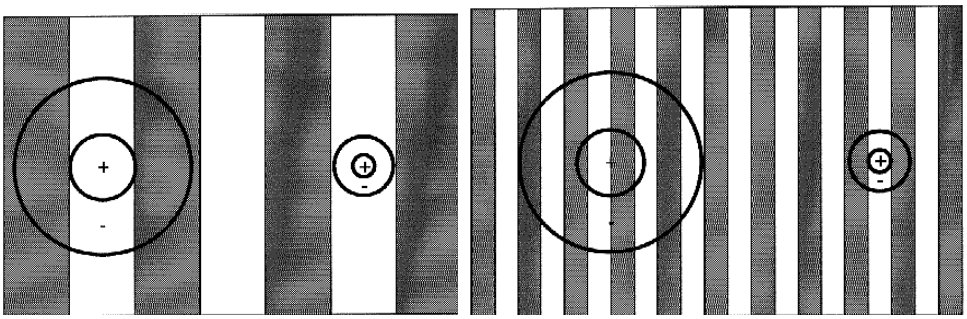

Figure 2: Origins of the scale adaptation mechanism of receptor cells (see text).

increased response. The right figure illustrates the opposite case. This scale-dependent encoding is referred to as a *size-adaptation* of visual perception [10].

It has to be specified that this spatial frequency selectivity should not be expected to be discrete, due to the analog nature of neural circuitry. Thus, one should expect some intermixing of data between different spatial frequencies, or a continuum of scales. In Computer Vision and Signal Processing, the analysis of signals at multiple scales has been studied by Scale-Space Theory [111, 58] and constitutes an important tool for feature extraction. The representation of visual content that is delivered to later visual stages is discussed in the next subsection.

2.1.3 Visual content representation

Visual data projected from both retinae to the LGN are there represented in a layered architecture and then forwarded to the V1 cortical area. It is there that a number of visual processes are initiated, including the detection of color and local form, as well as the extraction of three-dimensional depth information from binocular vision. The remainder of this subsection discusses the visual information generated by these processes.

The spectral information acquired by three cone systems has been reported to be represented in the LGN of primates, in a color-space that is roughly approximated by Hering's "opponent processes" theory [42]. Three types of color cells were observed to exist: the yellow-blue (YB), green-red (GR), and black-white (BW), all of them appearing in two versions, Y^+B^-,

B^+Y^-, G^+R^-, R^+G^-, B^+W^-, W^+B^-, with $^+$ and $^-$ referring to the excitatory and inhibitory surround respectively. The underlying physiological mechanism is speculated to be based on a re-parameterization process [27, 26]. The trichromatic representation efficiently extracts spectral information; the opponent process color space contributes to the determination of which lightness changes originate from changes in the level of illumination and which from changes in spectral surface properties. Additionally, color is psychologically perceived in a slightly different colorspace, that of hue, saturation, and value.

The visual data residing in the LGN are further processed after their projection to the striate cortex, through the optic radiation. The signal features extracted in the striate cortex neural circuitry can be classified under two categories: *(a)* temporal, and *(b)* spatial. Simple, complex, and hypercomplex cells provide spatial and temporal signal change detection at multiple scales [46]. Specifically, simple cells are correlated with the perception of edges, while the activity of complex cells is more related to motion detection. The function of hypercomplex cells is mostly related to the detection of line endings, which is further correlated with the existence of illusory contours. Computationally, the change detection in both the spatial and temporal domain is analogous with the computation of the spatial and temporal derivatives of image sequences. In Computer Vision, these features are commonly represented by two well-known structures, namely the image gradient and optic flow field. Based on spatial changes, simple and complex cortical cells perform the function of orientation selectivity and represent orientation information in different channels located at hypercolumns in the striate cortex.

An additional fact is that the spatial signal change (or gradient) of a color image is often computed on the gray-scale version of the original image. Although this roughly approximates the notion of perceived edges, color information can contribute to the formulation of a more perceptually compatible approach for two reasons. First, as previously mentioned, color can be used to extract information concerning the physical qualities (e.g. the change of reflectance spectrum) of surfaces in the visual environment. Second, the existence of gray-scale metamers (colors that have the same appearance after their conversion to gray-scale) restricts the detection of edges occurring where metameric surfaces meet. In the example below (see Figure 3), an image and its gray-scale version are used to illustrate this point. The colors of the shirt and trousers of the athlete are perceived as significantly different

Figure 3: A color image and its gray-scale version.

in the color image. In contrast, this color change information is reduced in the gray-scale version of the image. Finally, depending on the color space that is used, the computed gradient magnitude will vary. Color gradients are revisited later on in this chapter (see Section 2.2.3), where the estimation of primitive features is discussed.

Specific arrangements of primitive visual features give rise to the perceptual formation of more global features that may not be physically present in the image (the well known Kanizsa triangle is an indicative element of this class). The formation of *illusory edges* occurs in the visual cortex [83] and is observed to emerge from specific arrangements of certain primitive features, such as line endings, points, and corners. The patterns that induce illusory contour perception are often invariant to the type of primitive elements participating in contour formation.

Signals facilitating depth perception are encountered in the first visual cortical areas, namely V1 and V2. The estimation of disparities between corresponding visual features is thought to be the basis of stereoptical depth perception. However, other cues contribute as well to the complete perception of three-dimensional structure that can be encountered in single images. Such cues include shading, texture gradients, occlusion, perspective cues, and others. In this dissertation, the topic of stereo vision is not discussed further, due to the primary focus of this work on single images that are typically available in image databases.

2.1.4 Conclusions

In this section, a review of the primitive visual features generated in early visual stages was presented, highlighting components of the perceived visual content. The initial steps of formation of principal visual content components, such as color, local form, and optical flow field have also been outlined. These are the primitive visual features as considered in this work. In the next section, the significance of the scale component of primitive features and its application in the content-based retrieval of visual information are investigated in more depth.

2.2 Scale-Summarized Representation

In the previous section, an overview was presented of elements of the physiology of early vision related to the type of image acquired by photoreceptors, as well as the early processing of the visual stimulus. In this section, emphasis is placed upon the types of primitive features detected and their occurrence at different scales. In addition, the fact that visual features, which occur at different ranges of scales are handled separately by the human visual system is taken into account. This gives rise to the requirement of representing and describing primitive features with respect to scale. For this purpose, computational methods are borrowed from Scale-Space theory (see [58] for an overview of Scale-Space theory in Computer Vision).

The multiscale analysis of image features is also required because visual queries may be selectively targeted at image properties that occur at different scales. Given the observation that primitive image content[3] varies with respect to scale, the ability to represent and attribute primitive features with respect to scale is required in order to be able to refine queries with respect to the scale at which different primitive features occur.

Inspired from the LGN architecture, which consists of separate layers, interest is focused on the summarization of primitive image content over ranges of scales. Below, a representation is proposed that summarizes features over some range of scales, favoring scales at which features are actually observed. The proposed representation exhibits useful computational properties and provides the ability to focus interest on more than one ranges of scales.

[3]Henceforth, the phrase "primitive image content" will be used as an abbreviation for image content that consists solely of primitive visual features.

In the examples used to demonstrate the classification of primitive features with respect to scale, two broad ranges of scales are employed and are referred to as "fine scales" and "coarse scales". The selection of this partitioning of the scale-space is based on the need to provide end users with a comprehensible and also addressable characteristic of visual features. In fact, in Computer Vision terminology the terms "fine" and "coarse" in reference to scale are often encountered. In spoken language, expressions such as "image detail" or "abstract characteristics" are often utilized to refer to the same features. Thus, the coarse / fine feature classification contributes to the formulation of visual queries comprehensible by humans.

2.2.1 Scale-selection

Inspired from the scale-adaptation of the visual system, as briefly described in the previous section, the scale-classification of visual information is discussed and demonstrated in this subsection. Given the analog nature of a feature response, a continuous feature response function of scale is used to model feature presence at each scale. In a computer implementation, this function is discretely represented.

Given some primitive feature detector which is convolved with the image over some spatial neighborhood, the corresponding feature may be detected at the center of this neighborhood. Repeating the operation for all image points, feature presence at each pixel can be estimated. Thus, image features occurring at different scales can be detected by applying the detector at each image scale. The feature detector response F may be scale-normalized as in Equation (1), where $\tau = \log t$ is the logarithmic scale parameter, and $\vec{x}$ the pixel coordinates. The function F_S, will be referred to as the scale-normalized feature detector response (function):

$$F_S(\vec{x}, \tau) = tF(\vec{x}, \tau) \tag{1}$$

In [59], the maximum of the scale-normalized feature detector response function (F_S) is utilized to "indicate the scale at which feature presence is most intense and to reveal the spatial extent of the detected feature". This process will be referred as *explicit scale-selection*. For each image point, the maximum of the scale-normalized feature detector response may be represented in scale-space, indicating the scales at which feature presence is dominant. For example, Figure 4 shows an image and the representation of the maxima of a blob detector scale-normalized response function in scale-space,

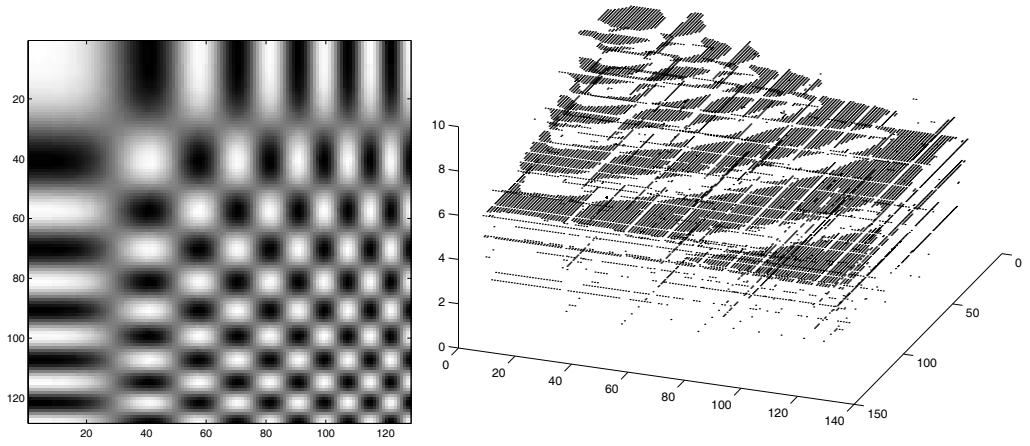

Figure 4: An image (left) and a plot (right) of the maxima of the blob response over scale, for each image point. In the plot, the oblique axes map image coordinates and the vertical the logarithmic scale parameter. Each point in the plot shows the scale at which the maximum of the blob detector occurs, for the corresponding image point.

for each image point. The equation of the blob detector is formulated using the absolute value of the Laplacian:

$$\mathcal{F}(\vec{x},\tau) = |\frac{\partial^2}{\partial x^2}L(\vec{x},\tau) + \frac{\partial^2}{\partial y^2}L(\vec{x},\tau)|, \tag{2}$$

where L is the image linear scale-space, given by $L(\vec{x},\tau) = G(\vec{x},t) * I(\vec{x})$. In the latter formula, G is a Gaussian centered at $\vec{x}$, with a standard deviation of t. In the figure, the order of scale enumeration is from fine to coarse and the vertical axis corresponds to the logarithmic scale parameter, indicating the scale at which the maximum appears. Below, two specific issues are addressed: *(a)* the interest in more than one scales of the scale-normalized feature detector response function F_S, and *(b)* certain computational difficulties in estimating the local maxima of the scale-normalized feature detector response function.

Selection of more than one scales To demonstrate the fact that more than one scales may be of interest, we consider the synthetic image shown

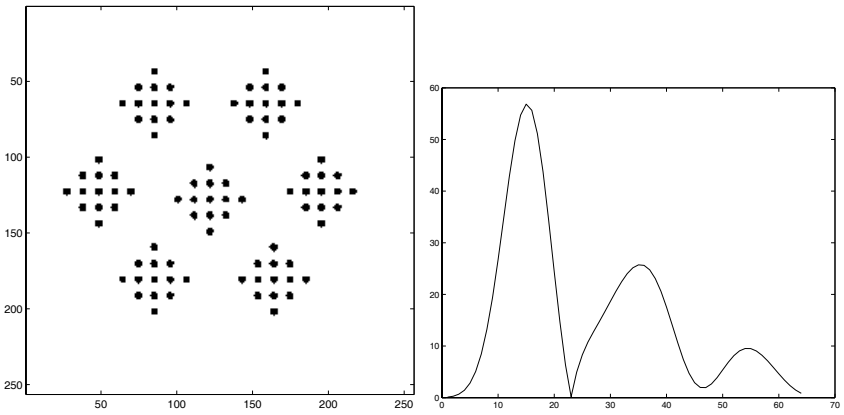

Figure 5: An image (left) and the plot of the scale-normalized blob detector response (right). In the plot, the horizontal axis maps the logarithmic scale parameter. The vertical maps the values of the blob detector for pixel $(121, 128)$, at each image scale.

in Figure 5. The graph on its right illustrates the scale-normalized response F_S of the blob detector F, in Equation (2), for pixel $\vec{c}$ with coordinates $(121, 128)$, which correspond to the center of the central black dot in the image (approximately the image center). In the plot of Figure 5, the values of logarithmic scale parameter τ are shown on the horizontal axis and it is observed that F_S has three local maxima, *approximately* at $\tau = 15, 32, 54$.

Using the maximum of F_S at $\tau = 15$, the leftmost image of Figure 6 is generated by selecting the image scale associated with $\tau = 15$. The other two images of Figure 6 are acquired, following the same procedure for $\tau = 32$, $\tau = 54$. These images represent two cases where image structure matches the structure of the feature (blob) of interest, at a scale other than the one that the scale-normalized detector response in maximized. Thus, several significant scales may be present at a single image point.

Local maxima estimation The discretization of the logarithmic scale parameter τ determines the accuracy with which the local maxima of the scale-normalized feature response function are localized. Thus, local maxima can be intractable or inaccurately estimated if the number of scales used is inadequate. In addition, the presence of noise may inhibit the accurate estimation of such maxima. Finally, explicit scale-selection [8, 58] requires

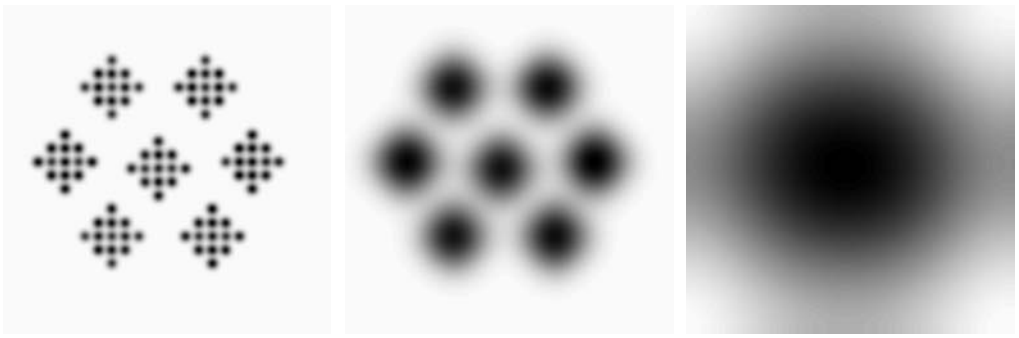

Figure 6: Three scales of the image in Figure 5, corresponding to $\tau = 15, 32, 54$ (from left to right).

tracking of feature detector responses over scale, which can be algorithmically complex. Thus, if an insufficient number of scales is used, this tracking could be practically impossible.

In the next subsection, a representation that facilitates the focusing of interest on multiple scales and avoids computational difficulties associated with the estimation of local maxima is proposed. In addition, the representation may be computed in parallel and exhibits reduced memory capacity requirements, when compared to the memory capacity that is required for the representation of the whole image scale-space. Furthermore, the proposed representation can be applied to a broad variety of features.

2.2.2 Scale summarization of visual content

The Scale-Summarized Representation (SSR) is introduced as a method to average the scale-normalized feature detector response over a range of scales. The representation may be derived from a variety of scale space images such as the linear scale-space consisting of Gaussian smoothed images, the family of corner detector response images [87] at various scales, the family of blob detector response images at various scales etc. The scale-summarized image is defined as a weighted sum over scale:

$$J(\vec{x}) = \sum_{\tau} w_f(\vec{x}, \tau) A_f[L(\vec{x}, \tau)], \tag{3}$$

$$L(\vec{x}, \tau) = G(\vec{x}, t) * I(\vec{x}) \tag{4}$$

$$\sum_{\tau} w_f(\vec{x}, \tau) = 1, \tag{5}$$

where $\tau = \log t$ is the logarithmic scale parameter, I the original image, A is a detector for feature f, and w_f is the probability of feature f being present. w_f will be henceforth be referred to as the *scale-selector*. Finally, L is the image linear scale-space derived using $G(\vec{x}, t)$, a $2D$ Gaussian centered at $\vec{x}$, with $t = \sigma^2/2$ and given by $G(\vec{x}, t) = (2\pi\sigma^2)^{-1} \exp(-|\vec{x}|^2/(2\sigma^2))$.

In order to focus interest on specific ranges of scales, Scale Focusing (SF) is introduced as the multiplication of the scale-selector function with the Gaussian function:

$$w'_{m,s}(\vec{x}, \tau) = \frac{1}{\sqrt{4\pi s}} \exp\left(-\frac{(\tau - m(\vec{x}))^2}{4s}\right) \tag{6}$$

where m is the scale of interest and s the width of the scale neighborhood. Typically, m is selected as the approximation of the scale at which some local maximum of the scale normalized feature detector response is encountered.

The implementation of the above representation can be carried out in a two-step parallelizable fashion. First, the feature response may be independently computed for each point of the image scale-space. Second, the accumulation and normalization of the feature response, for each point of the two-dimensional representation acquired, may be independently performed.

Before presenting scale-summarization results and demonstrating the effect of scale focusing, two scale-selectors that will be used are formulated.

Two cases of scale-selectors A simple scale-selector for image gradient related features, such as edges, orientation, or corners originates from the scale-normalized square gradient norm (denoted as Grad^2), which is given by $\mathrm{Grad}^2(\vec{x}, \tau) = t(L_x^2(\vec{x}, \tau) + L_y^2(\vec{x}, \tau))$, where $L_x(\cdot) = \frac{\partial}{\partial x}[G(\cdot) * I(\cdot)]$ and $L_y(\cdot) = \frac{\partial}{\partial y}[G(\cdot) * I(\cdot)]$, as in [58]. Normalizing with the sum introduces the following new scale-selector:

$$w_{\mathrm{edge}}(\vec{x}, \tau) = \frac{1}{k_{\mathrm{edge}}(\vec{x})} h\left(t\mathrm{Grad}^2(\vec{x}, \tau)\right), \tag{7}$$

where the function

$$k_{\mathrm{edge}}(\vec{x}) = \int_0^{\infty} h(t\mathrm{Grad}^2(\vec{x}, \tau))\, d\tau$$

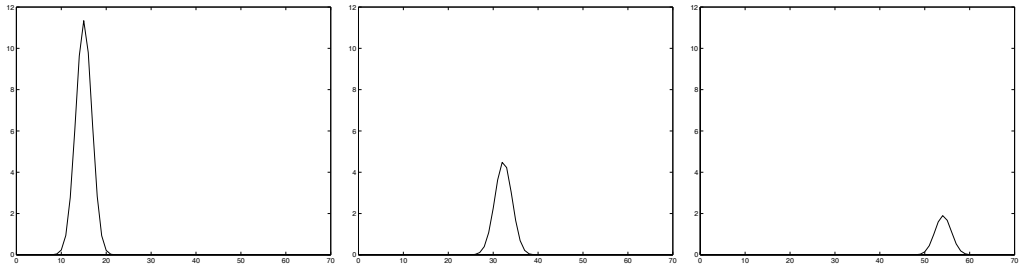

Figure 7: Three scale-selectors obtained by scale focusing on scales $\tau = 15, 32, 54$ (from left to right), of the blob detector scale-normalized response shown in Figure 5.

is the normalizing function at each spatial point and h is any strictly increasing function chosen according to the nature of the feature detector. For simplicity, the identity function $h(x) = x$ is used throughout this work.

A new scale selector for intensity blob-related features originates from (Equation 2):

$$w_{\text{blob}}(\vec{x}, \tau) = \frac{1}{k_{\text{blob}}(\vec{x})} h\left(t|L_{xx}(\vec{x}, \tau) + L_{yy}(\vec{x}, \tau)|\right) \tag{8}$$

where the normalization function is

$$k_{\text{blob}}(\vec{x}) = \int_0^\infty h(t|L_{xx}(\vec{x}, \tau) + L_{yy}(\vec{x}, \tau)|)\, d\tau.$$

In the above equations, function h could also be used to possibly rectify the scale-selector, given information about the physiology of feature extraction in the visual system of interest.

Focusing at ranges of scales In this paragraph, the process of summarizing the scale-normalized feature detector response at ranges of scales, around local maxima, is illustrated.

Given the image and scale-normalized feature detector response of Figure 5, three local maxima have been estimated at $\tau = 15, 32, 54$. The three scale-selectors produced by scale-focusing at the estimated maxima, using Equation (6) with $s = 2$, are illustrated in Figure 7. Figure 8 illustrates the three scale-summarizations of the set of scale-space images produced

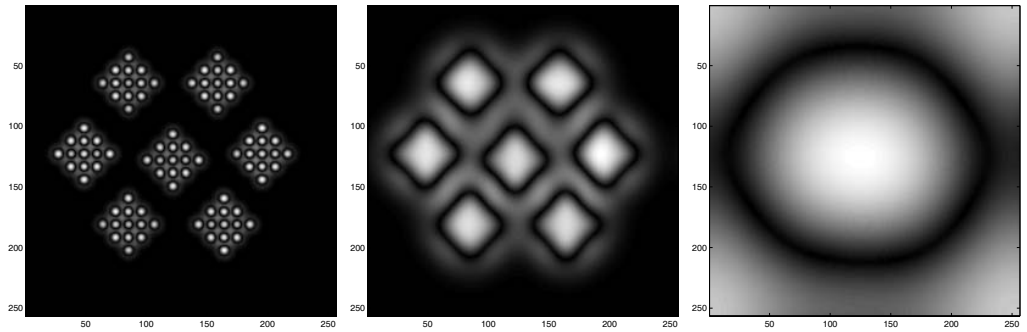

Figure 8: Three scale-summarized images of the set of the scale-normalized blob detector response images, using the scale-selectors of Figure 7.

using the scale-normalized blob detector of Equation (8) and the three scale-selectors of Figure 7. Using the identity function as the feature operator A_f, but retaining the same scale-selectors, three scale-summarizations are produced using the scale-selectors of Figure 7, which are illustrated in Figure 9. The three scale-summarized images obtained are approximately equal to those of Figure 6. The results indicate two applications of the SSR. First, by using local maxima instead of the (single) global maximum of the scale selector, more than one meaningful scales may be obtained from the image-scale space. Second, the explicit localization of local maxima over scale, as in the case of explicit scale-selection, may be avoided if the feature is known to occur at some range of scales R. Instead of explicitly selecting the scale where a local maximum occurs and then performing feature detection at that scale, the result of this process may be approximated by scale-summarizing the scale-normalized feature detector response over R. The latter application is demonstrated below. The first, is demonstrated in Section 2.2.3 for a variety of primitive visual features.

Let m be a local maximum of the scale-normalized feature detector response for some image point $\vec{p}$, whose exact location over scale is not known. Let R be the known range of scales at which this maximum occurs. Instead of computing the scale at which this maximum occurs and then performing feature detection at that scale for image point p, the SSR may be used to approximate the result, by scale-summarizing the scale-normalized feature detector response over R. Our interest will be focused on the detection and classification of features over broad ranges of scales and, in particular,

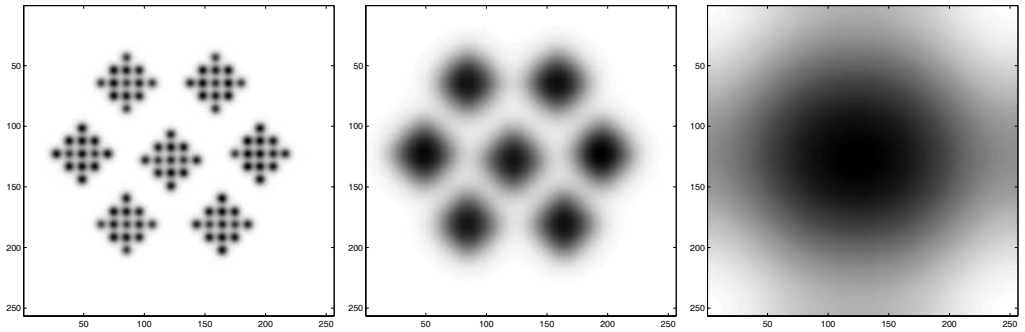

Figure 9: Three scale-summarized images of the image scale-space using the scale-selectors of Figure 7 and the identity function as the feature operator A.

at the classification of features into "coarse" and "fine". Thus, by scale-summarizing the scale-normalized feature detector response over two scale ranges R_{coarse} and R_{fine} the coarse / fine classification can be achieved in an algorithmically simple way. It is also noted that R_{coarse} and R_{fine} are selected to be partially overlapping, in order to capture features occurring approximately at the border of R_{coarse} and R_{fine}. In the example illustrated in Figure 10, only eight image scales have been used with R_{fine} consisting of scales $1, ..., 4$ and R_{coarse} of scales $4, ..., 8$. In the next section of this chapter, more such examples are presented.

Using scale-summarization over ranges of scale, features are classified with respect to scale. The granularity of scale-classification is proportional to the number of ranges in which the image scale-space is partitioned. Inspired by the magno / parvo LGN stimulus discrimination and motivated from reasons related to the comprehensiveness of this classification, examples of feature detection in Section 2.2.3 are classified into "coarse" and "fine".

In addition, due to the averaging nature of SSR, the effects of noise and scale-space discretization in the scale-classification process are reduced. This reduction is prominent in cases where a small number of image scales are used, such as in the example of Figure 10 (eight image scales were used). In such cases, errors in the localization of the local maxima over scale result in noticeable spatial discontinuities. Using the same image-scale space representation as above, the bottom-right image of Figure 10 was created by explicitly selecting the maximum of the scale-normalized blob detector

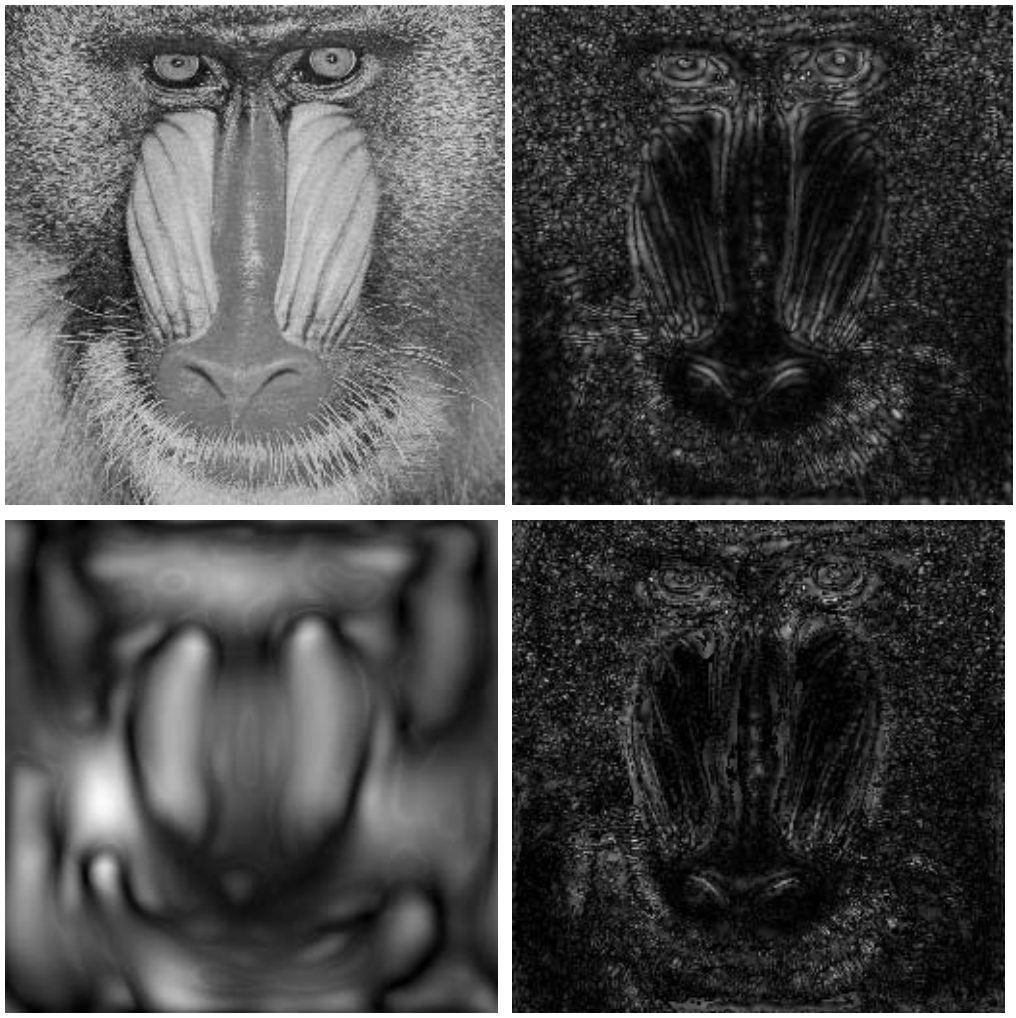

Figure 10: An image (top-left) and the scale-summarizations of its blob scale response for scale ranges R_{fine} (top-right) and R_{coarse} (bottom-left). The bottom-right image illustrates explicit scale-selection for the R_{fine} scale range.

Figure 11: An image (left) and the SSR scale-normalized feature detector response for edges (middle) and blobs (right).

response in the R_{fine} scale range, for each image point $\vec{x}$. Then, the scale $sc(\vec{x})$ at which this maximum occurs is derived and the presented image is computed as: $I_{blob} = L_{blob}(\vec{x}, sc(\vec{x}))$, where L_{blob} is the scale-space produced from the application of the blob detector at each, computed, image scale. Due to the small number of image scales used, several spatial discontinuities can be observed.

Scale-summarization over all image scales Finally, scale-summarization over all image scales is demonstrated. In order to acquire "all" image scales, the image is smoothed using Gaussian kernels whose size gradually increases varies from pixel level up to image size level[4]. This type of scale-summarization yields a single image that summarizes the features occurring at all scales. Figure 11 illustrates the SSR of the scale-normalized feature detector response, for edges and blobs, using the scale selectors of Equations (7) and (8). Notice that, in the images shown, there exist image regions in which both coarse and fine scale features can be observed. For example, this occurs in the SSR blob image region representing the eyes of the baboon, where both fine and coarse scale blobs can be observed.

The motivation underlying the process of scale-summarization over all image scales originates from the fact that different spatial frequency responses residing in the LGN are combined towards the perception of a single image. The scale-summarization of the scale-normalized feature detector response

[4]Henceforth, the expression *all image scales* will refer to this smoothing procedure.

function over all scales provides a method to represent features occurring at different scales in a single image.

2.2.3 Feature detection

In this section, multiple image scales are used in order to enhance the descriptive power of visual content description methods operating at a fixed scale. In this context, scale-summarization, over ranges of scale of the scale-normalized feature detector response function is utilized for the classification of features with respect to scale. In addition, the SSR is utilized to overcome computational difficulties related to the task of explicit scale-selection, discussed in the previous section of this chapter.

The framework of SSR is applied to two categories of feature detection tasks: *(a)* features derived from the image gradient, such as the detection of edges, linear feature orientation, and corners, and *(b)* intensity or color blobs.

Gradient-derived features By scale-summarizing the magnitude of image gradients, a "scale-less" edge detection result is obtained, representing edges at all scales. Using the scale normalized square gradient as the feature operator A and the scale-selector in Equation (7) as the weight function in Equation (3), for $h(x) = x$, yields:

$$\begin{gathered} J_w = \sum_\tau w_{\text{edge}}(\vec{x},\tau)(t\text{Grad}^2(\vec{x},\tau)) \Leftrightarrow \\ J_w = \sum_\tau \left(1/k_{\text{edge}}(\vec{x})\right)(t\text{Grad}^2(\vec{x},\tau))(t\text{Grad}^2(\vec{x},\tau)) \Leftrightarrow \\ J_w = (1/k_{\text{edge}}(\vec{x}))\sum_\tau \left(t\text{Grad}^2(\vec{x},\tau)\right)^2 \end{gathered}$$

Given that:

$$\begin{gathered} k_{\text{edge}}(\vec{x})\sum_\tau \left(w_{\text{edge}}(\vec{x},\tau)\right)^2 = k_{\text{edge}}(\vec{x})\sum_\tau \left((t\text{Grad}^2(\vec{x},\tau))/k_{\text{edge}}(\vec{x})\right)^2 = \\ = (1/k_{\text{edge}}(\vec{x}))\sum_\tau \left(t\text{Grad}^2(\vec{x},\tau)\right)^2, \end{gathered}$$

J_w may be written as : $J_w = k_{\text{edge}}(\vec{x})\sum_\tau \left(w_{\text{edge}}(\vec{x},\tau)\right)^2$

This representation results in an image that scale-summarizes edge information from all scales, as shown in Figure 11 (middle image). To illustrate the discrimination between edges at different scales, Figure 12 presents the results of scale focusing on fine (top-right) and coarse (bottom-left) scale

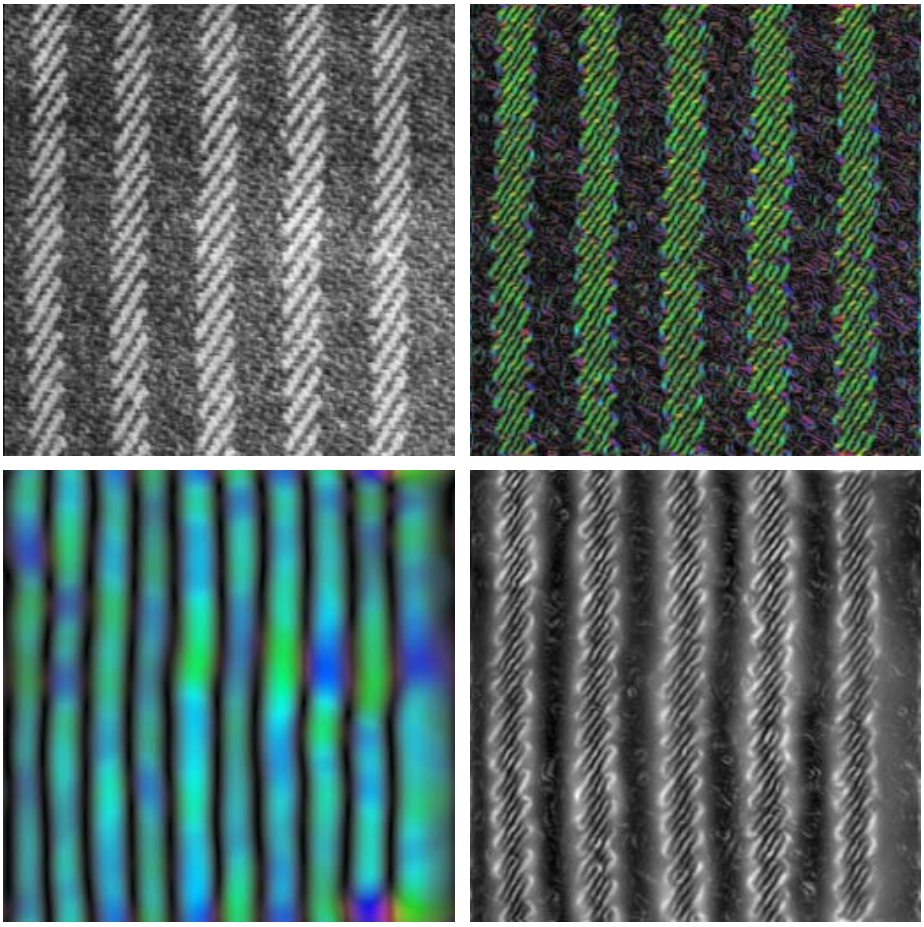

Figure 12: An image (top-left) and the scale-focusing of its gradient norm on a fine and coarse scale (top-right and bottom-left). The last image shows the scale-summarized gradient norms (bottom-right). In the top-right and bottom-left images, color represents local orientation.

ranges. The bottom-right image presents the result of SSR for all image scales.

