UAlg
UNIVERSIDADE DO ALGARVE

UNIVERSITY OF THE ALGARVE

FACE AND OBJECT RECOGNITION BY 3D CORTICAL REPRESENTATIONS

JAIME AFONSO DO NASCIMENTO CARVALHO MARTINS

THESIS

PhD in Computational Science and Engineering

Developed under the orientation of:

PROF. DOUTOR JOHANNES MARTINUS HUBERTINA DU BUF

PROF. DOUTOR JOÃO MIGUEL FERNANDES RODRIGUES

2013

DECLARATION

I hereby declare to be the author of this work, which is original and unpublished. Authors and works consulted are properly cited in the text and appear in the included reference list.

Declaro ser o autor deste trabalho, que é original e inédito. Os autores e trabalhos consultados estão devidamente citados no texto e constam da listagem de referências incluída.

COPYRIGHT

Faro, 2013

Jaime A. Martins

Ohana means family.
Family means nobody gets left behind, or forgotten.

— Lilo & Stitch

Dedicated to my wife and my parents,
who without their everyday support this thesis would not be possible;

To my supervisors and colleagues,
which always brought crucial expertise in the hardest times;

To my family and friends,
which were always ready to lend a hand and a smile,

THANK YOU!

We have seen that computer programming is an art,
because it applies accumulated knowledge to the world,
because it requires skill and ingenuity, and especially
because it produces objects of beauty.

— Donald E. Knuth (1974)

ACKNOWLEDGMENTS

The research published in this thesis was supported by the Portuguese Foundation for Science and Technology (FCT) through ISR/LARSyS pluri-annual funding, FCT PhD grant to Jaime Afonso Martins (SFRH–BD–44941–2008), the POS_Conhecimento Program (includes FEDER funds), FCT project SmartVision: active vision for the blind (PTDC/EIA/73633/2006), EU project NeuralDynamics (FP7–ICT–2009–6, PN: 270247), and FCT project Blavigator (RIPD/ADA/109690/2009).

I would also like to thank immensely for their contributions, to:

- Prof. Johannes M. H. du Buf for writing the Introduction of Chapter 2;
- Prof. João M. F. Rodrigues for the work done in Chapter 3 (texture boundaries in Section 3.3 and assembling the saliency map in Section 3.4), and also in Chapter 4 (developing the Line and Edge Disparity Model, Section 4.5.1);
- Prof. Roberto L. Lam for developing the code that allowed retrieval of normalised BU-3DFE face stereograms and OpenGL *Z-buffer* data in Chapter 6 (Section 6.2.1).

ABSTRACT

This thesis presents a novel integrated cortical architecture with significant emphasis on low-level attentional mechanisms — based on retinal non-standard cells and pathways — that can group non-attentional, bottom-up features present in V1/V2 into "proto-object" shapes. These shapes are extracted at first using combinations of specific cell types for detecting corners, bars/edges and curves which work extremely well for geometrically shaped objects. Later, in the parietal pathway (probably in LIP), arbitrary shapes can be extracted from population codes of V2 (or even dorsal V3) oriented cells that encode the outlines of objects as "proto-objects". Object shapes obtained at both cortical levels play an important role in bottom-up local object gist vision, which tries to understand scene context in less than 70 ms and is thought to use both global and local scene features.

Edge *conspicuity* maps are able to detect borders/edges of objects and attribute them a weight based on their perceptual salience, using readily available retinal ganglion cell colour-opponency coding. Conspicuity maps are fundamental in building posterior *saliency maps* — important for both bottom-up attention schemes and also for Focus-of-Attention mechanisms that control eye gaze and object recognition.

Disparity maps are also a main focus of this thesis. They are built upon binocular simple and complex cells in quadrature, using a Disparity-Enery Model. These maps are fundamental for perception of distance within a scene and close/far object relationships in doing foreground to background segregation.

The role of cortical disparity in 3D facial recognition was also explored when processing faces with very different facial expressions (even extreme

ones), yielding state-of-the-art results when compared to other, non-biological, computer vision algorithms.

KEYWORDS: Local gist, Saliency, Disparity-Energy Model, Proto-object classification, 3D Facial recognition.

RESUMO

A presente tese descreve uma nova arquitectura cortical integrada, com ênfase especial em mecanismos de atenção a baixo nível — baseados em conexões corticais que utilizam células retinais não-padronizadas — conseguindo agrupar diversas características visuais de baixo nível, ainda num estado pré-atencional, presentes nas áreas V1/V2, em formas específicas de "proto-objectos". As formas em questão são extraídas em primeira mão através de combinações de células especializadas que detectam localmente *cantos*, *rectas/arestas* e *curvaturas*, funcionando extremamente bem para a detecção de objectos com formas geométricas. Posteriormente, no lobo parietal (provavelmente no córtex Lateral Intra-Parietal), já podem ser extraídas formas arbitrárias, através de padrões de activação de populações de neurónios, presentes em V2 (ou até em V3-dorsal), que codificam a periferia de objectos como "proto-objectos" — representações básicas de categorias específicas de objectos no cérebro. Ambas as formas extraídas nos dois tipos de processamento cortical (utilizando células específicas ou uma codificação de formas arbitrária) desempenham um papel importante na visão *gist* local, que tenta compreender o contexto geral da cena apresentada ao sistema visual, em menos de 70 ms, sendo esperado que para tal se usem tanto características visuais *globais* como *locais*.

São também utilizados mapas de *conspicuicidade*, que permitem detectar linhas e arestas de objectos, atribuindo-lhes um peso baseado na sua saliência perceptual — utilizando para tal a codificação natural das células retinais, em que as cores são representadas por oponência: claro/escuro, vermelho/verde e amarelo/azul. Os mapas de conspicuicidade são fundamentais na construção posterior de mapas de saliência — importantes nos esquemas pré-atencionais de nível celular baixo e também para os mecanis-

mos de Foco-de-Atenção que controlam o movimento ocular e reconhecimento de caras e objectos.

Em paralelo, são também desenvolvidos os mapas de *disparidade* cortical, sendo estes também um dos maiores focos desta tese. Estes são baseados em células corticais binoculares simples e complexas, através de um processamento das últimas em quadratura — modelo denominado por "Disparity-Energy Model". Estes mapas de disparidade são fundamentais na percepção de distâncias dentro de uma cena visual e também para resolver o problema da segregação objecto/fundo.

O papel da disparidade cortical é também explorado no reconhecimento facial a 3D, em especial quando as faces a reconhecer apresentam expressões faciais de diversas formas e níveis de intensidade. O modelo utilizado apresentou resultados excelentes, atingindo o estado-da-arte, inclusivamente ficando acima de modelos de visão computacional não biológicos.

TERMOS-CHAVE: "Gist" local, Atenção, Saliência, Disparidade, Categorização de Proto-objectos, Reconhecimento Facial a 3D.

PUBLICATIONS

Some ideas and figures have appeared previously in the following publications (or are being prepared for submission):

- PUBLISHED IN CONFERENCE PAPERS:
  - "Region segregation and saliency using colour information," (Martins, Rodrigues, and Buf, 2008);
  - "An integrated framework for combining gist vision with object segregation, categorisation and recognition," (Rodrigues, Almeida, Martins, Lam, and du Buf, 2008);
  - "Object segregation and local gist vision using low-level geometry," (Martins, Rodrigues, and Buf, 2009);
  - "Focus of Attention and Region Segregation by Low-level Geometry," (Martins, Rodrigues, and du Buf, 2009);
  - "Disparity energy model with keypoint disparity validation," (Farrajota, Martins, Rodrigues, and du Buf, 2011);
  - "Local Gist Vision of Man-made Objects," (Martins, Rodrigues, and du Buf, 2011a);
  - "Disparity Energy Model using a Trained Neuronal Population," (Martins, Rodrigues, and du Buf, 2011b);
  - "Cortical multiscale line-edge disparity model," (Rodrigues, Martins, Lam, and du Buf, 2012);
  - "A disparity energy model improved by line, edge and keypoint correspondences," (Martins, Farrajota, Lam, Rodrigues, Terzic, and du Buf, 2012);

- "Biological Models for Active Vision: Towards a Unified Architecture," (Terzić, Lobato, Saleiro, Martins, Farrajota, Rodrigues, and du Buf, 2013);

- PUBLISHED IN JOURNALS:
  - "Local object gist: meaningful shapes and spatial layout at a very early stage of visual processing," (Martins, Rodrigues, and du Buf, 2012);

- SUBMITTED TO JOURNALS:
  - "Expression-Invariant Face Recognition using a Biological Disparity Energy Model," (Martins, Rodrigues, and du Buf, 2014a);

- FINISHING PREPARATION FOR SUBMISSION:
  - "Luminance, Color, Viewpoint and Border Enhancement Disparity Energy Model," (Martins, Rodrigues, and du Buf, 2014b);
  - "Proto-Object Categorisation and Local-Gist using Implicit Coding of Low-Level Spatial Layout Features," (Martins, Rodrigues, and du Buf, 2014c);

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

AIT	Anterior Inferotemporal Cortex
AUC	Area Under ROC
CMC	Cumulative Match Characteristic
CV	Computer Vision
DEM	Disparity-Energy Model
DET	Detection Error Trade-off
DoG	Difference-of-Gaussians
EER	Equal Error Rate
FAR	False Acceptance Rate
FFT	Fast Fourier Transform
FoA	Focus-of-Attention
FRR	False Rejection Rate
HTER	Half Total Error Rate
HV	Human Vision
IT	Inferior Temporal Cortex
L-DEM	Luminance Disparity-Energy Model
LCV-DEM	Luminance, Colour and Viewpoint Disparity-Energy Model
LCVB-DEM	Luminance, Colour, Viewpoint and Boundary enhanced Disparity-Energy Model
LCVE-DEM	Luminance, Colour, Viewpoint and Line/Edge enhanced Disparity-Energy Model
LEDM	Line and Edge Disparity Model
LGN	Lateral Geniculate Nucleus
LIP	Lateral Intraparietal Cortex
MST	Medial Superior Temporal Cortex

MT	Middle Temporal Visual Cortex (Visual Area V5)
PIT	Posterior Inferior Temporal Cortex
ROC	Receiver Operating Characteristic
ROR	Rank One Recognition
SC	Superior Colliculus
SVM	Support Vector Machine
V1	Primary (Striate) Visual Cortex
V2	Secondary (Prestriate) Visual Cortex
V4	Visual Area V4 of the Extrastriate Visual Cortex

1

INTRODUCTION

FACE AND OBJECT RECOGNITION IN THE CORTEX

ABSTRACT: This chapter introduces the scope of this thesis, namely an integrated cortical architecture supporting 3D face and object recognition, with special emphasis on a low-level local gist integrated framework, all relying on biological vision principles and properties.

1.1 SCOPE OF THE THESIS

Our brain is really a truly intricate and enigmatic machine. It is capable of quickly recognising each letter and word on this page, the *half-full* coffee mug on your desk, or the co-worker who just entered the room — all seamlessly and efficiently, as if they were simple and menial tasks. However, they are not. The complexity of any biological "computer" to achieve this kind of processing speed and power coupled with an extraordinary amount of input data is outstanding. Basic survival skills like hunting, scavenging and gathering have pushed brain evolution in the last 100,000 years in a way to sharpen and speed-up our sensorial recognition abilities to an extraordinary state (DiCarlo et al., 2012). Just focusing on vision alone, we are able to detect and classify objects from tens of thousands of choices (Biederman, 1987), all within a fraction of a second (Thorpe, 2009) and even when facing unfavourable odds, like huge discrepancies and/or deformations in

the object properties that reach our retinas (for a review see Logothetis and Sheinberg, 1996).

This thesis belongs to a domain of Computer Vision (CV) that focuses on Human Vision (HV) research — which tries to explore the unique opportunity of having access to a perfected biological machine (with thousands of years of evolution) that provides extremely robust solutions. However, the complexity of natural biological evolutive processes and "algorithms" extends much further than just the domain of pure vision, touching various disciplines — like psychophysics, optics, psychology, cognitive neuroscience, neuroanatomy, neurophysiology, computational neuroscience, computer vision, and machine learning — and consequently forcing the traditional boundaries between these fields to dissolve.

When analysing the performance of contemporary systems based on HV (often referred to as *biologically inspired* systems) one notices their tendency to fall somewhat behind in performance, efficiency and accuracy with respect to more direct CV systems, despite some very promising theoretical results. This is due to the fact that most HV systems are still on their infancy, racing to catch and understand thousands of years of brain evolution. However, we already know their extraordinary potential, simply because we use them every day. In addition, by studying how the cortical vision system works, HV research opens the doors to treating vision deficiencies and diseases, or even create specialised bio-implant prosthesis to compensate for those. For example, by pinpointing the neuronal circuitry responsible for face recognition, we might ultimately find ways to repair that circuitry in brain disorders that impact those perceptual systems (e. g., blindness, prosopagnosia, etc.).

Despite the above arguments, many questions remain open. For instance: how is the information really processed in the cortex? Is it done mostly bottom-up (also called data-driven) or are there also top-down feedback loops, and if so, only a few or many? Between which layers and/or which areas? What do they serve for? What about astrocytes, cells that were be-

lieved to only serve as support for neurons, but in fact respond to visual stimuli? Even the interconnections between neurons are not well known, i. e., there is no consensus between the principal groups working in this field about a neural architecture that could unify all processing steps into a single structure. For example, at the lowest level one can ask how and when each cell is activated in each layer, which cells combine within one single layer, and which cells activate the next layer. At the highest level the same questions are related to how object representations are stored in visual memory, to when and how scene and object recognition starts, etc.

Even considering all these questions, which are inherently very interesting from a viewpoint of various neurosciences, there are only a handful of groups focusing on cortical cell models, with even less trying to develop an integrated cortical architecture and almost none focusing on low-level cell data with bottom-up face and/or object recognition, especially relating those to fundamental perceptional contexts like gist. The lack of major research groups developing an integrated architecture has two main reasons: (1) it is an extensive interdisciplinary field where lots of biological cell data and cortical constraints have to be taken into account and (2) it does not return "excellent" results very fast (excellent in terms of data suitable for publications in a "publish or perish" academic society).

The current research groups working on bio-inspired cortical models or on HV are all looking for the Holy Grail: an approach, probably multi-scale, that can yield a complete "biological" characterisation of an image that rivals real-world cell data representations in the visual pathways. There are many practical applications that can benefit from advanced cortical models: object and face recognition, texture analysis, image segmentation, motion and depth prediction, and image enhancement and coding. There also are other scientific and technological areas where HV-related knowledge can be applied, like engineering (better ways to recognise persons), medicine (how to treat some brain injuries), education (more efficient ways to boost cognitive learning) and even arts (new ways to study paintings and painters).

Specifically, the main focus of this thesis is on computationally replicating specific visual cortex processes related to object and face recognition, especially in the context of gist vision, while exploring a possible integrated architecture and always having in mind practical applications. Main topics are:

- The computational implementation of functional cortical cell models for local gist and quick object and face recognition;
- Exploration of the role of Focus-of-Attention (FoA) and salience in local gist-based object detection and categorisation;
- Development of a cortical stereo-disparity model which incorporates simple and complex-cell energy models into a real-world disparity model;
- Combination of salience and disparity for figure/ground separation followed by object detection and categorisation;
- Assess the performance of a biological disparity model for precise face recognition tasks;
- Propose an integrated model or architecture of the cortex, relating *features* with *cells*, *cell layers* and with the information pathways of the visual system.

In the next sections, the most relevant aspects of these are discussed in detail and the structure of the thesis will be presented.

1.2 CORTICAL ARCHITECTURE AND FEATURE EXTRACTION

For anyone with a special interest in neuroscience, there is always a certain kind of awe when describing the brain — 10^{12} (a million times a million) neuronal cells, with each one receiving and transmitting information to

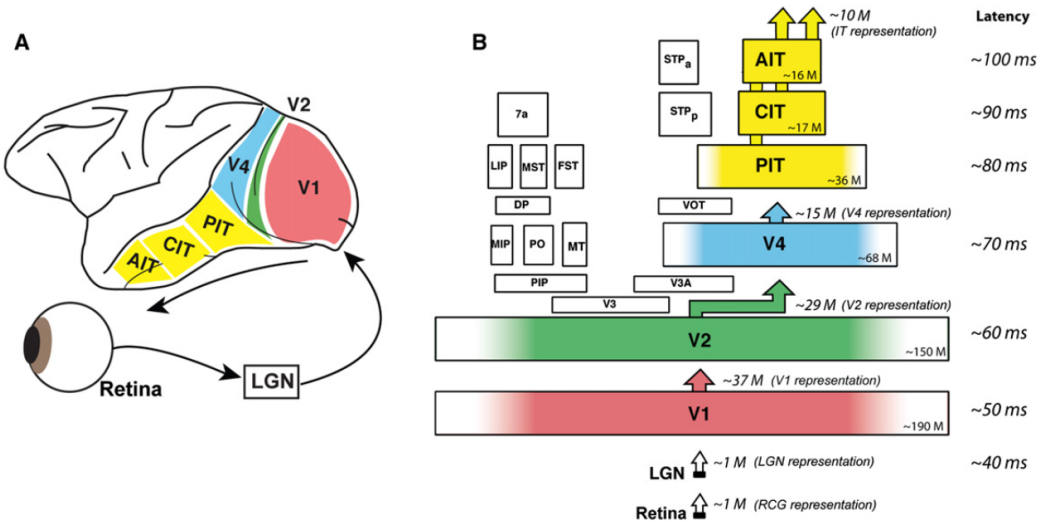


Figure 1.1: Information flow in the visual cortex. Figure from DiCarlo et al. (2012, pp. 419).

hundreds or even thousands of other neurons, with a total number of interconnections somewhere between 10^{14} and 10^{15} — nothing short of amazing! Even so, this is just the raw performance potential, telling nothing of the underlying, extremely optimised architecture, both *anatomically* and *functionally*, aspects which are very hard to quantify (Hubel, 1995). While individual neurons themselves are relatively simple cells, they do not see, reason or remember — but the brain as a whole does.

When we look at anything, the information captured in both retinas is propagated in the brain through the Lateral Geniculate Nucleus (LGN) of the Thalamus into the Primary (Striate) Visual Cortex (V1), in the cortical hyper-columns (shown in Figure 1.1), where most neurons display a property called *tuning* — they only respond to a specific subset of stimuli within their receptive field. This selectivity means that they can effectively work as *feature detectors* — e. g., in the earlier visual areas, some are tuned to simple patterns like corners, bars or gratings. However, neurons in the highest visual areas are tuned to much more complex patterns, e. g., in the Inferior Temporal Cortex (IT), a neuron may only fire when a certain face appears in its receptive field.

In V1, and the primary sensory cortex in general, neurons with similar tuning properties tend to cluster together in cortical columns, spatially arranged following two tuning properties: ocular dominance[1] and orientation (Hubel, 1995). Recent research has also shown them to be sensitive to several other features, such as colour (Horwitz and Hass, 2012), which was previously believed to be impossible (Hubel, 1995).

Both in CV and HV the term *feature extraction* is commonly used to define the process of detecting specific image landmarks or properties, by transforming raw visual data into a type of semantic representation. While CV disregards the method *per se* and focuses on the results, for HV both method and results are significant, mainly because it tries to recreate computational models as close as possible to existing biological ones. Regarding cortical feature extraction, this thesis focuses primarily upon *conspicuity* and *disparity*. We will address both of them in more detail in the next sections (and in their respective chapters). In the end, our cortical architecture relies heavily on both for detecting and recognising objects and faces. We should mention that these features can also be linked to other cortical features, like *lines and edges* and *keypoints*.

1.2.1 *Conspicuity and saliency features*

Attention plays a dominant role in modulating the visual system throughout its cortical layers. The classical representation of visual attention has been the *saliency map*, which represents the attentive or *salient* hot-spots in images or scenes. The creation of this map is mostly viewed as an exclusively higher-level parietal pathway function (Gottlieb, 2007), which is challenged in this thesis and by recent research from other authors who confirm the theory that V1 can elaborate a truly bottom-up saliency map (Zhang et al., 2012). One of the cornerstones of our approach is the use

1 Individual V1 neurons in primates and animals with binocular vision have ocular dominance, i.e. a preference for one of the two eyes.

of a cell feature denominated *conspicuity*, representing differences in luminance and colour between very small image patches, which naturally exploits retina colour-opponency coding in ganglion cells. Conspicuity can simultaneously serve as a measure of saliency and as an edge segmentation mechanism for object detection. Its role in FoA will be explored in depth in Chapter 3 (page 57).

1.2.2 *Stereo/disparity features*

Neuronal pathways devoted to disparity processing in the Primary (Striate) Visual Cortex (V1) and Secondary (Prestriate) Visual Cortex (V2) allow us to see the world in 3D, a hugely important survival skill. It plays an important role in many areas devoted to motor control, from walking around to precise eye-hand coordination, focus-of-attention and object segregation, even recognition with partial occlusions. The availability of disparity information at such an early stage plays a huge role in quick object segregation (even in complex backgrounds) and shape detection, making disparity a key factor for FoA/saliency prioritisation in low-level gist processing.

In CV there are numerous approaches for stereo vision (for a review, see Szeliski, 2011), but only few are biologically motivated. Of these, most have one common aspect — they are based on the widely accepted Disparity-Energy Model (DEM) (Ohzawa et al., 1997; Haefner and Cumming, 2008; Read, 2010; Martins et al., 2011b), which has become a *de facto* standard HV model for cortical stereo vision and was first introduced from research into the cat's visual cortex and pathways by Ohzawa et al. (1990). The usage and improvements made to the DEM are elaborated in Chapter 4 (page 73).

1.3 GIST VISION

One of the most important research fields in HV is to understand the biological fundamentals of gist vision. The brain is able to achieve a kind of "integrated exteroception" that could be defined as an *"almost instantaneous self-perception of spatial context awareness"* or *gist*, which happens within 50–70 ms after image/scene onset. This mechanism enables us to almost immediately ascertain our surrounding environment type (*"where am I?"*—e.g., classroom, mall, city, forest, beach, etc.). Research in human perception has shown that several of the more "global properties" of a scene are discerned at a very early stage of visual processing (e.g. coarse *spatial layout*, *naturalness*, *navigability*, *complexity*), and that brief exposure to a specific scene layout facilitates distance perception (for a review, see Ross and Oliva, 2010).

Recent work emphasises the role of disparity in the context of gist vision: global properties describing the three-dimensional layout of a scene, such as the dominant depth, openness, or perspective of an environment, can be perceived at the very beginning of a glance, and can influence scene categorisation (Greene and Oliva, 2009a,b). Also, gist seems to depend on two stages: an early, automatic stage, where local-contrast responses present in LGN or V1 seem to play a very important role (Groen et al., 2013), followed by a later, task-dependent stage. This was previously explained by Vogel et al. (2007)—humans rely on local, bottom-up information as much as on global, top-down information. We seem to integrate both types, as the brain makes use of scene information at multiple scales for scene categorisation.

Aside from a higher-level, task-dependent path, we propose that low-level, automatic scene gist can use two major complementary paths: (1) a "global gist" path, which is usually referred to as just *gist*, which is akin to more global scene properties (Ross and Oliva, 2010; Rodrigues and du Buf, 2011a), and (2) a "local gist" path, which aims to extract rough semantic object information and spatial layout as fast as possible, while also being able to simultaneously process object segregation using non-attentional "scene

schema," consisting of concurrent spatial-layout and gist subsystems which both drive attentional object recognition (Rensink, 2000; Martins et al., 2012). Basically, we envision both local and global gist as bottom-up processes that complement and prime each other. In this thesis we are focusing research efforts on the "local gist" part of the architecture, which is interestingly also the least researched one. This will be further explored in Chapter 2 (page 17) and Chapter 5 (page 109).