Orientation, as previously discussed, is recognized to be an intrinsic component of visual perception and quite a useful feature in many applications of image analysis. Local direction is derived from local changes of intensity or, synonymously, image gradients. The scale normalized gradient response function for a point located on the stripes of the original image in Figure 12, reveals two local maxima, one in the range of fine scales and another in that of coarse ones. Focusing on these two ranges of scale yields a representation for fine scales, in which the diagonal orientation dominates, and another at coarse scales, in which the vertical orientation dominates. In the two scale-focused images, local orientation is linearly mapped to the color hues, with green indicating the diagonal and blue the vertical orientation. When viewed in gray-scale, local orientation is represented by image intensity (luminance). In the above example, local orientation is estimated using the gradient direction.

Another primitive feature derived from image gradients is an image corner. In the Computer Vision literature, several methods exist for detecting corners in images. An overview of many of them may be found in [86]. Below, a corner detector is derived based on the work in [105]. The smallest eigenvalue of the structure tensor:

$$\mu(\vec{x},\tau) = \begin{bmatrix} L_x(\vec{x},\tau)^2 & L_x(\vec{x},\tau)L_y(\vec{x},\tau) \\ L_x(\vec{x},\tau)L_y(\vec{x},\tau) & L_y(\vec{x},\tau)^2 \end{bmatrix} \tag{9}$$

is used to formulate a feature response function, namely the "corner detector" $C(\vec{x},\tau)$.

With respect to the SSR framework, the scale-summarization formula for corners yields (for $h(x) = x$):

$$J_w = k_{\text{corner}}(\vec{x}) \sum_{\tau} \left(w_{\text{corner}}(\vec{x},\tau)\right)^2,$$

where

$$w_{\text{corner}}(\vec{x},\tau) = \left(k_{\text{corner}}(\vec{x})\right)^{-1} h\left(t\text{C}(\vec{x},\tau_1,\tau)\right)$$

and

$$k_{\text{corner}}(\vec{x}) = \int_0^{\infty} h(t\text{C}(\vec{x},\tau_1,\tau))\, d\tau$$

Figure 13 illustrates the detection of corners at a fine and a coarse scale, along with the SSR of the scale-normalized corner detector response.

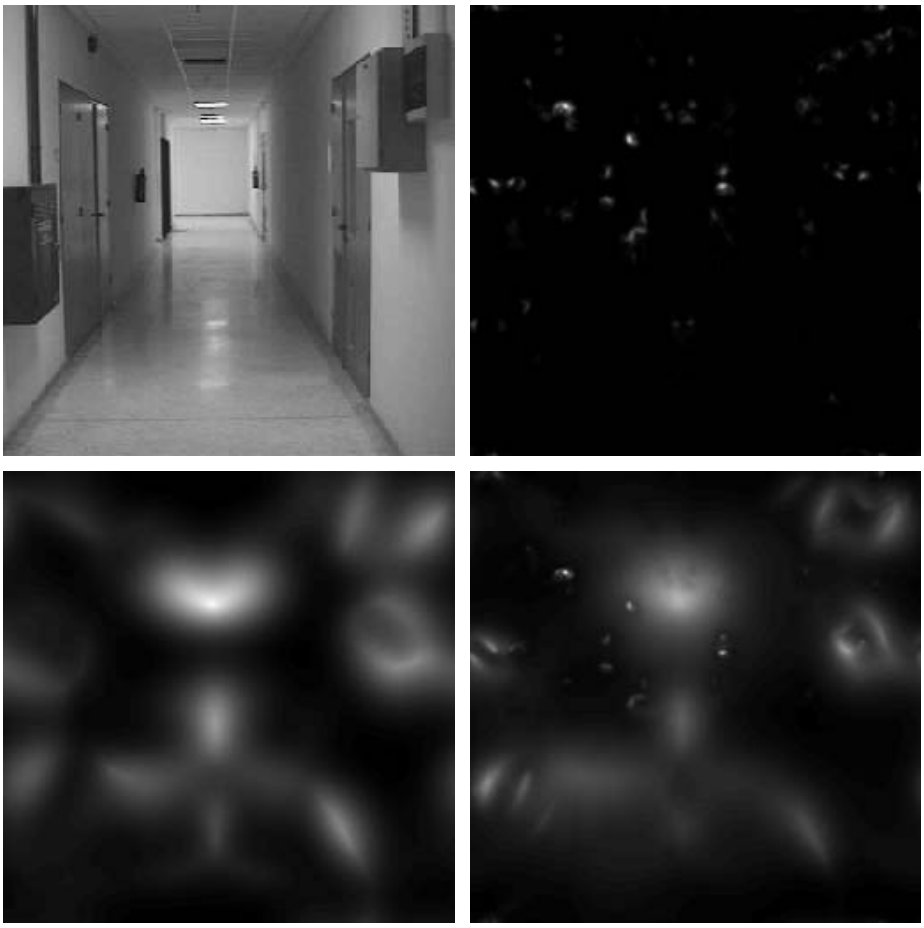

Figure 13: An image (top left) and the detection of its corners at fine (top right) and coarse scale (bottom left). Corner responses in the fine scale image correspond to structure determined by doors and wall-attached objects in the scene. The corner responses in the coarse scale image correspond to the large-scale corners formed by the corridor structure. The last image (bottom right) illustrates the scale-summarized corner information.

As mentioned in the review of early vision physiology, earlier in this chapter (see Section 2.1.3), edge detection in color images is different from edge detection in the corresponding gray-scale images. Specifically, the spatial structure of spectral images can be obtained by either converting the image into its gray-scale version, followed by standard spatial analysis, or by processing each channel independently and accumulating the results. These two approaches lead to quite different results. For a color image, the two gradient magnitudes are given as:

$$E_{gray} = |\bigtriangledown(R + G + B)| \tag{10}$$

$$E_{color} = |\bigtriangledown R| + |\bigtriangledown G| + |\bigtriangledown B|\,, \tag{11}$$

where $|\bigtriangledown \cdot| = \sqrt{(\frac{\partial \cdot}{\partial x})^2 + (\frac{\partial \cdot}{\partial y})^2}$. The color information can be used to extract changes in surface reflectance, thus contributing to the physical description of the environment. In the examples below, the application of SSR to the scale-summarization of color edge detection information is demonstrated. Independently of the color system used, color edge detection is carried out at multiple scales, and the SSR is extracted in order to summarize information from all of these scales. The increase in dimensionality introduced by color is seamlessly handled by SSR. The linear diffusion scale-space L, shown in Equation (3), is extended from the one-dimensional case:

$$L(\vec{x}, \tau) = \{LuminocityChannel\} \tag{12}$$

to the three-dimensional case

$$L'(\vec{x}, \tau) = \{ColorChannel1, ColorChannel2, ColorChannel3\}, \tag{13}$$

with $dim(L(x, y, z)) = 1$ and $dim(L'(x, y, z)) = 3$.

To generate the scale-space, each RGB color band is treated as a gray-scale image and linear diffusion is applied for each image, as in Equation (4). Each image scale consists of three of those images. The gradient magnitude is given by the color space distance for adjacent pixels. A "gradient magnitude" image is then computed for each scale. This set of scale-space images is then summarized by the SSR.

In the images of Figure 14 the SSR magnitude of the gradient vector, used for edge detection, is illustrated for a variety of color spaces. The experiments are carried out for the RGB, HSV, and Lab color spaces, using 8 image scales as in the previous edge detection examples. Furthermore, an invariant

to shading and surface orientation color transform for matte surfaces [31] was used to transform the original image and SSR edge detection was carried out in this color space, as well. The color transform used is given by: $c_1 = \arctan(\frac{R}{max\{G,B\}}), c_2 = \arctan(\frac{G}{max\{R,B\}}), c_3 = \arctan(\frac{B}{max\{R,G\}})$ where $c_i, i \in \{1,2,3\}$ are the bands of the transformed image. The first row of the figure illustrates the original image and the gradient magnitude E_{gray}. The other two rows illustrate E_{color} for different colorspaces. In particular, the second row demonstrates the use of the RGB (left) and HSV (right) color spaces, while the last row displays results given the Lab (left) and the color-invariant color space (right). As observed in the examples of Figure 14, the E_{color} results are quite different from the E_{gray} one. Attention is drawn to the last image of the shading invariant color transform, where the gradient magnitude indicates physical edges and can be used as an image segmentation cue. Also, the Lab gradient exhibits a discrimination of color blobs, based on visual impression.

The dimensionality of gradient vector components, for each scale, is triple. Thus, vector gradient information can be used to encode the "direction" in color space of the color change and can be taken into account for the discrimination of different types of region borders.

Intensity and color blobs When summarizing blob information over scales, a "scale-less" blob detection result is obtained, detecting blob-like intensity regions at all scales. Similar to image gradient derived features, using Equation (8) as the weight function in Equation (3) yields (with $h(x) = x$):

$$J_w = \sum_\tau w_{\text{blob}}(\vec{x},\tau)\left(t|L_{xx}(\vec{x},\tau) + L_{yy}(\vec{x},\tau)|\right) =$$
$$\sum_\tau \left(\left(1/k_{\text{blob}}(\vec{x})\right)\left(t|L_{xx}(\vec{x},\tau) + L_{yy}(\vec{x},\tau)|\right)\right)\left(t|L_{xx}(\vec{x},\tau) + L_{yy}(\vec{x},\tau)|\right) \Leftrightarrow$$
$$J_w = \left(1/k_{\text{blob}}(\vec{x})\right)\sum_\tau \left(t|L_{xx}(\vec{x},\tau) + L_{yy}(\vec{x},\tau)|\right)^2$$

Given that:

$$k_{\text{blob}}(\vec{x})\sum_\tau \left(w_{\text{blob}}(\vec{x},\tau)\right)^2 =$$
$$k_{\text{blob}}(\vec{x})\sum_\tau \left(\left(t|L_{xx}(\vec{x},\tau) + L_{yy}(\vec{x},\tau)|\right)/k_{\text{edge}}(\vec{x})\right)^2 =$$
$$\left(1/k_{\text{blob}}(\vec{x})\right)\sum_\tau \left(\left(t|L_{xx}(\vec{x},\tau) + L_{yy}(\vec{x},\tau)|\right)\right)^2,$$

J_w may be written as : $J_w = k_{\text{blob}}(\vec{x})\sum_\tau \left(w_{\text{blob}}(\vec{x},\tau)\right)^2$

This results in a representation that summarizes blob information over all scales, as shown in Figure 11 (right image). The SSR, based on the blob

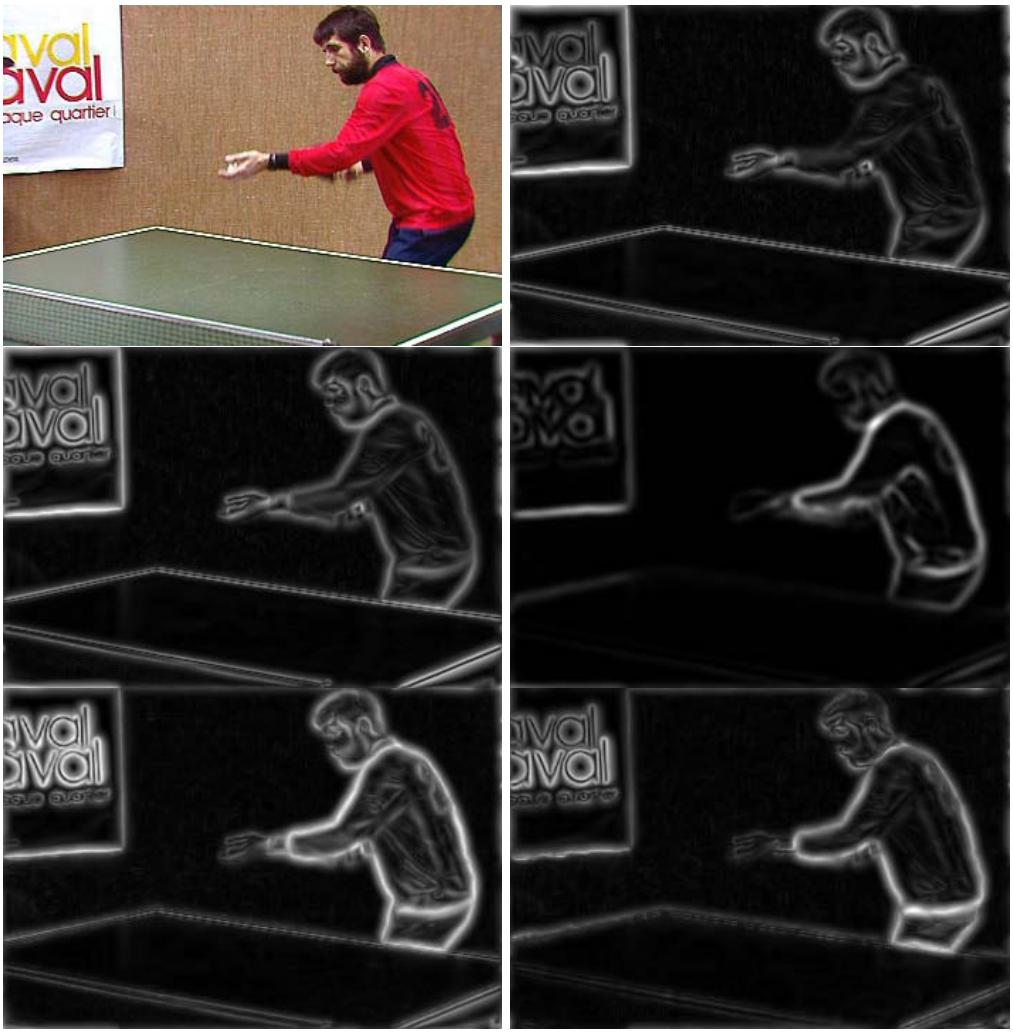

Figure 14: Demonstration of gray-scale and color gradient magnitude computation using the SSR (see text).

detector, favors scales that match local image structure size. As a result, the contribution of these scales dominates over others. In this category of applications the scale-selector is given by Equation (8). An example of blob classification with respect to scale was presented in Figure 10.

As with in case of color edges studied above, the blob feature detector can be extended to describe color blobs as well. In this case the Laplacian of the color image is three-dimensional and given by $L_{xx} + L_{yy} + Lzz$. Then, the color blob detector is expressed as

$$A = \sqrt{\sum_{i=1}^{3} L_i^2(\vec{x}, \tau)}, \tag{14}$$

where i refers to the color band and L to the original image at some scale. Similar to the color edge example, the color blob detector is applied to the same image and color space. Figure 15 illustrates the SSR for the blob detector, derived for the same set of color spaces as in the color edge example. Images are arranged in the same order as in Figure 14. In the experiment, 8 image scales were used and the order of images, with respect to the color space used, is identical to that of Figure 14.

Color is a powerful visual cue for image description. Its multiscale analysis contributes to the refinement of this description. Using the SSR, this procedure may be performed in a way that overcomes the computational difficulties of explicit scale-selection. The SSR may also be used to produce a result that accumulates both coarse and fine scale information, while the same framework can be used for the scale classification of color blobs by focusing the summarization on ranges of scale.

The summarization of visual content using the SSR framework, provides a simple and parallelizable computational method for the accomplishment of such tasks.

2.2.4 Image processing applications of the SSR

In this subsection, some indicative image processing applications of the SSR are presented. Emphasis is placed on applications of region-based processing in order to demonstrate the ability to adapt image filtering to local structure. In this context, gray-scale and color image smoothing, using the SSR, is initially considered. Then, iterative smoothing of images, based on the SSR, is used to define a structure retaining image scale-space. Finally, the enhancement of a known color constancy algorithm is demonstrated.

Figure 15: Demonstration of gray-scale and color blob detection (see text).

Knowledge of the scale of image structure has direct consequences for to image smoothing. If the purpose of smoothing is noise suppression, it is argued that the results are improved by adapting the spatial extent of the smoothing kernel to local image structure, rather than explicitly smoothing at one scale (an idea originally formulated in the Perona and Malik anisotropic diffusion approach [82]). If the blob detector is used as the scale-selector in the SSR of image content, then the resulting image retains its dominant structural features because SSR favors the contribution of scales that match local image structure. By favoring such scales, smoothing is restricted within intensity blobs, preserving edge structure without mixing the content of distinct image blobs.

Figure 16 presents an image (top-left) and the result of smoothing with a Gaussian kernel of constant size, $\sigma = 3$ (top-middle). Next (top-right), the smoothing result obtained with SSR, using the blob scale-selector given by Equation (8) and 8 image scales, is shown. It is observed that the image structure is retained, since SSR favors the contribution of those scales corresponding to local image structure. A comparison with the smoothing produced by the anisotropic diffusion approach proposed in [82]) (bottom-right) indicates that SSR yields a smoother result, as can be concluded by observing the background in the two images. Typically, anisotropic diffusion persistently retains the structure of fine-scale texture. In contrast, using the SSR the uniformity of the background at coarse scales is favored, due to the greater response of the blob detector at these scales. This effect is further emphasized by focusing on the range of scales around the maximum of the scale selector, at each pixel, using the same number of image scales and $s = 1$ in Equation (6) (bottom-middle). In this image the dominating-blobs' morphological features are intensely emphasized, while the variance of image intensities within their area rapidly decreases.

The instability of explicit scale selection is illustrated in the bottom-left image of Figure 16. In this case, the image scale space, using the same number of scales as before, is computed along with the scale-normalized blob detector response over scale at each image point. Then, the maximum $max_{blob}(\vec{x})$ over scale of this response is obtained for each image point $\vec{x}$ and the scale $sc(\vec{x})$ at which it occurs is derived. The image shown was computed as: $I_s(\vec{x}) = L(\vec{x}, sc(\vec{x}))$. It is interesting that an instability is mainly observed within image regions exhibiting fine-scale structure. The quality of the result at image regions containing fine scale structures is more sensitive to this instability than at those containing coarse scale structures.

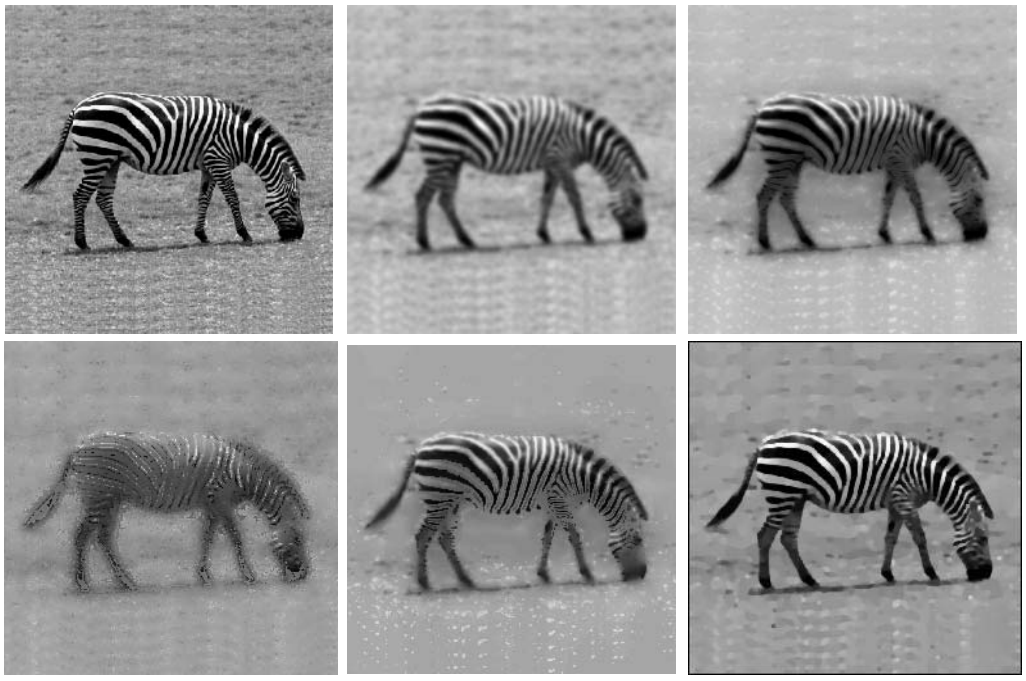

Figure 16: From top left and clockwise: Original image, Constant Scale Smoothing, SSR Smoothing, Anisotropic Diffusion, Dominant SF Smoothing, Explicit Scale-Selection Smoothing

The iterative application of SSR smoothing may be used to generate a scale-space biased in preserving image structure. This bias refers to the decrease of the variance of intensity values within blobs. Figure 17 illustrates such a scale-space by iteratively smoothing an image using the SSR. In the experiment, the scale-summarized image scale-space used consisted of 8 image scales.

In the next example, the expansion of the SSR smoothing for color images is demonstrated. In this case, the scale-normalized response of the color blob detector was utilized as the scale-selector. For the experiments, each RGB color band is independently scale-summarized. The resulting scale-summarized color bands are used to form the "SSR-smoothed" color image. In the following examples, eight image scales were again used (8) to generate each color scale-space image. In Figure 18, a color image is iteratively smoothed using the procedure described above (left column) and the results are compared to Gaussian smoothing (right column) [5].

As a final example of the ability to localize image processing to neighborhoods determined by image structure, the Retinex algorithm [54] is considered within the framework of SSR. The Retinex algorithm attempts the normalization of color images with respect to the illuminant. The restoration of the reflectance spectrum of a surface, based on samples taken over a constant sized neighborhood in an image, will eventually mix samples of different surfaces. This weakness of the Retinex algorithm has been previously reported [49] as a failure of the method near reflectance edges. On the other hand, using the SSR, the weighted average of all images scales is used, for the computation of the Retinex result. For simplicity and since the Retinex operates independently on each color band, the examples that are presented demonstrate results for the monochrome version of the algorithm. In particular, the Retinex formula is applied to gray-scale images, which are successively scale-summarized using the blob-scale selector. As demonstrated in the examples of Figure 19, the blob scale-selector prevents the sampling of surfaces with a different reflectance spectrum by responding with a low scale near edges. The middle column shows the results of Retinex image processing at a constant scale. The right column illustrates the results with the use of SSR. The Retinex algorithm is applied at all scales and then the results are merged with respect to the response of the blob detector. In the next

[5] Since a direct comparison of the two methods is difficult, the scales corresponding to $\sigma = 7$ and $\sigma = 21$ were empirically selected.

Figure 17: Image evolution in scale space created by SSR smoothing (top to bottom and left to right).

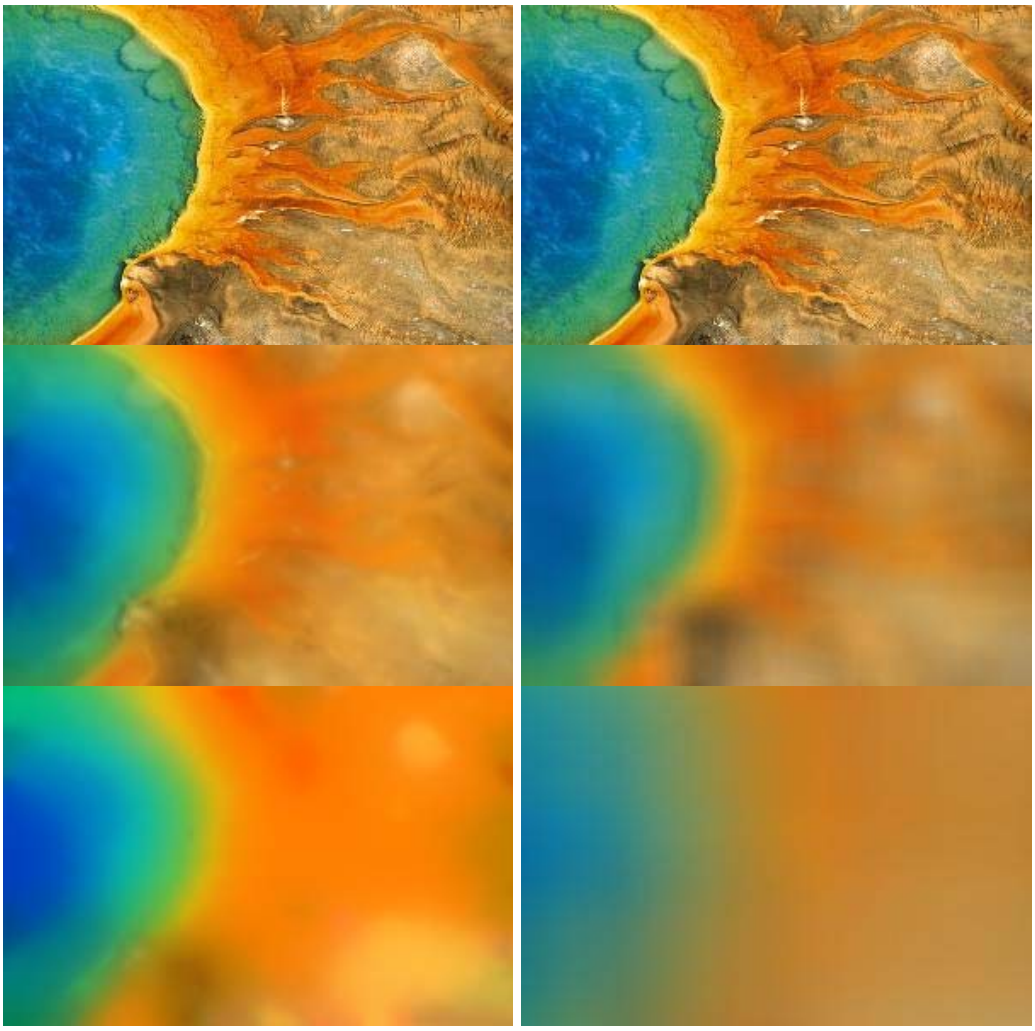

Figure 18: An image iteratively smoothed using the SSR (left column) with the color blob scale-selector, and using Gaussian smoothing (right column)

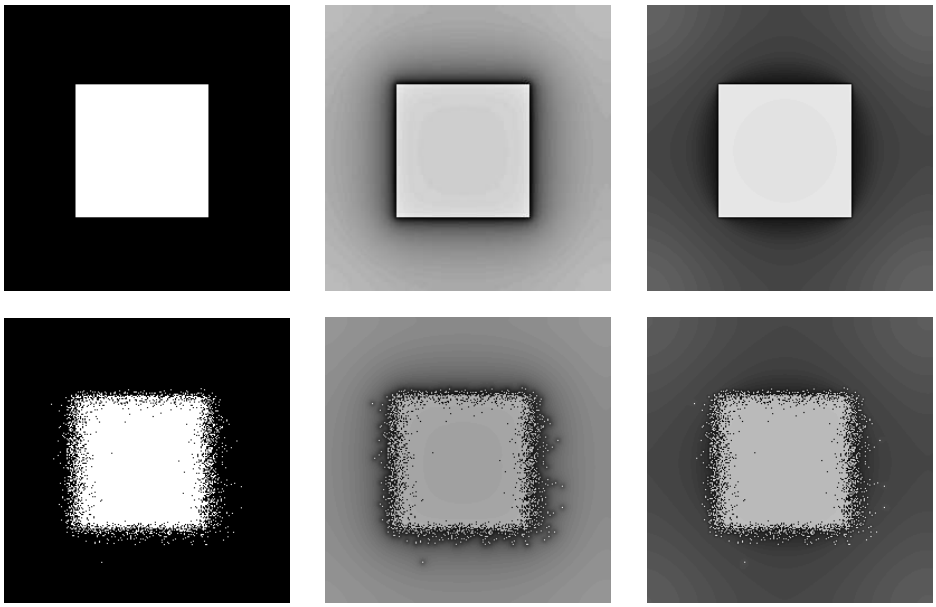

Figure 19: Square and locally orderless square (left column). Single scale Retinex results (middle column). SSR Retinex results (right column)

example (see Figure 20), the Retinex algorithm is applied to a real world image (top-left). The results from constant scale processing (top-right) and the SSR version of the algorithm (bottom-right) are shown to exhibit differences near physical edges as predicted. To indicate that differences between the two images occur mainly at physical edges a "difference" image is computed and presented at bottom-left.

2.2.5 Conclusions

In this section, a framework for scale-summarization of visual content was introduced. The purpose of applying the SSR for primitive visual feature detection and representation is twofold: *(a)* facilitates the focusing of attention to more than one scales of interest and *(b)* make possible the detection and classification of primitive visual features with respect to scale. In both cases (a) and (b), the computation is performed using the scale-summarization of image content in order to overcome computational difficulties, followed by the explicit selection of a single scale. Furthermore, a broad variety of primitive features can be represented and scale-classified by using the SSR. The ability to classify primitive features with respect to scale contributes to the refinement of content queries by attributing features with the scale at which they occur. In addition, the SSR provides a single framework for the described computation, for a variety of feature types.

Another application of the SSR in image content description is that it can be used to reduce the memory requirements of multiscale analysis of image content. Instead of using the whole image scale-space, a few characteristic "snapshots" (e.g. using scale-summarization over fine and coarse scales) of the image scale-space can be used to describe primitive feature image content.

Restricting the summarization of content within a range of scales yields a scale-normalized summary of visual content. Depending on the visual task, interest may be focused on coarse scales or image detail. Using the SSR, content extraction may be tuned for a coarse or a fine range of scales. The scale focused representation over a range of scales captures a larger portion of image content than that obtained from the analysis of one image scale in a given neighborhood.

The scale-adaptive description of visual content is also useful in visual information representation, since it facilitates the uncommitted processing of primitive image content, as well as the scale normalization of visual features occurring at different scales. The focusing of summarization on ranges of

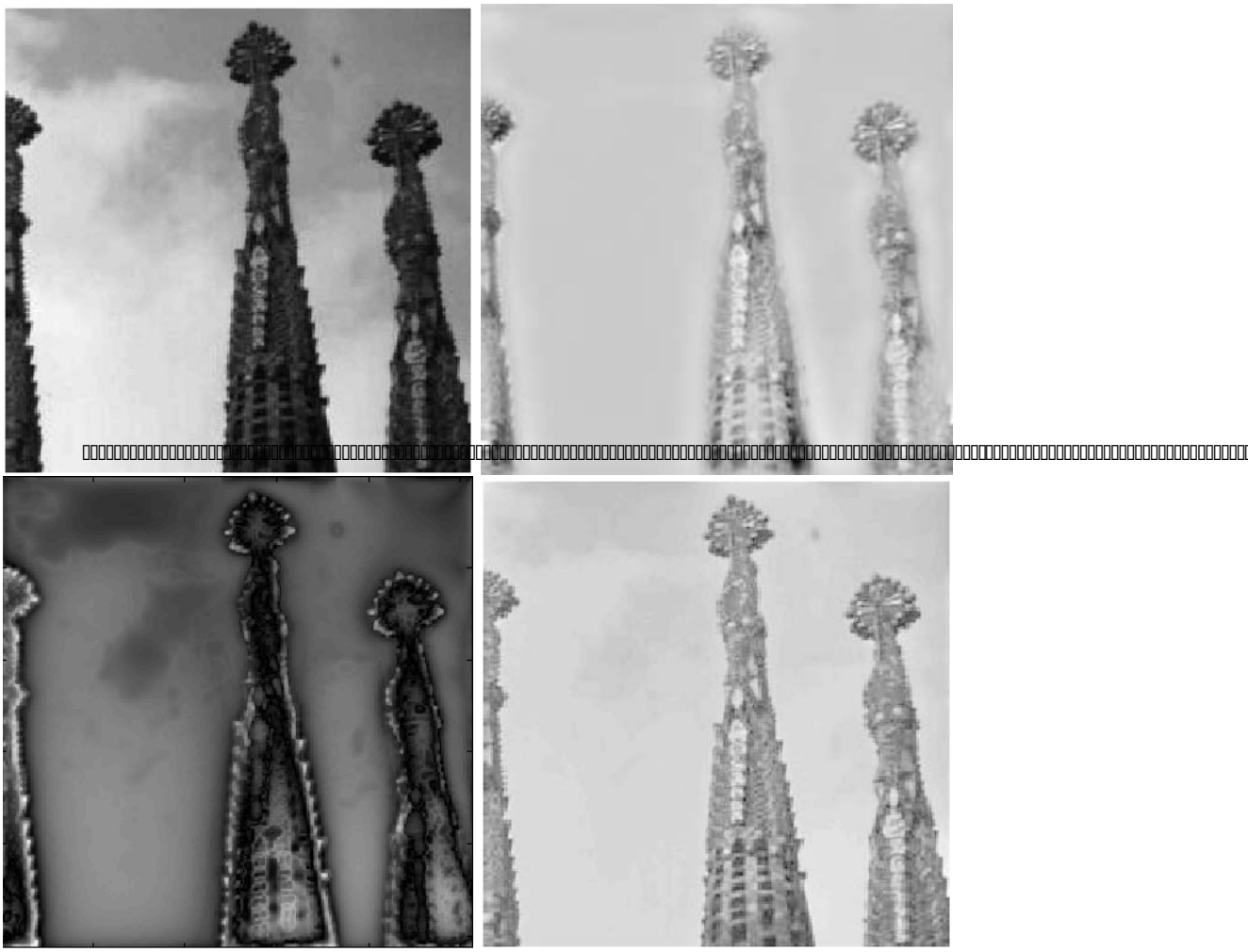

Figure 20: Clockwise, starting with top left image: Original image, constant scale Retinex result, SSR Retinex result, absolute difference of constant scale and SSR Retinex images.

scales, where local maxima of the scale-normalized feature detector response function occur, facilitates the classification and estimation of features with respect to scale.

The classification of features with respect to scale facilitates use of this scale of classification as a query predicate. In turn, the ability to use the scale attribute to form a query predicate facilitates the formulation of scale-specific or scale-independent visual content queries. The benefit of classifying visual features with respect to scale is that the query space of scale-specific visual queries can be restricted, targeting the search at only image scales of interest. Furthermore, as formerly demonstrated in Section 2.2.4, the ability to adapt image analysis tasks to the size of local structure with the use of a blob detector, can be useful in several image processing applications.

Finally, it is argued that the parallelizable nature of the computation proposed by the SSR framework, provides a method for capturing image content in real-time applications.

2.3 Utilization of primitive visual information

In this section, properties of the environment, derived from primitive visual features, are investigated and the role of primitive visual features in the general context of the description of visual content is discussed.

2.3.1 Properties of the environment

In this subsection, the relation between primitive visual features and properties of the environment, which can be extracted from them, is discussed. In this discussion, primitive features are classified in two broad categories: *(a)* those related to apparent surface reflectance properties and *(b)* those related to the geometry of surfaces represented in images, mainly observed through spatial changes of luminance or color in the image.

Visual features that reveal surface qualities are commonly used by the human visual system to extract information concerning the three-dimensional structure of the environment. In addition, the same information constitutes a strong visual cue for the perception of surfaces as concrete objects. Artists, have for long used color information to add the sense of depth and structure in paintings. The use of shading, reflections, and color change as three-dimensional cues in paintings, where much of the non-critical information is often abstracted, indicates their strong perceptual significance.

As discussed in the review of early vision physiology, perceptually compatible reparameterization of colorspace can yield a grouping criterion for image elements that belong to the same surface. Theoretically, given full information about the illuminant and the receptoral response, an observer can understand the absorption spectrum of an observed surface. Color constancy can simulated using a spatial filtering mechanism [54], in combination with the speculation that the human brain learns known patterns of illumination in order to yield a more efficient light interpretation [65, 14, 15]. This constancy assists observers in the recognition of an object from its color, despite changes in scene illumination. Thus, color information contributes not only to the segmentation of images into regions corresponding to different physical surfaces, but also to the recognition and retrieval of visual information. In Chapter 5, the significance of color as a cue for visual information retrieval is revisited.

Surface reflectance information provides structural cues concerning the three-dimensional structure of the observed surface. In Computer Vision, the topic of *Shape from Shading* refers to methods for extracting local surface geometry information from the change of apparent luminance of a surface (see [9] for an overview). Another well known approach that takes surface shading information into account is *edge labeling*, where such information is integrated with edge detection results. The visual tokens are accepted by a grammar whose predicates allow the deduction of realistic combinations, thus denoting which feature combinations correspond to physical objects. However, since the total understanding of such visual components requires information on spatial arrangements, of features, this topic is further discussed in the next chapter.