1.4 OBJECT AND FACE DETECTION

There are several CV-based object detection approaches that have gained some significance lately — we can highlight, e. g., methods using SIFT descriptors (Lowe, 2004), Random Forests (Gall et al., 2012; Baumann et al., 2013) or Histograms of Oriented Gradients (Vondrick et al., 2012), which are all heavily mathematically based. Looking at HV-based methods, object detection has been addressed mostly in the context of FoA and salience-based methods (Itti and Koch, 2000; Gao, 2005). In this thesis we also focus on the use of *conspicuity* and *disparity* as detection features, for segregating objects (or faces) from their background. This will be explored in depth in Chapter 5 (page 109).

1.5 OBJECT CLASSIFICATION AND RECOGNITION

Before moving on, it is necessary to define the term *recognition*. Classical "object recognition" research is one of the most tackled problems in CV, image processing and machine vision. The term, however, can be used with different interpretations, depending on the desired outcome. For example, we can refer to *recognition* as to: (1) identify one or several pre-specified or learned object *classes* (e. g., this is a coffee mug), (2) it may be the identification of a *specific object* inside a class (e. g., this is John's coffee mug), or even

(3) *detect* a specific object inside an image/scene (e. g., there are two coffee mugs in this image). It must follow from the context what is really meant, respectively, *categorisation*, *recognition* or *detection*.

Computationally, recognition is one of the most difficult tasks, but the difficulty depends on the task: it is quite easy for an average person to detect and recognise objects after seeing them only a few times, i. e., if we put a few and very distinct objects on top of a white table with good illumination. But if we put more and less distinct objects, with some objects partly occluding each other, and cover the table with a cloth with some complex texture, the task becomes more difficult, recognition performance decreases, and much more effort will be required to boost performance to an acceptable level.

Most real-world applications are not trivial, even the ones that appear relatively easy. For instance, counting how many people are waiting on a sidewalk to automate the control of a zebra crossing, or reading plates of cars passing a toll gate at high speed, are already quite complicated. Also, there are really very complicated applications, like recognising a person after a hair-style change, after growing or shaving a beard, or after having grown old and grey. The extreme case is spotting these in real time in low-quality CCTV video, searching for someone who does not want to be recognised, who therefore may use all concealment tricks and even plastic surgery.

In neuroscience the concept of object recognition is even more difficult since it involves several levels of understanding, from the information processing or computational level, to the level of circuits, cellular and biophysical mechanisms. After decades of research effort, neuroscientists working on functions in striate and extrastriate cortical areas have produced a huge and still rapidly increasing amount of data, and the emerging picture of how the cortex performs object recognition is in fact becoming too complex for any simple model (Serre et al., 2005). Recognition turns out to be a delicate compromise between selectivity and invariance. Therefore, the

key computational issue in object recognition is the specificity-invariance trade-off: the system must be able to finely discriminate between different objects or object classes, at the same time being tolerant to sometimes big object transformations which include scaling, translation and (2D) rotation; also changes of illumination, (3D) viewpoint, context and clutter, non-rigid transformations such as a change of facial expression and, in the case of categorisation, also shape variations within a class (Serre et al., 2005).

Another problem that increases difficulty in modelling "biological recognition" is the definition of the instant when it all starts. Psychologists and psychophysicists, who study how we perceive patterns and images, used to think that, before the processes of object categorisation and recognition could begin, the brain must first isolate a figure in the image — such as a tree or a piece of fruit — from its background: a process called "object segregation." However, recent research suggests that we actually categorise objects before we have segregated them, or that both processes occur in parallel. This means that by the time we realise that we are looking at something, our brain already knows what that thing is (Oliva and Torralba, 2006). This is the venue we are following in this thesis, integrating FoA and saliency (in the form of *conspicuity*) with proto-object detection and categorisation, an effort that has also been developed by other authors, e. g., Yanulevskaya et al. (2013) who focus on salient proto-object detection within an object-based attention theory.

Grill-Spector and Kanwisher (2005) tested three types of visual recognition by briefly flashing images before the eyes of human observers. The first type, object detection, was tested by showing images that may or may not have contained figures. Participants had to quickly judge whether or not there was a figure present against a background. The second type concerned categorisation, where participants were shown images of figures and they had to indicate what type of figure they saw, such as bird, car, or food. In the third part of the test, more specific images were shown in order to test identification. Participants had to identify figures within categories,

such as parrot or pigeon in the category "bird." It turned out that the participants were as fast and accurate in naming the category that an object belonged to as they were at saying whether or not they had seen an object at all. The ability of the subjects to process the images in such a short time proved that, by the time they knew an image contained some sort of object, they already knew its category. The authors concluded that "There are two main processing stages in object recognition: categorisation and identification, with identification following categorisation," also "Overall, these findings provide important constraints for theories of object recognition," and "Rapid categorisation obviously facilitates our survival and interaction with the environment on an everyday level." This process of rapid categorisation before identification restricts the brain's search for a match between the visual input (the picture you looked at) and internal category-relevant representations (stored images of other objects you have seen and identified prior to today). From these conclusions it follows that categorisation/identification should not be studied or modelled as a single-level task, but as a multi-level task where one or several levels of categorisation should be performed. In addition, categorisation should start at the same time as detection or segregation.

DiCarlo et al. (2012) explored this subject from a neuronal population coding perspective, with the notion that the ventral visual pathway gradually "untangles" information about object identity. When an object undergoes an identity-preserving transformation, such as a shift in position or a change in pose, it produces a different pattern of population activity, which corresponds to a different response vector. Together, the response vectors corresponding to all possible identity-preserving transformations (e.g., changes in position, scale, pose, etc.) define a low-dimensional surface in this high-dimensional space — an object identity manifold. For neurons with small receptive fields that are activated by simple light patterns, such as retinal ganglion cells, each object manifold will be highly curved. Moreover, the manifolds corresponding to different objects will be "tangled" together, like

pieces of paper crumbled into a ball. At higher stages of visual processing, neurons tend to maintain their selectivity for objects across changes in view, translating to a gradual "untanglement" of manifolds through nonlinear selectivity and invariance computations applied at each stage of the ventral pathway. This view fits nicely with the methods followed in this thesis, i. e., initial low-level proto-object detection and categorisation can be extracted just using *conspicuity* and *disparity* features, very quickly in V1/V2, biasing the "untanglement" of posterior information for precise categorisation or recognition, although we expect that the parietal cortex plays a very important role in gist vision.

Since we are mostly focused on what the brain is able to do in V1/V2 and in the dorsal pathway — from the Middle Temporal Visual Cortex (Visual Area V5) (MT) to the Medial Superior Temporal Cortex (MST) and Lateral Intraparietal Cortex (LIP) (all within 70ms) — we devised our local gist architecture using pathways that focus first on object and face *detection* (following low-level, attentional, saliency-based schemes) and afterwards on *categorisation* (what type/class of object does its shape represent?), as both provide the most important information needed for bootstrapping a gist vision system. "Recognising" *proto*-objects (i. e., prototype, representative shapes of object classes, that are usually very distinct from other classes and very similar inside each class) is effectively a form of object *categorisation*, since we are effectively "recognising" the generic shape of a class, so the term *object categorisation* will be used below.

1.6 FACE RECOGNITION

In computer vision, face processing consists of two tasks: (1) detecting faces in all types of scenes, and (2) recognising the persons associated with the detected faces. Significant problems are involved in both tasks, ranging from partial occlusions to dealing with different facial expressions, even extreme ones. These are huge hurdles for face recognition technology (Franco

and Nanni, 2009). Therefore, a robust system should employ techniques which give reliable results, regardless of any differences in acquired images. Most face recognition methods rely heavily on image processing techniques which are normally not related to models of cortical processing. Siagian and Itti (2004) proposed a biologically inspired face *detection* model based on saliency, gist and gaze data, to sequentially detect image regions containing faces. For a general survey on face recognition approaches we refer to Li and Jain (2011).

One of the more difficult problems of face recognition is to deal with facial expressions, where 3D structural information can help immensely. This is one of the main goals of this thesis. Face recognition in the context of large changes in facial expressions will be described in Chapter 6 (page 139).

1.6.1 *Biological differences between object and face recognition*

In terms of the neural pathways followed by either objects or faces, we envision our system as using a common path for most of the processing until reaching the MT/MST/LIP, where the recognition pathways split. The shape of a face is considered a special kind of "proto-object" that, when detected, is segregated and sent forward to a specialised cortical facial recognition pathway in IT (Biederman and Kalocsai, 1997).

1.7 OVERVIEW OF THE THESIS

This thesis is divided into seven chapters, each one corresponding to a specific subject (see below for a short summary of each one). Some of the figures and tables displayed are intentionally enlarged into their page's left or right margin space, when there is a justified benefit of sacrificing layout for content. Words printed in **blue** represent links to chapters, sections,

pages, figures and tables. Words printed in **brown** represent links to important terms that can be found in the acronyms list. Words printed in **green** represent links to author citations in the bibliography.

- Chapter 2 (page 17) describes our baseline model for Local-Gist Vision. It explores how low-level local object gist can be achieved for regularly-shaped geometric objects, with only a few cortical cell layers and with strictly bottom-up or data-driven processing. This allows for a simultaneous (i) quick categorisation of geometric-like object shapes and (ii) quick object segmentation from background clutter, based on conspicuity/saliency. This chapter was partially published in Martins et al. (2009, 2011a) and fully published in Martins et al. (2012).

- Chapter 3 (page 57) introduces Focus-of-Attention (FoA) and Saliency. It shows that these features deliver very important information for object and face detection. It also shows that saliency maps for FoA can be constructed from conspicuity features, and that such maps can be employed for directing eye gaze with inhibition of return. It also introduces how conspicuity can help to segregate objects in scenes. This chapter was partially published in Martins et al. (2008) and fully published in Martins et al. (2009).

- Chapter 4 (page 73) describes the Disparity-Energy Model (DEM) for processing real-world images/scenes. We used a new model for encoding disparity information implicitly by employing a trained binocular neuronal population. This model allowed decoding of disparity information in a way similar to how our visual system could have developed this ability, during evolution, in order to accurately estimate disparity in entire scenes. Also, a new disparity model based on multi-scale line and edge coding is presented, such that depth from stereo can be attributed to lines and edges. Both models are integrated for obtaining more accurate disparity maps. This chapter was partially published in Farrajota et al. (2011); Martins et al. (2012, 2011b); Ro-

drigues et al. (2012); Terzić et al. (2013) and is being prepared to be submitted to an appropriate journal.

- Chapter 5 (page 109) extends our baseline model of local gist vision (Chapter 2), in order to obtain a more general local gist architecture. A proto-object model is extended from geometric shapes to any kind of arbitrary shape, integrating our Disparity Energy Model and Local Object Gist implementations into a unified framework, supporting very fast object detection and categorisation. This chapter is being prepared to be submitted to an international journal.

- Chapter 6 (page 139) tests our disparity model in the context of face recognition. We tested the validity of our disparity data to effectively recognise faces based solely on disparity maps or in conjunction with image data, even when facing hard facial shape deformations due to extreme facial expressions. Our approach is very effective for a HV implementation, yielding better results than state-of-the-art CV approaches using the same databases. This chapter was submitted to *IEEE Transactions on Image Processing* on June 2013 and is currently under peer-review.

- Chapter 7 (page 167) integrates the previous chapters into a unified cortical architecture, and provides a summary of major achievements, concluding remarks and ideas for future research.

2

LOCAL OBJECT GIST

MEANINGFUL SHAPES AND SPATIAL LAYOUT AT A VERY EARLY STAGE OF VISUAL PROCESSING

ABSTRACT: High-level vision is based on semantic representations of scenes or context and important objects therein. The bottom-up data streams, from retina via Lateral Geniculate Nucleus (LGN) to the Primary (Striate) Visual Cortex (V1) and further, are devoted to moving objects and motor control in the dorsal "where" stream, and to invariant object recognition in the ventral "what" stream. They are steered, top-down, by attention and short-time memory in the ventral and dorsal regions of the pre-frontal cortex. However, these processes are bootstrapped, and probably continuously guided, by an extremely fast analysis devoted to scene gist and spatial layout, i. e., which types of objects are about where in the scene. Most research has been devoted to global scene gist, but in this paper we present an alternative approach which addresses local object gist for simultaneous object segregation, attention, and spatial layout. The proposed model only exploits colour information, although texture, motion and disparity information can also be integrated. Specifically, we focus on man-made objects which are dominated by a simple shape repertoire: squares, rectangles, trapeziums, triangles, circles and ellipses. It is shown that such shapes can be detected by a hierarchy of a few cell layers, with strictly bottom-up or data-driven processing. We argue that this processing may occur in very early vision, possibly only employing signals from non-standard retinal ganglion cells. Although proposed to play a role in the fast dorsal stream, similar process-

ing may occur in the slower ventral stream.

KEYWORDS: low-level vision, local gist, object segregation, shape extraction, spatial layout.

2.1 INTRODUCTION

In his introduction, Pinna (2010) quoted one of Wertheimer's observations: "I stand at the window and see a house, trees, sky. Theoretically I might say there were 327 brightnesses and nuances of colour. Do I have '327'? No. I have sky, house, and trees." This seems quite remarkable, for Max Wertheimer, together with Kurt Koffka and Wolfgang Köhler, was a pioneer of Gestalt Theory: perceptual organisation was tackled considering grouping rules of line and edge elements in relation to figure-ground segregation, i. e., a meaningful object (the figure) as perceived against a complex background (the ground). In general, the latter can consist of other, but bigger and partially occluded objects: a hierarchy of for example a bicycle left against a tree in front of a house in front of a forest.

At the lowest level — line and edge elements — Wertheimer (1923) himself formulated grouping principles on the basis of proximity, good continuation, convexity, symmetry and, often forgotten, past experience of the observer. Rubin (1921) formulated rules for figure-ground segregation using *surroundedness*, size and orientation, but also convexity and symmetry. Almost a century of research into Gestalt later, Pinna and Reeves (2006) introduced the notion of *figurality*, meant to represent the integrated set of properties of visual objects, from the principles of grouping and figure-ground to the colour and volume of objects with shading. Pinna (2010), went one important step further and studied perceptual meaning, i. e., the interpretation of complex figures on the basis of past experience of the observer. Re-establishing a link to Wertheimer's rule about past experience,

he formulated five propositions, three definitions and seven properties on the basis of observations made on graphically manipulated patterns. For example, he introduced the illusion of *meaning* by comics-like elements suggesting wind, therefore inducing a learned interpretation. His last figure shows a regular array of squares but with irregular positions on the right side. This pile of (ir)regular squares can be interpreted as the result of an earthquake which destroyed part of an apartment block. This is much more intuitive, direct and economic than describing the complexity of the array of squares. Indeed, Pinna noted that such a formulation of what might have happened to the array of squares, a "happening," relates to "affordance" as introduced by Gibson. James J. Gibson, as for Roger G. Barker, was a pioneer of ecological psychology, also known as environmental psychology. They preferred to study behaviour in the real world instead of in the artificial environment of a laboratory. One of their assumptions was that perception not only influences behaviour, but that perception and behaviour form a closed loop such that behaviour can also influence perception. In this sense, vision should be studied by considering real problems in the real world, preferably while performing real tasks. Of course, this does not automatically exclude studies with precisely defined conditions, which can normally be best controlled in a laboratory.

Our brain has evolved, during hundreds of thousands of years, for perceiving and interpreting the natural world and, during only a few thousands of years, increasingly also a man-made world. It should therefore be no surprise, as Pinna (2010) found, that "everything has a meaning," because our brain has extensively learned to extract meaning. And we are thinking in terms of semantics, not in terms of low-level syntax, although it is now clear that part of our brain, at least the primary visual cortex, is devoted to low-level syntax. The link between low-level syntax and high-level semantics remains subject to research, and an inexhaustible source for amusing and often puzzling visual illusions which demonstrate how particular brain regions have developed in order to solve particular vision

problems. This is not to say that Gestalt Theory is an obscure branch of psychology with sometimes amusing visual effects. It should be taken seriously, but in proper context.

Pinna (2010) built on Gestalt and went one step further in the direction of less abstract but meaningful objects. We, here, take the opposite direction. By asking what is necessary to extract meaningful objects, it should be possible to go down to the level of Gestalt Theory. However, this only concerns principles such as good continuation and not past experience nor learned interpretations, because we also address abstract and therefore man-made objects, but only those without ambiguity and possibly at a very early stage in the visual system. For example, for detecting a square traffic sign — which is an explicit application of our research — one needs four edges connected at four corners, whereas a triangular one requires three edges and three corners. Occlusion of one of the square sign's corners does not necessarily lead to an interpretation of a triangle, nor does occlusion of one of the square's corners leads to a triangle. At a very early stage in visual processing, much below conscious reasoning on the basis of past experience, our neural circuits have also learned: to use and combine all available information in order to obtain the most robust and reliable solution. Although we may not be aware of this, it is crucial while driving a car at 120 km/h. In this sense, our research represents, as for Pinna's, a further step in Gestalt, but, unlike Pinna's, not necessarily at the level of conscious report. Although there is no doubt that geometric shapes like squares, triangles and circles are important at high semantic level, we will argue that they may also be important at a much lower level, for example in the dorsal "where" data stream for directing attention and eye/head control. In order to be able to understand why this may be the case, we need to explain some concepts of the visual system in more detail.

2.1.1 *Attention, spatial layout and local gist*

Our visual system is a very hierarchical one. The brain is divided into front and back parts, roughly at the central sulcus, with the front part "looking at" the back part (Crick and Koch, 2003). In the back part there is fast and massively parallel processing, from low-level syntax to invariant object representations, whereas the front part works at a slower pace, with serial processing involving covert and overt attention at semantic level. For example, experiments with different types of "snake" patterns revealed that some may "pop out" instantaneously, which is an indication for parallel processing, whereas others require effortful scrutiny and serial processing (Houtkamp and Roelfsema, 2010). In this article we focus on the transition between low-level syntax and low-level semantics, using very elementary information such as colour. The goal is to develop a system for local gist vision: which types of objects are about where in a scene. This is necessary to bootstrap and guide, even alleviate, the processing in the ventral and dorsal data streams. These streams are known to serve two goals: the dorsal stream, also called the where or vision-for-action stream, is devoted to optical flow and stereo disparity, whereas the ventral stream, also called the what or vision-for-perception stream, is devoted to object recognition (Konen and Kastner, 2008; Farivar, 2009).

One problem is that precise object recognition in the ventral stream requires object segregation, but object segregation is only possible if the system already knows what the object is, assuming of course that objects are seen against complex backgrounds. Another problem is that object recognition is a sequential process: while fixating one object, its features must be routed to normalised object templates held in memory. This routing blocks the system until recognition has been achieved, after which the system is released for dealing with another object. Therefore Rensink (2000) proposed a non-attentional "scene schema" consisting of concurrent spatial-layout and gist subsystems which both drive attentional object recognition, all employ-

ing "proto-objects" resulting from low-level vision. However, gist vision addressed so far concerns global gist of entire scenes (Bar, 2004; Siagian and Itti, 2007; Ross and Oliva, 2010; Rodrigues and du Buf, 2011b).

Global scene gist can be used to bias — select or exclude — object templates in memory in the matching process: when in a classroom it is not very likely that we see a horse. But global gist lacks localisation. On the other hand, when seeing a horse it is not very likely that we are in a classroom. Local object gist has the advantage of solving, or at least contributing to, Rensink's spatial-layout subsystem. Although both global and local gist can determine context, probably with a straight relation between them, local gist can solve important problems like a first and fast object categorisation, localisation and segregation, the latter being related to figure-ground organisation (Craft et al., 2007).

In what follows, one should keep in mind that global gist vision is very fast: our brain is able to pre-categorise scenes in as few as 19-67 ms, but final scene recognition takes 100-200 ms, whereas object recognition takes 200-300 ms (Oliva and Torralba, 2006; Bar et al., 2006; Greene and Oliva, 2009b). Local gist vision, also assumed to be very fast, is not meant for precise object recognition with conscious report; it may only be one preprocessing stage for guiding attention to meaningful locations while performing a specific task. For example, man-made objects with a simple shape repertoire include traffic signs, which is one application that will be tested. The basic shape of a sign — circular, triangular or square — implies a certain function. Hence, the visual system of a car driver can already be alerted and biased when still being far away from a sign. After this, an attention window can be created and updated for covert or overt attention with eye fixations driven by the dorsal stream for determining, in the ventral stream, which sign it actually is when approaching it. This general organisation of the processing involved may appear intuitive, but in practice there are some complications, like the roles of episodic and procedural memories, i. e., learned observation and driving behaviour on often-driven roads, and,

as we will see below, problems when dealing with multiple traffic signs which have been mounted together.

We know that the dorsal data stream for stereomotion (Peng and Shi, 2010) and motor control is faster than the ventral stream. Average activation latencies in the case of geometric shapes — which we address here — are 62 ms in the Lateral Intraparietal Cortex (LIP) (dorsal area of the posterior parietal cortex) and 101 ms in the Anterior Inferotemporal Cortex (AIT) (ventral area of the inferior temporal cortex) (Lehky and Sereno, 2007); these areas are described below. Two key questions therefore are whether analysis of geometric shapes can occur in the dorsal stream, which is *not* devoted to object recognition, and whether this can occur at a very early stage, for example by directly employing simple information from retinal ganglion cells instead of much more complex information present in cortical areas V1 and V2, but in both cases in areas which are very far from the frontal part of the brain, as referred to above, where semantic representations are handled. We must keep in mind that extraction of geometric shapes can be seen as a first object categorisation, and the latter can also be achieved at a higher level with coarse-to-fine-scale processing by combining information in the dorsal and the ventral streams (Rodrigues and du Buf, 2009b). For answering the two questions we need to go into more detail.

2.1.2 *The ventral and dorsal pathways*

For a very recent and good overview of early processing, from retinae to Lateral Geniculate Nucleus (LGN) and to early visual cortical areas, we refer to Troncoso et al. (2011). After area V2, the dorsal "where" stream continues to the Middle Temporal Visual Cortex (Visual Area V5) (MT), the Medial Superior Temporal Cortex (MST) and other intermediate areas up to the posterior parietal cortex. MT neurons are selective to direction of motion, speed and binocular disparity. MST neurons convey more global information about a scene's structure and spatial relationships (Smith et al., 2006),

including egomotion (Wall and Smith, 2008). The later authors suggested that MST has a central role in guiding heading in macaques. All these processes would benefit from an early object segregation such that motion and disparity are integrated and attributed to meaningful items in visual space, especially when also motion prediction is applied for estimating where objects are expected next. Motion prediction is a form of adaptation which can explain the motion after-effect, for example our illusion that a railway station moves after the train in which we sit has stopped. This may occur in area MT (Kohn and Movshon, 2003). These are indications that attention is not only a static process directed by complexity in the visual field, for example colour conspicuity, but a dynamic one involving motion and motion prediction. If these are processed at an early level, they can (i) control processing at a very low level, even down to the LGN (Lehky and Sereno, 2007), and (ii) have a bottom-up connection to the prefrontal cortex with two top-down attentional components from the prefrontal cortex, i. e., from PF46d (d from dorsal) to the posterior parietal cortex (see above) and from PF46v (v from ventral) to the inferotemporal cortex (see below) (Deco and Rolls, 2005).

Concerning the posterior parietal cortex, the highest visual area in the dorsal stream, Creem-Regehr (2009) presented an overview of the functionalities of its different sub-areas for sensory-motor planning and online control of eye, head, arm and hand, which includes pointing, reaching, grasping and even correct handling of tools. These functionalities are closely related to cognitive skills like imagining, gesturing and pantomime. For example, correctly grasping tools by their handles requires interaction between cognitive and action systems, where objects have sensory-motor affordances which guide behaviour on the basis of object structure, relevant goals, and the tools' known functions. It was even proposed to extend the dichotomous what/where organisation to a trichotomous what/where/how one, involving semantic memory (which tool), procedural memory (how to handle it), and attention (where is it).