Another cue about the structure of observed scenes comes from the observation of color across a scene. In particular, apparent color yields a depth cue for large distances. For example, the observer is assisted in the coarse estimation of the distance of the mountains illustrated in the left image of Figure 21, if the change in color saturation is correctly interpreted. *Aerial* or *atmospheric perspective* refers to certain systematic differences in the contrast and color of objects that occur when they are viewed from great distances. Distant surfaces typically exhibit reduced contrast in images, due to additional atmosphere through which they are viewed (also evaporated water, dust, or pollutants participate in this phenomenon). Differences in the chromaticity of the atmosphere are related to the absorption spectrum of air and non-uniform scattering of light with respect to its wave-length.

Figure 21: In the original image (left), the reduced contrast and increased presence of blue color components, at image regions depicting distant surfaces, add to the impression of depth in this picture. The middle image illustrates color saturation and the last image shows the *Blue* channel.

The two rightmost images of Figure 21 show the color saturation (middle) and the "blue" (RGB) channel. As expected, the farther the surface the less saturated and more bluish it appears. Thus, the saturation and blue channel in combination can provide an empirical cue towards depth perception.

Visual features derived from local structure rarely capture visual content alone. However, features of higher dimensionality may be derived from their grouping. The grouping of edgels into edges is one of the most common structure descriptors used in the morphologic description of images. An indication of its high relevance to the formation of visual perception is its derivation from an early visual stage by simple, complex, and hypercomplex cells in the striate cortex. The analysis of structural information gives rise to the detection of lines, curves, and corners in images, which are primitive features widely used in image processing and understanding. A framework for the computation of some of these tokens can be found in [66].

Furthermore, the grouping of edgels that form the boundary of an object yields a contour which represents the silhouette of a visual entity. The boundary of this silhouette is often used as a powerful cue in image retrieval by content. In addition, the spatial arrangement of image structure and orientation characterizes image *texture*, which can yield cues for surface recognition and three-dimensional information extraction. Feature grouping is explored further in Chapter 4, while the spatial arrangement of primitive features is considered in the next chapter.

2.3.2 Visual information description and management

In this subsection, the conclusions drawn from Section 2.3.1 are taken into account in a discussion of visual content description. Descriptions of image content are at the core of content-based image retrieval systems and the information represented such descriptions is crucial, as it forms the basis for an accurate response to similarity queries. In this subsection, primitive features or *tokens* of content description are discussed and scale, as an attribute of such tokens is considered.

Description tokens The visual tokens derived from the class of primitive features fall under two main categories. The first is derived from the organized perception of picture elements with similar color properties and the second from the grouping of points contributing to the definition of form. It should be noted that the use of the term "token" does imply or not require that there exists a physiological symbolic representation of this information. The adopted perspective is utilized in the context of machine assisted visual information browsing and retrieval, for which an analytic formulation of related visual components is often preferred as discussed in Section 5.2.2.

The description of image regions with respect to their *apparent* reflectance properties, or color, is related to image segmentation. A large number of color image segmentation methods exist in the literature (see [75] for an overview). The assumption underlying this approach to image description is that a region of picture elements exhibiting similar color properties belongs to the same physical surface, and thus to the same visual entity. Taken a step further, this assumption is commonly adopted by content-based image retrieval techniques, in the pursuit of images containing regions of similar color and preferably in a similar layout. An investigation of the conditions under which these assumptions hold, as well as the limitations of the derived methods, yields results that are useful in the refinement of query formulation based on color features.

Color features may be classified using one of two approaches:

- the first is based on surface reflectance information and, therefore, color features are classified in terms of their surface absorption spectrum.
- the second is based on a phenomenological description of perceived color.

In this subsection, only the local properties of color are examined. The visual information related to spatial arrangements of color features are discussed in the next chapter.

The distinction of physical surfaces in terms of their spectral properties requires that image pixels corresponding to a given surface are classified in the same set and that this set does not contain other members. Theoretically, if the extraction of this information from a 3-color band image is possible, then surfaces of the same type can recognized in images acquired under different conditions of illumination. However, if the scene illumination is unknown, then color information alone is not enough for the computation of the surface's absorption spectrum.

The study of color invariants shows that the effect of shadows, reflections, or non-uniform illumination can be eliminated from an image description by transforming the colorspace into a one that exhibits color constancy, or otherwise is invariant to such phenomena. An example of such a case was illustrated in Section 2.2.3, for the detection of color edges and blobs. Nevertheless, the problem of retrieving surfaces of similar spectra from a database cannot be solved without the definition of scene illumination for each image in the data set. The perceptual mechanism that enables the approximately constant perception of surfaces under different illumination conditions, referred to *color constancy*, supports the recognition of objects under varying illumination patterns. Although methods that perform color normalization exist, they do not work for all possible illuminations. Approximations of color constancy can be devised if a set of familiar illumination or reflectance spectra is assumed [65].

A phenomenological approach to the same problem would represent the color information as encountered in the image. Image pixels are represented by their nominal colorspace values and color is compared given a color distance metric. However, if color characterization aims at a representation that is relevant to human perception, some color contrast phenomena should also be taken into account. As discussed earlier, color perception cannot be defined on the basis of image elements, but also requires a number of other surrounding cues. In an accurate simulation of color perception, such factors should be taken into account.

Primitive features related to the geometry of surfaces represented in images (or gradient-derived primitive visual features) are derived from spatio-temporal changes in apparent color or intensity. Although features such as edges, corners, line endings etc. cannot capture structure on their own,

higher level features, emerge from specific arrangements of primitive features. Examples of higher level features are contours and perceptual groups (see Chapter 4).

Scale attribution of visual tokens Primitive features of image structure constitute a fundamental component of visual content. If the scale of these is known, the problem of estimating properties of observed structures, and thus obtaining a more expressive content representation, is simpler to solve.

Since primitive visual features can be observed at different scales, it is expected that classifying them, with respect to scale will improve the expressiveness of their representation and facilitate the refinement of visual queries. The study of the stimulus representation in the LGN, implies the rapid perception of motion and coarse scale image structure and the relatively slower perception of color and detailed image structure. The SSR framework introduced in this chapter contributes to the extraction of scale information about primitive features. Furthermore, using the SSR information extracted from coarse and fine ranges of scale can be individually processed, stored, and retrieved in a simple way and at a low computational cost.

3 Spatial Arrangements of Primitive Visual Features

The central theme of this chapter is the spatial arrangement of primitive visual features. Specifically, the information contained in spatial arrangements of primitive visual features is discussed and its role in the perceptually relevant description of visual content is considered. The derived descriptions are applied to the task of content-based image retrieval.

In the first section, properties of spatial arrangements of primitive features are presented. In addition, it is shown that image regions that exhibit a constant spatial arrangement of primitive features have descriptive value.

In the second section, perceptually relevant representation of spatial arrangements of primitive features is considered. First, the required properties of such a representation are discussed. Second, a framework for the representation of spatial arrangements of primitive features is proposed and modeled from a computational point of view, with emphasis on its storage capacity requirements. Finally, the scale-summarization of the discussed representation is proposed as a method for the reduction of memory capacity requirements, which also facilitates the classification and normalization of spatial arrangements of primitive features with respect to scale. The perceptual relevance of the proposed framework is based on the scale-normalization and uniform representation of scale-varying spatial arrangements of primitive features.

The third section deals with the generic description and similarity comparison of spatial arrangements of primitive features. Conclusions are drawn that are used to formulate methods for the extraction of image regions that exhibit a constant spatial arrangement of primitive features, by grouping local descriptors of such arrangements. The perceptual relevance of the extracted regions is based on the scale-summarized representation introduced in the previous section.

In the fourth section, higher level descriptors of spatial arrangements of primitive features are proposed and their descriptive power is demonstrated. Emphasis is placed on the mapping of attributes of such descriptors onto image properties that correspond to a description, of the spatial arrangement of primitive features in the image, that is comprehensible to a human observer.

Finally, the application of the derived methods to the task of visual information browsing is described and demonstrated in a content-based image retrieval experiment.

Figure 22: Some visual interpretations of the term *stripes*.

3.1 Introduction

In this section, the contextual dependence of the information contained in spatial arrangements of primitive visual features is shown and the role of non-visual cues in the acquisition of such information is discussed. Next, a phenomenological approach towards the description of image content is considered. In this context, it is argued that the extraction of image regions that exhibit a constant spatial arrangement of some feature is of significance, in a perceptually relevant description of image content.

Context-related description The descriptional value of spatial arrangements of primitive features is supported by the existence of specific linguistic terms. Explicit terms exist for the description of structural feature arrangements. Examples are: *(a)* repeated, sharp, with stripes, with dots, etc. for the description of patterns, *(b)* warm, cold, etc. for the description of color, *(c)* smooth, rough etc. for the description of gradients, *(d)* horizontal, vertical, diagonal etc. for the description of orientations, and many others. Often these terms are generic and several visual samples, consisting of different primitive features, could fit the description. For example, Figure 22 illustrates some examples fitting the description "stripes". Thus, the characterization of spatial arrangements of primitive features should not be generic. It is possible that specific and contextually-related known patterns are required in the goal-driven description and retrieval of spatial arrangements of primitive features, instead of generic formulations that cover the whole spectrum of possible arrangements (see also Chapter 5). Furthermore, an exhaustive representation of the elements of spatial arrangements of primitive features would support the extraction of any type of characterization,

but would be rather unrealistic given finite computational resources.

Phenomenological description A context-free or, otherwise, phenomenological description of visual content is of interest, since a generic approach towards visual information modeling would theoretically be applicable to all types of images. Although such an approach is quite applicable in the study of sensory information, its adoption for the purpose of studying perceptual features raises a number of issues. These are:

- The perception of spatial arrangements of primitive features is produced by perceptual means, which are context-related. Furthermore, the perception of such arrangements is not defined solely with respect to appearance and often depends on image type and observation goal.
- Context-free phenomenological description of visual content can be more abstract than often required. A typical example of this is the description of images with respect to the layout of some feature (e.g. color). Many images that are intuitively dissimilar may match a certain description.
- Most often, context *is* required in visual information related applications, in order for visual information to be combined with other types of information. In this case, the significance of each component is task-dependent.

Thus, it is not a trivial task to discriminate spatial arrangements of primitive features based on purely visual characteristic properties. Another reason for this is that the intuitive notion of phenomenological, or context-free, description of visual content still inherits perceptual cues. For example, a characterization of image content with respect to color or texture typically focuses on image regions of coherent feature expression in the image (regions of approximately constant color or texture). This characterization is strongly related to surfaces or objects encountered in the three-dimensional environment. A discussion about an information-theoretic approach to the description of visual content that makes possible the extraction of coherent image regions, with respect to some feature, is presented below.

Information theoretic description In the previous paragraph it was argued that the pure phenomenological description of spatial arrangements

of primitive features is subject to several constraints. However, such a phenomenological description is of interest for the following reasons:

- Regardless of the approach adopted for the description of image content, issues of representation economy arise from the need to efficiently manage (visual) information. Even in the absence of perceptual insight, the storage, compression, and coding of images typically benefit from an information-theoretic study of their content.

- Given that little is known about several physiological and conceptual aspects of visual information representation information theory provides an objective approach to the description of visual content.

- Although a pure phenomenological description does not capture context-dependent image content, it can be intelligently used in content-based image retrieval applications. In many cases, the comprehension of visual information involved is not a requirement for the retrieval of similar visual content[6].

The notion of order or constancy is understood to be intrinsically related to the formal definition of information (see [99]). Intuitively, order is inversely proportional to description length. Thus, given a data set, the most expressive description is the minimum in length. Consequently, constancies in a data set can be exploited towards an expressive or length-efficient description. From this perspective, measures of order, such as entropy, can be used to reveal ordered information "entities" in images. To illustrate this point, a classification of image regions, with respect to information order, was carried out and is described below.

In the following example, an image was "described" by a machine learning method that attempts to hierarchically describe a data set, making use of classifiers. The image was hierarchically partitioned in such a way that each cut would be optimal in decreasing the global information entropy of the description. Qualitatively, the partitioning process attempts to minimize the distance between two points in the same cluster and maximize the distances between points in different classes. This process yields a *partition or categorization tree* corresponding to the selected partitioning classifier at each

[6]In the same manner that a text retrieval system does not necessarily *comprehend* keywords.

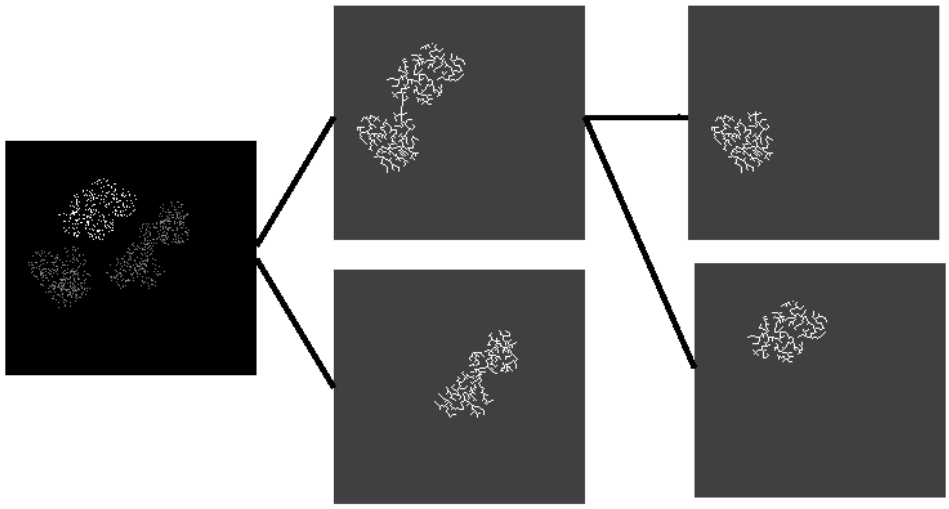

Figure 23: Segmentation of an image into coherent intensity regions (see text).

node. The partition tree constitutes a hierarchical image description. Similar approaches are encountered in unsupervised machine-learning methods and conceptual clustering. In this example, the information entropy-based partitioning criterion or *category utility* was selected, as in [28].

The segmentation was achieved through the following procedure: *(a)* the image was represented as a graph (one node per pixel), in order to capture the spatial relationships between picture elements, *(b)* close neighbors were linked to each node and the minimum spanning tree of that graph was computed, *(c)* the tree-structured graph was recursively partitioned using a branch and bound procedure, which selected each time the cut that would generate the most ordered partitioning. When the order stops increasing, the process is terminated for that recursion node. Figure 23 illustrates the acquired partitioning result, corresponding to a hierarchical classification of visual information with respect to information order. The leftmost (original) image of the Figure 23 consists of three "clouds" of points, the members of each one illustrated using a different grayscale value.

Analogous segmentation results can be achieved with other types of features, yielding a quantitatively coherent spatial arrangement. Figure 24 illus-

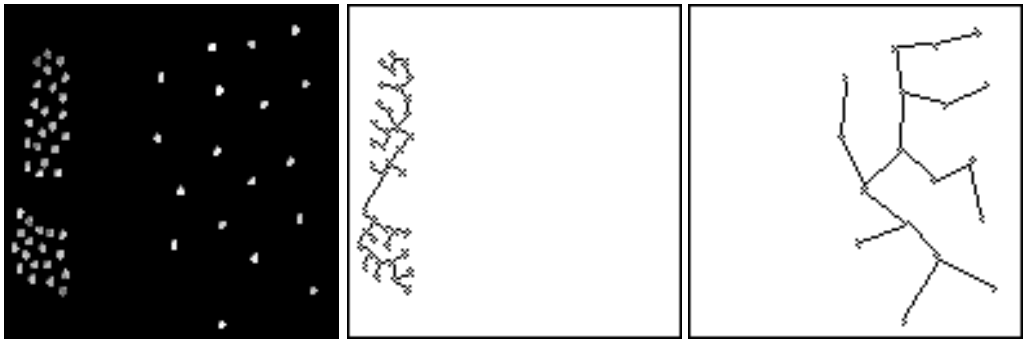

Figure 24: Segmentation of an image into coherent local scale regions.

trates the segmentation of image regions with respect to local scale, defined by local structure and quantified using the blob detector of the previous chapter, as formulated in Equation (8). The image is segmented into regions of coherent nominal node values, representing the elements of the lowest informational order.

The adoption of this approach, as a generic methodology of spatial clustering in image segmentation problems, is prohibited by its computational ($O(N^2 \log N)$, where N is the number of pixels). Later on in this chapter, more efficient computational methods for spatial clustering of visual information are discussed.

The lack of disorder can also be characterized by the constancy, with respect to its nominal expression, of some feature in the image. As there is no nominal difference between elements of a coherent region, they cannot be discriminated. Thus, another segmentation, resulting in more ordered segments does not exist. The experiment indicates the descriptional value of coherent structures, which are typically encountered at coarse scales in images. This indication complies also with the segmentation results based on minimum length descriptions [60] and color [55]. These results indicate that the most informational description classifies features into nominally coherent subsets of curvature.

Discussion The relation of the constancy of spatial arrangement of some feature to a perceptually relevant description of visual content provides insight into the visual system's "strategy" of obtaining a raw visual percept. The rapid estimation of content at coarse scale and its subsequent refinement

based on fine structure is currently a speculation in physiology and psychology. Its application has boosted the computational efficiency and increased the effectiveness of related artificial visual competences. Some characteristic examples are found in the computation of optical flow [106] and the tracking of points in image sequences [105], where coarse scales are utilized in the extraction and anchoring of features.

The description of visual content using informationally coherent regions of interest, or visual entities, is also interpretable from an evolutionary perspective. Patterns met in natural environments exhibit order in various forms, such as symmetry, fractals, and constancies [1]. The speculation that visual systems evolve towards the optimal perception of the environment justifies attempts to describe visual content in terms of visually constant regions, since such regions are correlated with elements in the environment. In this context, the description of visual entities can be seen as a process of constancy estimation. From the Gestalt perspective, gathering of visual features with respect to their appearance (or apparent movement) and location constitutes a form of perceptual grouping that leads to the perception of visual entities.

Emphasis is given to the fact that although an information-theoretic description of visual content seems to be compliant with the intuitive tendency to describe visual information in terms of nominally constant elements, both quantitative and qualitative knowledge, concerning the characterization of visual elements, is missing. Specifically, not all of the features involved in such constancies are known. Also, the metric properties of feature representation and similarity remain mostly not estimated. In addition, context-based knowledge might be involved in this task.

3.2 Representation

In this section, the representation of spatial arrangements of primitive features is considered. A representation framework for such information is formulated and computationally optimized with respect to memory capacity. In addition, the suitability of the proposed framework for the representation of spatial arrangements of primitive features with respect to scale is demonstrated.

3.2.1 Requirements

In this subsection, the requirements for the representation of spatial arrangements of primitive features are discussed. Initially, qualitative requirements are discussed dealing with the type of represented information. Next, quantitative requirements are considered. The qualitative requirements that are presented concern the type of the representation. Quantitative requirements concern the memory capacity that is required for the proposed representation.

Representation type From the discussion in the first section of this chapter and the review of primitive features in the previous chapter, the following qualitative requirements for the representation of spatial arrangements of primitive visual features, are derived:

Multiple channels Independent information, such as color and directional spatio-temporal change, originate at the retina and is propagated to the primary visual cortex, where different "information channels" are independently represented. The components of visual content, derived from such information, can be perceived in isolation by observers. Thus, two image regions may be similar with respect to a visual component (e.g. color), but dissimilar with respect to another (e.g. orientation). For example, in Figure 25, two pairs of spatial arrangements of primitive features are illustrated. Both of them differ with respect to one visual component, but are similar with respect to another. More specifically, the left pair differs in the arrangement of local orientations, but is quite similar in the arrangement of intensity values. The right pair differs in the intensity values, but is similar with respect to their arrangement.

Given the task of content-based image retrieval, it is required that a perceptually relevant representation be capable of capturing differences and similarities, such as the ones described above. Therefore, such a representation should support the separation of different visual components corresponding to spatial arrangements of primitive features. In order to achieve this goal, a multi-channel representation is proposed, where each channel corresponds to the description of a certain arrangement of a particular primitive feature.

Multiple scales The visual information extracted from the observation of spatial arrangements of primitive features is dependent on scale in two

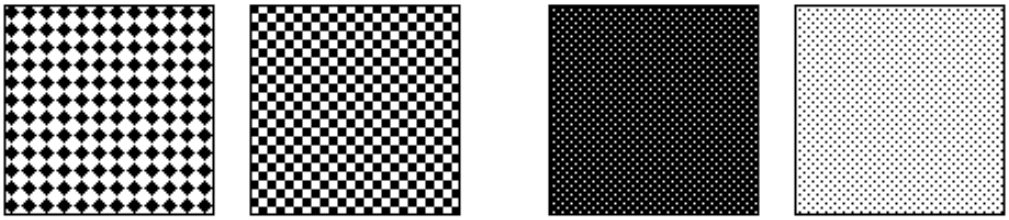

Figure 25: Two pairs of spatial arrangements of primitive visual features. The members of each pair exhibit differences and similarities (see text).

ways. First, the primitive elements of an arrangement are dependent on the range of scales within which they may be observed. As discussed in the previous chapter, an informative description of primitive features requires the consideration of multiple image scales. Second, the spatial arrangement of primitive features is intrinsically dependent on scale. In content-based image retrieval, regions of interest ought to be characterized as similar when they differ solely with respect to scale (e.g. texture gradients [33]). Thus, one may conclude that spatial arrangements of primitive features should be represented at multiple scales.

Locality The computation of certain primitive features requires the local (over a small spatial neighborhood) consideration of feature expression in the image. For example, the perception of edges, corners, and orientation is intrinsically defined within the extent of some spatial neighborhood. In addition, color perception is also determined within some spatial extent.

Thus, local descriptors of primitive feature arrangements are required in order to represent the arrangement of primitive features at each image point.

Conclusion For the reasons stated above, a representation for spatial arrangements of primitive features is proposed that is based, on local descriptors. Each image point is associated with a local descriptor, which is characterized by its sampling aperture. By varying the size of the aperture a "scale-space" of local descriptors is defined.

Representation requirements In this paragraph, a rough estimation of the memory capacity required to store a multiscale local descriptor represen-

tation in memory is presented.

Let some visual field occupying an area A be expressed in some area measurement unit u. Let D be the density of photoreceptors in some visual system, for visual field A. D is defined as the number of photoreceptors per unit area u. An acquired image then consists of $A \cdot D$ picture elements. If the signal is simultaneously represented at S scales then the number of data elements required for the multiscale representation of the signal is $S \cdot A \cdot D$.

If a non-uniform spatial arrangement of photoreceptors is considered (e.g. as a function of eccentricity), then the previous estimation can be reformulated as: $S \int^{\vec{x}} A \cdot D(\vec{x}) d\vec{x}$, where $\vec{x}$ is the location of an image point. If a non-uniform density of photoreceptors over scale is also considered, then the number of the multiscale representation elements can be estimated as: $\int_0^S \int^{\vec{x}} A \cdot D(\vec{x}, s) d\vec{x} ds$. Each point $(\vec{x}, \tau)$ in this "scale-space" $\mathcal{X}$, will be denoted by $\mathcal{X}(\vec{x}, \tau)$, where τ is the logarithmic scale parameter, defined in the previous chapter.

The information complexity of each channel depends on the type of feature represented. The dimensionality of color is 3, while change (motion and gradient) is described in two dimensions (intensity and direction). As an example, if the features represented are color and luminosity gradient, which require three and two data elements for their representation respectively, the memory capacity of $\mathcal{SS}$ would be:

$$(3d_{color} + (d_{orientation} + d_{grad\_values})) \int_0^s \int^{\vec{x}} A \cdot D(\vec{x}, s) d\vec{x} ds, \qquad (15)$$

where d_{color} is a representation element capable of acquiring as many different states as the color-band tones, and $d_{orientation}$ the number of orientation selectivity channels. d_{grad_values} is a data element capable of acquiring as many different states as the granularity of gradient-magnitude value representation.

If instead of the raw primitive feature information, at each image point $(\vec{x})$ and for all scales (S), the local arrangement of primitive features were to be represented, then for each point $(\vec{x}, \tau)$ of $\mathcal{X}$, a data structure $\mathcal{DS}$ would be required. This data structure would store the representation of the description provided by the local descriptor, for each image point $\vec{x}$ and at some scale S. The description would provide information about the feature arrangement within the local descriptor's spatial neighborhood or *sampling aperture*.

As in the case the (standard) linear Scale-Space, where each scale is associated with some value of τ, the area covered by the sampling aperture of

a local descriptor at some scale s is defined as a function of τ, and, thus, this area can be denoted as $\alpha(\tau)$. The memory capacity of this data structure for one element, over a spatial neighborhood of area $\alpha(\tau)$ would be $\mathcal{DS}(\alpha(\tau))$. This yields:

$$\int_{\tau=0}^{S} \int^{\vec{x}} \mathcal{DS}(\alpha(\tau)) \cdot A \cdot D(\vec{x}, s) d\vec{x} d\tau \tag{16}$$

for the total memory capacity of $\mathcal{X}$.

Given image sizes typically encountered in image databases (e.g. 512×512), the total memory capacity of $\mathcal{X}$ grows considerably. In the case of database images of great variety, most often the image resolution is constant. In the (standard) linear Scale-Space all scales exhibit the same resolution. Thus, function $D(\vec{x}, \tau)$ will be considered as constant.

Conclusion The required memory capacity for the proposed representation is substantial and, therefore, unsuitable for typical computational resources. The ability of biological organisms to cope with such computational and memory requirements, resulting in a remarkable speed of comprehending visual information, is due to several factors:

- The neural circuitry in LGN facilitates the parallel acquisition of the elements of $\mathcal{X}$.
- The acuity of representation and processing of visual features is not constant with respect to location $\vec{x}$ and scale / LGN layer S. Thus, a different computational effort is required for the analysis of elements of $\mathcal{X}$ at each scale.
- The scale-selectivity and individual processing of separate LGN layers indicates that, for each feature type, feature extraction is restricted to a subset of layers.

The representation of spatial arrangements of primitive features with respect to scale raises several issues besides that of memory optimization. In the next subsection, such issues are discussed and the SSR is utilized for the optimization of memory requirements.

3.2.2 Scale-summarization

In this subsection, scale-related representational requirements of spatial arrangements of primitive features are considered and a framework for repre-

senting this information is proposed, based on local descriptors. The framework is then optimized with respect to memory capacity requirements. The suitability of the representation with respect to the requirements stated in the previous subsection, is also demonstrated.

Problem Statement - Goals In this paragraph, the representation requirements concerning the generation of a perceptually relevant description of spatial arrangements of primitive features are discussed. These are classified as descriptional and computational requirements:

Descriptional requirements This class of requirements concerns the expressiveness of the acquired description. Requirements related to feature scale are: *(a)* scale-invariance and *(b)* classification of visual content with respect to scale. The interest in a scale-invariant representation of visual content is twofold. First, scale varying patterns could be described in a scale-invariant fashion and, thus, more easily grouped. Second, comparison and matching of similar patterns observed at different scales in images could be achieved. The classification of visual content with respect to scale achieves to the following: *(i)* the attribution of visual features with the property of scale can facilitate scale-specific queries that focus on abstract or detailed visual content. *(ii)* individual processing and description methods can be applied to different scale ranges, if required by some application.

Computational requirements The computational requirements concern mostly the optimization of the representation space, computational time, and reduction of complexity. In specific, the large representation space coarsely estimated in Section 3.2.1 strictly prohibits the simple generation of such an exhaustive representation. Finally, the ability of parallely implementing the representation process, would reduce execution time.

Framework Formulation In this paragraph, a framework for representing spatial arrangements of primitive features at multiple scales is introduced as an extension of the SSR, introduced in Section 2.2.

Let $h_s(\vec{x})$ be some local descriptor over a sampling aperture of s. The data structure DS, which represents the spatial arrangement for each image point, consists of an "image" with dimensionality equal to $2 \cdot dim(h_s)$. By varying the sampling area size, a scale-space of such images is defined as

$\mathcal{SS}(\vec{x},\tau)$, where $\tau = \log t$ is the logarithmic scale parameter, and $\vec{x}$ the spatial coordinates. If $w(\vec{x},\tau)$ is the sampled feature response function over scale for an image point, or in the context of the SSR the scale-selector, then by summarizing content contribution of neighboring scales, given

$$\mathcal{X}_{\mathcal{SSR}}(\vec{x},\tau) = \int_{\tau} w(\vec{x},\tau) DS(\vec{x},\tau) d\tau, \tag{17}$$

a dimension reducing, representation of content over a neighborhood of scales is obtained. The proposed accumulation is an extension of the SSR for the scale-space $\mathcal{X}$. Variations, such as Scale Focusing, are transparently applied by transforming the scale normalized response w as in Equation (6). There are several points to clarify in the above framework formulation. These are: *(a)* the definition of the nature of the local descriptor h_s, *(b)* the sampling of features by local descriptors, and *(c)* the summation of local descriptors.

Local descriptor The selection of the local descriptor h_s is closely related to the feature represented. In a generic approach to the representation of spatial arrangements of primitive features, a statistically unbiased descriptor is required. In this discussion, local histograms of features are used. In [53], the expressiveness of color and local intensity histograms is demonstrated, along with the ability of this representation to implicitly encode structure. Naturally, the histogram does not capture the spatial arrangement of elements inside the spatial neighborhood. Although it certainly can be extended to do so, e.g. in [40], it is observed that the local description of intensity and orientation can capture significant descriptional aspects of local image structure. Furthermore, the independent representation of primitive feature matches the "multiple-channel" representational requirement, discussed in the previous subsection.

The proposed formulation facilitates the transparent application of the representation to features of arbitrary dimension. The spatial arrangement of elements can also be captured by other types of histograms, such as the ones of filter responses, and spatial frequency coefficients. Since these descriptions are also based on local samplings, the extension of SSR that is presented forward, will still be compatible with such alternative local representations if: the function of description summation is defined, as it is discussed below.

Feature sampling Sampling of features for the generation of local descriptions is Gaussianly weighted, with respect to distance from the central

pixel. Furthermore, the arrangement of neighboring local descriptions is overlapping and dense (a local descriptor per image point). An observation concerning the "images" of such a scale-space is that they exhibit smooth spatial variation, since neighboring histograms overlap.

Descriptor summation Finally, in order to complete the formulation of the proposed scale summarizing framework, the summation of two descriptors has to be defined, in order for the integral in Equation (17) to be computable. If local descriptors can be expressed in vector format then their linear accumulation is used for the application of the SSR framework. For the case of local histograms, this accumulation is defined as: $w_1 \cdot h_1 + w_2 \cdot h_2$ where h_i is a local histogram, and w_i a weighting that will be used in a similar fashion as the "scale-selector" in the SSR. Depending on the dimensionality of the sampled feature (e.g. 3 for color, 1 for intensity and orientation), each histogram bin will consist of $dim(h_s)$ vector components. The accumulation is defined as for weighted vector accumulation. For example, if $h_{color} = \{(v_{00}, v_{01}, v_{02}), (v_{10}, v_{11}, v_{12}), ..., (v_{n0}, v_{n1}, v_{n2}))\}$ and $h'_{color} = \{(v'_{00}, v'_{01}, v'_{02}), (v'_{10}, v'_{11}, v'_{12}), ..., (v'_{n0}, v'_{n1}, v'_{n2}))\}$ two color histograms, then their accumulation is: $\{(w \cdot v_{00} + w' \cdot v'_{00}, w \cdot v_{01} + w' \cdot v'_{01}, w' \cdot v'_{02} + w \cdot v_{02}), ...\}$, where w and w' values of the scale selector.

By using an appropriate scale-selector, the scale-summarization of local descriptors shall yield a result for each image point, in which the contribution of each scale is proportional to the existence of the sampled feature at this scale. The scale-summarized representation favors feature samplings which correspond to scales of salient feature existence.

Framework Instantiation In this paragraph, the instantiation of the proposed framework for the primitive features of intensity, color, and orientation is presented. The descriptional properties of this representation scheme are also discussed.

A significant, difference between the variation of image scale and the histogram sampling aperture is primarily noted. When an image is Gaussianly smoothed, in order to create the linear Scale-Space of the image, the pixels values are *changed*. In contrast, when varying the histogram sampling aperture, pixel values are retained. Thus, primitive features are sampled as encountered in the original image.

Representing orientation arrangements over scale would require a his-

togram for each scale-space element. Thus, for an image of N pixels, and an analysis of S scales, the total representation cost would be $N \cdot S \cdot O_{bins}$ floating point memory elements, where O_{bins} the number of histogram bins. The orientation histogram represents a discretization of orientation angle in the interval 0 and π, with the number of bins typically varying from to 8 to 16. A corresponding representation for color would require three-dimensional histograms for each data element. An upper bound, with respect to typical imaging capabilities of modern computers, for the discretization of each histogram axis is 256, resulting in a number $C_{bins} = 2^{24}$ floating point elements of for each scale-space element, and thus $N * S * C_{bins}$[7].

Figure 26 presents an image in which a horizontal scan line is marked and portions of the multiscale histograms representation $\mathcal{X}$ for the intensity feature are shown. In particular, Gaussian weighted histograms were computed for each image point residing on the marked horizontal scan line, for all scales. Histograms of a small sampling aperture (a fine scale of $\mathcal{X}$) are illustrated perceptively tiled, in the left graph. In this graph, the horizontal axis indicates spatial coordinates along the marker line and the oblique axis maps intensity histogram bins. As observed, the small sampling aperture accurately captures the intensity value change in the image. The procedure is repeated for histograms of broader (a coarse scale of $\mathcal{X}$) sampling aperture capturing a coarser and more constant result, but with a decrease of precision[8]. Here the term "more constant" refers to the fact that the histograms corresponding to zebra points have more similar structure, than the previous case. The same holds for the background points as well. In the above example, 8 sampling apertures, or image scales, were used. Their radius was determined by the value of the, exponentially increasing, logarithmic scale parameter.

Although the coarse sampling of the image abstractly discriminates between the two image intensity patterns (the zebra vs. the background), it is neither scale invariant nor accurate. Furthermore, if a pattern exhibits scale varying behavior, two equal-area spatial samplings can be fundamentally different. The intensity of this dissimilarity is maximized when the sampling aperture of the histogram is narrower than the local spatial scale, e.g. smaller than the local texton size. A straightforward method for the achievement of

[7]Typically, this size can be reduced using a representation that takes advantage of empty bins [71].

[8]Note that in the graphs the maximum histogram value is not 1 for reasons related to the graphical presentation of this data set and not conceptual ones.

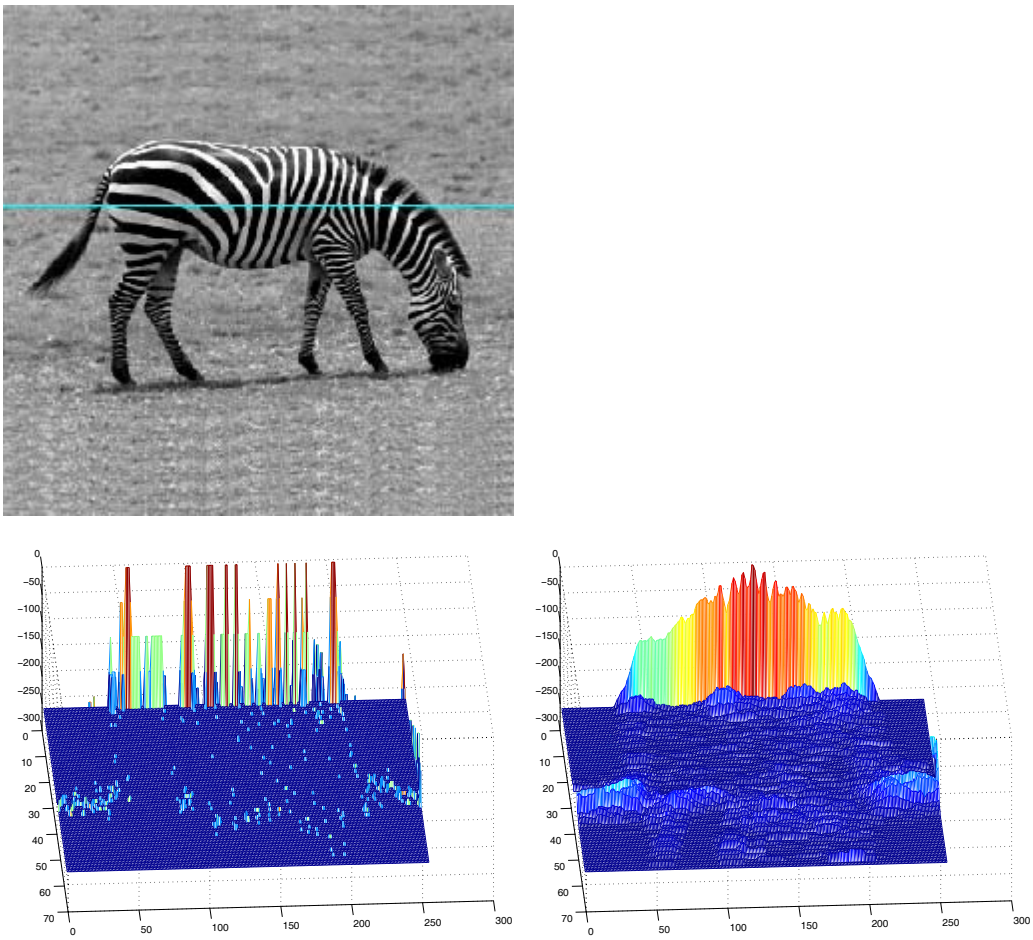

Figure 26: Fine and coarse local histograms of intensity computed along a scan line in an image (see text).

scale-invariance would be to vary the histogram sampling aperture, proportionally to local spatial scale. In other words, to *explicitly select* a histogram sampling aperture size. However, this approach inherits the computational problems discussed in the previous chapter.