In general, attention and action can be related to three types of behaviour: (1) reflexive in the case of sudden events like a moving object or an unexpected sound, e. g., a cat's arousal, (2) covert or automatic in the case of frequent or repetitive tasks, and (3) overt or consciously controlled when a task requires close scrutiny. If motor actions are controlled by the Superior Colliculus (SC) (with binaural source localisation via the inferior colliculus to the SC), these behaviours may be based on three pathways: (1) from retina straight to the SC, (2) from intermediate areas MT/MST to the SC, and (3) from the highest area, the posterior parietal cortex, but with input from area PF46d, to the SC. As we will see below, this is still speculative, but early object-centred segregation, localisation and attention through local gist vision can be very useful for steering most processes, especially in case of man-made objects with simple shapes.

The ventral "what" stream continues after area V2 to area V4 and other intermediate areas, up to the inferotemporal cortex. Many V4 neurons code colour, also orientation, width and lengths of bars, and curvilinear as well as linear gratings. The main purpose of this stream is object recognition. Also in this stream it would make sense to extract geometric shapes at an early stage, for a first object categorisation like man-made items, for example traffic signs which are always circular, triangular or square, with perspective projections to elliptic and trapezoidal shapes, *vs.* natural items like persons, animals, plants and trees.

Given the role of the ventral data stream, it can be no surprise that different objects are coded there, but at a higher level. For example, Kiani et al. (2007), who studied 674 neurons in a monkey's inferotemporal cortex, clustered the neurons' responses to about 1100 natural and artificial object images belonging to 23 intuitive categories. They found that different categorical structures are represented by different subsets (or populations) of the neurons. Animate and inanimate objects are represented by different subsets. In the case of animate objects, different subsets are devoted to bodies, hands and faces, the latter being divided — in other subsets —

into human and monkey faces. Bodies of humans, birds and four-limb animals are clustered together, whereas lower animals like fish and reptiles formed another cluster. Interestingly, in case of artificial objects like furniture, lamps, kitchen utensils and home appliances, these categories were not represented in the monkey's inferotemporal cortex, with the exception of cars, despite the fact that the animals were raised in human houses and later in zoos.

Kiani et al. (2007) also performed a similar cluster analysis on the basis of responses of (simulated) simple and complex cells. This analysis revealed no such clusterings, which means that categorical structures are formed after area V1, i. e., in higher areas which group V1 elementary features into meaningful items. This grouping can be based on Gestalt rules like good continuity, and can resemble the grouping into simple geometric shapes of many man-made objects as will be explained below.

In earlier work we developed a framework for invariant object categorisation and recognition, assuming multiscale representations in the two streams: keypoints in the dorsal stream and lines and edges in the ventral one. Starting at a coarse scale, keypoints are used to route lines and edges of an unnormalised input object to those of normalised object templates in memory. This yields a first, fast, but coarse categorisation, after which information at progressively finer scales is added to refine categorisation until final recognition has been achieved (Rodrigues and du Buf, 2009b). Line, edge and keypoint detection are based on models of simple, complex and end-stopped cells in V1 and V2. One questionable assumption was that keypoints have a dominant role in the dorsal stream, and lines and edges in the ventral stream. This assumption was based on the now abandoned idea of a strict dichotomous organisation: in the meantime there is substantial evidence that the two streams are communicating at many if not at all levels, and that some processing may be common to both (Konen and Kastner, 2008; Farivar, 2009). The latter is supported by the processing of geometric

shapes in both areas LIP (dorsal) and AIT (ventral), although activation in LIP is faster: 62 ms *vs.* 101 ms (Lehky and Sereno, 2007).

One of the two key questions has been answered: the same geometrical shapes are processed in both data streams and the dorsal one is faster. The later only makes sense if the dorsal stream also goes, via areas MT and MST, to the Superior Colliculus for eye and head control. But we wrote above that area MT is fed by area V2, where lines, edges and keypoints may be extracted and processed for both data streams. That those two streams exist is now generally accepted, also the fact that the border between them may be rather fuzzy, yet how about other data streams? For answering this question we must take a closer look at the retina.

2.1.3 *Non-standard retinal ganglion cells*

Cones (also rods) are photoreceptors which sample the image projected onto the retina. *Horizonal* and other cells combine the samples into concentric bandpass functions, as ON and OFF signals, often visualised as Mexican hat or DoG functions. Retinal ganglion cells transmit these signals to the LGN in the thalamus, which relays them on to the Primary (Striate) Visual Cortex (V1) where simple, complex and hypercomplex cells reconstruct anisotropy for building multi-scale line, edge and keypoint representations. Most of the ganglion cells feature high spatial resolution for the ventral data stream, less cells are devoted to motion for the dorsal stream, and both types are called standard retinal ganglion cells. It is now known that there also are, although even less, non-standard ganglion cells devoted to other functions: these are *direction-selective* cells, *local edge-detection* cells and *suppressed-by-contrast* cells. All these project not only on the SC for direct motor control, but also on the LGN and higher areas MT, V2 etc. (Masland and Martin, 2007); see Figure 2.1. In other words, the retina is much more intelligent than thought before, although much work remains to be done in order to better understand the role of the non-standard cells. Such cells

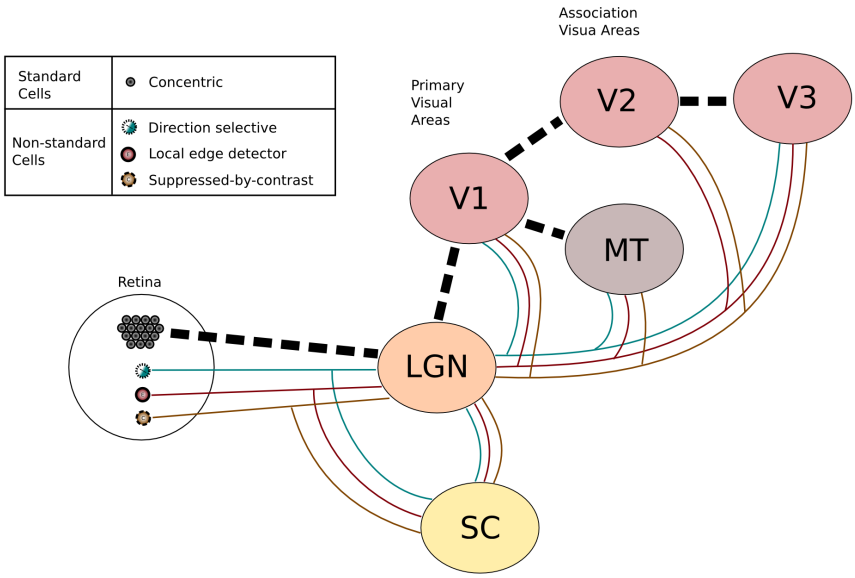


Figure 2.1: Non-standard retinal ganglion cells and pathways. Figure adapted from Masland and Martin (2007).

and their pathways may explain the phenomenon of blind sight: some persons or primates without visual cortex, who are effectively blind, can nevertheless avoid obstacles. A monkey named Helen was even able to detect, localise and discriminate visual objects (Stoerig and Cowey, 1997). Hence, there must exist more pathways which complement the dorsal and ventral ones. Another role could be for face detection and recognition, because faces are not normal 3D objects which can be arbitrarily rotated and they play an important role in social contexts. Their processing must be very fast and may circumvent the pathway for normal 3D objects (Biederman and Kalocsai, 1997).

The second key question has now also been answered: if static and moving edges are already detected in the retina and this information is conveyed by non-standard retinal ganglion cells, either directly or indirectly, to areas like MT and SC which are involved in motion prediction and eye-head control, it makes sense to assume that bottom-up attention focuses on entire (moving) objects. Such areas probably have sufficient computational resources to apply some principles of Gestalt Theory, like good continuation in case of edges which are partially occluded. Also for figure-ground,

since motion and disparity may easily separate a moving object from its background, and for the extraction of elementary geometric shapes like squares, triangles and circles which are abundant in the man-made world. Hence, local gist vision of meaningful objects may be obtained or at least prepared at a very early stage, and even in the dorsal data stream.

2.1.4 *Behavioural studies involving geometric shapes*

The geometric shapes used by Lehky and Sereno (2007) include a square, a triangle and a circle, which in most if not all countries are used for traffic signs. One would expect that studies related to traffic safety underpin the importance and meaning of such shapes, including the speed of detection of and discrimination between them. Indeed, there exist studies, but mainly behavioural ones addressing eye movements and fixations. For example, Luoma (1992) found that, while driving a car, traffic signs are analysed with an average glance duration of 500 ms, and glances as short as 100 ms may be enough to identify sign shape and colour. The average fixation time of 500 ms was confirmed by Martens and Fox (2007), but: (a) total fixation times including repeated fixations of the same sign were 500 ms with a standard error also close to 500 ms, such that total time varied between 0 and 1 s; (b) the total number of fixations ranged from 0 to 2; (c) there was a large variability between signs with different shapes, between individually mounted signs and signs mounted next to one or two other signs; (d) there was almost no difference between real driving and looking at a video; and (e) fixation times became shorter when the road became more familiar to the driver. The last effect may be due to the roles of procedural and episodic memories while simultaneously observing the road and controlling the car. Interestingly (or alarmingly?), shortest fixation times were measured for round and triangular signs with a red border (pedestrian; speed limit) when these were mounted next to two other signs, and for an individual white round sign (end speed limit), all also with the small-

est number of fixations. Fixations were measured at distances smaller than 250 m, with a normal speed of perhaps 50 km/h which means a maximum visibility of 18 s. This may imply (1) that most drivers were overloaded and simply ignored multiple signs, paying more attention to individual signs, or (2) for most drivers one brief glance at some distance was enough to grasp shape and meaning. If the ventral data stream is required to separate and analyse combined signs with one or two fixations, this stream may be too slow (option 1), whereas the dorsal stream may analyse fast individual signs (option 2).

Unfortunately, Martens and Fox (2007) only used signs with a symbol in the centre, which is a complication when one is only interested in the effect of a sign's shape. Karttunen and Häkkinen (1982) studied the discrimination of traffic signs in peripheral vision. They used 10 signs, only two of which without a symbol in the centre: a triangle and a circle. In all experiments these scored highest, i. e., completely correct discrimination of shape, colour and symbol, the latter lacking of course. The size of the signs was 4 degrees, which in practice corresponds to an observation distance of about 10 m, and the presentation time was 125 ms. Hence, the size was comfortable, but the presentation time was too short to identify also the symbol, if present. These results were obtained at peripheral angles beyond 30 degrees where retinal resolution is reduced; at angles below 30 degrees the differences between the scores were less. If flanked by two other signs, one above and one below, the same two signs also scored highest, but a bit lower if compared to the non-flanked condition. The discrimination results of Karttunen and Häkkinen (1982) seem to confirm the eye-fixation results of Martens and Fox (2007), namely that symbols in the centres of traffic signs complicate recognition. This effect might be due to the fact that fast and low-level gist of geometric shapes in the dorsal pathway is a possibility, but it has limitations because it is not intended for the recognition of more complex patterns. The later require (para-)foveal vision and probably processing in the ventral pathway. This is supported by Lehky and Sereno

(2007), who found that activation in the dorsal area LIP is faster than that in the ventral area AIT, but AIT neurons are more selective to different patterns than LIP neurons.

There are also studies concerning other abstract shapes. Solving a Tangram puzzle with seven pieces, three small and two big triangles plus a square and a parallelogram, average fixation times while only observing the pieces — not solving one of two puzzles — was about 245 ms with a standard deviation of about 55 ms, hence with most times between 200 and 300 ms (Baran et al., 2007). A task involving reproduction of the "Rey–Osterrieth Complex Figure (ROCF)," a neuropsychological assessment test which consists of many geometric shapes in a complex line drawing, revealed fixation times between 260 and 440 ms with a median of 320 ms (Manor et al., 1995). Manor and Gordon (2003) evaluated two temporal fixation thresholds of 100 and 200 ms, i. e., the minimum time gaze must be stable at the same position to count as a fixation, where 200 ms is standard practice in reading experiments. Using a free viewing condition, fixation times were measured while observing the ROCF and a photograph of a human face with neutral expression. With a threshold of 100 ms the median fixation times were shorter — with a 200 ms threshold, shorter fixations are excluded of course — and the ROCF yielded shorter fixation times (346 $\pm$ 114 ms) than the face image (396 $\pm$ 108 ms). The difference can be explained by the fact that the ROCF requires serial processing of elementary shapes, perhaps in the dorsal stream, whereas faces are processed holistically, possibly in a pathway dedicated to faces (Biederman and Kalocsai, 1997).

2.1.5 *From Gestalt Theory to application*

Gestalt Theory has been and still is a useful paradigm to understand *how* and *why* we perceive certain structures, from very simple and elementary patterns to rather complex ones which reveal meaning on the basis of learned interpretations (Pinna, 2010). Since our visual system is still much

more efficient and reliable than most systems developed in computer vision, for example in robotics, it makes sense to apply our knowledge of the visual system, via development of advanced models, to real-world problems. Autonomous service robots, for example, must be able to deal with complex and cluttered scenes, often containing complex objects which must be recognised and manipulated. Apart from scene and object complexity, digital images captured by modern cameras may have a good resolution, but they still contain noise due to digitisation and environmental factors like edges caused by non-uniform illumination. We will mention only two approaches. The first one is a nice example of the explicit application of the grouping rules of Gestalt Theory. The second does not apply such rules, but it represents an ideal case which could profit from advanced models of local gist and attention.

In their recent paper, Pugeault et al. (2010), who already applied grouping rules for the detection of lines and edges in their previous work, combined line and edge detection in 2D images with stereo disparity. Using Gestalt's grouping rules as constraints, they showed that 2D detection combined with 3D information leads to much more robust detection, especially in image regions where many features are very close and provide ambiguous information caused by local complexity. In other words, the application of constraints based on grouping rules is able to disambiguate such local information, leading to more consistent and complete 3D object and scene representations. Given the facts that the cortical hypercolumns in area V1, originating from retinotopic projections of the left and right eyes, are very close there, and that simple and complex cells in V1 serve to code lines and edges, it is very likely that our visual system extracts 3D shape information already in V1 and attributes depth to (mainly vertical) lines and edges. In addition to exploiting optical flow, this facilitates segregation (figure-ground) into meaningful and 3D objects.

Faubel and Schoner (2009) developed a system for simultaneous localisation and recognition of objects in a scene. They apply a circular Gaus-

sian "receptive field" (RF) to extract a colour histogram of pixels with saturated colours, which is invariant to 2D rotation, and an edge-orientation histogram, which is cyclically rotated after 2D rotation. These histograms are complemented by a shape descriptor resulting from maximum pooling. First, objects are learned by positioning them in the RF, where each object can be represented by multiple views and therefore multiple histograms.

Then, a scene is analysed by covering the entire image with partly overlapping RFs. For the sake of simplicity we will explain the analysis here using only one histogram, for example colour. In the first step, the histograms of all RFs are multiplied by weight factors and summed, which yields one histogram. This histogram is correlated with the histograms of all objects in memory, and the correlation factors are subjected to competition in order to suppress unlikely objects. Then all histograms in memory are multiplied by the reduced correlation factors and summed, which yields again one histogram. This histogram is correlated with all histograms of the RFs in the image, and the correlation factors are subjected to competition, as before, but now to suppress unlikely object positions. The reduced correlation factors are used as weight factors in the first step. Hence, the analysis is done in a closed loop from input space to memory and back to input space. This entire process is controlled by a dynamic neural field system of the parameters over space and time, such that the solution can converge to one object at one position.

The closed loop resembles the processing in the visual system, with bottom-up and top-down projections which converge to a stable solution on the basis of adaptation or plasticity at different levels. First localisation is obtained (where) and then precise object recognition with pose (view) estimation (what), in which different weight factors of the histograms and shape descriptors are applied. The main problems, of course, are that only one object can be dealt with at any time, that objects may be partially occluded, that the sums of the histograms of two different objects may resemble the histograms of one other object, and that a complex background may

lead to false positives. However, our visual system must deal with the same problems. As explained before, our visual system serially fixates regions on the basis of saliency, conspicuity and attention, the latter also driven by gist vision. Therefore, the system as developed by Faubel and Schoner (2009) could be modified such that it first applies object categories to an entire scene — which types of objects are about where — and then localised attention windows for identifying objects in those windows instead of the entire scene, but, as for the visual system, this step could be done serially.

2.1.6 *Summary and outlook*

In summary, there is evidence for differences but also similarities of the processing in different pathways in our visual system. The average fixation times of 245, 320 and 346 ms mentioned above are longer than the times of 62 and 101 ms in areas LIP and AIT as measured by (Lehky and Sereno, 2007), but the latter are *activation* times, i. e., onsets of neural activities. After the onsets, activities of neurons in both areas reach a peak and then decay, but they remain active until about 500 ms after stimulus onset. This indicates that the sets of measured neurons are part of bigger populations which serve to process the input patterns, but it is not yet quite clear what these populations do and how they do it, especially in the posterior parietal cortex in the dorsal stream. Different populations in the inferior temporal cortex in the ventral stream code categorical structures like bodies and hands, in principle the basic information which must be combined to detect and identify entire objects. In any case, it seems that local gist vision, at least involving elementary geometric shapes of man-made objects, is possible in early vision and also in the dorsal data stream. Specifically, below we focus on man-made objects which are dominated by a simple shape repertoire: squares, rectangles, trapeziums, triangles, circles and ellipses. It is shown that such shapes can be detected by a hierarchy of a few cell layers, with strictly bottom-up or data-driven processing. As we will see, straight bars

and curves known from Gestalt Theory must be complemented by corner information for such shapes, and all information must be combined by a few grouping rules which specify each shape.

The rest of this chapter is organised as follows: the next section deals with different cell layers to obtain low- and mid-level geometry. Section 2.3 explains shape retrieval using mid-level geometry. In Section 2.4 we discuss our approach and lines for future research.

2.2 LOW AND MID-LEVEL GEOMETRY

In this section we explain the process of preparing shape retrieval by low- and mid-level geometry. This is a two-fold process: we first construct a hierarchical *cell layer map* which encodes local geometric primitives by grouping cells: the primitives' *type* and *orientation*. Then, this information is used to detect geometric shapes based on spatial relationships between grouping cells.

2.2.1 *Cell-layer map construction*

We postulate a bottom-up hierarchy of cell layers in which each layer serves a specific purpose: (1) colour normalisation and boundary enhancement which mimic double-opponent cells (Bomberger and Schwartz, 2005), (2) detection of salient image points and regions, (3) enhancement of the most salient features, (4) determination of feature properties like orientation, aperture and curvature, (5) feature type assignment, (6) corner grouping cell condensing, and (7) object shape identification. Below the layers are explained in detail.

2.2.1.1 *Light source normalisation layer*

The input image $I_{in}(x, y)$ is first colour corrected ($I_{cc} = f_{cc}(I_{in})$), taking into account the geometry and temperature of the light sources. Let each pixel P_i of image $I_{in}(x, y)$ be defined as (R_i, G_i, B_i) and (L_i, a_i, b_i) in the RGB and Lab[1] colour spaces, with $i = \{1...N\}$, N being the total number of pixels in the image.

We first process the input image I_{in} using the two transformations described by Finlayson et al. (1998) and Martins et al. (2009), as shown below in (2.1) and (2.2), both in RGB colour space. This method applies iteratively steps A and B ($P_i^A \rightarrow P_i^B \rightarrow P_i^A \rightarrow P_i^B \rightarrow \ldots$), with step A being *local* and step B being *global*, until colour convergence is achieved, usually after 4–5 iterations. Each individual pixel is first corrected in step A for illuminant geometry independency (i. e., *chromaticity*),

$$P_i^A = \left(\frac{R_i}{R_i + G_i + B_i}, \frac{G_i}{R_i + G_i + B_i}, \frac{B_i}{R_i + G_i + B_i} \right), \tag{2.1}$$

followed in step B by global illuminant colour independency (i. e., *grey-world normalisation*),

$$P_i^B = \left(\frac{N \cdot R_i}{\sum_{j=1}^{N} R_j}, \frac{N \cdot G_i}{\sum_{j=1}^{N} G_j}, \frac{N \cdot B_i}{\sum_{j=1}^{N} B_j} \right). \tag{2.2}$$

After the process is completed, the resulting RGB image is converted to Lab colour space and the a_{cc} and b_{cc} components, where subscript cc stands for colour-corrected, are combined in I_{cc} together with the *unmodified* L_{in} channel from the input image I_{in}, as depicted in Figure 2.2. The main idea for using the Lab space is that it mimics double-opponent colour cells found in human vision, making it more useful for determining the conspicuity of borders between regions. The reason for using the L_{in} component instead of the L_{cc} one is that, as observed by Finlayson et al. (1998), the simple and fast repetition of steps A and B does a remarkably good job.

1 We use Lab as an informal abbreviation for the CIE 1976 (L*, a*, b*) colour space.

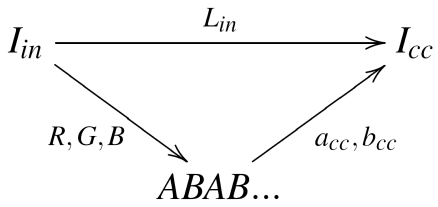


Figure 2.2: Light source normalisation layer. The output image I_{cc} consists of the L_{in} component from the input image I_{in} and the a_{cc} and b_{cc} components after applying the algorithm of Finlayson et al. (1998).

In fact, it does the job too well because all grey pixels (with values R=G=B from 0 to 255) end up having R=G=B=127. In other words, all information in grey image regions will be lost. Hence, we use only the colour corrections of the method, while keeping the lightness channel of the original image for maintaining all the image's details.

Figure 2.3 shows three results of colour correction applied to the traffic sign image, *from top-left to top-right:* original image, modified image with a blue tint (R -12%, G +4% and B +50%), and modified image with a warm white balance. The three results are shown below the input images. As can be seen, colour correction yields very similar images despite the rather large differences in the input images. Colour correction as explained above simulates colour constancy as employed in our visual system (Hubel, 1995).

Summarising, the initial I_{in} image in RGB is normalised to I_{cc} and then converted to the colour space $L_{in}a_{cc}b_{cc}$.

2.2.1.2 *Adaptive colour filtering layer*

After colour correction, image regions are smoothed for removing redundant information which is not necessary for shape detection, while preserving the region's boundaries. The smoothing is done using an adaptive filter $\Gamma(x, y)$, with separable and equal $\Gamma^H = \Gamma(x)$ and $\Gamma^V = \Gamma(y)$ components for horizontal and vertical filtering. Each component consist of a centred Difference-of-Gaussians (DoG)

$$F_{1,2}(x) = N_1 \cdot \left\{ \exp\left(\frac{-x^2}{2\sigma_1^2}\right) - \exp\left(\frac{-x^2}{2\sigma_2^2}\right) \right\}, \tag{2.3}$$

Figure 2.3: *Top and middle:* colour illuminant and geometry normalisation; input images *(top)* and respective results *(middle)*. *Bottom, left to right:* adaptive colour-region filtering, border saliency by colour conspicuity, and non-maximum suppression.

which is split into $F_1(x < 0)$ and $F_2(x > 0)$, and a centred Gaussian which is *not* split,

$$F_3(x) = N_2 \cdot exp\left(\frac{-x^2}{2\sigma_2^2}\right), \tag{2.4}$$

taking $\sigma_1 \gg \sigma_2$. N_1 and N_2 are normalisation constants which make the integrals of all three functions equal to one.