In the following experiment, the SSR is utilized in the adaptation of the histogram representation to local pattern scale. Figure 27 illustrates the histogram representation of the same image line (the marked horizontal line of the image in Figure 26) for scale-summarized histograms In the experiment. eight image scales were used for the summarization. In this case, summarization was executed using the ranges of coarse scales, in particular scales 4 to 8, out of the total of eight scales. It is empirically observed that histogram sampling areas adapt to the coarse structure pattern (the zebra). As observed in the second graph of Figure 27, the representation: *(a)* represents intensity patterns in a spatially constant fashion, *(b)* normalizes the representation of local descriptors with respect to scale, *(c)* is more accurate compared to the coarse representation of Figure 27, with respect to the discrimination of the foreground and background arrangement. In the next subsection, this result is further evaluated through the clustering of similar histograms.

In order to demonstrate the generic framework formulation, the summarized histogram description is computed for the orientation feature. The next example illustrates scale-summarization over all scales. Despite changes in the scale of gradient observation, the derived representation remains smooth. In the graphs illustrated in Figure 28, each vertically tiled histogram was weighted with respect to the magnitude of scale-summarized gradient, in order to highlight structure-defining arrangements. In the graphs, the highest histogram value (1) corresponds to the darkest gray level (black) and the lowest (0) to the brightest (white). Other gray levels are linearly distributed in this interval. For the experiments eight orientation histogram bins were used. Furthermore, it is noted that the graph illustrated should be conceived as cylindrical, rather than a flat surface, since orientation histograms are cyclic. In the graphs, the vertical axis maps eight orientation values from 0 to π (from top to bottom).

Finally, the scale-summarization of color histograms can be also performed using the proposed framework if colorspace and color mixture function are determined. In general, any type of vector encoded feature may be represented, and scale-summarized using the same method, if the summation of two descriptors is defined.

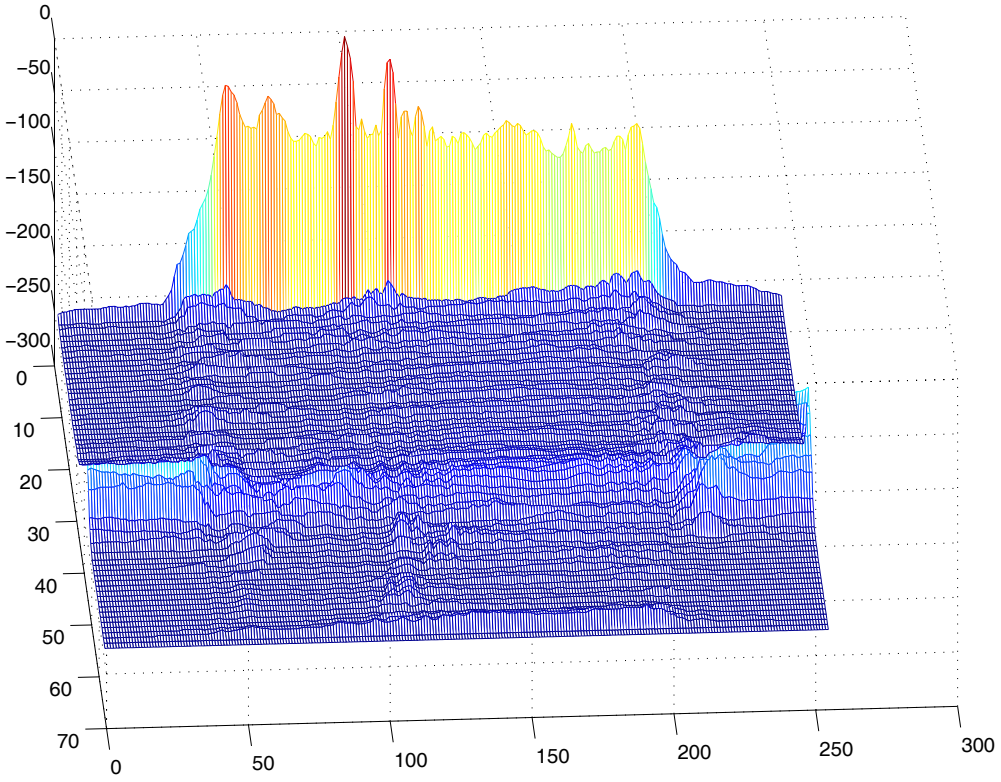

Figure 27: Scale-summarized and scale-focused local histograms of intensity computed along the marked scan line of the image presented in Figure 26 (see text).

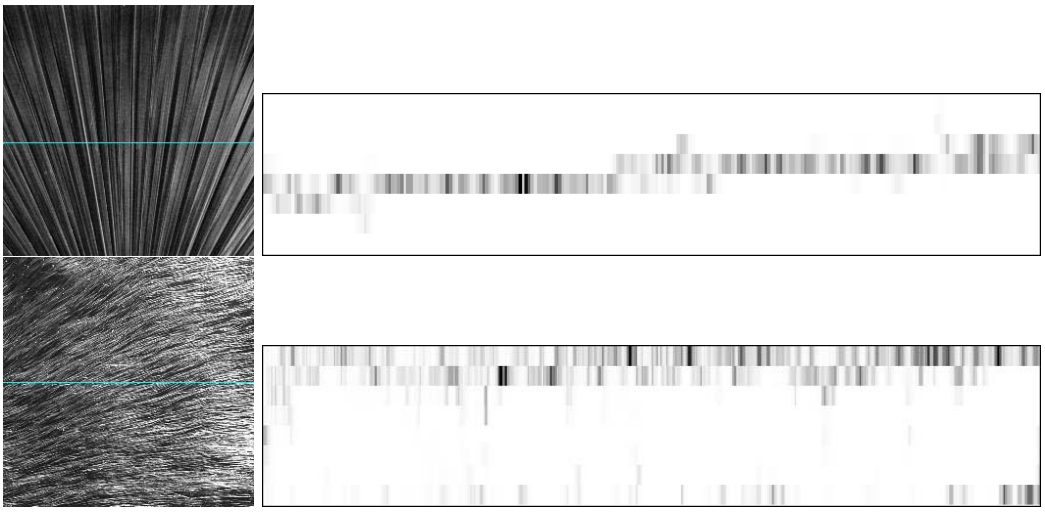

Figure 28: Scale-summarized histograms of orientation, computed along the scan lines shown in images on the left (see text).

Conclusions From the discussion in this subsection it can be concluded that the scale-summarization of local descriptors contributes to the:

- *dimensionality reduction* of the multiscale arrangement representation space, which preserves informationally and perceptually significant aspects of visual content, while reducing memory capacity requirements,
- *classification* of spatial arrangements of primitive features with respect to scale, as discussed in the previous chapter, and
- *normalization* of feature of spatial arrangements of primitive features it with respect to scale.

Regarding the dimensionality reducing property of scale-summarization, the SSR can be used in the computation of a few characteristic "snapshots" of this representation space, by summarizing local descriptors over certain scale ranges, as in the previous chapter. In addition, the method is computationally simpler than explicitly selecting an optimal histogram sampling aperture, and can also be implemented in parallel, as discussed in the previous chapter.

3.3 Description

In this section, the scale-summarized representation formulated in Section 3.2 is utilized in the extraction of image regions that exhibit constant primitive feature arrangement. This extraction is found useful in the description of image content.

In order to extract image regions of constant spatial arrangement of primitive features using local descriptors, similar ones are clustered. However, in order to estimate the similarity of local descriptors a metric is required and, thus, formulated. This clustering mechanism is utilized in the demonstration of the ability of the SSR, to scale-normalize local descriptors.

The remainder of this section is organized as follows: in Section 3.3.1, descriptional requirements are discussed. Next, in Section 3.3.2 the similarity of local descriptors is discussed and a dissimilarity metric is formulated. In Section 3.3.3 region extraction, through clustering of local descriptors, is demonstrated.

3.3.1 Requirements

In order to automatically identify image regions of constant spatial arrangement of primitive features, a method for their extraction has to be formulated. Thus, a formalization of this constancy is required that is based on the available representation for spatial arrangements of primitive features.

Given a representation which is based on local descriptors, a similarity metric is required, in order to estimate the visual resemblance of two spatial arrangements. Two issues to consider in this effort are: *(a)* the use of a representative, with respect to the type and dimensionality of information, similarity metric and *(b)* the perceptual relevance of the adopted metric.

Representativeness It is required that the dissimilarity estimation method should be compatible with the type and dimensionality of the represented information. A generic and commonly encountered approach to the problem of the dissimilarity estimation of two distributions is to compute the distance of the descriptions in vector state space. For example, by representing histograms as vectors of dimensionality equal to $bins$, their Euclidean distance would be $d = \sqrt{\sum_i^{bins}(h_1(i) - h_2(i))^2}$, with $i \in [1, ..., bins]$. Other distance measures may be used for this task, that may incorporate information about human perception or domain knowledge. For example, if the grouping of

color features is required a more perceptually relevant choice would be the utilization of distance in the *Lab* color space. See also [21], for the case of angular histograms.

Perceptual relevance Using the Euclidean distance for the estimation of arrangement dissimilarity, for all types of local descriptors, consists a generic approach. However, indications from vision [88] and other senses, such as hearing [114], or even temporal perception [98], point to other non-linear dissimilarity metrics. Besides the metric properties of a similarity assessment method, qualitative issues are probably even more significant. For example in Figure 27 the scale summarized local representations of intensity can be intuitively discriminated into those corresponding to the foreground entity and background, simply from the *structure* of the histogram.

In [90] several types of distances are reviewed with respect to their perceptual relevance, however the issue is currently open. Some reasons are:

- The perceived similarity of local distributions can be influenced by contextual factors, which are not represented in the local distribution (e.g. surrounding or global image content).
- Similarity estimation can temporally vary.
- Similarity estimation can be observer and task specific.

It is thus understood that *(a)* a similarity estimation that overlooks such factors would result in rather coarse, or even counter-intuitive dissimilarity estimations and *(b)* that the identification of visual entities solely based on local descriptor information is incomplete and approximate.

Conclusion From the discussion above it is concluded that the formulation of some perceptually relevant metric, for the dissimilarity estimation of local arrangements, is inhibited due to missing information. Thus, in order to demonstrate the enhancement of the description of spatial arrangements of primitive features, due to the SSR and independently of the dissimilarity metric, the following procedure is performed: the Euclidean distance is used for the dissimilarity estimation, although inapproximate. The beneficial effect of scale-normalization is demonstrated using this distance together with another, statistically unbiased, method. Thus, if knowledge is provided

that casts the refinement of the dissimilarity metric possible, then the results could be updated using the new metric.

In the next subsection, the Euclidean vector distance is used as a local descriptor dissimilarity metric. In addition, the gradient magnitude, based on this metric, is formulated in order to visualize the estimated spatial change of primitive feature arrangement. Subsequently, in Section 3.3.3 the estimated dissimilarity is used in the clustering of local descriptors.

3.3.2 Similarity estimation

In this subsection, the spatial change of spatial arrangements of primitive features is estimated, based on the Euclidean distance of local descriptors. In addition, the effect of the scale-normalization of local descriptors is presented.

Local descriptor gradient Consider the task of weakly segmenting an image with respect to pixel intensities, typically resulting in image regions of approximately constant brightness. In this case, the image gradient magnitude is a visualization of the local dissimilarity of image brightness and, also, an indication of which picture elements should be grouped together. Segmentation algorithms exploit local dissimilarity information in various ways, in order to extract the pursued regions. The underlying principle, is that neighboring and dissimilar descriptions signify a segment of the separating border in-between two adjacent segments.

Let $\mathcal{H}$ be a representation of the spatial arrangements of primitive features of an image, for which one local descriptor is associated with each image point. Such a representation can be an "image" for which a local histogram of some image feature is computed for each point, as well as the scale-summarized data structure $\mathcal{X}_{\mathcal{SSR}}$, defined in Equation (17). An analogous to image gradient magnitude indication of local dissimilarity, but for the case of the local descriptors of spatial arrangements of primitive features, is the magnitude of the local description gradient magnitude, defined as:

$$|\Delta\mathcal{H}| = \sqrt{\left(\frac{\partial\mathcal{H}}{\partial x}\right)^2 + \left(\frac{\partial\mathcal{H}}{\partial y}\right)^2} \tag{18}$$

which extends the notion of image gradient magnitude, for local descriptors that are represented in vector format.

Due to the scale-dependence of spatial arrangements of primitive features, arrangement disparities may occur at different scales. Thus, using histograms of constant aperture may not be adequate for capturing such disparities. In the following, some examples illustrate the local description gradient magnitude for feature histograms of various aperture sizes. Subsequently, the local description gradient magnitude for the scale-summarized local descriptor representation ($\mathcal{X}_{\mathcal{SSR}}$) is defined and utilized in the visualization of the effect of scale-normalization.

In the following experiment, the multiscale local descriptor representation $\mathcal{X}$ is computed for local histograms of intensity and orientation using seven aperture sizes, or scales. Each scale is derived from the exponential increase of the sampling aperture of local histograms. For each scale, the local description gradient magnitude is computed. Finally, the scale-normalization of local descriptors is demonstrated by visualizing the local description gradient magnitude for $\mathcal{X}_{\mathcal{SSR}}$. The first goal of the experiment is to exhibit that, due to the scale variation of spatial arrangements of primitive features, image regions of constant primitive feature arrangement cannot be captured using histograms of constant aperture. The second goal of the experiment is to indicate that the scale-normalization of local histograms is a useful tool for this capture.

In Figure 29, the local histogram gradient magnitudes for the seven scales of $\mathcal{X}$ and for the feature of image intensity are illustrated. In the images gray values are linearly mapped to values of $|\Delta\mathcal{H}|$. The order of images is from left to right and top to bottom. For the intensity histograms, 64 bins were used. In Figure 30 the local descriptor gradient magnitudes for orientation histograms are illustrated. The same scales and the same image were used as in the previous example, but for orientation histograms of 8 bins. The images are presented increasing in scale from left to right and top to bottom. In the images, gray values are linearly mapped to values of $|\Delta\mathcal{H}|$.

In Figure 31, the local descriptor gradient magnitude of intensity and orientation histograms and for scale-summarized local descriptors is illustrated. The intensity and orientation histograms were composed of 64 and 8 bins, respectively, and seven scales were used for $\mathcal{X}$. The weighting of scale-summarization was performed using the scale-selectors in Equations (8) and (7) for intensity and orientation histograms, respectively. In the images, gray values are linearly mapped to values of the local descriptor gradient magnitude. Through this example, the effect of scale-summarization is demonstrated to scale-normalize local descriptors. This effect is visualized by the

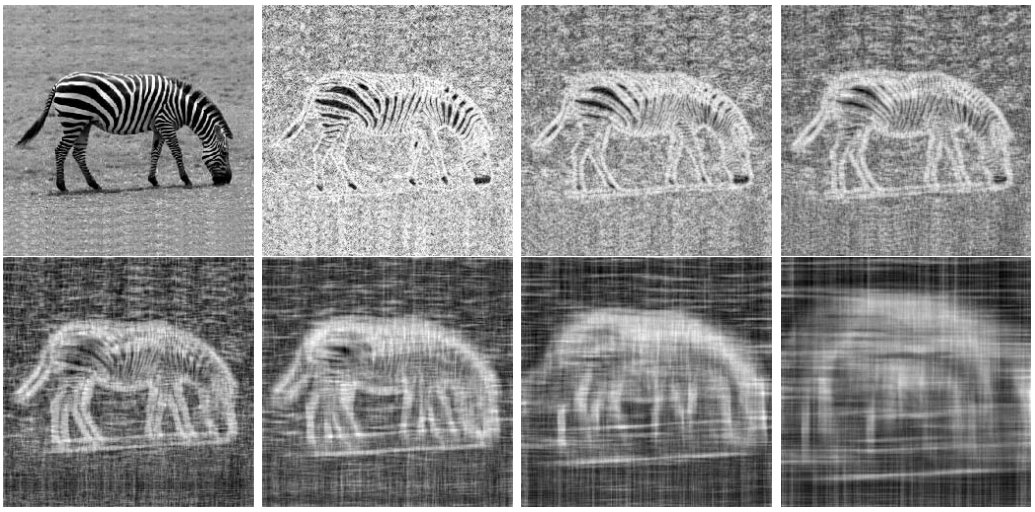

Figure 29: From left to right and top to bottom: Original image and local descriptor gradient magnitudes for image intensity histograms. Results are presented in increasing scale order and gray values are linearly mapped to values of $|\Delta\mathcal{H}|$.

low values of the local descriptor gradient magnitude for image regions that a exhibit a feature arrangement that varies solely at scale. The quantitatively constant description of such, scale-varying, patters simplifies their detection and extraction.

Finally, Figure 32 illustrates the local descriptor gradient magnitude for color histograms. In this figure, original images and the local descriptor gradient magnitude are illustrated, for color histograms using 64 bins. Eight image scales were used for the computation and gray values are linearly mapped to values of the local descriptor gradient magnitude. The results demonstrate the scale-normalization of local descriptors through the low values of the local descriptor gradient magnitude, for scale-varying spatial arrangements of color features.

Conclusion In this subsection, the scale-summarization of local descriptors of spatial arrangements of primitive features was demonstrated to scale-normalize such descriptors. The local descriptor gradient magnitude was utilized for the visualization of spatial arrangement dissimilarities. The

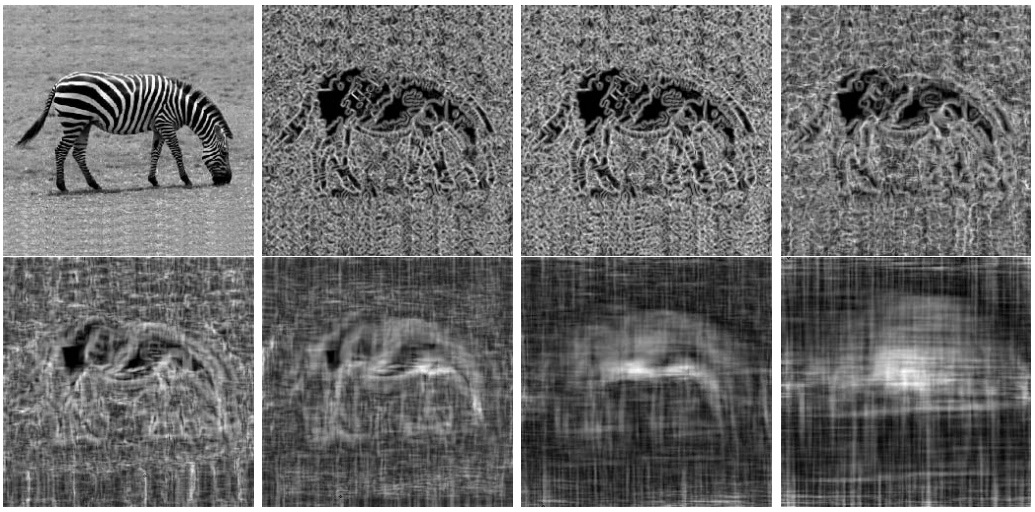

Figure 30: From left to right and top to bottom: Original image and local descriptor gradient magnitudes for orientation histograms. Results are presented in increasing scale order and gray values are linearly mapped to values of $|\Delta\mathcal{H}|$.

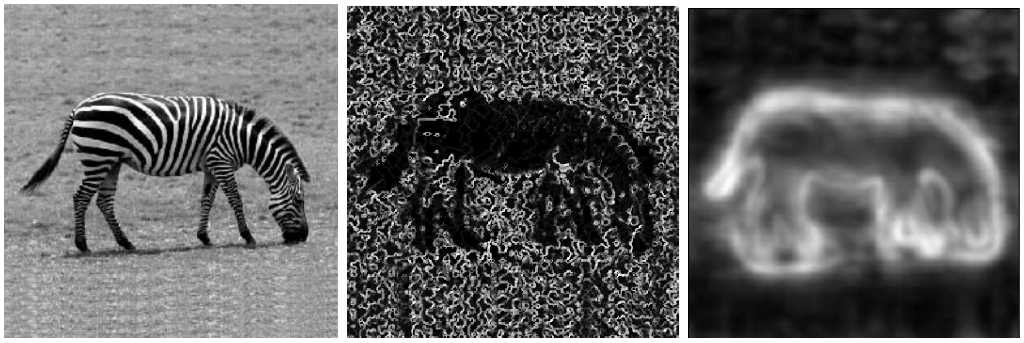

Figure 31: Original image (left) and scale-summarized local descriptor gradient magnitudes for orientation (middle) and intensity (right). Gray values are linearly mapped to values of the local descriptor gradient magnitude.

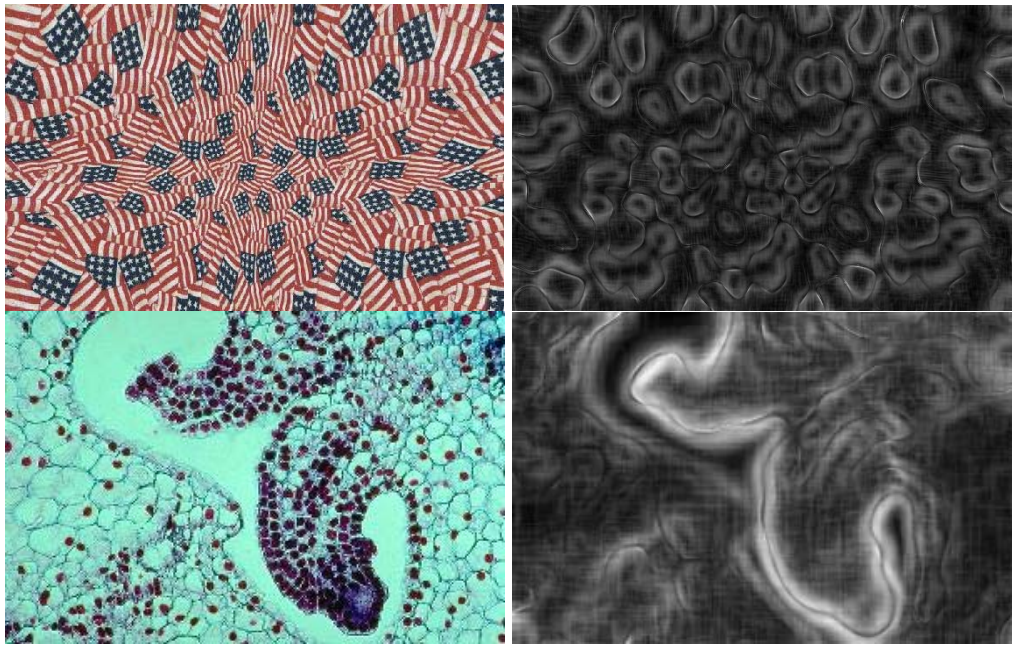

Figure 32: Original images (left column) and the local descriptor gradient magnitudes (right column), for scale-summarized color histograms. Gray values are linearly mapped to values of the local descriptor gradient magnitude.

benefit from this process is that a simple descriptor of the change of spatial arrangement of primitive features is obtained, which can be used in the identification of scale-varying feature arrangements, as shown in the next subsection. The advantage of this approach is that the result is obtained in a algorithmically simple and computationally inexpensive fashion, which can be parallely implemented.

3.3.3 Spatial grouping

In this subsection, the extraction of image regions that exhibit a constant spatial arrangement of primitive features, based on the clustering of local descriptors, is demonstrated. In order to cope with the scale-dependent nature of the arrangements and, thus, extract image regions corresponding even to scale-varying arrangements the scale-normalization of local descriptors is utilized.

Introduction The ability to identify image regions that exhibit constant spatial arrangement of primitive features yields a descriptive and perceptually relevant competence for the description of visual content. In content-based image description and retrieval, the image is often segmented by grouping pixels in object silhouettes, clusters of points, or point-sets. The surface- or object-based, otherwise *strong*, segmentation of the image in a perceptually relevant fashion additionally requires non-visual information and is also dependent on varying factors such as duration of observation, scene illumination, observer knowledge, and other. The difficulty of achieving a strong segmentation may be compromised by weak segmentation, where grouping is based on data-driven properties, which partitions the image in regions that are internally homogeneous according to some criterion [102]. In this subsection, the weak image segmentation, based on the extraction of image regions that exhibit constant spatial arrangements of primitive features, is demonstrated. The extracted regions are significant for the description of image content, not only in terms of raw visual entity identification, but also in terms of formation of higher level features. Such can be the boundary of a visual entity, the formation of regions of interest in an image, and others.

In this subsection, it is argued that the scale-normalization of local descriptors of spatial arrangements of primitive features is a useful pre-segmentation procedure, that contributes to the extraction of scale-varying patterns. The purpose of the presented experiments is to demonstrate that the performance

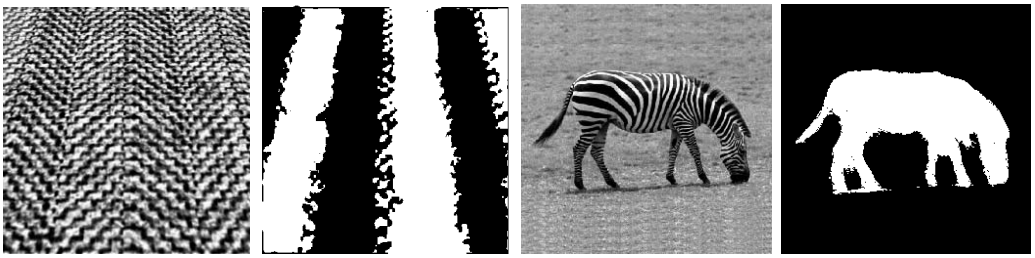

Figure 33: Clustering of scale summarized local histograms for orientation (left) and intensity (right) spatial arrangements.

of segmentation algorithms that operate based on the dissimilarity of local descriptors, can be enhanced through scale-normalization. Thus, in order to demonstrate the benefit of scale normalization, with respect to the task of image region extraction, a generic and statistically unbiased clustering algorithm is used, namely the k-means algorithm.

Experiments Image segmentation based on local descriptor clustering can be generalized for a variety of feature types, if some local descriptor dissimilarity metric is known. For example in Figure 33, the original images have been clustered with respect to the spatial arrangements of intensity and orientation, after computing the SSR histograms for all image scales. As observed in the images, the scale of feature observation varies. The result provides a weak segmentation of the image, composed of regions of qualitatively constant feature arrangement. Clustering was carried out using the k-means clustering algorithm without taking spatial layout into account. For the summarization, eight image scales were used. The same number of scales was also used for the rest of the experiments presented, in this subsection.

The benefit of scale-summarizing local histograms is demonstrated by the performance of the same operation for the case of intensity histograms. In Figure 34 the right original image of Figure 33 is clustered at several scales, without achieving the same accuracy of result, which is yielded by the SSR version of the process.

Figures 35 and 36 illustrate examples of image segmentation through local descriptor clustering for a variety of images. In all of the examples, intensity histograms of 32 bins were used. In the figures below, the results are arranged into columns. From left to right, the first column displays the original image.

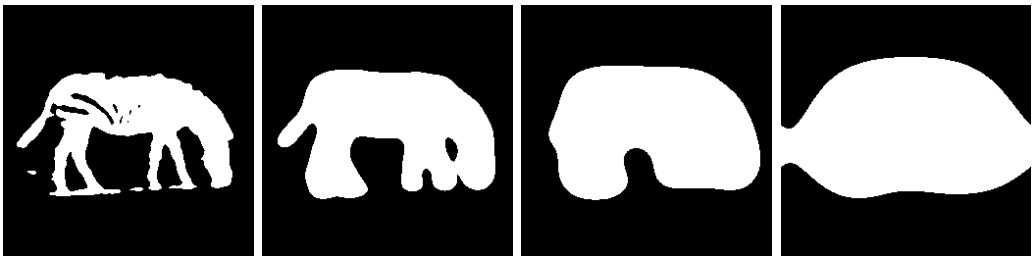

Figure 34: Clustering of local intensity histograms at several scales.

The second column shows the clustering result using histograms of a small sampling aperture of radius equal to 1 pixel. The third column illustrates the clustering result using histograms of a medium size, of radius equal to 6 pixels. The rightmost column, illustrates the clustering result for scale-summarized local descriptors, over all scales of $\mathcal{X}$. The radii of the sampling apertures of $\mathcal{X}$ obtained values that were logarithmically increasing over the interval of $[1, ..., 64]$ (measured in pixels).

Conclusion It is possible that the weak segmentation results presented above may be upgraded using other, more sophisticated, algorithmic schemes that take spatial relationships into account, if a formulation of local descriptor dissimilarity, such as the local descriptor gradient, is available. Using a grouping method such as region growing [2], graph partitioning [101, 30] or level-set segmentation [97, 96] based on histogram vector values, will result in the grouping of spatially-neighboring and similar descriptors into image regions of coherent feature arrangement. Such algorithms typically require an estimation of the dissimilarity between two neighboring local descriptors. This estimation can be provided from several dissimilarity metrics, such as the local descriptor gradient magnitude. The contribution that was highlighted in this subsection is the effect of scale-normalization of local descriptors. In order to support the claim that the scale-normalization of local descriptors can be utilized within a broad context of segmentation algorithms a generic and statistically unbiased clustering algorithm was used.

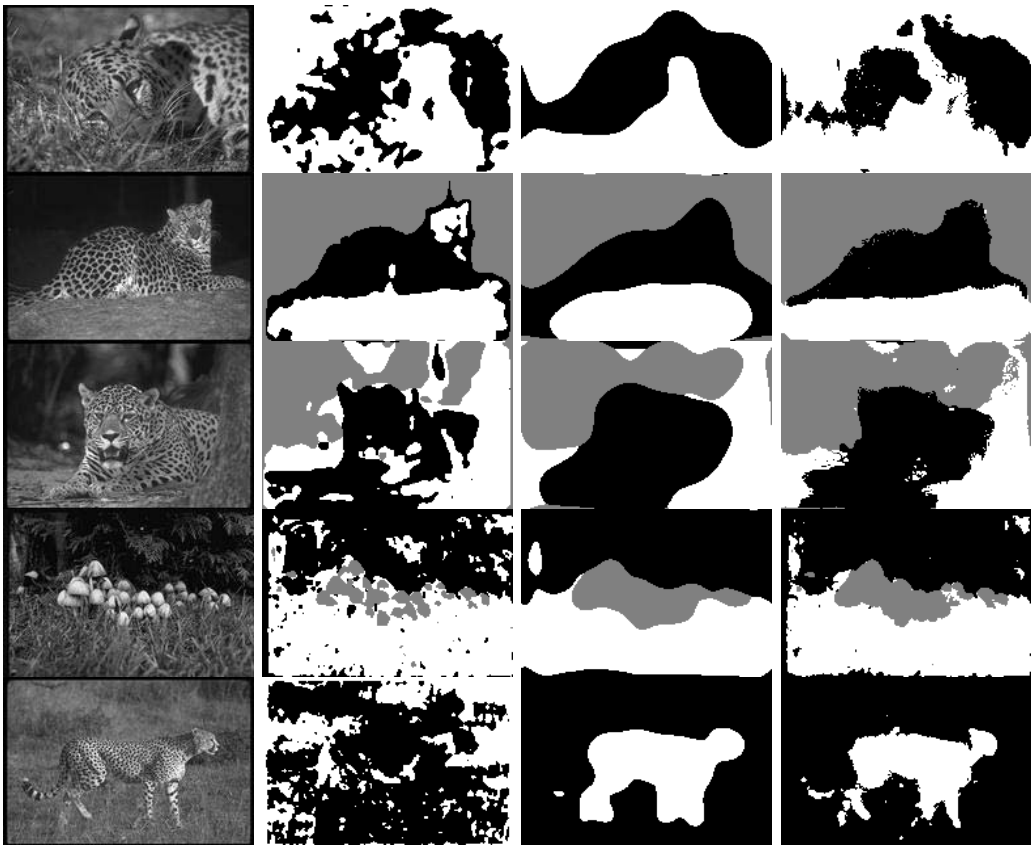

Figure 35: Clustering of local descriptors. Column order from left to right: *(a)* Original image and clusterings of *(b)* fine scale descriptors, *(c)* coarse scale descriptors, *(d)* scale-summarized descriptors.

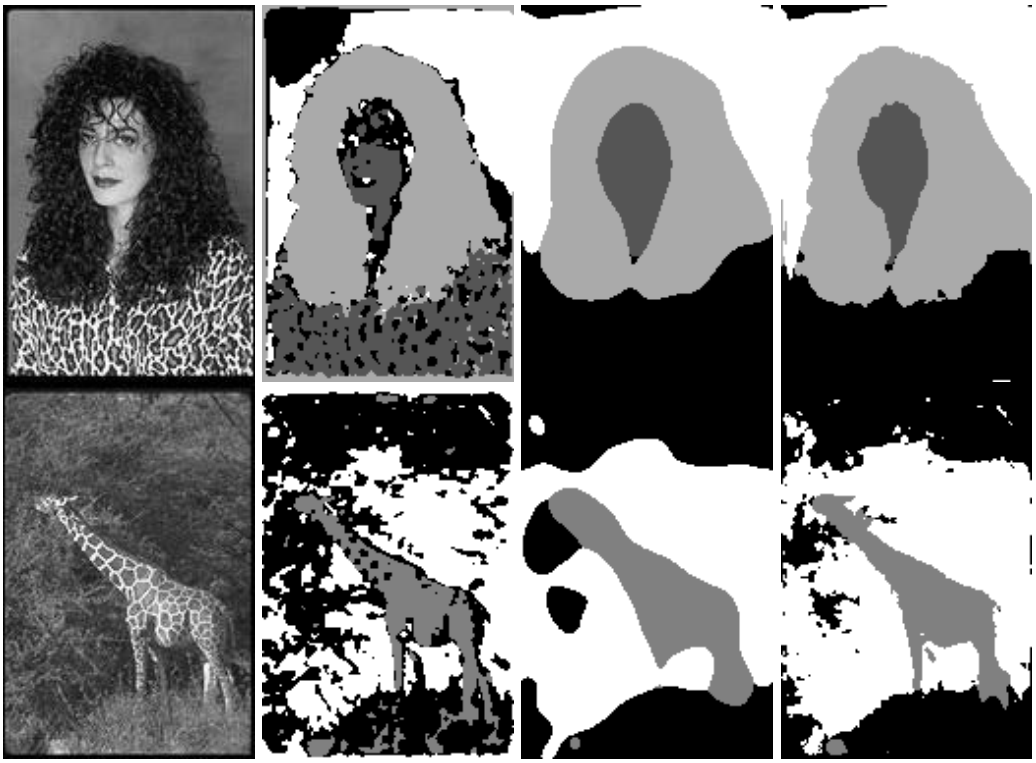

Figure 36: Clustering of local descriptors. Column order from left to right: *(a)* Original image and clusterings of *(b)* fine scale descriptors, *(c)* coarse scale descriptors, *(d)* scale-summarized descriptors.