The three functions implement a group of three summation cells at the same position, but with different dendritic fields in the colour-opponent

channels a and b of Lab colour space. F_3 yields the excitatory response of the cell with an *on-centre* dendritic field, whereas $F_{1,2}$ yield the excitatory responses of the two cells with *off-centre* dendritic fields. From the three cell responses $R_{1,2,3}$ we first compute the contrast C between the left (R_1) and right (R_2) responses,

$$C = \left| \frac{R_1 - R_2}{R_1 + R_2} \right|. \tag{2.5}$$

Then, using the contrast C and the minimum difference between the on-centre response R_3 and the left and right off-centre ones R_1 and R_2, the response R of an output cell is determined by

$$R = \begin{cases} C\,R_1 + (1 - C)\;R_3 & \textit{if} \quad |R_1 - R_3| < |R_2 - R_3| \\ C\,R_2 + (1 - C)\;R_3 & \textit{otherwise}. \end{cases} \tag{2.6}$$

In words, if the local contrast is low, as in almost homogeneous regions, the filter support is big, but if the contrast is high, at the boundaries between regions, the filter support is small. The adaptive filtering is applied to I_{cc} at each pixel position (x, y), first horizontally with Γ^H and then vertically with Γ^V,

$$I_{ci}(x, y) = \Gamma^V \left[\Gamma^H \left[I_{cc}(x, y) \right] \right], \tag{2.7}$$

where subscript ci stands for colour-improved. Results after both steps, for the illustration image, are shown in Figure 2.3 *(bottom–left)*. It can be seen that the filter not only preserves boundaries, but also sharpens blurred ones. In our experiments we obtained good results with $\sigma_1 = 7$ and $\sigma_2 = 3$, and adaptive filtering in horizontal and vertical directions was sufficient to sharpen blurred boundaries even with oblique orientations. Furthermore, the processing is very fast because the three filter functions need only be computed once.

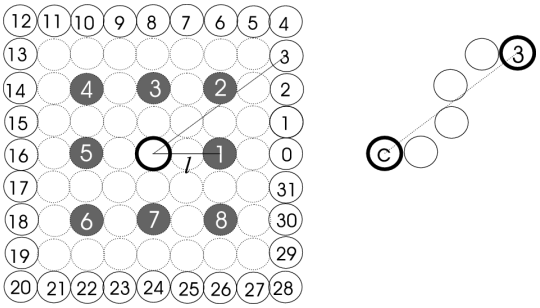


Figure 2.4: *Left:* conspicuity cell clusters, in grey: the four clusters of gating cells at positions (1,5), (2,6), etc. used for colour conspicuity; in white: a cluster of 32 cells for orientation layer Θ. *Right:* an example of a possible path between the centre cell (c) and cell 3.

2.2.1.3 *Colour conspicuity layer*

Following the idea of Martins et al. (2009), *conspicuity* $\Psi(x, y)$ is defined as the maximum difference between colours in I_{ci} at four pairs of symmetric positions at distance l from (x, y), i. e., on horizontal, vertical and two diagonal lines. However, here we apply a new concept of conspicuity directly to I_{ci}, effectively discarding the need for a previous edge-filtering step.

Figure 2.4 *(left, in grey)*, shows the positions of the clusters of gating cells. If the gating cells are called G_i, opposite pairs are (G_i, G_{i+4}), with $i = \{1, \ldots, 4\}$, for example (G_1, G_5) and (G_4, G_8). We define conspicuity Ψ as the maximum Euclidean distance between the pairs of colour triplets (L, a, b) of the four pairs of opposite cells around (x, y),

$$\Psi_{Lab}(x, y) = \max_{i=1}^{4} \left(\sqrt{\sum \left(I_{ci}^{L,a,b}(\vec{x}_i)^2 - I_{ci}^{L,a,b}(\vec{x}_{i+4})^2 \right)} \right). \tag{2.8}$$

The result of this layer as shown in Figure 2.3 *(bottom–middle)* was obtained by using a distance $l = 1$ and a threshold at 0.4 of $\max(\Psi_{Lab})$ in the entire image.

2.2.1.4 *Non-maximum Suppression Layer*

In this cell layer Ω non-maximum suppression is applied in order to extract the positions where $\Psi(x, y)$ has a local maximum in horizontal, vertical and

diagonal directions, in 3×3 neighbourhoods. As in the previous layer, this is achieved by four oriented cell clusters plus one grouping cell at the output. Mathematically,

$$\Omega(x,y) = \begin{cases} \text{ON} & \text{if } (\Psi(x_i,y_i) > \Psi(x_{i-1},y_{i-1}) \wedge \Psi(x_i,y_i) > \Psi(x_{i+1},y_{i+1})) \\ & \vee (\Psi(x_i,y_i) > \Psi(x_{i-1},y_{i+1}) \wedge \Psi(x_i,y_i) > \Psi(x_{i+1},y_{i-1})) \\ & \vee (\Psi(x_i,y_i) > \Psi(x_{i-1},y_i) \wedge \Psi(x_i,y_i) > \Psi(x_{i+1},y_i)) \\ & \vee (\Psi(x_i,y_i) > \Psi(x_i,y_{i+1}) \wedge \Psi(x_i,y_i) > \Psi(x_i,y_{i-1})) \\ \text{OFF} & \text{otherwise.} \end{cases} \tag{2.9}$$

Results for the illustration image are shown in Figure 2.3 *(bottom–right)*.

2.2.1.5 *Local feature layers*

In order to extract meaningful information from layer Ω it is necessary to analyse local geometric relations between adjacent activated cells. We use three parallel cell layers Θ, Υ and Λ, which are dedicated to orientation, curvature and connectivity, respectively.

The **orientation layer** Θ encodes edge orientations in local neighbourhoods. Each active cell ($\Omega(x,y) = \text{ON}$) triggers a cluster of 32 cells, each with two dendrites: one at the Ω cell's position (x,y) and one at a (discretised) distance of four cells (pixels) around that position, for a total of 32 orientations. This is illustrated in Figure 2.4 *(left, in white)*, with orientations n numbered 0 to 31. Those of all 32 cells with dendritic input equal to 2 are excitated and the others are inhibited. Excitated cells provide the output of the Θ layer, the cells themselves implicitly coding all detected local orientations in the Ω layer. The dimension of the Θ layer is 32 times that of the

Ω layer to accomodate all possible local edge orientations. Mathematically, the orientation equals

$$\phi_n = \begin{cases} \arctan\left(\frac{\theta}{4}\right) & \text{if } 0 \leqslant n \leqslant 4 \\ \arctan\left(\frac{4}{|8-\theta|}\right) & \text{if } 4 \leqslant n < 8 \\ \frac{\pi}{2} & \text{if } n = 8 \end{cases} \tag{2.10}$$

in the case of the first quadrant ($0 \leqslant n \leqslant 8$) and similarly for the other quadrants.

The **curvature layer** Υ is composed of clusters of curvature detection cells. These cells are also triggered by active output cells of the Ω layer and they also analyse active output cells of the Ω layer at a distance of about four cells (pixels). However, instead of combining the centre position (x, y) and one on a circle around it as in the Θ layer, they combine pairs of positions at near-opposite orientations on the circle (active Ω cells at exactly opposite orientations from the centre indicate zero curvature). In addition, since evidence for different local curvatures must be combined by grouping cells which determine the average curvature, evidence for curved edges on, for example, the left and right sides of (x, y) cannot be grouped because the average may be close to zero. Therefore, information on all semi-circles is grouped and the output of layer Υ is composed of 16 times 2 cells. Mathematically, the curvature model resembles computing the cluster curvature index $C_i(x, y)$ of all intersections of lines perpendicular to lines between all point pairs on a semi-circle, i. e., the mean Chebyshev distance between the N intersections (x'_k, y'_k) and the centre (x, y):

$$C_i(x, y) = \frac{1}{N} \sum_{k=1}^{N} \left[\max\left(\left|x - x'_k\right|, \left|y - y'_k\right|\right)\right] \ . \tag{2.11}$$

The **connectivity layer** Λ also analyses the output of the Ω layer, but it employs the outputs of the Θ and Υ layers. Active or excitated output

cells in those layers trigger clusters of grouping cells in the Λ layer: all detected orientations (Θ) and curvatures (Υ) trigger grouping cells which check whether there are active output cells in the Ω layer which connect the centre position (x, y) with the corresponding active cells on the circle around the centre (Figure 2.4, *at right*). If so, output cells of the Λ layer are activated and these signal connectivity in the corresponding orientations.

In addition to the three layers described above, there is on top of Θ a $\hat{\Theta}$ layer which uses information of the connectivity layer Λ. This layer groups all detected orientations (active Θ cells) with confirmed connectivity (active Λ cells) for determining the *average orientation* on the entire circle. If there are opposite orientations, the average orientation can be orthogonal, but this information will be combined with other information at a higher level for distinguishing between corners and bars. The $\hat{\Theta}$ layer is complemented by a $\check{\Theta}$ layer which determines the *angular aperture*, i. e., the spread of all orientations around the average orientation. If the aperture is small, this is evidence for a corner, but a large one indicates a continuous structure like a bar. Mathematically, the angular aperture $\check{\phi}(x, y)$ and the average orientation $\hat{\phi}(x, y)$ are defined by

$$\check{\phi}(x,y) = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} \left|\phi_i - \phi_j\right| \wedge i \neq j \wedge 180^\circ \geqslant \left|\phi_i - \phi_j\right| > 2 \cdot \frac{360^\circ}{32}, \tag{2.12}$$

$$\hat{\phi}(x,y) = \frac{1}{N} \sum_{i=1}^{N} \phi_i \,, \tag{2.13}$$

with ϕ_i being the angles of active cells in the Θ layer.

2.2.1.6 *Mid-level geometry*

This layer (Ξ), serves to translate the local low-level features into meaningful geometric primitives: straight bars, curves and corners. This is achieved by combining the information in the three previous cell layers Θ, Λ and Υ, more specifically, in $\hat{\Theta}$, $\check{\Theta}$ and Υ. But the processing is still local: layer Ξ

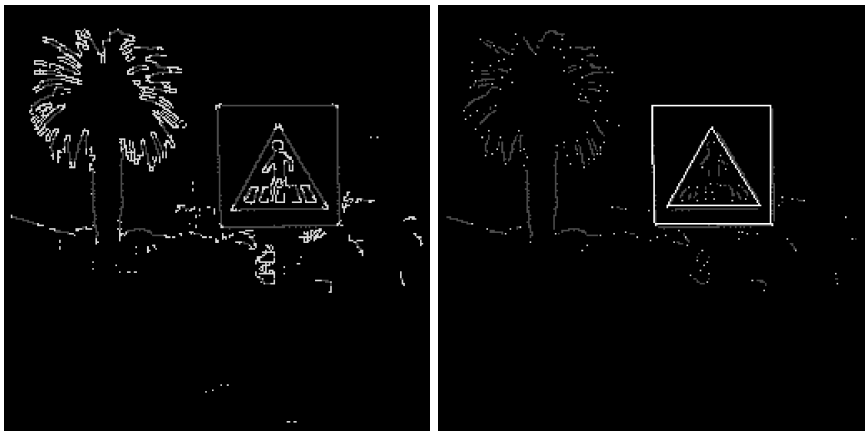


Figure 2.5: *Left:* cell layer Ξ^{BCo} for the illustration image before corner condensing. *Right:* after corner condensing and shape retrieval.

assigns a geometric primitive to each active Ω cell by a one-to-one mapping. Layers $\hat{\Theta}$ and $\check{\Theta}$ provide orientation angles and apertures for corners and straight bars, whereas layer Υ provides curvature information. There are two parallel Ξ layers, Ξ^{BCo} and Ξ^{BCu}, the first for *bar–corner* cell clusters and the second for *bar–curve* cell clusters, with the purpose of fast shape processing in the further steps.

Mathematically, the angle aperture $\check{\phi}(x,y)$ and curvature index $C_i(x,y)$ are used to determine which feature type $\{t_1, t_2\}$ will be assigned to cells $\xi^{t_1}(x,y)$ and $\xi^{t_2}(x,y)$,

$$\xi^{t_1}(x,y) = \begin{cases} \text{Bar} & \text{if } \check{\phi}(x,y) \geqslant 120^\circ \\ \text{Corner} & \text{if } \check{\phi}(x,y) < 120^\circ \end{cases} \tag{2.14a}$$

$$\xi^{t_2}(x,y) = \begin{cases} \text{Bar} & \text{if } \check{\phi}(x,y) < 120^\circ \ \vee \ C_i(x,y) \leqslant 3 \\ \text{Curve} & \text{if } \check{\phi}(x,y) \geqslant 120^\circ \ \wedge \ C_i(x,y) > 3\,. \end{cases} \tag{2.14b}$$

An example of the assignment (layers Ξ^{BCo} and Ξ^{BCu}) is shown in Figure 2.5 *(left)*, with white pixels being corner cells, grey pixels being curve cells and dark grey pixels being bar cells.