3.4 Higher order descriptors

The description of spatial arrangements of primitive features can be enhanced using descriptors that capture significant visual aspects of the arrangements. Due to the fact that the proposed descriptors operate on the already computed description of the spatial arrangements of primitive features they are referred to as *higher order descriptors*. Using appropriate descriptors characteristic and perceptually significant visual properties of spatial arrangements of primitive features, which are not always representationally evident, can be extracted.

In this section, two classes of higher order descriptors are proposed. The first deals with the elementary statistical analysis of local histograms and the second with the identification of qualitative components of them. Emphasis is placed on the mapping of descriptor attributes onto image properties yielding, thus, a human comprehensible description of the spatial arrangement of primitive features in the image. The image features that are extracted using the methods described in this subsection, are utilized in Section 3.5 for the formulation of visual queries and for the extraction of perceptual groups in Chapter 4.

3.4.1 Statistical descriptors

Some commonly used statistical characteristics of histogram distributions include the expected value, mean, median, variance, and information entropy. By replacing the local description with such a representative, the description is condensed, retaining only a portion $(1/bins)$ of the original information.

Typical applications of the mean and median descriptors are encountered in image filtering. Replacing the histogram description with its mean or median results in a noise-suppressed result. The replacement of a local histogram by its expected value is equivalent to filtering the image with a constant smoothing kernel of equal shape and size with that of the sampling, for both color and intensity images. Furthermore, if histogram samples are Gaussianly weighted with respect to their distance from the center of the sampling area then Gaussian smoothing is implemented. By varying the sampling aperture of the histogram with respect to local structure scale the filtering may adapt to image structure. The SSR may be used for this purpose in order to overcome computational difficulties of explicit scale selection. Examples of this process appear in the previous chapter (see Section 2.2.4),

for both color and grayscale images. However, the expected value of the distribution is not compatible with human perception in all types of features, as discussed at the end of this subsection.

Descriptors such as the variance or information entropy (given by $-P \log P$, where P is the probability of a particular value) of the histogram encode certain aspects of the histogram which can be used to describe spatial arrangements of primitive features. Histograms centered around a single value exhibit small variance. Similarly, the entropy of a histogram increases with the "spread" of the histogram. The example shown in Figure 37 demonstrates the detection of highly ordered scale-summarized orientation histograms, which are characterized by low information entropy. The middle image shows the scale-summarization of gradient magnitude for all image scales. The right image shows the entropy of scale-summarized orientation histograms. As opposed to the scale-summarized gradient magnitude image, in the entropy image, regions exhibiting prominent orientational order stand out. The effect is characteristically observed in the image region that corresponds to the building, due to the presence of parallelism in the local image structure. In the scale-summarized gradient magnitude image parallel line segments that correspond to the building are not clearly observed due to the low gradient magnitude value. In the brightness-inverted entropy image the same region is highlighted, due to the decreased entropy of local orientation histograms. To generate the results eight image scales were summarized both for the gradient and entropy case. For the scale-summarization the scale selector of Equation 7 was used. The orientation histograms were composed of 8 bins. In addition, the expected value of the orientation histogram was computed for each scale-summarized histogram and encoded with a color hue in the figures. The reason that the expected value of the orientation histogram is perceptually relevant, in this case, is that it is illustrated for histograms centered around one specific value (low information entropy descriptions).

The low information entropy indicates the concentration of orientation values around some histogram value. Mapped onto image properties the lack of entropy can be interpreted as parallelism, since all orientation components of the arrangement shall share the same direction. This notion of parallelism can be extended to non-straight line segments as well, as illustrated in Figure 38. In this example the same histogram and scale-summarization parameters were used an in Figure 37 (right image). Regions may be extracted by thresholding of the result and grouping the resulting image components. A demonstration of this process appears in Section 3.4.2.

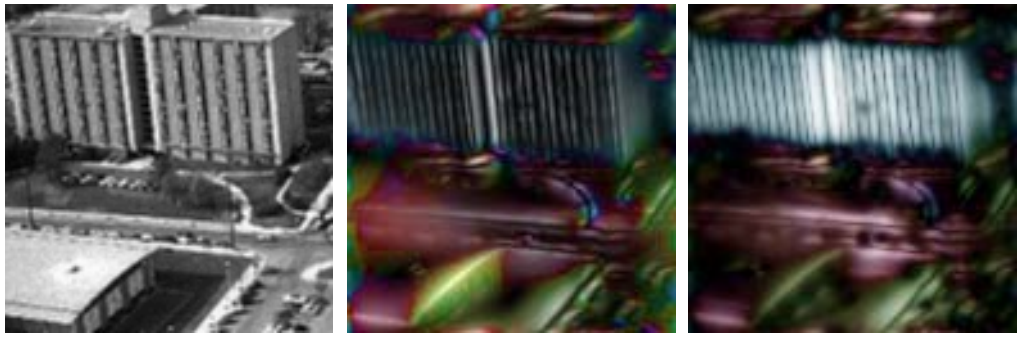

Figure 37: Left to right: Original image, scale-summarized gradient, and brightness-inverted entropy of scale-summarized orientation histograms.

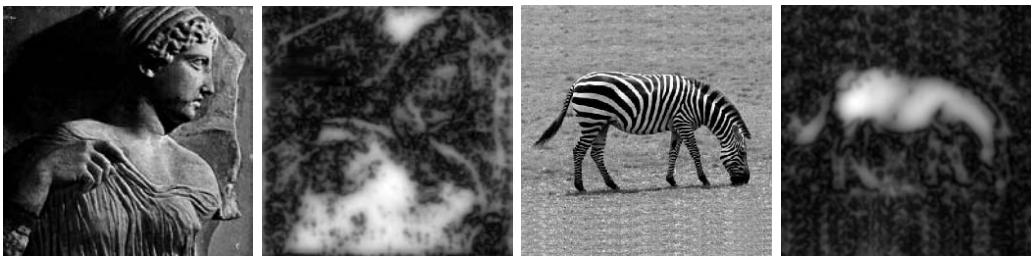

Figure 38: Original images and brightness-inverted entropy of scale-summarized orientation histograms.

It should be noted that if the goal is to encode the "parallelism" of the arrangement the information entropy of a histogram should be preferred over its variance. Since the variance is computed with respect to the expected value of the histogram, the extracted description would be of poor descriptive power, if the expected value is not a representative descriptor, such as in the case of orientation.

3.4.2 Qualitative descriptors

In this subsection, attributes of local descriptors are utilized in the identification of characteristic image properties. Such attributes are related to the "structure" of the histogram. In particular, local maxima of the histogram are used to identify the principle components of the histogram distribution and, thus, of the feature arrangement in the image.

A qualitative description of a feature arrangement can be obtained based on the structure of its local histogram. The next example (Figure 39) illustrates the descriptional significance of the principal components of a histogram. Corners, crosses and junctions can be identified in images from the number of principal components, of the local orientation histograms. The formed angles can be estimated as the relative angle of these components, given by $min(|pc_1 - pc_2|, \pi - |pc_1 - pc_2|)$, where $pc_{i=1,2}$ is the angle corresponding to the principal component. In the images local orientation histograms with 8 bins were used. Multiple histograms (8) of varying sampling aperture were centered at each point and scale-summarized, using Equation (7) for the scale-selector. In addition, the scale-summarization was focused on fine scales in order to capture texture-like arrangements (instead of coarse-scale structures). In the images, pixels are color coded with respect to the cardinality of principal orientation components at each point. The internal image legend illustrates the correlation of color with the cardinality of principal components. The background color is associated with the zero value. The other colors are associated with values $1, 2, ...$ in order of appearance in the internal legend (from "cold" to "warm" color, or from dark to bright if viewed in gray scale).

Solely the dominant principle components can be used to describe certain spatial arrangements of primitive features. Figure 40, illustrates the dominant orientation in color code. To generate the image, the same histogram and scale-summarization parameters were used as in the example of Figure 39. Especially in the case of orientation the use of the expected value

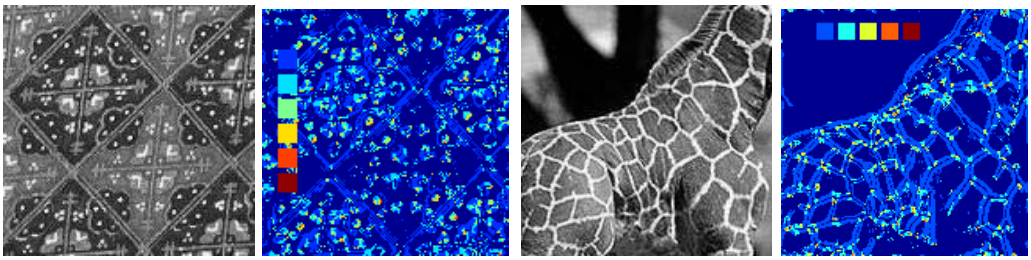

Figure 39: Original images and cardinality of principal orientation components at each point. Internal legend: correlation of color with the cardinality of principal components ($background = 0$).

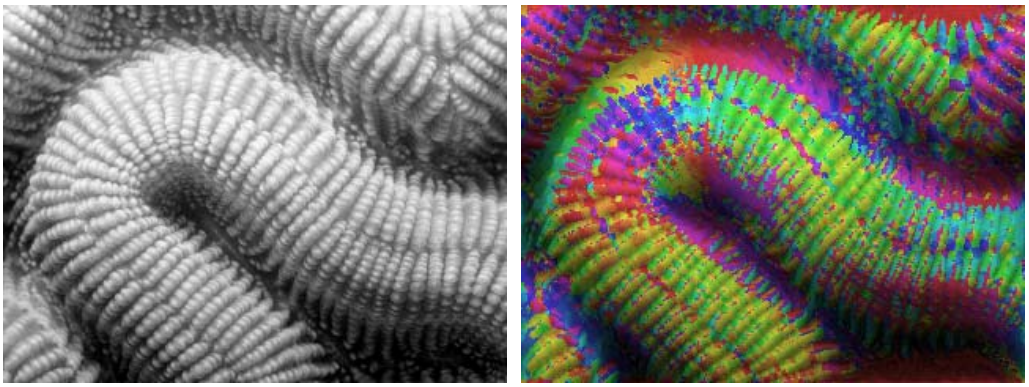

Figure 40: An image and its dominant orientation at each point, color coded.

of the histogram would be a rather inappropriate choice. Computing this value for histograms that exhibit two principal orientation components does not correlate with the perception of local orientation.

Figure 41 illustrates local scale-summarized descriptors that exhibit one principal orientation component. The same histogram and scale-summarization parameters were used as in the example of Figure 39. In order to group spatially neighboring description elements, the connected-components algorithm was used. Local descriptors with elements were grouped with respect to distance and expectation value of the orientation histogram. The grouping parameters of the grouping criterion were arranged so that immediately neighboring descriptors (considering 8 neighbors for each image point) were

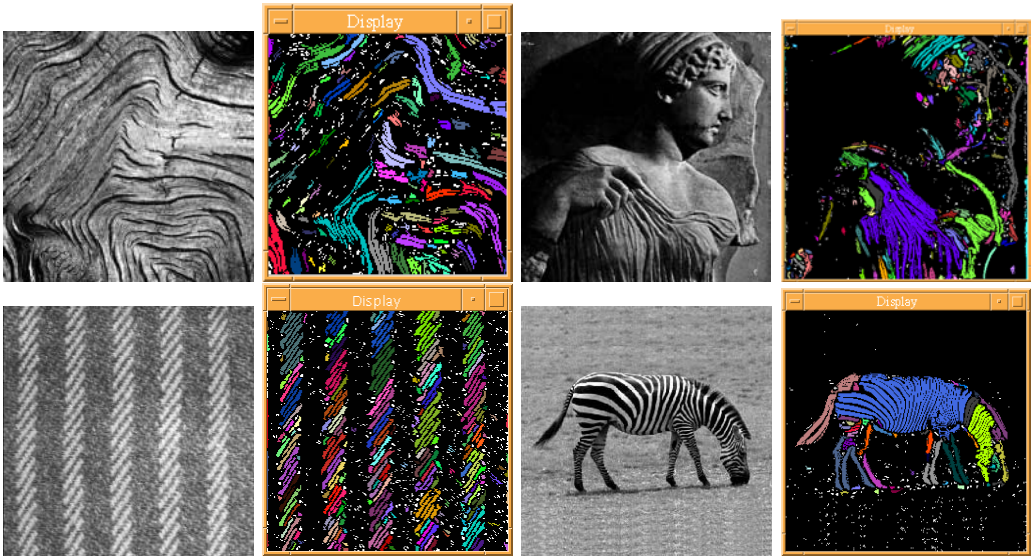

Figure 41: Extraction of image curves that parallely evolve.

grouped, if the histogram expectation value was not greater than two histogram bins. In the presented images, a different color is used to indicate each group, while non-grouped elements are colored white. The result of the computation yields a description of curves that parallely evolve in the image.

3.5 Image retrieval based on descriptions of spatial arrangements of primitive features

In this section, the application of the derived methods for the description of spatial arrangements of primitive features is demonstrated in the task of content-based image retrieval. The representation and matching of spatial arrangements of primitive features in images is carried out using arrangement descriptors. The mapping of arrangement descriptor attributes onto image properties, as described in the previous section, contributes to the comprehensibility of visual queries by end users.

The spatial arrangement of visual features constitutes a fundamental and intrinsic component of visual information. From a phenomenological point of view, regions of constant spatial arrangement of primitive features define and attribute characteristic image regions. Thus, the adoption of perceptu-

ally descriptive attributes is critical, regarding query formulation and result appreciation. The descriptional value of scale-inspecific querying is demonstrated in the following example of visual information retrieval, based on some of the distribution descriptors mentioned in the previous section of this chapter:

The spatial arrangement of primitive features defined by the white cluster of pixels of the rightmost image of Figure 33 was used as a query sample for the feature-based retrieval of image regions. Using scale summarization, the following description was scale-invariantly extracted from the local distribution descriptions of the sample: *(a)* exactly one principle orientation distribution component, *(b)* exactly two principle intensity distribution components, their approximate values defined by the sample, *(c)* intensity histogram of the arrangement. For all image pixels of the query set, a metric was formulated, that expresses the similarity of the feature arrangement at each pixel, with the sample. The similarity metric uses the description criteria (a) and (b) as logical operators and criterion (c) as a metric operator, its value given by the Euclidean distance of intensity histograms. The metric was given by the multiplication of the three operators. Figure 42, illustrates the results. In the presented images, gray values are linearly mapped to the values of the similarity metric for all images.

The presented images are not in any particular order. Instead image intensity in the "similarity response" images is used to represent feature arrangement similarity. It is observed that retrieval results match the given description and, in some cases, contextually similar objects (zebras) are highlighted. Most insightful though are counter-intuitive results, that are also presented. The image located in the fifth row at the third column matches the given description, but the property of orientational order has been overlooked. The retrieved pattern exhibits almost no variance of orientation distribution, otherwise interpreted as highly organized structure, just as in the previous example (Figure 37). The consideration of orientational organization is required for the understanding of the counter-intuitiveness of the result. Similarly, in the tiger image (third row, first column), color information that would trivially distinguish a black-white pattern from a black-orange one was not taken into account. In the example illustrating a human (third row, third column), the retrieved pattern appears to be visually similar to the sample, however contextual reasons (object recognition, semantics) discriminate that image from the zebra one. From such examples, it is understood that the set of features that constitute visual impression is often not

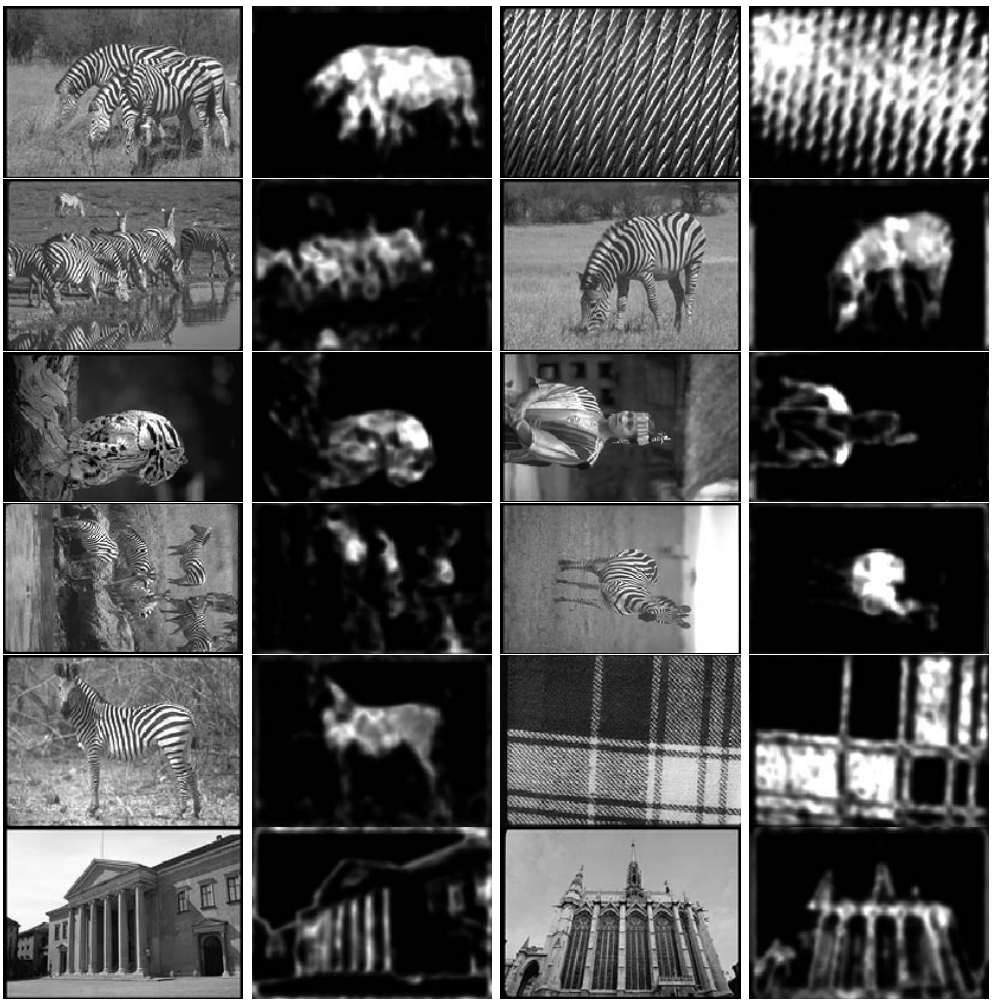

Figure 42: Intensity mapped similarity response for a given pattern.

clearly understood. In many cases, the identity of features that are relevant to the visual entities that are intended to be retrieved from images may not be prominently noticed, such as in the present case regarding the directional components of the retrieved patterns. Analogous query formulations may be devised for the distribution descriptors described in this chapter, as well as for others found in literature. It is of considerable importance though which of them appeal to perception and most important which are relevant towards the goal of retrieving intended visual content.

4 Perceptual Organization

This chapter is related to the topic of *perceptual organization*, referring to the grouping of pieces of visual information into larger units of perceived objects and their interrelations. Such *perceptual groups*, which are typically composed of structural visual features, constitute dominating components of perceived visual content. Thus, their detection, description, and utilization in the task of content-based visual information browsing and retrieval is crucial.

The first section of this chapter briefly reviews known laws of perceptual organization. In this section, emphasis is placed at the review of certain properties of the perceptual groups that are related to the methods that are presented in the second and third sections of this chapter. Furthermore, certain other aspects of perceptual organization that are considered relevant to the description of visual content are presented.

In the next section, the issue of grouping visual elements into perceptually relevant entities is considered and an approach towards the perceptual grouping of line segments is presented. The approach concerns the grouping of line segments that are parallel in the three-dimensional world and occur as converging in the two-dimensional image matrix. The resulting perceptual groups are subsequently utilized in the content-based retrieval and classification of images.

In the third section, the discussion is focused on the study of contours, motivated by their characteristic relevance to object recognition and informationally rich nature. In particular, a perceptually relevant approach towards the description of object boundaries is proposed, based on the hierarchical and piecewise parsing of shapes into primitives (pieces). The approach exploits the salience of high curvature boundary points towards their segmentation. This segmentation is subsequently demonstrated to be of use in the description and matching of silhouette boundaries.

4.1 Introduction

The visual perception of a scene is populated with large-scale objects instead of a confetti of primitive features such as the ones discussed in Chapter 2. The LGN-encoded retinal stimulus, forwarded to the V1 area, follows different *visual pathways* [61] that process this encoding, leading to the perception of stereo-vision, color, form, and motion. The individual processing of these components seems counter-intuitive, since visual perception of objects ap-

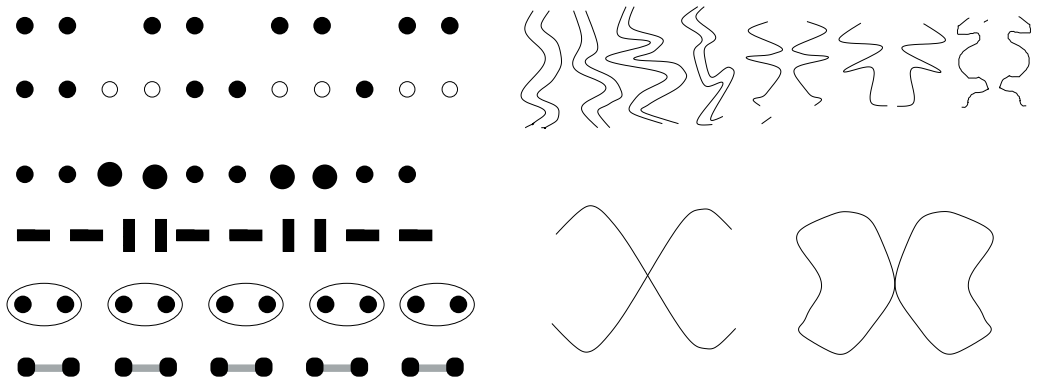

Figure 43: Some typical examples of perceptual grouping (see text).

pears to be completely integrated. However, observers are able to distinctly refer to color, form, and motion as orthogonal visual properties.

The concept of perceptual organization originated from Gestalt psychologists early in the 20^{th} century [110], indicating multiple perceptual grouping criteria that are based on feature appearance and spatial arrangement. According to these criteria, the spatial arrangement of motion, texture, and color, determines the perception of distinct visual entities. In addition, context related factors such as familiarity of arrangement, figure / foreground discrimination, and prior knowledge are also observed to affect perceptual grouping. Traditional laws of *perceptional grouping* that describe such organizing behaviors originate from Gestalt psychology, while novel ones have been more recently formulated by vision scientists [76, 78]. In Figure 43, some examples of perceptual grouping of static visual content are presented. In the left part of the figure and from top to bottom, the first four rows of structures exhibit the perceptual laws of *proximity*, *color similarity*, *size similarity*, and *similarity of orientation*. The last two rows exhibit the rules of *common region* and *element connectedness*. In the right part of the figure, perceptual groupings originating from *parallelism* and *symmetry* (top row left and right, respectively) as well as *continuity* and *closure* (bottom row left and right, respectively) are demonstrated.

In the next paragraph, certain properties of perceptual groups originating from the grouping of structural, or gradient derived, image elements are reviewed.

Boundary-related perceptual grouping The grouping of structural, otherwise gradient-derived, elements contributes to the formation of figures, contours, and three-dimensional volumes. Primitive, boundary-related structure physiologically originates from spatio-temporal change detection. However, in the context of this dissertation only spatial changes are considered and discussed. In addition, illusory edges are observed to be characteristic elements of contour perception. Large-scale structure can emerge from the proximal, collinear, and / or parallel spatial arrangement of these boundary-related features.

Another important aspect of perceptual organization in the segmentation of single structural image elements into parts is referred to as *parsing* [77]. Parsing appears to be an influential process of perceptual organization because it determines what subregions of a perceptual group are perceived together. The examination of regions at which the division seems natural are contour regions of deep concavities: points at which the contour undergoes a sharp bend towards the interior of the region [43].

Approaches towards the perceptual grouping of structural elements can be classified into local and global, as in the case of region based grouping. Local or bottom-up approaches target at the grouping of adjacent or loosely neighboring structure elements that exhibit specific arrangement properties, such as collinearity, good continuation etc. [62, 48, 63]. Top-down approaches target at the confirmation of some model or distribution through supporting evidence in the image [94, 13]. For a detailed overview of perceptual grouping in Computer Vision see [93].

Chapter outline In the remainder of this chapter, two methods for the description of structural elements are proposed. The first concerns the perceptual organization of line segments. The novel element in the proposed approach is that the perspective convergence of line segments is taken into account. The resulting *linearly perspective* groups are subsequently utilized in the content-based retrieval and classification of images. In addition, the method is extended for illusory line segments, formed by the collinear arrangement of boundary-related visual features. The second method is related to the description and retrieval of elements of a class of, already organized, perceptual groups. In particular, the class of silhouette boundaries is considered. In Section 4.3 of this chapter, a method for the perceptually relevant description and matching of such boundaries are proposed, based on the

detection of their salient points.

4.2 Perceptual grouping of line segments

In this subsection, a method of detecting perceptual groups originating from the parallel, in the 3D world, spatial arrangement of line segments, is proposed. Henceforth, this convergence of line segments to a vanishing point, will be referred to as perspective convergence. Utilizing the perspectively converging appearance of such line segments, the detected perceptual groups may be utilized as depth cues. With respect to the task of content-base image retrieval, the component of visual content that arises from the perceptual organization of line segments is utilized as an image similarity cue.

Taking account of perspective convergence yields a more complex problem than several treatments of it that can be found in the literature [62, 48, 63, 94, 13, 93]. In the description of the perceptual grouping method below, the standard issues of grouping suitability, digital image noise, and computational requirements are raised.

The proposed approach adopts a *hypotheses formulation and justification* model-matching strategy. Evaluation of grouping hypotheses is based on deviations of the image data from the hypothesis. The issue of computational cost which arises from the exhaustive search for an optimal grouping ($O(2^N)$) is tackled with the adoption of a commonly encountered perceptual bias. This bias indicates the increased perceptual significance of coarse-scale, contrast-salient structure elements (see Section 2.1.2).

4.2.1 Grouping Method

The algorithm presented in this paragraph consists of three steps, which are sequentially described in the next paragraphs. First, structural tokens (line segments) are selected from the image. Second, a set of hypotheses is formed, based on the selected tokens. Finally, the members of the set of hypotheses are validated or rejected, with respect to supporting or contradicting evidence found in the image. The result of the process is the extraction of perceptual groups that are composed of line segments.

Token Selection Initially, points of salient gradient magnitude (edgels) are selected from the image, using the Canny edge detector [19] (including hysterisis thresholding), and successively line segments are extracted utilizing

the Hough transform [45]. The set of these line segments will be henceforth denoted as $\mathcal{LS}$. In order to avoid detecting edgels originating from image noise the image is initially smoothed. In the effort to retain structure during this process, the SSR smoothing that was introduced in the second chapter is utilized for this purpose. In particular, image smoothings originating from the range of fine scales are considered. In all of the examples below, a linear Scale-Space of 8 image scales was taken into account. The range of fine scales, which were summarized to provide the smoothing, consisted of the three first scales. For the determination of scales, the logarithmic scale parameter τ was exponentially increased as described in Chapter 2. Next, a subset of $\mathcal{LS}$, containing the ones of dominant size is formed. (e.g. a subset with a cardinality of 10% of the original groups, containing the longest elements). The resulting subset of line segments consists of contrast-salient line segments that exist in the image, since weak segments were filtered-out at the edge detection stage.

In summary, the parameters that are used in this step of the algorithm are: *(a)* the range of scales used for the SSR smoothing (described above), *(b)* the higher and lower intensity thresholds utilized by the edge tracking process of Canny edge detector (values 0 and 255 were used for this purpose), *(c)* the resolution of the ρ, θ Hough transformation matrix (a 360×360 matrix was used), and *(d)* the percentage of the line segments which are characterized as size dominant (the value of 10% was used, in the examples below).

Hypotheses formulation From the resulting set of line segments, a number of hypotheses are formulated through the selection of line segment triplets that approximately converge to a single point, or infinity. The process results in a subset of all possible triplets (the hypotheses) whose convergence could be the result of the perspective observation of parallel, in the 3D environment, line segments.

In order to form the set of triplets from which the hypotheses will be formed, the set of candidate triplets is reduced by rejecting triplets that include: *(a)* intersecting line segment couples, *(b)* line segment couples in which: the extension of either one of the two line segments, results in an intersection extension with the other segment (without extending the latter segment), and *(c)* triplets in which the same line was taken more than one times into account (e.g. (l_1, l_2, l_1)). The members of the set of remaining triplets, let $\mathcal{R}$, are evaluated as to their convergence through the following

procedure:

Let (l_1, l_2, l_3) a line segment triplet, member of the set $\mathcal{R}$. Each possible line-segment couple ($\{l_1, l_2\}$,$\{l_2, l_3\}$,$\{l_3, l_1\}$) of the triplet is considered as a "weak-hypothesis". The *mismatch error* of the third segment with respect to this weak-hypothesis is quantified (see below for definition of mismatch error). An arrangement that exhibit a small mismatch error (with respect to a certain threshold, see below) is considered as "non-accidental" and assumed as to be composed of parallel, in the 3D world, line segments. Subsequently, a hypothesis is formed, based on the (approximate) convergence of the three line segments. If more than one hypothesis exhibit a small error then the one with the least is selected. The definition of the mismatch error and description of the process performed for its computation are presented below.

The mismatch error of the third line segment of a triplet, with respect to the weak hypothesis that is formed by the other two members of the same triplet, is defined as: the deviation (error) from the ideal case, where the three line segments are parallel in the 3D world and due to their perspective observation are apparently converging at a single point (the vanishing point).

Let $\{l_1, l_2\}$ be a non-intersecting (even if one is extended) pair of line segments which form a *weak-hypothesis* of perspective convergence. The extensions of these segments meet at vanishing point A, as illustrated in Figure 44A (see below, for the case that the two segments are parallel). The mismatch error of l_3 to this weak-hypothesis is quantified as the angle of the rotation of l_3 around its midpoint, so that its extension will pass through A (in the figure, when A and A' coincide). The estimation of the angle formed between two line segments is computed from the formation of an angle by three points (p_1, p_c, p_2) and by use of the vector dot product as (see Figure 44B):

$$\arccos\left(\frac{p_1.x \cdot p_2.x + p_1.y \cdot p_2.y}{\|p_1 p_c\| \cdot \|p_2 p_c\|}\right) \tag{19}$$

In the equation above, $p_1.x, p_2.x, p_c.x$ denote the horizontal coordinates (in the image) of points p_1, p_2, p_c, respectively. Similarly, $p_1.y, p_2.y, p_c.y$ denote the vertical coordinates of points p_1, p_2, p_c, respectively. As shown in Figure 44B, the computation of the described angle is always performed using the line segment endings, that reside farthest from the vanishing point, in order to minimize pixel discretization error. In the special case that l_1 and l_2 are parallel the vanishing point is considered to lie at infinity and the mismatch error is defined as: the relative angle of l_3 with l_1 or l_2. In order to

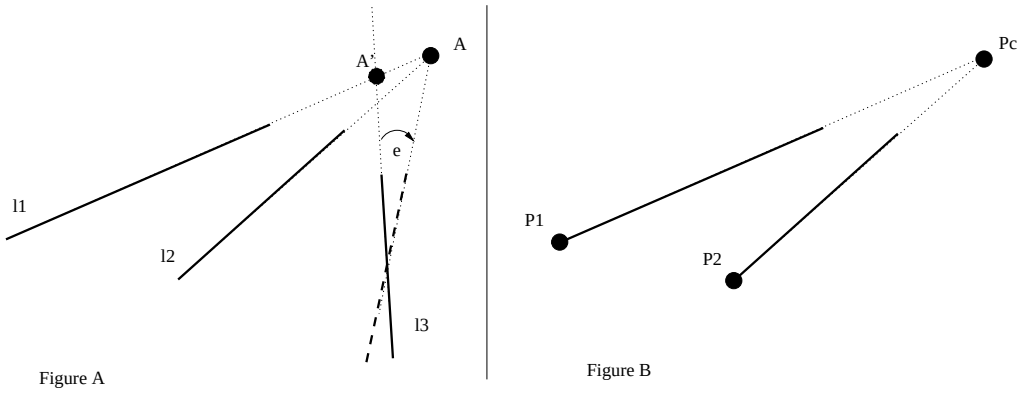

Figure 44: *(A)* Hypothesis formulation and computation of the mismatch error for l_3, as angle e. *(B)* Angle computation (see text).

compute the mismatch error for this case, l_3 is translated so that the midpoint of l_3 coincides with the midpoint of l_1. Then the mismatch error is given by the least angle formed by the intersection of the two line segments.

Finally, the set of hypotheses exhibiting a small error, let $\mathcal{LE}$, with respect to some threshold, are passed to the next step of the algorithm. The threshold, let et, that was used in the experiments presented below was 6 degrees, as previously described.

Group Extraction Here, each member of $\mathcal{LE}$ is regarded as a grouping hypothesis, which is either accepted or rejected, based on supporting or contradicting evidence that is found in the image.

For each hypothesis $q \in \mathcal{LE}$, members of the set $\mathcal{LS}$ are correlated to q. This correlation is performed by computing the mismatch error of each $ls \in \mathcal{LS}$ with respect to q. In this case, the mismatch error is computed for ls and the two line segments that formed the weak hypothesis, based on which q was formed. The ones that exhibit a mismatch error smaller than et, are grouped together. Thus, for each line segment that is finally grouped a mismatch error is computed. The mean value of these errors for all segments is also computed and henceforth denoted as Er. It is observed that the resulting groups of line segments may have common members.

In order to discriminate and finally obtain dominant perceptual groups of an image, an optimal grouping is selected with respect to a suitability

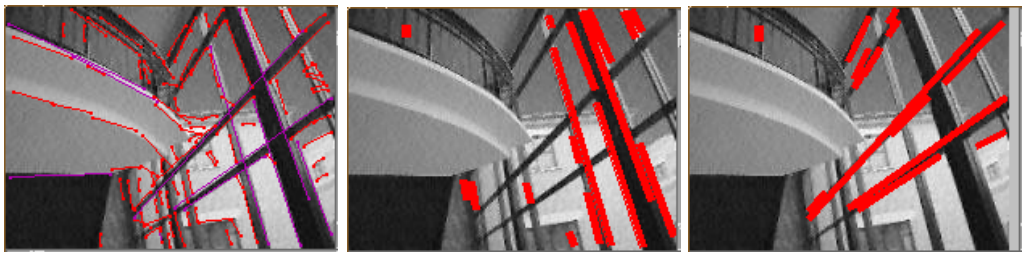

Figure 45: Iterative extraction of perceptual groups from an image.

function (see below for its definition). Once the group G that maximizes the suitability function has been detected, groups that have common members with G are removed from $\mathcal{LE}$. Then, the next group that maximizes the suitability function is extracted. The process is repeated until all independent groupings are extracted. Figure 45 illustrates the iterative extraction of two perceptual groups from an image. The first image illustrates the original image, with line detected segments superimposed. The next two present the two perceptual groups extracted, in the same order and from left to right.

The grouping suitability quantification is given by Equation (20), which encodes independent grouping criteria as a product and emphasizes the existence of dominant segments. In (20) Er is the mean group angle error, D the set of dominant segments participating in the grouping, E the set of their edgels, T the set of segments participating in the groups, and Tol an angle error tolerance. Maximization of the suitability function will ultimately group all segments complying with the, finally dominating, hypothesis, favoring large groups that include salient segments. In order to find the group that maximizes $\mathcal{S}$, each group is addressed and the value of $\mathcal{S}$ for that group is computed. The group that yields the maximum value is then selected.

$$\mathcal{S} = (Tol - Er) \cdot (1 + card(D))^2 \cdot card(E) \cdot card(T) \tag{20}$$

Results and discussion Some indicative results of the grouping procedure are presented in Figure 46. The left two images illustrate typical results from perceptual groups originating from surfaces oriented in depth. The third from the left illustrates the case, where the vanishing point lies at infinity.

Figure 46: Grouping of projectively converging line segments.

The rightmost image illustrates a false positive result[9], in the sense that the horizontal line does not appear to belong to the group. The crucialness of cue integration is noted, since stereo, depth, or knowledge about the environment could have canceled out the false positive result, if exploited. Such false positive results originate from the fact that the attempt to retrieve 3D information from converging lines is ill-posed, due to perspective metamers.

The next example proposes an extension of the discussed grouping method for subjective lines. Salient patterns of collinearly arranged gradient-related features in images are perceived as illusory lines, which perspectively converge. In the examples shown in Figure 47 the same grouping method was applied but instead of edgel detection, corner detection was performed. The corner detection process was carried out by scale-summarizing the response of the Harris corner detector [41]. The Hough transform which was earlier used for line detection was applied directly on the scale-summarized corner response image. Similarly the response of other point-like feature detectors may be used for the detection of linearly perspective feature arrangements in images. In combination with scale-summarization, such arrangements can be detected even if they are composed of features that occur at multiple scales.

[9]Notice that, in this case, extending any of the line segments of the group will not intersect the horizontal line.

Figure 47: Grouping of projectively converging, subjective line segments that consist of the collinear spatial arrangement of corners. From left to right: The first and third image show original images and groups of line segments which were extracted. The second and fourth images present the scale-summarized corner detector response.

4.2.2 Information content and management

The information content of linearly perspective perceptual groups about the observed scene is related to the perception of distinct visual entities, surface geometry [62], and scene perspective. The perspective cues yielded by perceptual groups are observed to also participate in the perception of size, distance, and shape, as suggested from the phenomena of size and shape constancy. In addition, perspective cues play a role in the characterization of the visual impression determined by the observation of some scene. Description of such groups may refer to low level features, e.g. cardinality and orientation of a set of parallel lines. Other, context-related interpretations of such perceptual groups may also describe the grouping as a surface in 3D space, deduced from perspective cues (e.g. linear perspective). However, in order to obtain such knowledge, information that is not included in the image is required (e.g. viewpoint).

Next, the utilization of perceptual groups in a content-based image retrieval query is demonstrated, through a retrieval experiment.

Retrieval Experiment Figures 48 and 49 illustrate the retrieval, classification, and sorting of perceptual groups in images.

In the experiment, images containing a set of converging line segments in an approximately horizontal direction were extracted from a pool of images. The pool of images that was used was composed from two image categories: scenes containing man made structures (urban scenes) and wildlife scenes. Each category consisted of 500 images. Finally, the images containing man-made structures were further classified, with respect to the properties of the detected perceptual groups.

Initially images containing man-made structures were detected using the occurrence of a perceptual group of line segments as a characteristic feature. This method has been formerly used in [64, 48], but without taking the perspective component into account. The classification of images and results from the process of the detection of images containing man made objects are discussed in the fifth chapter.

The retrieval of the set of images containing linearly perspective perceptual groups, was followed from their classification into subclasses. Two classes of images were selected from the images that contain man-made objects. The selection of these two classes was related to the orientation of the perceptual groups contained in the images (thus if an image contained more than one perceptual groups it could appear in both classes). In particular, the two subclasses were selected to contain perceptual groups of line segments that converged from right to left and vise versa. The images were classified into two subclasses utilizing an approximate orientation descriptor. The orientation descriptor that was used was defined by the line segment dividing in half the angle formed by the pair of outmost line segments (intersecting at the vanishing point). Finally, groupings were sorted within each class with respect to the orientation value of the same descriptor. Figures 48 and 49 illustrate these two classes and the sorting of images with respect to the response of the orientation descriptor. The images that are shown in the figures correspond to the top matches of the image detection process, which targeted images containing man made objects.

With respect to the retrieval goal, other attributes of a perceptual group of parallel lines may be employed in the formulation of visual queries such as the cardinality of elements, their salience, the angle formed by the line segment pencil etc. Often queries may target at subsets of the perceptual group or integrate other cues such as color, texture etc. Such queries are demonstrated in the fifth chapter.

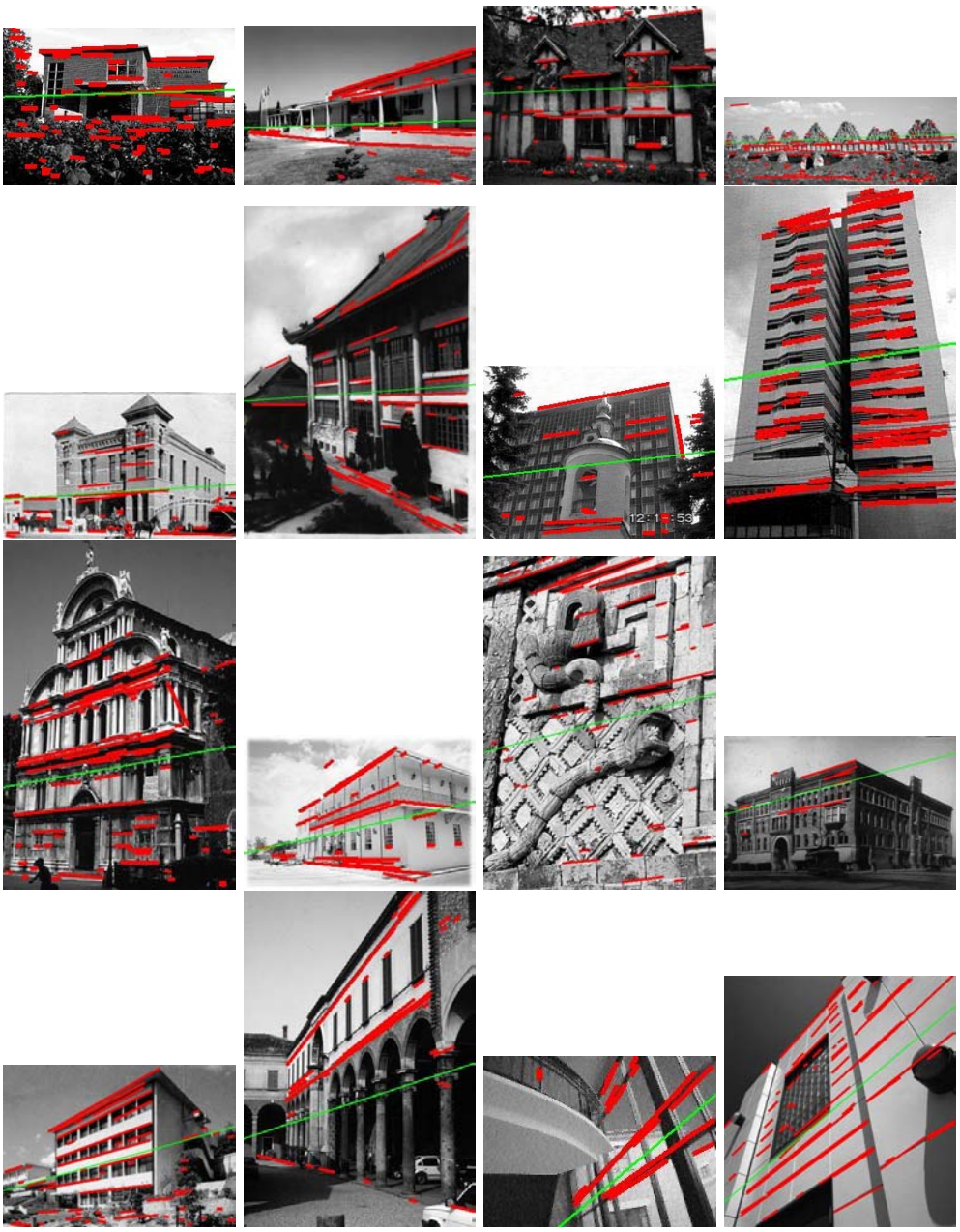

Figure 48: Groupings of projectively converging line segments classified with respect to orientation

Figure 49: Groupings of projectively converging line segments. classified with respect to orientation

4.3 Piecewise description and matching of silhouette boundaries

In this section the information content of two-dimensional closed curves, otherwise silhouette boundaries[10], is discussed and a method for their piecewise description and matching is proposed.

Motivation The apparent boundary of solid entities has been used in content-based image retrieval systems as a powerful content-based similarity criterion. Its descriptive information content stems from its contextual relevance as a visual cue. The boundary of a solid object is often used as an outline or abstract description of its structural properties. For example, the fact that specific objects can often be recognized solely from their boundary indicates the significance of such information. This latter observation is typically exploited by visual queries to retrieve similar images based on global "shape" of objects. However, the comprehension of the role of boundaries in object recognition and shape matching is related to other factors such as familiarity, observation goal, and salience. In addition, the notion of a "silhouette" can be not only related to the structural properties of environment entities, but also with the identification of "object parts" and their recognition.

Apparent figures are observed to originate from the perceptual grouping of gradient and edge derived tokens, illusory contours, motion, texture segregation and other. In this study, emphasis is placed at the description and matching of boundaries after their extraction from the image.

Constraints Most often the pursuit of similar shapes, in image databases, encounters the following constraints:

- *Noise / Sampling* Image acquisition introduces noise concerning pixel values and affects subsequent feature localization. Although the shape extraction mechanism may smooth out noise, minor differences have to be expected even between "identical" shapes.

- *Occlusion* Physical objects may reside in a variety of arrangements in the environment. In many cases, only a portion of an object's boundary

[10]The term silhouette boundary refers to a closed contour without holes. In this section, the term contour will also refer to the same type of visual structure.

may be depicted.

- *Pose* Depending on viewpoint the same object may form quite different two-dimensional shapes on the image plane. The issue of recognizing objects from their boundaries regardless of pose requires some knowledge about the three-dimensional structure of the prototype. In this work, a phenomenological approach towards the similarity assessment of boundaries is adopted, thus requiring the identification of similar shapes under rigid transforms such as the similarity, Euclidean, affine, and projective transforms[11].

Key points of approach In the remainder of this section, related work on shape representation is reviewed and a perceptually relevant, piecewise boundary description is proposed that is later on applied to the similarity-matching of boundaries. Prior to the presentation of related work, some keypoints of the theoretical approach which is adopted are outlined:

Salient boundary regions Information-theoretic approaches [56, 17, 107] towards boundary description fail to encode the perceptual significance of certain shape attributes, such as parts. In addition, the perception of contours in parts evidences the increased perceptual significance of specific boundary regions. A psychological evaluation of the salience of contour segments [6] indicates the strong appeal of intensely curved boundary regions to human perception. The intense form-descriptive property (for perception) of such regions is taken into account in the effort of deriving a perceptually relevant description for the similarity-matching of boundaries.

Scale of observation In the boundary description method, which is formulated in Section 4.3.2), the salience of boundary regions, is observed to be dependent on observation scale. More generally, shapes are considered as two dimensional functions permitting the analysis of boundaries in scale space, which has known descriptional benefits concerning the importance of some feature with respect to its apparent size [111].

[11]Usually the latter is computationally unstable and difficult to solve, and is approximated by the affine.

Boundary as a cue The content of boundary information is considered as a structural cue, since there is no one-to-one relation between objects and boundary representations, even if the later are devised to be invariant to projective transformations. Often, the interpretation of the boundary cue is related to context. A compelling example is that of multistability figures (e.g. the Necker cube, the rabbit-duck illusion, vase / face figures etc.) in which the represented figures can be comprehended as two different objects, depending on context.

4.3.1 Related work

A considerable number of methods can be found in the literature for the analysis of shapes by their boundary. A review that emphasizes topics related to the proposed contributions is presented below. In this review, the presented methods are classified into three categories: Voting methods, curvature descriptors, and scale-space related methods.

Voting methods Voting methods statistically try to merge the influence of more than one feature or shape descriptor. The joint statistics of global attributes such as area, perimeter, and compactness, may be used to compare shapes [24, 18]. Similarly, the Hough transform has been used to detect both primitive shapes (lines, ellipses) and arbitrary ones [7]. Joint Histograms have been used for a similar purpose in [5]. Finally, plain measures of point distances have been used in a similar manner such as the Procrustes [12] and the Hausdorff distance [44].

Curvature descriptors (e.g. [70, 79, 52]) are related to theoretical approaches that target the formulation of semi- or total- invariant signature functions under groups of transformations with focus on Euclidean and affine transformations. Typical signature functions use length or area ratios between contour points.

Signatures In curvature descriptor methods, shape is typically associated with a set of invariant (or semi-invariant) functions that define the specific notion of shape over arclength, otherwise *signatures*. Dissimilarity is then encoded as a difference of signature functions. Although such an invariant or semi-invariant function is a useful framework for the comparison of shapes, three issues are raised:

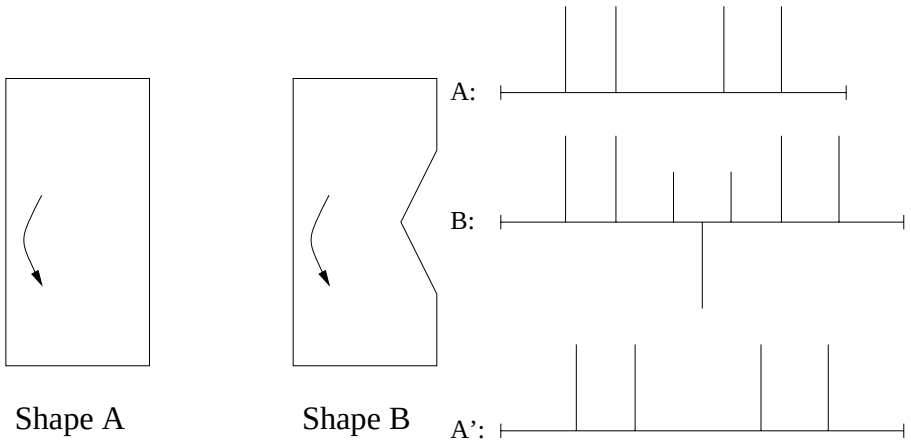

Figure 50: Two curves and their corresponding idealized curvature functions.

1. When comparing invariant functions, the shape must be assigned an origin for point to point comparison. A straightforward choice would be to assign the origin in an intrinsic manner, e.g. as the highest value of the invariant function. Thus, the selected representative should be some not trivial and stable candidate. However, selecting however the maximum value for this purpose is not adequate, due to the presence of noise.

2. Silhouette boundaries do not typically exhibit the same circumference. Therefore the invariant function of one is typically stretched to make point to point comparison with another. In Figure 50, the two shapes are almost identical except that one shape has an additional dent-like structure. The dent causes its circumference to be longer. A linear stretching of the shorter boundary to match the length of the longer shall cause the high values of the invariant functions to be dislocated with respect to each other.

3. Difference of curvature functions does not coincide with the intuitive notion of shape difference. In Figure 51 two curve pieces are considered. The curvature function is in essence a second order derivative and, thus, taking the difference between two curvature functions is equivalent to the computation of a third order derivative, which implies noise sensitivity.

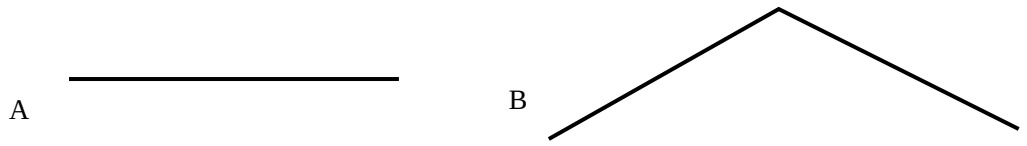

Figure 51: Two similar curve pieces. Curve A is straight and curve B is bent in a single point.

Finally, features may be computed that capture essential characteristics of a boundary. Such features could be Fourier coefficients [113, 36], singular value decomposition coefficients of the stiffness matrix [81], extreme curvature points [84], the method of Curvature Primal Sketch [4], shock [51], and the medial axis [11].

The Procrustes distance is further described here, due to its forward invocation in 4.3.3. Let the points (x, y) in the plane be given as a complex number $p = x + iy$, where $i = \sqrt{-1}$, and $\sum_j p_j = 1$ and $\sum_j p_j \bar{p}_j = 1$, with $\bar{\cdot}$ as the conjugate operator. Two such point sets p_j and q_j have the squared Procrustes distance [12]:

$$D^2 = 1 - \left(\sum_j p_j \bar{q}_j \right)^2 ,$$

This is equivalent to the minimal pointwise sum of squared Euclidean distances under rotation about the origin. The optimal rotation of q_j into p_j is given by $\tilde{q_j} = \left(\sum_j p_j \bar{q}_j \right) q_j$. Euclidean or affine action can be removed by normalizing the point set before calculating Procrustes distance.

Scale-Space Representations A multiscale representation for 1D functions was first proposed by Iijima [47], and later developed by a number of authors, see [111, 108] for an overview. Mokhtarian *et al.* have developed algorithms that use the so-called fingerprints of Mean Curvature scale-spaces and Affine Curvature scale-spaces to identify objects [68, 91, 67]: A scale-space of curves is used to generate successively coarser representations of a shape boundary. At each scale the zero-crossings of the curvature function (the inflection points) are detected and the curve in between two successive

inflection points is stored in a database. Variants of the method use an area preserving scale-space [92, 29] .

4.3.2 Boundary representation

In this subsection, a method of selecting salient boundary points is presented. Next (in Section 4.3.3), a piecewise alignment of boundaries based on the selected points is introduced. This piecewise alignment is subsequently used in the content-based retrieval of boundaries. In the remainder of this subsection, a scale-summarizing scheme for the curvature feature is proposed. The integral of scale-normalized curvature over scale is introduced as a quantification of the salience of a contour point. It is also argued that the description of contour points responding with a high value to this summation is more stable to noise and, thus, they can be utilized for form description and contour matching [104]. These issues are further discussed below, starting with an analysis of the stability of boundary segment description with respect to scale.

Scale-space representation A curve evolution is a scale-space if it fulfills a number of properties [3], indicating that the key property of scale-spaces is structure reduction. The scale-space representation of boundaries has been thoroughly studied in literature, also along with the required invariance of description. For example, the Mean Curvature scale-space [68] is invariant under Euclidean transformation and is non-increasing in the number of extrema and inflection points of the curvature function. Likewise, the Affine Curvature scale-space [91] is invariant under affine transformation and non-increasing in the number of extrema and inflection points of the affine curvature function.

In Figure 52, a random shape is shown together with snapshots of the Mean Curvature and the Affine Curvature scale-spaces. The Euclidean and Affine Curvature extrema are shown on the respective snapshots. It is observed that the Mean Curvature scale-space tends to a circle, while the affine Curvature scale-space tends to an ellipse. In both cases, the resulting curve shall continue to exhibit exactly four zero-crossings, even if further smoothing is applied. Another way of representing the evolution of the extreme curvature points is by the *fingerprint* images shown in Figure 53. As it can be observed, certain extrema "survive" up to coarser scales. These are typically defined as the stable extrema and are correlated with boundary regions

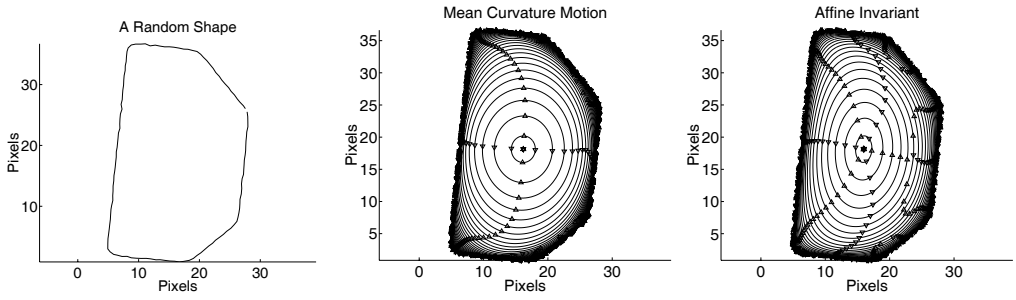

Figure 52: The scale-space evolution of a shape (left) in different scale-spaces. Snapshots from the Mean Curvature (middle) and Affine Curvature (right) evolution. Triangles denote location of Euclidean and affine curvature extrema.

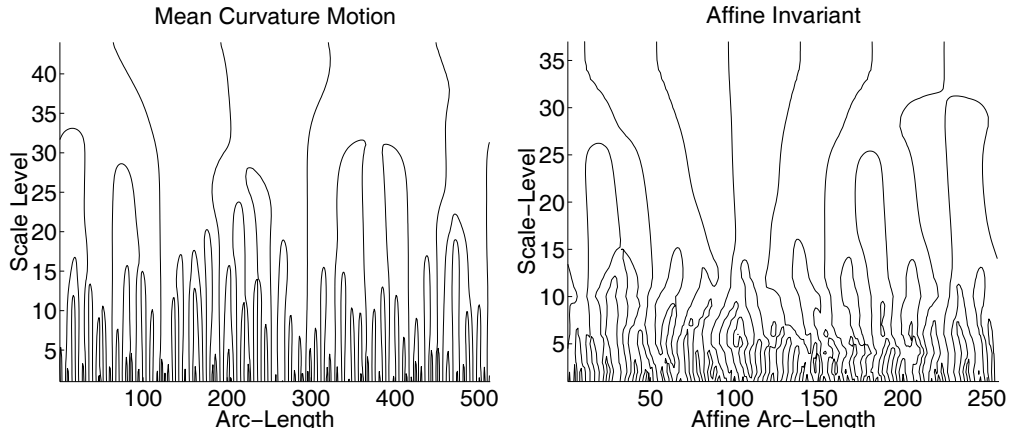

Figure 53: The fingerprint images of the evolutions in Figure 52 MIDDLE and RIGHT respectively.

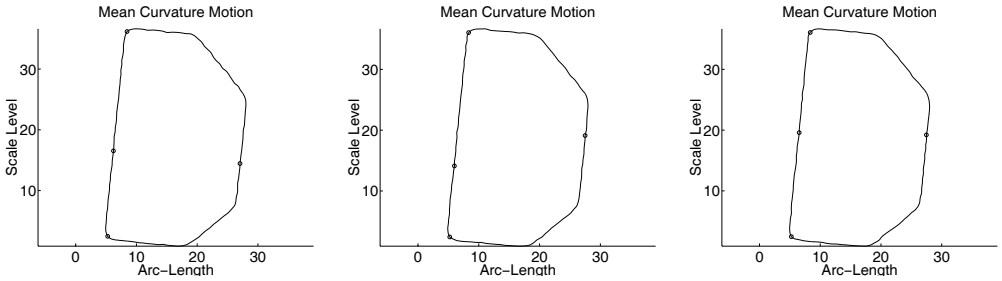

Figure 54: The accuracy of a trace depends on the curvature and the neighboring structure

that retain significance at coarse scales.

In both scale-spaces, the stable extrema can be backwardly tracked to zero scale. The sensitivity of this process is observed to depend on the type of structure. Normally distributed noise was added to the coordinate functions of the shape shown in the left of Figure 52, with a standard deviation of 0.05. Successively, the four most stable extrema of the Mean Curvature scale-space were backwardly traced to the original scale. The process was repeated three times and results are presented in Figure 54. From the statistics of this experiment, it is observed that the locations of the two approximately 90 degree angles are the most certain, followed by the rightmost location. The most uncertain point is the extremum on the longest straight piece of the curve, which in the fingerprint image yields the varying structure from arclength approximately 150–250. Next, a more formal study of the stability of curvature extrema tracking is elaborated.

A study of stable curvature extrema of ellipses Let the ellipse given by:

$$\begin{bmatrix} x_e(s) \\ y_e(s) \end{bmatrix} = 10 \begin{bmatrix} \frac{1}{e}\cos(s) \\ e\sin(s) \end{bmatrix},$$

where the free parameter e is referred to as the eccentricity of the ellipse. The major and minor axes of the ellipse are given by $10e$ and $\frac{10}{e}$, their area is 100π (independently of e) and their circumference increases with the absolute value of e. The ellipse has 4 curvature extrema $s^* = \{0, \frac{\pi}{2}, \pi, \frac{3}{4}\pi\}$, and the curvature values in these points are given by $K(s^*) \in \{\frac{1}{10e^3}, \frac{e^3}{10}\}$. Thus, $K(s^*) \to \frac{1}{10}$ for $e \to 1$ and $K(s^*) \in \{0, \infty\}$ for $e \to \infty$.

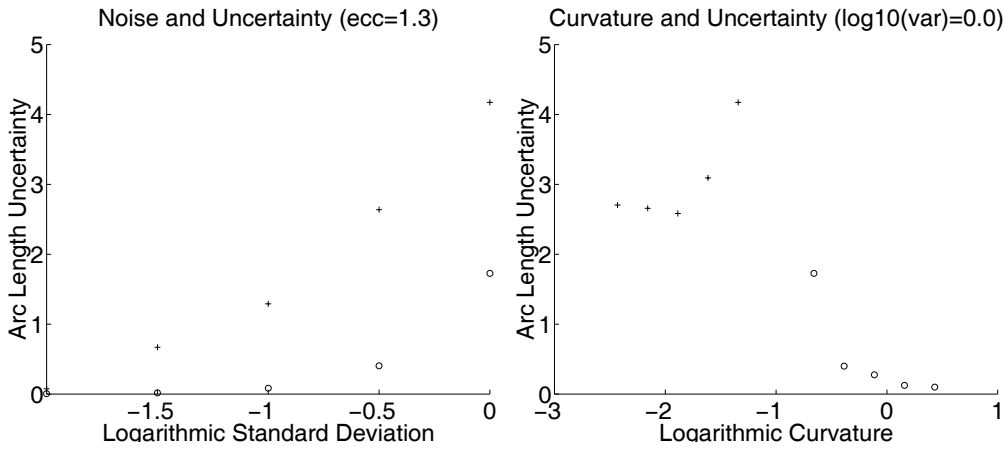

Figure 55: Noise makes curvature-extrema localization more uncertain, whereas their sharpness reduces uncertainty. Left: standard deviation of the noise (log-scale) vs. standard deviation of extremum localization. Right: curvature (log-scale) vs. standard deviation of extremum localization .

In practice, a boundary representation is subject to noise. A simple model of additive noise normally distributed perpendicular to the curve was utilized:

$$\begin{bmatrix} \tilde{x}_e(s) \\ \tilde{y}_e(s) \end{bmatrix} = \begin{bmatrix} x_e(s) \\ y_e(s) \end{bmatrix} + \mathcal{N}(0,\sigma) \begin{bmatrix} -y_e'(s) \\ x_e'(s) \end{bmatrix},$$

where prime denotes differentiation with respect to arclength and $\mathcal{N}(0,\sigma)$ is an independent normally distributed stochastic noise source of standard deviation σ.

Curves given by $[\tilde{x}_e(s), \tilde{y}_e(s)]^T$ have been studied in the Mean Curvature scale-space as follows: for a given curve the four most stable corners were tracked to zero scale and their locations in arclength were recorded. This experiment was repeated for a range of eccentricities and noise levels. In Figure 55, the standard deviation of the arclength location outputted by the algorithm is plotted, versus noise level and curvature. Circles represent low-curvature extrema (located on the "flat" part of the ellipse) while crosses mark the high- curvature ones. Each data point in the figure is based on 1000 experiments.

In the left graph of Figure 55, the vertical axis maps representation certainty, in terms of the variance of extremum of localization. The horizontal

axis maps the logarithmic standard deviation of noise. In the graph, the uncertainty monotonically increases with respect to noise and low-curvature points are observed to be more sensitive to noise than high-curvature ones. In the right graph of Figure 55, the same mapping for the vertical axis is retained, and curvature is mapped on the horizontal using a logarithmic scale. There, uncertainty tends to zero as curvature tends to infinity. In both extrema-type cases, the uncertainty is larger for the low-curvature extrema, as also observed in the previous graph.

The conclusion of this empirical analysis points to the speculation earlier formulated by the example of Figure 54. High curvature points tend to provide of a more robust to noise description of contour points in the given representation.

Salient boundary points Similarly to most local features, local curvature, given by

$$\frac{x' \cdot y'' + x'' \cdot y'}{((x')^2 + (y')^2)^{\frac{3}{2}}},$$

is defined with respect to scale (in this case the arc length over which the derivation is performed). The application of this observation gave rise to the scale-space methodologies of boundary description previously discussed. Since the curvature function scales inversely proportional with contour (or arc) length, curvature may be scale- normalized as follows:

$$\kappa^*(s) = \sigma^*(s)\kappa(s) \tag{21}$$

in order to be able to compare curvature at different scales, where parameter s denotes a parameterization of the fingerprint line. Qualitatively, s can be correlated with scale, since the larger the value of s the sparser the sampling of the contour. The parameterized points are subsequently interpolated (using cubic splines) in order to result in a smooth function and be able to compute curvature. Thus, increasing the value of s yields a scale-space of the contour. Equation 21 resembles the scale-normalized feature response function, discussed in Chapter 2. The value

$$\mathcal{K} = \max_s \text{sign}(\kappa(0))\kappa^*(s) \tag{22}$$

relates to the spatial extent of the contour point of high curvature [58], correlating the notion of optimal scale of observation with the maximal mode

of the feature detector. The sign($\kappa(0)$) operator in this function (as well as in the one that is next presented) is used to provide with a positive value of $\mathcal{K}$ and $\mathcal{K}_{SSR}$, since curvature may exhibit a negative value. In such a case, the sign of the series of values $\kappa(i)$ (with $i \in$ the range of s's values) will be constant and, thus, multiplying with sign($\kappa(0)$) will always yield a positive result.

The scale-summarization of scale-normalized curvature:

$$\mathcal{K}_{SSR} = \text{sign}(\kappa(0)) \int_s \kappa^*(s) ds \tag{23}$$

provides a description of the curvature contribution of a boundary point over all scales. As discussed in the following example (see Figure 56), the magnitude of $\mathcal{K}_{SSR}$ may be used to detect significant components of a boundary, independently of their spatial scale.

The following example offers insight of the proposed method and indicates the significance of observation scale in the characterization of the salience of a boundary segment. Figure 56 (right) plots the scale-normalized curvature response over scale for three boundary points of the shape on the left: one of very high spatial frequency (observed as noise rather than structure and at coordinates $220, 140$), one of a medium scale (which is a sharp peak at $330, 70$) and one of a large-scale (the mild curve at $250, 270$). In the right figure, the horizontal axis maps the logarithmic scale parameter and the vertical the scale-normalized curvature response. The noise dent corresponds to weakest response, while the sharp curve to the one that takes the maximum value among the three. The large scale curve corresponds to the longest surviving in scale response. Qualitatively, Equation (23) favors boundary points that are either intensely persisting for a first scales or more subtly existing but for the majority of resolutions.

The empirical study conducted in the previous paragraph, indicates that high curvature boundary points exhibit greater certainty when tracked in scale space and, thus, are more qualified candidates for contour description. Given the relation of curvature to scale, a salience metric for contour points (in Equation 23) was proposed. The performance of the metric is demonstrated in Figure 57. In the figure, several shapes are presented. For these shapes, curvature extrema are tracked and those that survive for the largest amount of scales are selected. The salience descriptor given in Equation (23), is applied to the selected extrema. In order to illustrate the response of the descriptor the following procedure was performed: a circle with radius pro-

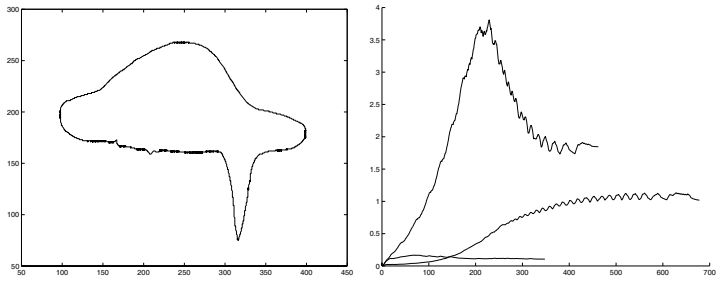

Figure 56: A shape and the scale normalized curvature response over scale for three of its points.

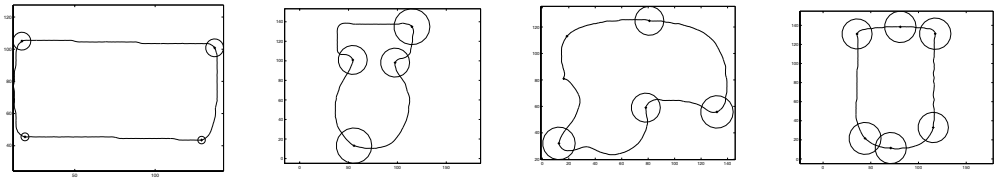

Figure 57: Demonstration of the behavior of the salience metric (see text). In the figures, the estimated salience of boundary points is illustrated by plotting circles centered at such points. The radii of those circles are proportional to the estimated salience.

portional to the response of the descriptor is plotted, centered at the coordinates of each extremum. As observed in the figure, points of high curvature and coarse scale are correlated with a high response of the descriptor.

4.3.3 Description and matching of boundaries

In this subsection, the discussion is focused on the description and matching of silhouette boundaries. In this effort, the dissimilarities between boundaries are quantified in order to provide with a similarity measure for contours. Matching results from experimentation with a small database (Coil-100) [72].

Description of shape boundaries It is argued that the number of points chosen to represent the boundary is proportional to the quality of representation. As implied by the previous study, if some salient points of the bound-

ary are chosen, then a perceptually relevant description is obtained. In this paragraph, a classification of boundary points with respect to their salience is proposed. This classification is used in the next subsection, for the indexing of shape boundaries, suited for their content-base retrieval.

Description length In principle, the description of a boundary using a few significant points given by Equation 23 casts possible for a whole family of "structural metamers"[12] to fit the description. The description of a shape's boundary using representative points is of interest for a number of reasons:

- The family of boundaries that matches a, perceptually relevant, description may be similar.
- Boundaries may be indexed with respect to a set of representatives.
- The identification of perceptually relevant representative points finds application in the piecewise decomposition, description, and matching of shapes.
- Representation space may be possibly optimized using a characteristic selection of significant boundary points.

However, if too few contour points are selected then the utilization of the resulting boundary description can be problematic. For example, data points sampled from a circle are considered. Using three points, the circle can not be discriminated from a triangle and in general with n points the points might as well come from an n^{th}-ordered polygon. Naturally, the polygon shall increasingly resemble a circle as n is increased and for some large n it would be more *expressive* to describe the points as a circle rather than a polygon. Expressiveness is understood by means of lossless compression: in the example, the choice of describing the points as a circle depends on if this yields a shorter description of the data set.

Comparing data in terms of minimal compression requires the comparison of the coding cost of the model against the deviation of the model from the data [85]:

$$L(D, M) = L(M) + L(D|M),$$

[12]Given a set of points, this term is refers to all the shapes that have these points in common.

where D is the data set and M the model parameters. The optimal model is one that minimizes $L(D, M)$.

In the retrieval task from a database of boundaries, it can be assumed that a given shape is to be described using a prototype from the database. Therefore, the codelength of the model consists of identifying which shape from the database is being used and how many salient points are used in the description:

$$L(M) = -\log(card(\{db\ elements\})) + \log^*(card\{description\ points\}).$$

The code can be designed so that all shapes in the database are equally probable, and that the number of confident points is coded by the Universal Distribution of Integers, $\log^*(i) = c + \log(i) + \log(\log(i)) + \ldots$ [85]. The summation is performed over all positive terms.

In order to estimate the deviation from the model, a segmentation of the boundary defined by the description points is assumed. The two boundaries are point to point aligned (see next paragraph) and individual polygons, or "pieces", are sampled equidistantly. Differences from the prototype model are represented as deviations, with respect to the prototype, of corresponding sampled points:

$$L(D|M) = \sum_i \log^*(N_i) + \log^*(10\sigma) - \sum_i \log\Big(G(M_i, D_i, \sigma)\Big),$$

where N_i is the number of sample points for piece i. The deviations are coded as a two dimensional Gaussian distribution,

$$G(M_j, D_j, \sigma) = \frac{1}{2\pi\sigma^2} \exp\left(\frac{(X_j - x_j)^2 + (Y_j - y_j)^2}{2\sigma^2}\right).$$

The point (x_j, y_j) is coded as the displacement from the corresponding point (X_j, Y_j) from the shape in the database.

It is emphasized that in contrast to [79, 52] the difference between two curves is not coded by the difference in their corresponding curvature functions, but by the difference of spatial coordinates of "pieces". There are two reasons that the difference between two curvature functions does not coincide with the notion of shape dissimilarity as earlier discussed in Section 4.3.1, problem 2. In contrast, the point to point displacement vectors $((X_j - x_j), (Y_j - y_j))$, that define the displacement between corresponding points of aligned contours, notably penalize the global difference between these curves.

Matching of shape boundaries Using reference points, boundary regions may be corresponded, in order to compare curve segments. To do so, contours must be piecewise aligned, prior to comparison. In this paragraph, a method for the similarity matching of boundaries is proposed, based on the description of salient boundary points proposed earlier. The method is initially implemented for matching invariance under the Euclidean transform and retrieval results are presented for a small contour database. In addition, a normalization of the description for affine contour matching is formulated and its compatibility with known invariants is discussed. Finally, the suitability of the proposed piecewise shape description is discussed with respect to other real-image content-based retrieval requirements.

Matching algorithm outline Qualitatively, the piecewise matching algorithm operates as follows: Initially, the two compared shapes are centered, size-normalized, and rotationally aligned, which results to invariance to the Euclidean transform. Then a correspondence between "pieces" of the two shapes is defined and a one-to-one correspondence between the points of each pair of corresponding pieces is determined. The sum of the total distances between sampled points from the corresponding pieces is regarded as the dissimilarity of the two shapes.

Let two boundaries b_1 and b_2, denoted as $b_i = \{p_1, p_2, ..., p_N\}$, where p_i is a boundary point and N is their population. In order to align two shapes all $N-1$ cyclic permutations of the indexes of b_2 are considered. The one that yields the least dissimilarity in terms of a geometric distance metric is selected. In order to retain the piecewise property of alignment and to speed-up the computation, the dissimilarity with respect to only the reference points is compared. The Procrustes distance [12] is used as a distance measure between reference points since it uses an intrinsic size-normalized description of points.

The same number of reference points is chosen for both silhouette boundaries. A variety of cardinalities of the reference point set are considered by the algorithm and the one that yields the minimum distance is selected. Typical cardinality values of this set would be 3, 4, 5, ... etc. as required.

Matching algorithm formulation The algorithm is formulated as following:

Given is a shape as a list of points on its boundary and a database:

1. Compute the scale-space until only 4 curvature extrema are present.

The number of curvature extrema (4) that defines the termination criterion is chosen due to the following fact: regardless of the amount of smoothing that is applied to a boundary, the minimum number of extrema is 4 (see also Section 4.3.2).

2. Track the extrema along scale.

3. Choose the most stable extrema, defined as the ones that exhibit the maximum value of the integral of scale-normalized curvature.

4. Find the maximum scale-normalized curvature values for extrema tracked along scale.

5. Sort the extrema at zero (original) scale by their maximum scale-normalized curvature values.

6. For each of the $n = 2, 3, \ldots$ points with highest maximum scale-normalized curvature, sort them according to location on the $s(0)$-parameterized curve.

The subsets of points are considered different segmentations of the boundary. Two shapes are then compared by aligning their segmentations. For this, the Procrustes (scale and translational invariant) distance is used.

7. Generate all cyclic shifts of one subset and calculate the Procrustes distance to the other. Select the shift minimizing this distance.

At this point the boundary is encoded, relatively to a shape from the database. Once an alignment of two segmentations is obtained, a piecewise, linear point to point correspondence between the boundaries can be defined.

8. Generate the piecewise, linear point to point correspondence between the segments of the boundaries.

9. Calculate the code length using the correspondence and minimize this code length with respect to translation, rotation, and scaling.

The minimal code length is taken as the distance between the curves. The distance between two curves is symmetric, when the curves have the same segmentation and number of points.

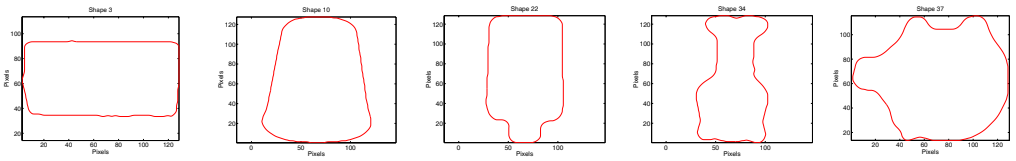

Figure 58: Some boundaries from the database (from left to right): A gum package, a reusable food container, a tube with a screw-lid, a glass, and a pig.

Given n points of a shape, t scale levels and m stable extrema, the computational complexity of the above algorithm is $\mathcal{O}(nt\log(nt) + m^2\log m^2)$. The first term is typically the largest and refers to the scale-space computation while the later refers to Step 6 in the algorithm.

Comparing real shapes under Euclidean invariance An implementation of the algorithm has been tested on a small database of 100 boundaries. The shapes have been selected from automatic segmentations[13] of images from the Columbia University Image Library (Coil-100) [72]. In Figure 58 some of the shapes are shown. The distance of each of these has been calculated to the other 99 shapes and in Figure 59 the 3 closest matches are presented. These results were selected as a representative of the shape similarity ordering produced by the algorithm. The shapes in the left column refer to the best match in the database. From top to bottom, the codelength of the best match increases downwards (codelength is shown on top of graphs, denoted as "its per arc length") and as observed correlates with the quality of the match. The same observation is made for each row individually.

Affine Invariance The requirement for affine-invariant matching of boundaries indicates two possible implementations of the task, based on the algorithm that was proposed above. The first option is to use an affine invariant parameterization of the shape and then Affine Invariant scale-space. However, as indicated in Section 4.3.2 tracking of extrema in this scale-space is computationally complex and inrobust. Another option is to estimate the affine transform (R) connecting the compared shapes, find its inverse (R^{-1})

[13]Since the objects in this database are displayed over a black background, segmentation was simple to perform using image thresholding.

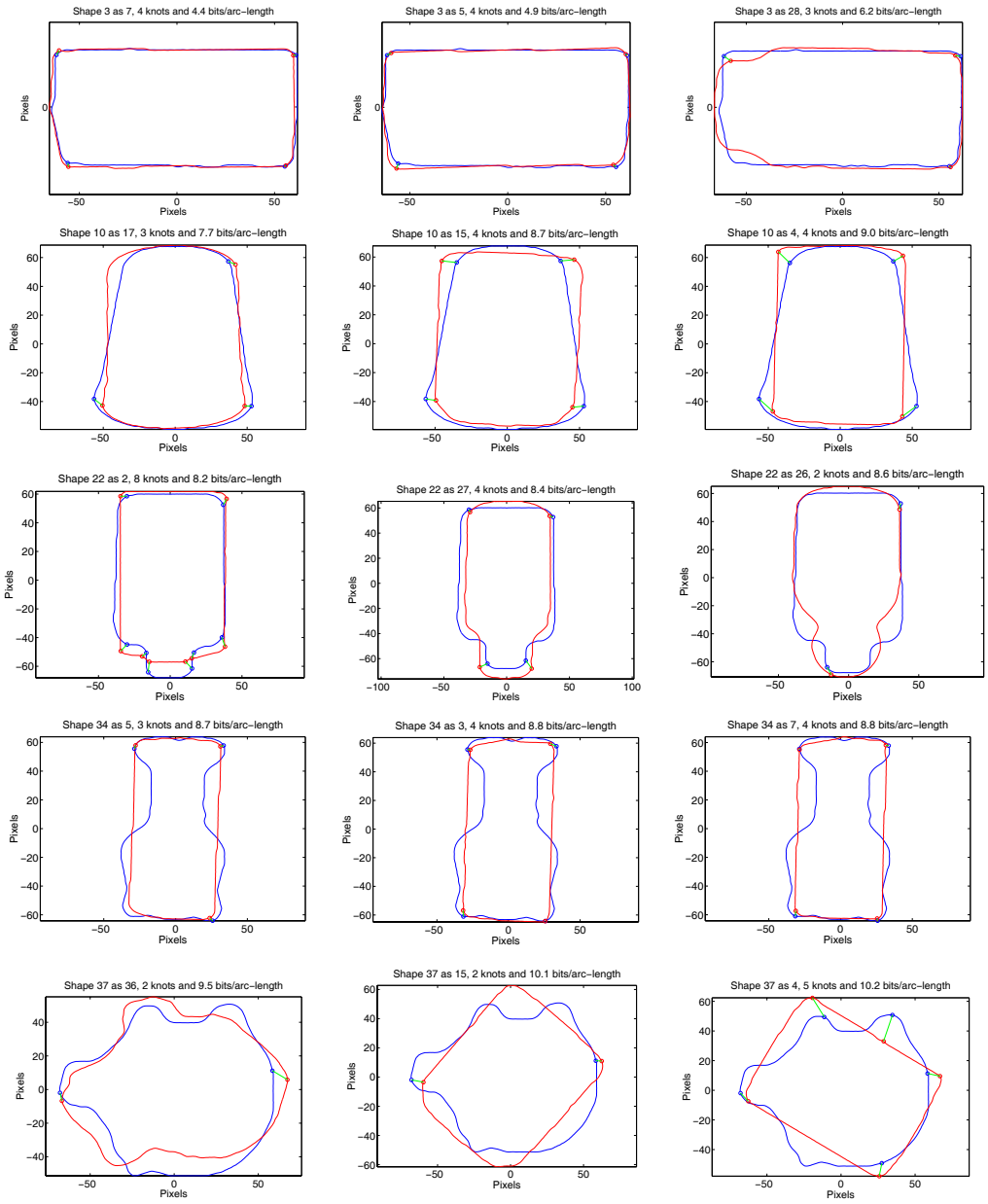

Figure 59: Three best matches for each of the shapes in Figure 58. Best match is shown in left column and worst (out of 3) in right column.

and transform one shape using R^{-1}. Registering is performed as already described; a robust implementation of the R^{-1} estimation process can be found in [105].

It is emphasized that in the latter case the estimation of R^{-1} would robustly transform one shape to some that perfectly matches the prototype only if the two shapes are indeed similar. Consequently, matching using the described method would be found probably useful by a shape recognition process rather than a similarity assessment one.

4.3.4 Conclusion

In Section 4.3.3 a hierarchic description of shape boundaries based on salient summarized-curvature points was presented. The proposed description is hierarchic in the sense that the value of the salience measure defines a hierarchy. The description was exploited in the content-based retrieval of boundary information. The detection of salient boundary points finds application in the definition of reference points on boundaries as well as in their piecewise description.

From the discussion in Section 4.3.1 and the development of an algorithm for rotationally invariant boundary comparison in Section 4.3.3, it is understood that alignment of arbitrary arclength-parameterized descriptions requires the exhaustive cyclic permutation of contour points in $O(N^2)$. Instead, if permutations are restricted to $M \ll N$ possible alignment points, then computation time is reduced. In order for the operation to be successful, the reference-point selection process is required always select the same type of points chosen. Signature singular points have been used in literature (such as the maximum of Euclidean or affine curvature [52]). However, this choice can be affected by noise. The selection of salient contour points, extracted by curvature scale-summarization, constitutes a better candidate for this choice.

4.4 Summary

This chapter was related to component of visual content that is related to the perceptual organization of visual features. In this context, two methods of content-based description and matching of perceptual groups were presented. The first one is related to the detection and description of linearly perspective perceptual groups. The second one is related to the description

of, already grouped, silhouette boundaries. Both of the resulting descriptions were utilized in content-based retrieval of visual information.

The first method that was introduced targets the grouping of parallel line segments in the three-dimensional world, based on their parallelism or perspective convergence in the image. The algorithm was initially formulated for linearly perspective line segments and then generalized for subjective ones. These subjective line segments could be composed of collinearly arranged point-like features, such as corners. The novel aspect of the introduced method, as compared to other perceptual grouping methods of line segments that can be found in the literature, is the ability to take into account the perspectively transformed occurrence of parallel line segments in images.

The second method that was introduced in this chapter was related to the description of silhouette boundaries, due to their significance in various aspects of visual information retrieval. In particular, a method for selecting corresponding (anchor) reference points when comparing shapes was introduced, founded on the analysis of curvature with respect to scale. The method utilized the accumulation of scale-normalized curvature to attribute such points with a salience metric, resulting in the selection of high curvature points of dominating spatial extent. The resulting selection of points was argued to be perceptually relevant and descriptionally significant, based on psychophysical observations and computational experiments respectively. The selected salient boundary points were further utilized in the decomposition of boundaries into pieces, by segmenting the contours at the location of each point. Using this decomposition the alignment of boundaries was made possible, which was the basis for a similarity matching algorithm.

5 Visual information browsing and retrieval

This chapter is focused on the provision of the general ability to browse and retrieve images, based on their content. In the previous chapters, cases of similarity matching of visual information have been demonstrated based on a single type of visual feature. In this chapter, emphasis is placed upon the integration of features in the content-based browsing and retrieval of images. In the introductory section, the motivation underlying the chosen approach is discussed. In the second section, methods for the content-based visual information browsing and retrieval are presented.

5.1 Introduction

In this section, the motivation underlying the proposal of methods that are presented in the next two sections is discussed and the key points of the adopted approach towards the content-based retrieval of visual information are outlined.

Visual query evaluation A major difficulty associated with the design of efficient methods for content-based image retrieval is the evaluation of the performance of a tested content-based retrieval criterion. In other words, it is not obvious which content-based retrieval criteria to adopt when designing a content-based image retrieval system, since there exists no formal description of the efficiency of each. The reason for this is that currently there is not any generic, with respect to observation goal and image context, objective method to estimate the apparent similarity of images.

In particular, the application of the traditional measures of *precision*[14] and *recall*[15] [23], which originated from in field of textual information retrieval, to the case of image databases is considered as problematic. While the mentioned measures are found to be quite useful in the evaluation of textual information retrieval queries, their effectiveness cannot be appreciated in the case of image databases mainly for two reasons. First, the determination of which of the retrieved images are relevant to the query is quite subjective.

[14]The number of retrieved images that are relevant over the number of the images that were retrieved.

[15]The number of retrieved images that are relevant over the number of the relevant images in the database.

The observation goal, individual observer, and image context, can introduce different judgments about the similarity of two images. Second, visual queries often target the similarity rather than the identity of visual subjects and, thus, an ordering with respect to similarity value is required in the presentation of results. Consequently, the query result $A = \{a_1, a_2, ..., a_n\}$ consists of an ordered set, which contains all the elements of the database. The choice of the element k through which A can be divided into "relevant" ($\{a_1, a_2, ..., a_k\}$) and "irrelevant" ($\{a_{k+1}, a_{k+2}, ..., a_n\}$) subsets is critical and may lead to ambiguous results.

Domain knowledge Literal visual similarity, such as the existence of identical or similar colors in images, is not found to be sufficient given the requirements of visual information browsing and retrieval applications. The comprehension of the visual information residing in an image is intuitively linked with domain knowledge that does not reside within the image itself. Thus, although that the comparison of images based on the literal similarity of their features is a useful tool for the content-based retrieval of images, it is found not to be sufficient for the total fulfillment of the requirements of this task. Consequently, applications of content-based image retrieval often implicitly include requirements related to *(a)* physical laws, e.g. which can detect whether physical surfaces exhibit same or different properties describe the equal and different properties of physical surfaces (e.g. [32]), *(b)* geometric and topological rules (e.g. [103]), as well as *(c)* category-based rules, which encode context-related class characteristics. In the approach described below, the general inability to derive purely visual and generic criteria that fulfill the mentioned factors is compensated by the modular description of visual content components. In this way, physical, geometrical, and context-related information can in each case be integrated with visual content descriptors, given the application requirements.

Description of approach As outlined in the introduction chapter, the adopted approach to, visual information browsing and retrieval, relies on a collection of generic description and matching mechanisms. The reasons motivating this choice of approach are:

- The activation of individual mechanisms for the description of specific components of visual content may be tuned to formulate explicit and specific queries that match specific application requirements.

- The description of visual content may be fused with domain knowledge that reflects application requirements.
- The psychophysical evaluation or tuning of description and matching behaviors can be performed through controlled experiments, by regarding a single similarity factor (rather than complex composition of several of those).
- Interaction with the queried collection of images can be enabled though experimentation with the activated mechanisms, which may be tagged with weighted significance. Given some machine learning or relevance feedback method, computational models of the targeted information may be learned from experience.
- From a software engineering perspective, the existence of a modular collection of generic content description mechanisms contributes to the simple assembly of goal-oriented description and matching competences. By restricting this collection to the required components, with respect to the specific application requirements, the assembly of goal-oriented image retrieval system becomes simpler than designing a new system each time.

In the sections below, the visual content description mechanisms presented in the previous chapters will consist the modules of the described collection. The description yielded from each module consists of a "visual cue" providing a basis on which browsing, classification, and retrieval of images is based upon. Methods for the integration of visual cues with browsing and querying mechanisms are proposed and indicative experimental results are demonstrated.

5.2 Visual Content Querying and Browsing

In this section, content-based management of visual information is discussed from the perspective of browsing images with respect to their visual similarity. Towards this objective, two methods are contributed to the two corresponding subsections below. The first one aims to provide high specificity in visual queries, by prompting the user to explicitly identify the visual content components that are intended to be taken into account in the formulation of the query. The second one, is related to the support of the content-based

browsing task, by using the hierarchical classification of images for the task of interactively specifying the targeted visual content.

5.2.1 Visual query formulation

Often, in visual search engines, a whole image is given as input to a query which will somehow retrieve similar images, while the user's interest is focused on specific objects or features of the image. In the best case, the comparison mechanism will be efficient enough to select these features or objects, among others present in the image, and will then evaluate the query, taking into account the similarity of all features. Clearly, query precision shall decrease as irrelevant features are included in the query, since the retrieval result will contain matches for the non-interesting features as well. In the worst case, the comparison mechanism will not select all or even any of visual information components in which the user is interested in. In the same context, the spatial arrangement of features in the image may be of interest to the end user. If the retrieval mechanism does take spatial layout into account, then the previous ambiguity is re-encountered in the arrangement domain (e.g. non-interesting layouts will be queried for).

Explicit feature selection The ambiguity of user interest concerning the components of visual content, which is encountered in typical queries by example, can be partially compensated by an explicit specification of the visual content components that are of interest in an image. The determination of visual features and spatial relationships, based on a visual interface, facilitates the explicit specification of user interest about visual features and their spatial arrangement. For this reason, detected features in the query image and their attributes are proposed to be specified in a query formulation process. In order to match this requirement, extracted features such as image blobs, contours, or perceptual groups are proposed to be available for online selection and specification. However, even this representation may be ambiguous, since these visual entities may include more than one visible feature. For example, image blobs can be attributed by texture, color, shape, structural entropy, orientation principal components etc.

In the proposed approach, blob-like image regions of potential interest, originating from some weak or complete segmentation, are presented as selectable regions superimposed on the image. In such a case, the image region is defined by the boundary of the image segment. In other cases, the bound-

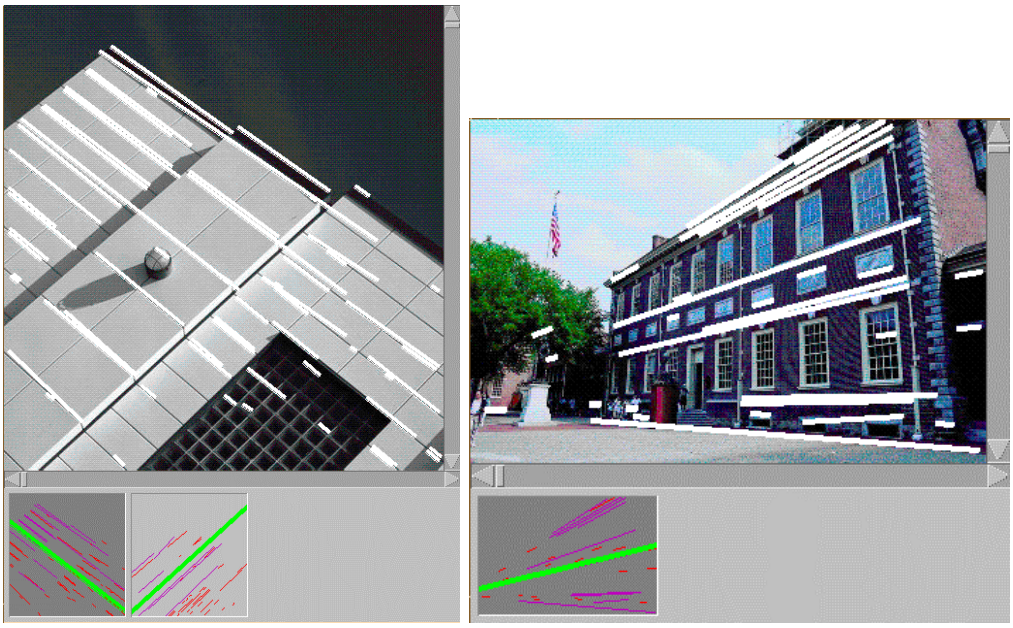

Figure 60: A user interface which facilitates the selection of visual entities.

ary of the region is not primitively available, such as in the case of perceptual groups. In those cases, the image region occupied by the perceptual group is approximated by its convex hull. Given this presentation, the query formulation mechanism illustrated in Figure 60, uses a feature representation space (attached below the graphical image display interface component) to select among available features. For the displayed visual entities, the attributes of interest may be specified in a two-step procedure: *(a)* entity selection and *(b)* presentation and user-selection of interesting feature(s) from those associated with the entity.

Similarly, spatial relationships can be represented, if of interest, within the context of some query. Relative location, intersection, or containment are automatically defined from the selection of two or more visual elements and the evaluation of their spatial relationship. The result of this evaluation yields a spatial relation predicate, which can be encoded by various representations, such as the 2D string [22], R-trees [37], or topological graphs. Figure 61 illustrates such a procedure. The described formulation, facilitates the representation of absolute positions within an image, or spatial relation-

ships.

The result of the described visual query formulation process is a set of features and spatial relationships that can be directly compiled into some formal database query language, such as Image-SQL, to name one. Feature similarity may be encoded by logical predicates such as $(A =_f B)$, $(A \approx_f B)$, $(A \neq_f B)$, where f refers to the compared feature, and A, B refer to the prototype and candidate features. The meaning of the mentioned above mathematical operators refers to strict equality, approximate equality, and inequality, respectively. Similarly, spatial relationships can be encoded as $(A * B)$ or $(NOT(A * B))$, where $*$ refers to the spatial relationship. An example of such a query follows:

select `*`
from `images`
where `(`$A \approx_f B$`) AND (`$A \neq_f C$`) AND (NOT(`A `northwest_to` C`))`
and `(`B `in_center_of_image);`

In addition, symmetry may be taken into account, if spatial relationship operators are inverted (e.g. operator "northwest" would be substituted by "southeast") and new predicates are added to the query using the logical `OR` operation as: $A * B \rightarrow ((A * B)\texttt{OR}(B * A))$.

Finally, an important aspect of the proposed method for query formulation is that users often engage the task of image retrieval without explicitly knowing what visual properties of the query image they are interested in. Using a query formulation method that presents visual features as tokens, which can subsequently be used in the formulation of query predicates, contributes to the resolution of such ambiguities. The resolution of the ambiguity is performed by the presentation of the image features and properties that are available to be used in queries. Furthermore, the existence of features that are often not compellingly perceived but play an important role in visual impression, such as perceptual groups, can be indicated to the user in order to be included in the query. For example, the organized structure of a perceptual group, often not regarded as an individual feature, can be pointed out e.g. as in Figure 60.

Discussion Current trends [102] in visual information retrieval explore ways of bridging the gap between formulated queries and the intended content to be retrieved, by activating or not the description of individual content

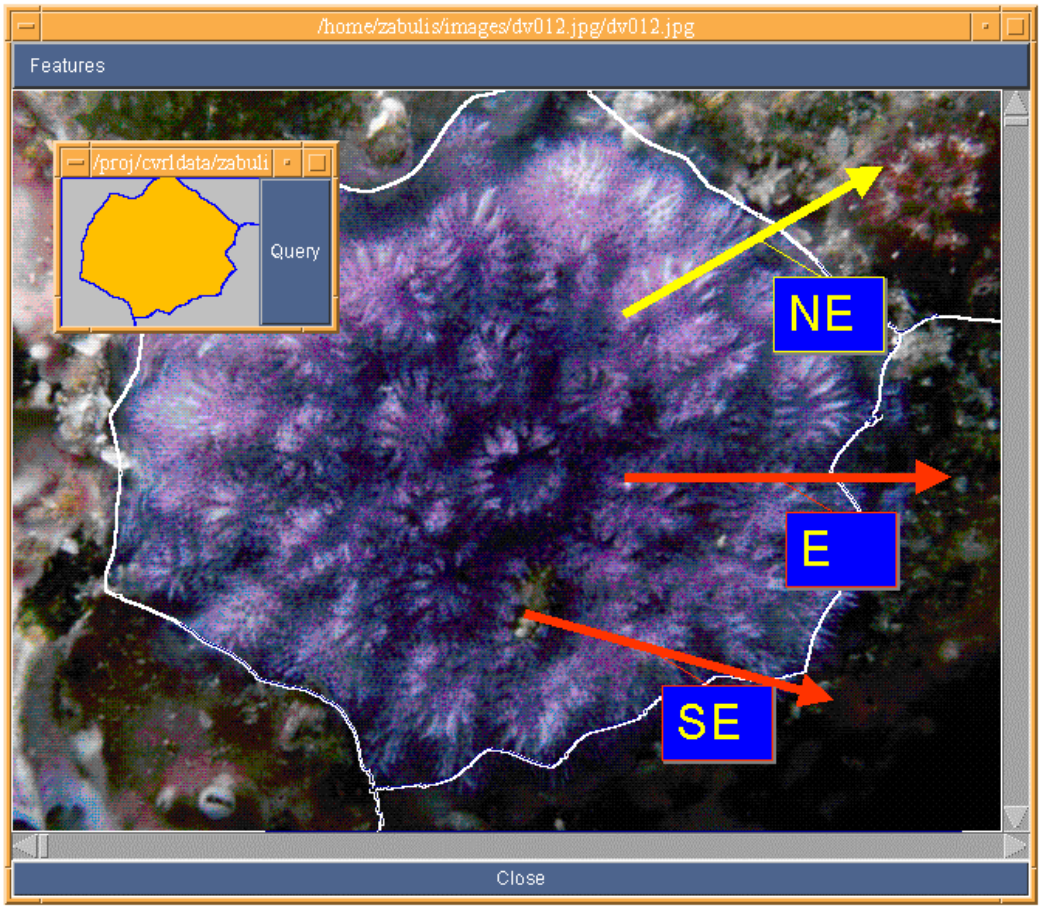

Figure 61: A user interface which facilitates the definition of spatial relationships ($NE = northeast$, $E = east$, etc.)

features and providing feedback through the user interface [89]. In this context, an approach supporting the ability to explicitly reference and attribute visual content components within a query, was presented in this subsection. The proposed approach aims at the specification and indication of visual features and spatial relationships that can be utilized in the formulation of a query.

5.2.2 Hierarchical classification of image collections

There exists a discrepancy in the attempt to retrieve visual content in a perceptually relevant way by discretely specifying image features and performing image similarity matching based on the similarity of each feature. In particular, it can be argued that the visual system utilizes some more complicated strategy towards the similarity matching of images, that simultaneously takes more than one matching criteria into account. On the other hand, when people are engaged in the task of organizing data, concepts, information etc., often this results in the use of a discrete hierarchical mental organizations in order to easily manipulate concepts, or mental tokens. For example, books are composed of chapters, sections, subsections, paragraphs etc., words, names, and digital documents are listed in directories, subdirectories etc., just to name two. Clearly, hierarchical organization is a useful conceptual tool for organizing knowledge or information. This subsection is oriented towards the computer-aided support of image browsing through hierarchical classification.

The proposed method aims at the interactive refinement of targeted content through the interactive classification of features. Initially, motivation and the theoretical basis of visual content classification are discussed and an example of visual query refinement using classification of visual features is demonstrated. In the second paragraph of this subsection, indicative experimental results of image collection classification are presented.

Visual content classification In queries targeting the phenomenological information of visual collections, image features and their interrelations are of primary interest. Which features are significant, how they are mentally conceived, integrated, and matched is not known for all cases. In this paragraph, the application of hierarchical classification of image collections, in the support of content description as well as query formulation and refinement, is discussed.

In Section 3.5, a content-based image retrieval strategy was presented based on the combined feature similarity of some prototype distribution. The conclusion was that subsequent categorization of the matched results with respect to additional criteria was required (color, orientation distribution entropy), in order to discriminate matching but counter-intuitive cases. Although such criteria could have been initially incorporated in the query, it is often the case that query refinement is more plausible than the re-execution of the initial query.

Classification procedure In the simplest case of some visual query, a description capturing image content in terms of features can be adequate for the discrimination of a piece of visual information from others. Typically, only part of these features is known or beforehand understood, such as in the example of Section 3.5. In that example, the extraction of a visual entity was made possible, based on weak image segmentation, . Subsequently, a description was extracted for that entity through the selection of features, which were all used in the formation of query predicates. The execution of the query yielded results that contained cases in which matching with respect to the description was correct, but were not fully compatible with intuition. It was speculated that with the use of a finer classification of results, these cases would have been rejected. The case described above can be formulated as: $S = \{c(D)\}$, denoting that a binary *classifier*[16] c is applied on the data set D, yielding an ordered set S, with two elements; the sets of similar and dissimilar elements of the original data set. The case discussed in Section 3.5, where part of the initial features were unknown or not well understood, can be formalized as follows:

Let a similarity matching strategy which consists of a set of criteria $S = \{c_1 \circ c_2 \circ ... \circ c_n(D)\}$ and c_{n+1} be a new criterion that is intended to be evaluated as more discriminative (the symbol $\circ$ refers to function synthesis, as $f(g()) \equiv (f \circ g)()$). It is assumed that the strategy has been already applied to some query space yielding the ordered set S. A straightforward solution would be the re-execution of the retrieval process including the new criterion (as $S' = \{c_1 \circ c_2 \circ ..., c_n \circ c_{n+1}(D)\}$). However, this is not always possible or at least computationally expensive. A simple method in order to empirically reason if c_{n+1} refines query results, is the classification of S with respect to $S' = \{c_{n+1}(S)\}$. Figure 62 illustrates the discussed classification of images

[16] A classification criterion that sorts elements into two classes.

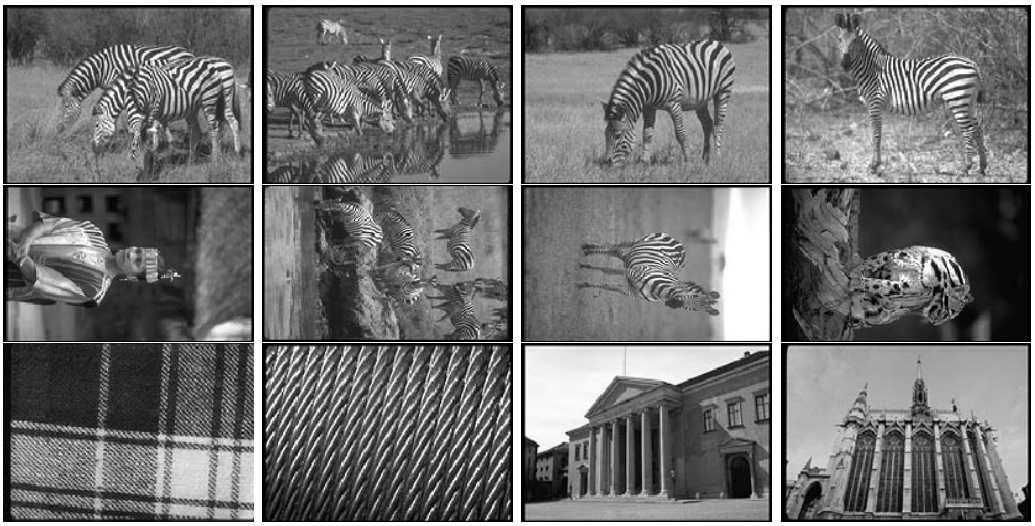

Figure 62: The results of the retrieval experiment of Section 3.5 classified with respect to orientation distribution entropy.

with respect to the added feature (c_{n+1}) of orientation distribution entropy (as discussed in Section 3.4.1). In the figure, the bottom row illustrates images for which the entropy of the distribution exhibited a low overall value. In the two rows above, the opposite case is shown. It is observed that after the introduction of c_{n+1}, the image collection was classified into two classes: one that exhibits low variance in orientation values of the targeted pattern (the black and white stripes) and another which exhibits higher variance, with respect to the same feature. In this case, the first class corresponds to straight / parallel stripes and the second to locally parallel stripes which change directions over space.

Experiments The heterogeneity of retrieval goals in combination with the lack of knowledge about how visual similarity is determined cast difficult the development of a generic, with respect to goal and image type, visual content similarity matching method. As a compromise, the development of utilities and competences for the partial management of visual content was proposed (see discussion at the introduction of this chapter and Section 1.5). In this paragraph, the hierarchical classification of images is discussed in the context of the development of utilities for the browsing and classification of image

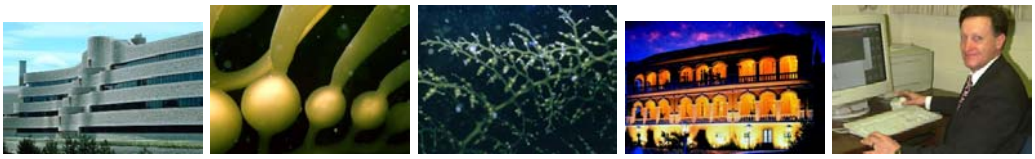

Figure 63: A random subset of images from a collection.

Figure 64: A random subset of images from the classified subset MM.

collections.

The ability to hierarchically browse and classify images with respect to similarity of visual features is proposed as an interactive method for multi-feature query formulation. To illustrate this, a demonstration of hierarchical classification of visual content is presented that facilitates *(a)* the discrimination of visually similar image content *(b)* the evaluation of the suitability of visual classifiers with respect to the goal of classification.

In the following example, a random image set of 800 images was used, taken from a larger collection[17], of which a random subset is illustrated in Figure 63. Initially, the existence of perceptual groups of perspectively converging long line segments was used as a visual attribute of man-made structures (class MM). A first classification into two image classes (man-made structures and other) was quite successful, as already reported in the literature [64]. Some elements of class MM are presented in Figure 64. Subsequently, the convex hull of perceptual groups in MM was computed and color features were detected within these regions. For each one of these image regions, the HSV color histogram was computed for the pixels within them. Elements in MM were sorted with respect to color distribution similarity for some selected samples. The top matches with respect to the quadratic color histogram distance [38] for two such cases, are shown in Figures 65 and 66

[17]The $IMSI$ Master Photos commercial collection was used. See `http://www.imsisoft.com` for more details.

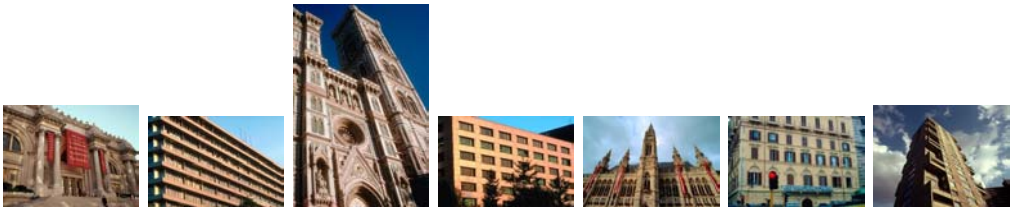

Figure 65: Image classification based on color features of perceptual groups.

Figure 66: Image classification based on color features of perceptual groups.

presenting the color distribution prototype image first, on the left.

A second classification of elements of class MM was iteratively performed, using a heuristic feature which was computed with the help of the method, discussed in [57]. Repeated elements were detected within the bounds of the convex hull of the perceptual groups in MM. The feature devised was $F = C \cdot A$, where C was the mean compactness of elements[18] and A the ratio of element area over the spatial extent of the perceptual group (determined by the convex hull of the points composing the line segments). The result of the classification of images in Figure 65 is illustrated in Figure 67. In the top row of the figure, the elements shown are the ones that exhibit a strong response to the heuristic criterion, while the second row illustrates the opposite case. As observed, the criterion selects images that exhibit compact repeated elements.

Similarly to the classification demonstrated in Section 4.2, Figure 68 illustrates another possible classification of elements of class MM. In the figure, the browsed images exhibit a vertical orientation of the linear perspective perceptual group. Further classification may be performed based on color features or repeated elements as previously demonstrated.

Another example of visual content classification is demonstrated in Figure 69. In this example, size and compactness of repeated elements was taken into account. The images illustrated all contain repeated elements of

[18]Estimated as the ratio of the squared perimeter and the area of the element's boundary.

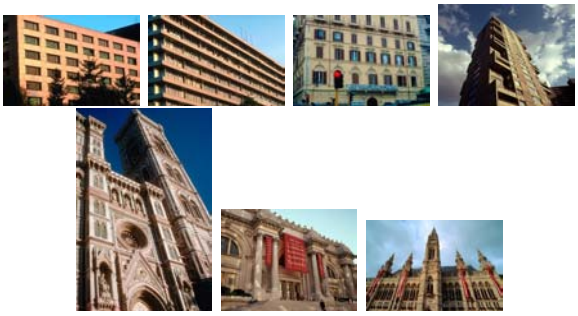

Figure 67: Image classification based on the relative size of repeated elements in comparison with the area occupied by perceptual group earlier detected.

Figure 68: Image classification based on the vertical orientation of the linear perspective perceptual group.

Figure 69: Browsing of images that contain repeated elements of large compactness that are scattered over the whole image area. Further classification was carried out with respect to the size of the elements. Images containing relatively small elements are shown in the top row.

high compactness[19] scattered over the image. Further classification was performed with respect to the actual size of the repeated elements. Repeated elements of minor size reside on the top row of the figure, while the bottom row presents some elements occupying a larger area in the image. As in the previous cases, further classification of images may be performed for other features such as color, which in this case could also have a semantic interpretation (e.g type of vegetable etc.).

The classification of content with respect to scale, which was discussed in Chapter 2, is illustrated in Figure 70. In this example, content was initially classified with respect to scale and images that exhibited a low entropy of orientation distribution are presented. As already discussed, lack of orientation entropy corresponds to increased directional order in the observed image, as observed in the browsing result.

In some cases, database queries may be formulated so that they would initially restrict search for similar distributions at coarse scales, where the image is more abstract and dominant spatial features prevail. Images that do not have a similar visual content at coarser scales may be excluded from the query space, while the search could be continued for the rest of the images at a finer scale. This would result in a reduced computational cost for queries, for which coarser scales exhibit a dominant significance. Inversely, and depending on the field of application (e.g. texture similarity), detail may be

[19]High compactness is exhibited by shapes that are approximately identical to the perfect circle.

Figure 70: Browsing of images exhibiting coarse scale structure and orientational order.

primarily investigated by focusing on fine scale structure. The classification of visual content may be used in this procedure. In general, the use of color, shape, texture etc. can be arbitrary combined in such browsing sessions, however the open issue that remains is the lack of knowledge concerning the psychophysical significance of each similarity component.

Conclusions In the examples presented, images were classified utilizing an interactive procedure of feature selection, which facilitates the empirical evaluation of the suitability of visual classifiers with respect to the goal of classification. The demonstrated procedure may be used for the development of browsing and classification utilities for image collections. Using such a procedure, characteristic image classes that contain similar visual content can be distinguished. From a certain perspective, the successive restriction of the data set using classifiers may be regarded as a method for interactive query formulation, which yields a result by iteratively restricting the query search space. In addition, it was demonstrated that the classification may focus not only on "traditional" visual features, but can be also performed with respect to feature scale of observation (as also discussed in Section 2.2.5), thus contributing to the refinement of queries.

The development of goal oriented classification strategies is claimed to contribute to the formulation of a palette of content matching behaviors. The selective activation of the latter is argued to be capable of founding the basis of a content-based visual information system.

5.3 Summary

This chapter was focused on the ability to browse and retrieve images based on their content, by integrating features and description methods. In particular, two methods for the content-based image browsing and retrieval were

proposed that integrate more that one content descriptor.

The first method introduced consists of a proposal for visual query formulation. The key points of the approach are that it makes use of individual visual content components, facilitates the integration of features, and can be compiled into a structured database query language. An implementation demonstrated in this chapter consists of a user interface of high specificity, which enables the presentation of image descriptors and visual entities that are available to be used in queries and provides the end user with the ability to analytically specify the predicates of a visual query. Such predicates may concern the visual entities that were detected in an image as well as the spatial relationships between them. This selection of predicates is the key point of this method and contributes to the increase of specificity of the visual queries.

The second method utilizes the sequential introduction of image classification criteria, in order to interactively yield a hierarchical classification of an image collection. This type of classification corresponds to an interactive query formulation, which permits the end user to sequentially narrow the image search space, until it contains only the targeted images. The iterative introduction of criteria facilitates their integration, as well as their interactive selection. The ability to introduce one criterion at a time make possible the evaluation of each individually and, thus, the ability to interactively select such criteria. The method was demonstrated by presenting experimental results, based on the description methods and similarity criteria introduced in the previous chapters.

6 Discussion

In this chapter, conclusions and research contributions presented in this dissertation are summarized, followed by the proposals for future work.

6.1 Summary

In this subsection, the conclusions and research contributions that were presented in this work are summarized, by order of appearance.

Description of approach In the introductory section of this dissertation, the approach towards the content-based retrieval of images which was adopted in this work was described. The key points of the approach were the modular design of a system that would integrate several specialized content-based image description and retrieval mechanisms, which could be integrated at a later stage. Another significant point that was discussed was the independence of such mechanisms from context-related knowledge, which was motivated by the ability to selectively activate and tailor the derived mechanisms to match specific application requirements. Most important though, this independence stems from the goal to derive content-based description and matching mechanisms that are relevant to human visual perception and appropriate for its further study.

Primitive visual features In the second chapter, issues related to the perceptually relevant description of the simplest components of visual content, namely primitive visual features, were addressed. Interest in this domain of image properties originates from the typical appearance of this feature family in almost all visual comprehension tasks. In content-based browsing and retrieval of images, meaningful features are composed from the perception of primitive content elements. A common factor, recognized from the review of related work on the topic, is the image scale at which primitive features occur. Special emphasis was given to the consideration of this factor, motivated by its perceptual relevance.

More specifically, a study of the image scale-space demonstrated that visual features occur in different scales. Based on this observation and inspired from the physiology of early vision, the Scale Summarized Representation (SSR) framework was introduced in the second chapter. The proposed representation utilizes the scale-normalization of feature detection response

functions to accumulate visual feature information from multiple scales. This accumulation weights the contribution of each scale proportionally to the value of the scale-normalized feature detection response function, independently for each image point. The resulting representation summarizes the contents of a scale-space over some range of scales into a single image, for a variety of feature types.

The SSR was demonstrated to be applicable in the following problems: *(a)* the consideration of more than one scale in the description of image content, *(b)* the classification of image content with respect to the scale of observation and, thus, the refinement of image description, *(c)* the enhancement of the quality of scale-selection results, *(d)* the reduction of memory capacity requirements for the multiscale analysis of images, and *(e)* other image processing applications. An added-value contribution is that the SSR provides a generic, with respect to type of primitive feature, framework for the representation of primitive visual features with respect to scale. Several types of primitive features were used to demonstrate the above cases, such as edges, corners, grayscale blobs, color blobs, etc. Finally the SSR can be computed in parallel and is, thus, a qualified candidate for use in real-time applications. A more detailed summary of the contributions on these issues is presented below.

The consideration of multiple scales (a) in image description was based on the observations that *(i)* the visual content of an image may vary with respect to the scale of observation and that *(ii)* given some image, more than one scale may be of interest to describe for each point of that image. Thus, in order to be able to acquire a description that captures all meaningful cases, the whole image scale-space has to be taken into account. By focusing the scale-summarization on specific ranges of scales, the content of those was independently acquired. In addition, by sorting such representations with respect to the range of scales, from which they were obtained, the classification of image content with respect to scale was possible (b). This classification can be used for the refinement of image description by attaching the attribute of scale of observation to the image features detected. Thus, visual queries can potentially focus on specific ranges of image scales, e.g. targeting image detail or coarse scale features.

Another issue that was discussed in the second chapter was the fact that the process of explicitly selecting a single scale is subject to the effect of noise. Typically, this effect is inversely proportional to the number of scales utilized for the scale-space image description. Using the SSR it was possible

to acquire smooth results, using a just few scales (c), due to the averaging nature of the method. Also, the fact that the result of scale-summarization collapses a whole range of scales into a single image contributes to the reduction of memory capacity requirements of multiscale image description algorithms (d). Finally, the SSR was demonstrated to be applicable in several image processing algorithms, where processing the image data with respect to scale, or size, of local structure was of interest (e).

Finally, in the last part of the second chapter, environmental properties that can be derived from primitive image features were investigated and their role in the general context of the description of visual information was discussed.

Spatial arrangements of primitive visual features In the third chapter of this dissertation, the study focused on the description of the spatial arrangement of primitive visual features.

Initially, the descriptive importance of image regions of coherent spatial feature distribution was highlighted, showing that the ability to describe and extract such regions is informative concerning the description of image content. In addition, the requirements of a representation of the spatial arrangements of primitive features were estimated, emphasizing its storage capacity requirements. Next, the SSR framework for primitive visual content representation with respect to scale, as proposed in the second chapter, was extended for spatial arrangements of primitive features. This extension inherits the computational and descriptional properties of the SSR that are related to the execution time and memory optimization, as well as scale selection.

In the third section of this chapter, the proposed representation was utilized in the region extraction and similarity matching of arrangement descriptors. In particular, the ability to describe the spatial arrangement of primitive features with respect to scale led to the formulation of a scale-normalization method for local descriptors of such arrangements, based on the SSR framework. Using this scale-normalization method, descriptors that were associated with visually similar spatial arrangements, but which occurred at different scales, yielded similar descriptions. This similarity of description was utilized in the clustering of scale-normalized local descriptors, which resulted in the extraction of image regions that exhibit scale-varying, but otherwise constant arrangement of primitive features.

The description of spatial arrangements of feature distributions was furthermore enhanced, by extracting attributes of the local descriptors of the spatial arrangement of primitive features. In the experiments, the local descriptors that were used were local histograms of intensity and orientation. Using attributes such as the principal components and the information entropy of the feature distribution within a local description, the extraction of characteristic and perceptually significant visual properties of spatial arrangements of primitive features was demonstrated. In addition, the attributes of the local descriptors that were extracted were directly mapped to visual image properties. For this reason, the enhanced description of spatial arrangements of primitive features that was derived by attributes of local descriptors was also comprehensible by humans and, thus, characterized as perceptually relevant. In addition, attributes of local descriptors were utilized in the formulation of analytic content-based visual queries and relevant results were presented.

Perceptual Organization In the fourth chapter of this dissertation, the discussion was focused on the topic of perceptual organization, referring to the grouping of low-level pieces of visual information into larger units of perceived objects and their interrelations. In particular, two methods were presented, the first dealing with the detection of perspectively converging, gradient-derived image features and the second with the description of, already grouped, silhouette boundaries . The resulting descriptions from both methods were utilized in content-based retrieval of visual information.

More specifically, the method that was introduced for the grouping of perspectively converging gradient-derived image features was initially formulated for linearly perspective line segments. Furthermore, it was generalized for subjective line segments formed by collinearly arranged point-like features, such as corners. The method utilized the detection of the approximate convergence to a single point of three dominant, with respect to length and contrast, line segments in order to form a hypothesis. Subsequently, the set of line segments that was extracted from the image was investigated for supporting or contradicting evidence to the formed hypothesis. The hypotheses that were supported from such evidence were used to cluster perspectively converging line segments, which were characterized as perceptual groups and assigned with appearance-related attributes. Such perceptual groups were utilized in the content-based retrieval of images and the assigned attributes

in the further classification of the retrieval results. The contribution of this grouping method, in comparison to others found in the literature, is the detection of parallel segments under the projective transform.

The second method that was introduced in this chapter is concerned with silhouette boundaries, since they are a class perceptual groups of significant descriptive power, in visual content description, object recognition, and visual impression. The problem of selecting corresponding (anchor) reference points when comparing shapes was addressed and a salience metric was proposed for such a selection. Such points were selected among the curvature extrema of silhouette boundaries, since they are reported to be perceptually significant and were shown to be computationally more reliable. The analysis of curvature with respect to scale was used to locate such boundary points, which were then used for the perceptually relevant, piecewise decomposition of silhouette boundaries. This decomposition was further utilized in the piecewise alignment of such boundaries, which provided a tool for their similarity matching. In particular, a dissimilarity metric was formulated that accumulates the displacement of aligned (corresponding) boundary pieces. This metric was utilized in the introduction of a silhouette boundary matching algorithm, which was tested on a small database.

Visual information browsing In the fifth chapter, the discussion was focused on the ability to browse and retrieve images based on their content. In this chapter, emphasis was placed on the integration of features and description methods. In particular, two methods for the content-based image browsing and retrieval were proposed that integrate more than one content descriptor.

The first method consists of a proposal for visual query formulation that makes use of individual visual content components, facilitates the integration of features, and can be compiled into a structured database query language. A user interface of high specificity was proposed that *(i)* enables the presentation of the image descriptors and visual entities that are available to be used in queries and *(ii)* provides the end user with the ability to analytically specify the predicates of a visual query. Such predicates may concern the visual entities that were detected in an image as well as the spatial relationships between them. The benefit from such an approach is that the end user explicitly defines the features of an image that are of interest, and based on this, controls which images should be retrieved.

The second method utilizes the iterative introduction of image classification criteria in order to interactively yield a hierarchical classification of an image collection. This type of classification corresponds to an interactive query formulation, which permits the end user to sequentially narrow down the image search space until it contains only the targeted images. In this method, the integration of features was achieved by iteratively introducing the classification criteria. The method was demonstrated by presenting experimental results based on the description methods and similarity criteria introduced in the previous chapters.

6.2 Future work

In this section, intentions for future work are discussed. Initially, extensions to the methods presented in the previous chapters are proposed and, subsequently, future research directions are outlined.

6.2.1 Extensions

In this subsection, work that would extend the methods that were introduced in the previous chapters is proposed.

Concerning the scale-summarization of primitive features the study of novel feature detectors is proposed. In particular, gradient-derived feature detectors could be devised for the detection of visual features related to properties of the environment, such as e.g. image gradient due to cast shadows. In addition, it would be interesting to test the applicability of the method and experiment with other types of "feature detectors" or filter banks that are encountered in the visual systems of primates, such as Gabor functions or spectral response functions in cones. The work proposed would in these cases deal with the description of such features with respect to scale.

Furthermore, the study of novel feature detectors could be applicable to the scale-summarization and description of spatial arrangements of primitive features. In particular, the study of filter banks exhibiting a derivative like structure is proposed for experimentation in order to be able to detect complex structures, instead of deriving such descriptions from the integrated use of intensity and orientation descriptors. Thus local descriptors could be devised which are able to yield similar descriptions for complex texture patterns, which vary at scale. In addition, the clustering of scale-normalized local descriptors of feature arrangement is proposed to be tested and demon-

strated with more efficient algorithms than the K-means image segmentation algorithm, which are based on graph-partitioning.

Regarding the perceptual grouping of features it is argued that future work should be generally focused on the integration of several perceptual grouping rules. With respect to the presented algorithm, which deals with the grouping of linearly perspective image features, future work could involve the extraction of depth cues concerning the structure of the environment. In addition, the study of such cues in the perception of shape could be also of interest. More specifically, rules of shape constancy could be explored in order to derive a perceptually relevant description of shape given the context within it appears. Thus shapes or contours could be compared after normalization with respect to pose, instead of trying to estimate a transform that could possible correlate two shapes that are candidates for matching.

Finally, ample room exists for future work concerning the integration of visual content descriptors. In addition, ways of utilizing context-related knowledge, as well as similarity metrics that represent human perception would be quite useful to explore. Such issues, are further discussed in the next subsection.

6.2.2 Research directions

In this subsection, future research directions that would extend the work presented in this dissertation are proposed. Our interest for future research is focused on the ability to propose more effective methods of integrating visual cues and similarity criteria, integrating knowledge about the functionality of the visual system acquired from psychophysical experiments, and integrating top-down information into content-based image retrieval methods.

Voting A voting approach is proposed as an experimental tool for the problem of image similarity estimation. The interaction and merging of different similarity modules can be studied through autonomous agent modeling of different similarity matching behaviors, casting their votes regarding the similarity of images. Each agent's vote is based on its specialized knowledge of image content and some similarity assessment method. Depending on both image content and the comparison task, different behaviors should be activated. An experimental platform featuring voting procedures using a

variety of voting systems was implemented[20] for the study of such behaviors.

Ways of integrating information sources include:

- *dominance*, where one agent dominates over all others and which is implemented with various versions of majority voting,
- *compromise*, where the solution may not be necessarily consistent with the majority of votes, but attempts to accommodate the preferences of all voters at the highest possible degree (this type of integration was implemented based on certain variations of preferential voting, in which each voter casts a prioritized list of votes with the possibility of weighting their importance as well).
- *interaction*, where the result is reached after the convergence of information sources (this type of voting was implemented using consecutive voting rounds).

The voting infrastructure may be used in the evaluation of different content similarity evaluation strategies, exploiting the dynamic formulation of the set of voters and thus yielding a system where a variety of experiments may take place, with minor technical effort. Most importantly, it is intended to be used as a tool in order to approximate human similarity perception by selecting a voting method that would select the same similarity correlated images as subjects would do in psychophysical experiments.

Psychophysical experimentation Controlled psychophysical experiments are proposed as a method to objectively evaluate the effectiveness of similarity matching methods. More specifically, the ability to measure the opinion of human subjects about the similarity of images or specific image features under controlled experiments is proposed as a way derive "laws" of visual similarity. In such experiments, the role of context should be de-emphasized or optimally isolated, probably with the use of unidentifiable visual stimuli (meaning that they are not recognized as familiar objects). Furthermore, in order to be able to measure the impact that specific image properties have on perceived visual similarity, the stimuli that are to be presented to subjects should always isolate the visual property of interest. However, this

[20]Georgios Ch. Chalkiadakis: "An Agent-Based Architecture for the Conduction of Voting" , Master of Science Thesis, Department of Computer Science, University of Crete, October 1999.

would not be adequate for investigating the similarity judgment derived by taking more than one similarity criteria into account and, thus, further experiments should then be conducted in order to investigate the integrated effect of such criteria. In addition, the groups subject to such experiments should be appropriately selected so that the biased or misleading conclusions are avoided.

The conduction of psychophysical experiments is also proposed as a method to: *(a)* design or approximate perceptually relevant similarity measures, *(b)* supervise the training of machine learning methods and in this way learn a similarity matching strategy instead of trying to design one, and *(c)* explore the visual components that are relevant with respect to context or semantics as well as about their significance.

Grammars Research concerning the integration of visual content descriptors could be conducted in terms of pursuing a "visual grammar" that would provide information about how perception combines visual cues into meaningful percepts. If such a grammar becomes available then more meaningful and descriptive visual entities could be extracted from an image.

Context Since context-related knowledge and observation goal are key factors to the appreciation of content-based image retrieval results, further research that would lead to the derivation of methods for integrating such knowledge into visual queries is proposed. As a first step towards this direction, the design of software agents which are specialized in the performance of specific tasks and also exhibit the capability of adapting their behavior to the preferences of specific end users or user groups is proposed.

References

[1] C. Adami. *Introduction to Artificial Life.* Springer Verlag, 1998.

[2] R. Adams and L. Bischof. Seeded region growing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(6):641–647, 1994.

[3] L. Alvarez, F. Guichard, P.-L. Lions, and J.-M. Morel. Axioms and fundamental equations of image processing. *Archive Rational Mechanics and Analysis*, 123(3):199–257, September 1993.

[4] H. Asada and M. Brady. The curvature primal sketch. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(1):2–14, January 1986.

[5] A.P. Ashbrook, P.I. Rockett, and N.A. Thacker. Multiple shape recognition using pairwise geometric histogram based algorithms. In *IC-IPA95*, 1995.

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

[7] D. H. Ballard. Generalizing the hough transform to detect arbitary shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8:111–122, 1981.

[8] F. Bergholm. Edge focusing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 9:726–741, 1987.

[9] K. Berthold, P. Horn, and M. J. Brooks. *Shape from Shading.* The MIT Press, Cambridge, MA, 1989.

[10] C. Blakemore and P. Sutton. Size adaptation: a new aftereffect. *Science*, 166:245–247, 1969.

[11] H. Blum. Biological shape and visual science (part i). *J. Theoretical Biology*, 38:205–287, 1973.

[12] F. L. Bookstein. Shape and the information in medical images: A decade of morphometric synthesis. *Computer Vision and Image Understanding*, 66(2):97–118, 1997.

[13] K. L. Boyer and S. Sarkar. Integration, inference, and management of spatial information using bayesian networks: Perceptual organization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(3):256–274, 1993.

[14] D. H. Brainard and B. A. Wandell. *A bilinear model of the illuminant's effect on color appearance.*, pages 1433–1448. MIT Press, Cambridge, Massachusetts, 1991.

[15] D. H. Brainard and B. A. Wandell. Asymmetric color matching: how color appearance depends on the illuminant. *Journal of the Optical Society of America.*, 9:1433–1448, 1992.

[16] B. Brookes. The foundations of information science. *Journal of Information Science*, 2:125–133, 1980.

[17] H. Buffart, E. Leeuwenberg, and F. Restle. Coding theory of visual pattern completion. *American Journal of Experimental Psychology : Human Perception and Performance*, 7:241–274, 1981.

[18] A. Califano and R. Mohan. Multidimensional indexing for recognizing visual shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16:373–392, 1994.

[19] J. Canny. A computiational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(6):679–698, November 1986.

[20] C. Carson and et al. Blobworld: A system for region-based image indexing and retrieval. In *Third Int. Conf. on Visual Information Systems*, Amsterdam, June 1999.

[21] S. Cha and S. N. Srihari. Distance between histograms of angular measurements and its application to handwritten character similarity. In *Proceedings of 15th ICPR 2000*, pages 21–24, Barcelona, Spain, September 2000.

[22] S. K. Chang, Q. Shi, and C. Yan. Iconic indexing by 2-d string. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 9(3):413–428, 1987.

[23] C. Cleverdon, J. Mills, and M. Keen. *Factors Determining the Performance of Indexing Systems: ASLIB Cranfield Research Project. Volume 1: Design.* ASLIB Cranfield Research Project, Cranfield, 1966.

[24] P. E. Danielsson. A new shape factor. *Computer Graphics and Image Processing*, 7:292–299, 1978.

[25] R. De Valois and K. De Valois. *Spatial Vision.* Oxford Science Publications, Oxford, 1988.

[26] R. De Valois and K. De Valois. A multi-stage color model. *Vision Research*, 33(8):1053–1065, 1993.

[27] R. L. De Valois and G. H. Jacobs. Primate color vision. *Science*, 162:533–540, 1968.

[28] D. Fisher. Conceptual clustering, learning from examples, and inference. In *Proceedings of the 4th International Workshop on Machine Learning*, 1987.

[29] P. Forte and D. Greenhill. A scalespace approach to shape similarity. In *Scale-Space Theory in Computer Vision, Proc. 1st International Conference*, Lecture Notes in Computer Science, Utrecht, The Netherlands, July 1997. Springer-Verlag.

[30] Y. Gdalyahu, D. Weinshall, and M. Werman. Stochastic image segmentation by typical cuts. In *Proceedings of the Computer Vision and Pattern Recognition*, Fort Collins, Colorado, 1998.

[31] J. Geusebroek, D. Koelma, A. Smeulders, and Th. Gevers. Image retrieval and segmentation based on color invariants. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Hilton Head Island, South Carolina, June13–15 2000. IEEE Computer Society Press.

[32] T. Gevers and A. Smeulders. Pictoseek: A contentbased image search system for the world wide web. In *Proceedings of Visual 97*, pages 93–100., Knowledge Systems Institute, Chicago, 1997.

[33] J. J. Gibson. *The Perception of the Visual World.* Houghton-Mifflin, Boston, 1950.

[34] J. J. Gibson. *The ecological approach to visual perception.* Houghton Mifflin, Boston, 1979.

[35] A. G. Goldstein. Judgments of visual velocity as a function of length of observation time. *Journal of Experimental Psychology*, 54:457–461, 1957.

[36] G. Grandlund. Fourier processing for handprinted character recognition. *IEEE Trans. on Computing.*, C-21:195–201, 1972.

[37] A. Guttman. R-trees: A dynamic index structure for spatial searching. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 47–57, Boston, MA, June 1984.

[38] J. L. Hafner, H. S. Sawhney, W. Equitz, M. Flickner, and W. Niblack. Efficient color histogram indexing for quadratic form distance functions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(7):729–736, 1995.

[39] A. Hampapur, A. Gupta, B. Horowitz, C. F. Shu, C. Fuller, Bach J. R., M. Gorkani, and R. Jain. Virage video engine. In *Storage and Retrieval for Image and Video Databases (SPIE)*, pages 188–198, 1997.

[40] R. M. Haralick. Statistical and structural approaches to texture. In *Proceedings of IEEE, 67(5)*, pages 786–804, 1979.

[41] C. Harris and M. Stephens. A combined corner and edge detector. In *Proc. 4th Alvey Vision Conference*, pages 147–152, 1988.

[42] E. Herring. *Outlines of a Theory of the Light Sense.* Harvard University Press, Cambridge, MA, 1878/1964.

[43] D. D. Hoffman and W. A. Richards. Parts of recognition. *Cognition*, 18:65–96, 1984.

[44] J. Hong and X. Tan. Recognize the similarity between shapes under the affine transformation. In *Proceedings of the 2nd International Conference on Computer Vision*, pages 489–493, Tarpon Springs, Florida, December 5–8 1988. IEEE Computer Society Press.

[45] P. V. C. Hough. Methods and means for recognizing complex curves. U.S. Patent 3 069 654, 1962.

[46] D. Hubel and T. Wiesel. Receptive fields of single neurones in the cat's striate cortex, 1959.

[47] T. Iijima. Basic theory on normalization of a pattern (in case of typical one-dimensional pattern). *Bulletin of Electrotechnical Laboratory*, 26:368–388, 1962. (in Japanese).

[48] Q. Iqbal and J. K. Aggarwal. Applying perceptual grouping to content-based image retrieval: Building images. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 42 – 48, Fort Collins, Colorado, June 1999.

[49] D. J. Jobson, Z. Rahman, and G. A. Woodell. A multi-scale retinex for bridging the gap between color images and the human observation of scenes. *IEEE Transactions on Image Processing: Special Issue on Color Processing*, 6(7):965–976, 1997.

[50] P. Kelly and M. Cannon. Query by image example: the candid approach, 1995.

[51] B. B. Kimia, A. R. Tannenbaum, and S. W. Zucker. Shapes, shocks, and deformations i: The components of two-dimensional shape and reaction-diffusion space. *International Journal of Computer Vision*, 15:189–224, 1995.

[52] M. Kliot and E. Rivlin. Invariant-based shape retrieval in pictorial databases. *Computer Vision and Image Understanding*, 71(2):182–197, 1998.

[53] J. J. Koenderink and A. J. van Doorn. The structure of locally orderless images. *International Journal of Computer Vision*, 31(2/3):159–168, 1999.

[54] E. H. Land and J. J. McCann. Lightness and retinex theory. *Journal of the Optical Society of America*, 61(1):1–11, January 1971.

[55] Y. C. Leclerc. Constructing simple stable descriptions for image partitioning. *International Journal of Computer Vision*, 3:73–102, 1989.

[56] E. Leeuwenberg. A perceptual coding language for visual and auditory patterns. *American Journal of Psychology*, 84(3):307–349, 1971.

[57] T. Leung and J. Malik. Detecting, localizing and grouping repeated scene elements from an image. In *Fourth European Conf. Computer Vision*, Cambridge, England, 1996.

[58] T. Lindeberg. *Scale-Space Theory in Computer Vision*. The Kluwer International Series in Engineering and Computer Science. Kluwer Academic Publishers, Boston, USA, 1994.

[59] T. Lindeberg. Feature detection with automatic scale selection. Technical Report ISRN KTH/NA/P--96/18--SE, Dept. of Numerical Analysis and Computing Science, KTH, May 1996.

[60] T. Lindeberg and M. X. Li. Segmentation and classification of edges using minimum description length approximation and complementary junction cues. *Computer Vision and Image Understanding*, 67(1):88–98, 1997.

[61] M. Livingstone and D. Hubel. Segregation of form, color, movement, and depth: anatomy, physiology, and perception. *Science*, 240:740–749, 1988.

[62] D. G. Lowe. *Perceptual Organization and Visual Recognition*. Kluwer, Boston (MA), 1985.

[63] H. Q. Lu and J. K. Aggarwal. Applying perceptual organization to the detection of man-made objects in non-urban scenes. *Pattern Recognition*, 25(8):835–853, 1992.

[64] H.Q. Lu and J.K. Aggarwal. Applying perceptual organization to the detection of man-made objects in non-urban scenes. *Pattern Recognition*, 25:835–853, 1992.

[65] L. T. Maloney and B. A. Wandell. Color constancy: a method for recovering surface spectral reflectance. *Journal of the Optical Society of America-A*, 3(1), January 1986.

[66] G. Medioni, M. Lee, and C. Tang. *A Framework and for Segmentation and Grouping*. Elsevier Science, New York, 2000.

[67] F. Mokhtarian and S. Abbasi. Curvature scale space for shape similarity retrieval under affine transforms. In *8th International Conference*

on Computer Analysis of Images and Patterns, Ljubljana, Slovenia, September1–3 1999.

[68] F. Mokhtarian and A. Mackworth. Scale-based description and recognition of planar curves and two-dimensional shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(1):34–43, January 1986.

[69] J. D. Mollon and P. G. Polden. Post-receptoral adaptation. *Vision Research*, 19:35–40, 1979.

[70] T. Moons, E. J. Pauwels, L. J. van Gool, and A. Oosterlinck. Foundations of semi-differential invariants. *International Journal of Computer Vision*, 14:25–47, 1995.

[71] N. M. Nasrabadi and R. A. King. Image coding using vector quantization: A review. *IEEE Transaction on Communications*, 36(8):957–971, 1988.

[72] S. A. Nene, S. K. Nayar, and H. Murase. Columbia object image library: COIL-100. Technical report, Department of Computer Science, Columbia University, 1996. Technical Report CUCS-006-96.

[73] W. Niblack, R. Barber, W. Equitz, M. Flickner, E. Glasman, D. Petkovic, P. Yanker, Faloutsos C., and G. Taubin. The qbic project: Quering images by content using color, texture, and shape. In *SPIE Conference on Storage and Retrieval for Image and Video Databases*, pages 173–187, San Jose, CA, February 1993.

[74] H. Nyquist. Certain topics in telegraph transmission theory. *AIEE Trans.*, 47, 1928.

[75] N. R. Pal and S. K Pal. A review on image segmentation techniques. *Pattern Recognition*, 26:1277–1294, 1993.

[76] S. E. Palmer. Common region: A new principle of perceptual grouping. *Cognitive Psychology*, 24:436–447, 1992.

[77] S. E. Palmer and I. Rock. On the nature and order of organizational processing: A reply to peterson. *Psychonomic Bulletin and Review*, 1:515–519, 1994.

[78] S. E. Palmer and I. Rock. Rethinking perceptual organization: The role of uniform connectedness. *Psychonomic Bulletin and Review*, 1(1):29–55, 1994.

[79] E. J. Pauwels, T. Moons, L. J. van Gool, P. Kempenaers, and A. Oosterlinck. Recognition of planar shapes under affine distortion. *International Journal of Computer Vision*, 14:49–65, 1995.

[80] A. Pentland, R. Picard, and S. Sclaroff. Photobook: Tools for content-based manipulation of databases, 1994.

[81] A. Pentland, R. W. Picard, and S. Sclaroff. Photobook: Content-based manipulation of image databases. *International Journal of Computer Vision*, 18(3):233–254, 1996.

[82] P. Perona and J. Malik. Scale-space and edge detection using anisotropic diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(7):629–639, 1990.

[83] E. Peterhans and R. von der Heydt. Mechanisms of contour perception in monkey visual cortex. ii. contours bridging gaps. *Journal of Neuroscience*, 9(5):1749–1763, 1989.

[84] W. Richards, B. Dawson, and D. Whittington. *Encoding Contour Shape by Curvature Extrema*, pages 83–98. MIT Press, Cambridge, MA, 1988.

[85] J. Rissanen. *Stochastic Complexity in Statistical Inquiry*. World Scientific, Singapore, 1989.

[86] K. Rohr. Modelling and identification of characteristic intensity variations. *Image and Vision Computing*, 10(2):66–76, 1992.

[87] K. Rohr. Localization properties of direct corner detectors. *Journal of Mathematical Imaging and Vision*, 4:139–150, 1994.

[88] W. S. Rorgerson. Multidimensional scaling of similarity. *Psychometrika*, 30:379–393, 1965.

[89] S. Santini. Exploratory interfaces for visual information systems. In *Proceedings of Vision Interface '99*, Trois Rivieres, Quebec, CA, 1999.

[90] S. Santini and R. Jain. Similarity matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(9):871–883, 1999.

[91] G. Sapiro and A. Tannenbaum. Affine invariant scale-space. *International Journal of Computer Vision*, 11(1):25–44, 1993.

[92] G. Sapiro and A. Tannenbaum. Area and length preserving geometric invariant scale-spaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(1):67–72, January 1995.

[93] S. Sarkar and K.L. Boyer. Perceptual organization in computer vision: A review and a proposal for a classificatory structure. *IEEE Transactions on Systems, Man, and Cybernetics*, 23:382–399, 1993.

[94] S. Sarkar and K.L. Boyer. Computing perceptual organization in computer vision. In *World Scientific*, pages ISBN: 981–02–1832–X, 1994.

[95] E. L. Schwartz. Spatial mapping in the primate sensory projection: Analytic structure and relevance to perception. *Biological Cybernetics*, 25:181–194, 1977.

[96] J. A. Sethian. *Level Set Methods.* Cambridge Monograph on Applied and Computational Mathematics. Cambridge University Press, 1996.

[97] J. A. Sethian. Fast marching methods. *SIAM Review*, 41:199–235, 1999.

[98] M. Shallis. *On time.* Burnett Books, London, 1982.

[99] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423 and 623–656, July and October 1948.

[100] R. Shapley and V. H. Perry. Cat and monkey retinal ganglion cells and their visual functional roles. *Trends in Neurosciences. Specias Issue: Information processing in the retina.*, 5(9):229–235, 1986.

[101] J. Shi and J. Malik. Normalized cuts and image segmentation. In *Proc. of IEEE CVPR Puerto Rico*, pages 731–737, June 1997.

[102] A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta, and R.C. Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, 2000.

[103] J. Smith and S. Chang. *Intelligent Multimedia Information Retrieval*, chapter Querying by color regions using the VisualSEEk content-based visual query system. IJCAI, AAAI Press, 1997.

[104] J. Sporring, X. Zabulis, P. E. Trahanias, and S. C. Orphanoudakis. Shape similarity by piecewise linear alignment. In *ACCV , Taipei, Taiwan*, pages 306–311, January 2000.

[105] C. Tomasi and T. Kanade. Detection and tracking of point features. Technical Report CMU-CS-91-132, Carnegy Mellon University, School of Computer Science, Pittsburgh, PA 15213, April 1991. Part 3.

[106] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: a factorization method. *International Journal of Computer Vision*, 9:137–154, 1992.

[107] P. Van der Helm and E. Leeuwenberg. Accessibility: A criterion for regularity and hierarchy in visual pattern code. *American Journal of Mathematical Psychology*, 35:151–231, 1991.

[108] J. Weickert, S. Ishikawa, and A. Imiya. On the history of Gaussian scale-space axiomatics. In Jon Sporring, Mads Nielsen, Luc Florack, and Peter Johansen, editors, *Gaussian Scale-Space Theory*, chapter 4, pages 45–59. Kluwer Academic Publishers, Dordrecht, The Netherlands, 1997.

[109] R. B. Welch. *Handbook of perception and visual performance*, volume 1, chapter Adaptation of space perception, pages 24.21–24.45. Wiley, New York, 1986.

[110] M. Wertheimer. *A sourcebook of Gestalt phychology*, chapter 1. Gestalt Theory, pages 1–11. The Humanities Press, New York, 1924.

[111] A. P. Witkin. Scale–space filtering. In *Proc. 8th Int. Joint Conf. on Artificial Intelligence (IJCAI '83)*, volume 2, pages 1019–1022, Karlsruhe, Germany, August 1983.

[112] R. B. Yates and B. R. Neto. *Modern Information Retrieval.* Oxford University Press, 1999.

[113] C. Zahn and R. Roskies. Fourier descriptors for plane closed curves. *IEEE Trans. on Computers*, C-21:269–281, 1972.

[114] E. Zwicker and H. Fastl. *Psychoacoustics. Facts and Models.* Springer Verlag, Berlin Heidelberg, 1999.