An overview of the processing is illustrated in Figure 2.6, which shows the different cell layers in the case of a corner detection. Two major parts

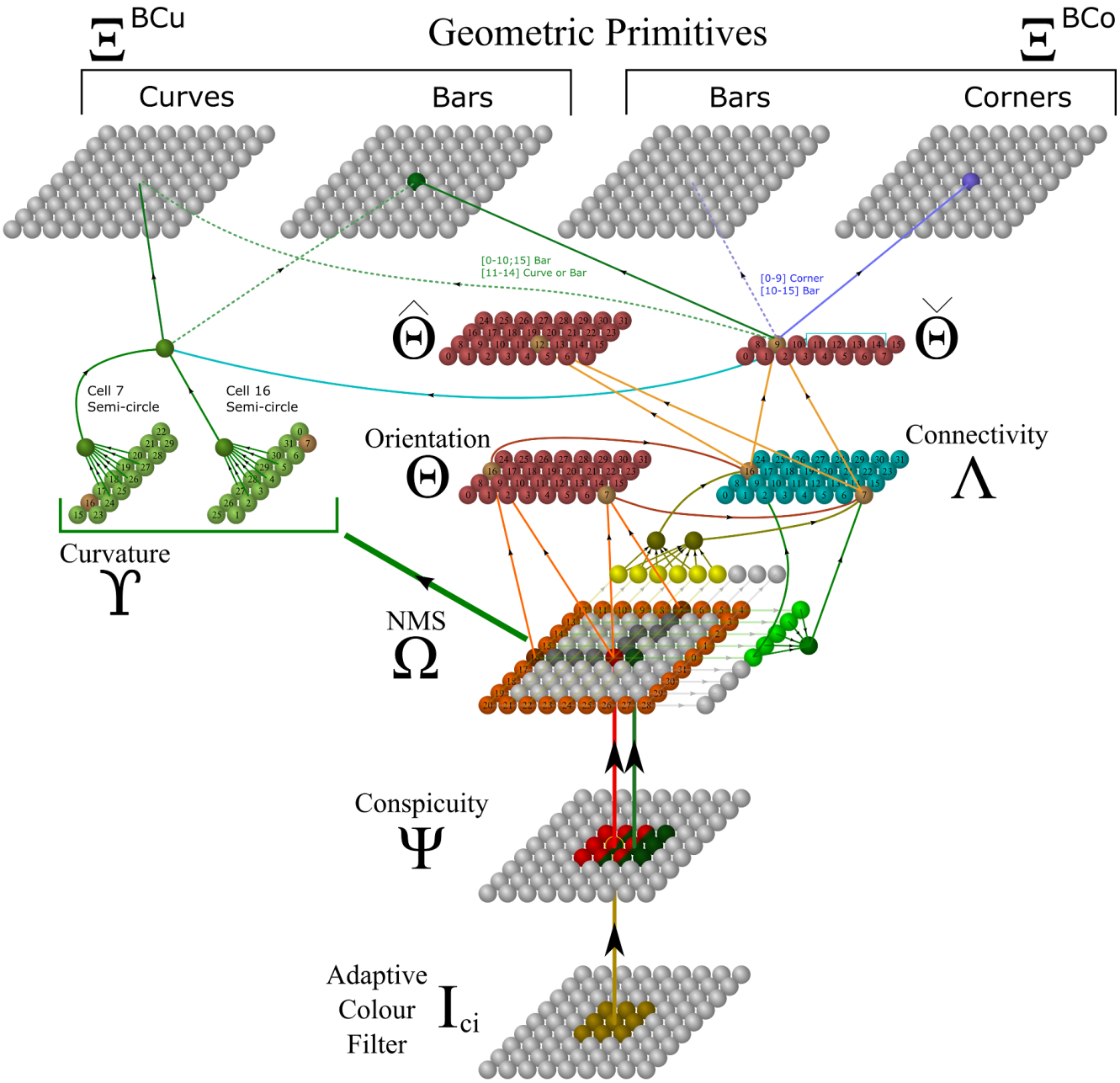


Figure 2.6: Overview of the whole process with the different cell layers, for a corner detection example. Conspicuity layer Ψ is a 9:1 mapping of layer I_{ci}. Similarly, non-maximum suppression layer Ω is a 9:1 mapping of layer Ψ. The corner in layer Ω is checked for curvature in Υ, for orientation in Θ, and for connectivity in Λ. It is then represented as a Bar in layer Ξ^{BCu} and as a Corner in layer Ξ^{BCo}.

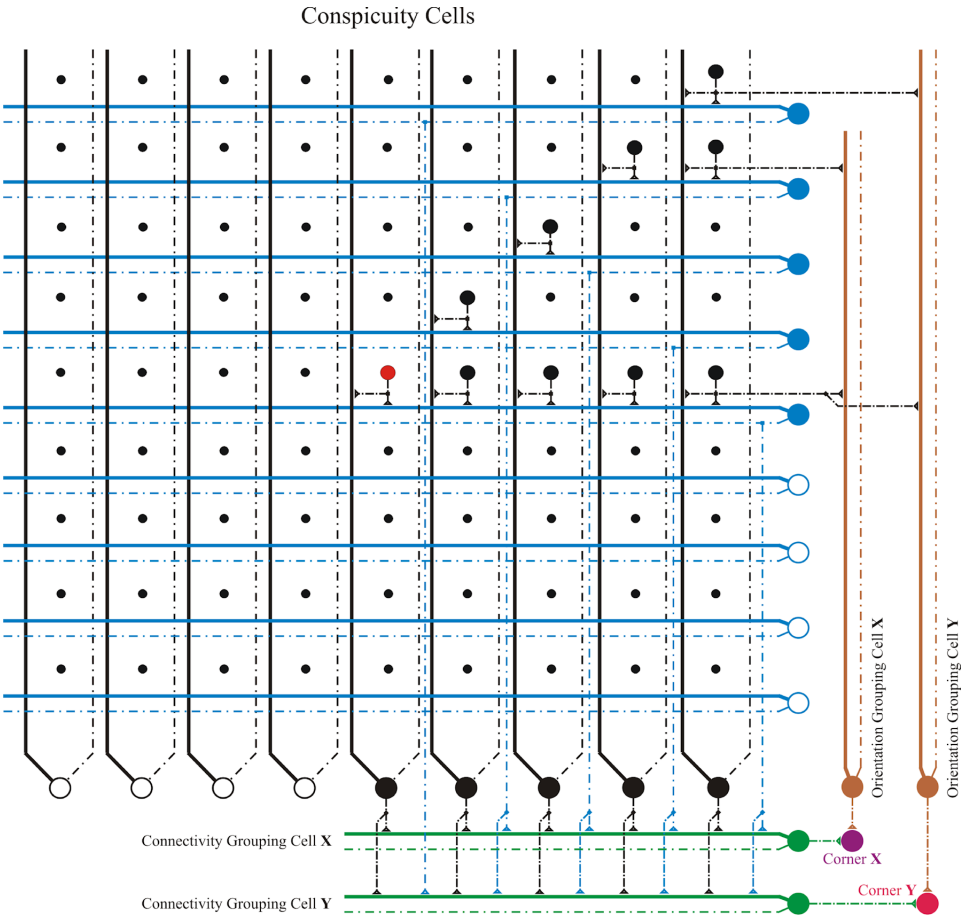


Figure 2.7: Cell matrix showing a corner cell.

of the processing are illustrated in more detail in Figures 2.7 and 2.8. In Figure 2.7, the Ω cells are the small black dots. Large black dots at the bottom are Λ *layer one* column connectivity cells and large blue dots at the right are Λ *layer one* row connectivity cells. For each active Θ orientation cell (brown), a specific Λ *layer two* connectivity cell (green) must also be active to detect a corner. In Figure 2.8, the required connectivity cells and orientation cells are different, in order to detect a bar.

Corner cells are subjected to one more processing step, *condensing*, similar to Krüger et al. (2003). First, all active corner cells that have six or seven inactive neighbouring cells are inhibited. Second, all groups of cells $\xi(x, y)$ in layer Ξ^{BCo}, in a 7×7 neighbourhood, are condensed into a single "centre-of-gravity" cell, with orientation and aperture angles equal to the averages of the group. This facilitates and speeds up further processing of corners

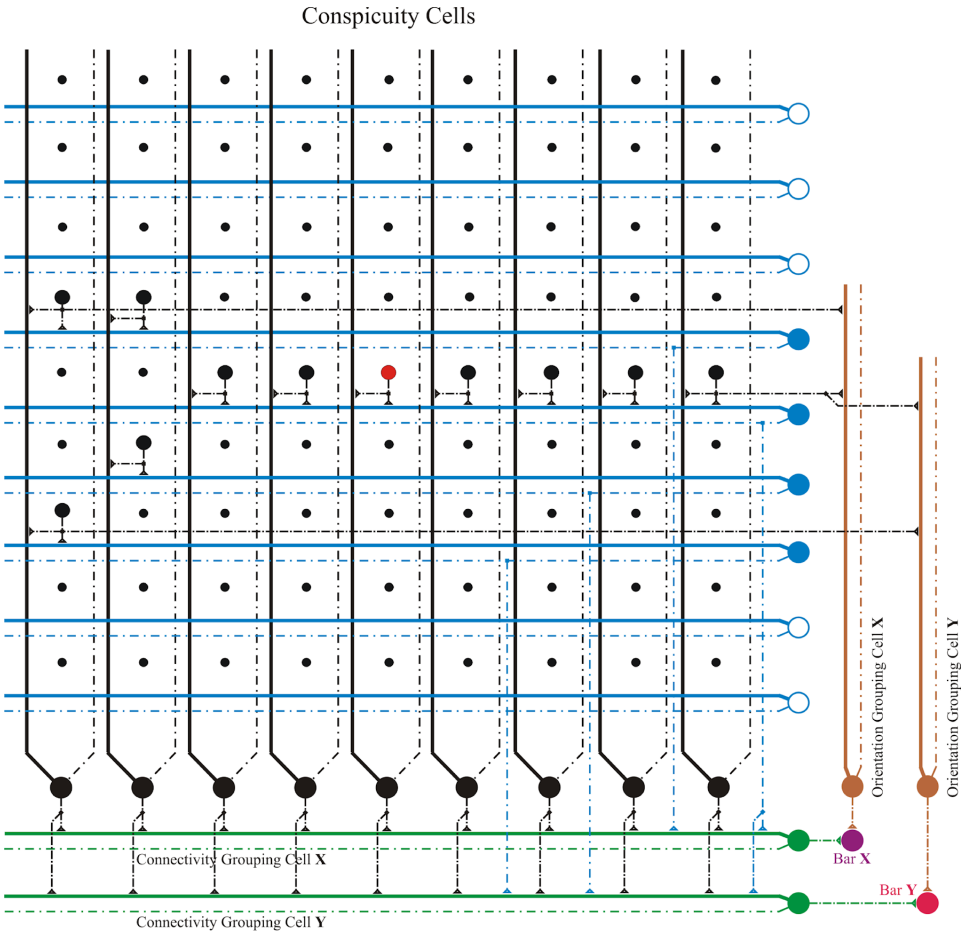


Figure 2.8: Cell matrix showing a bar cell.

for final shape recognition. An example, for the illustration image, is shown in Figure 2.5 *(right)*, where groups of corner cells have been replaced by a single cell *(in white)*.

In the practical implementation, condensed corner and curve cells have their geometric information stored in three arrays, for all possible cell pairs (ξ_i, ξ_j) in Ξ^{BCo} and Ξ^{BCu}, i. e., three arrays for corner pairs and another three for curve pairs, where i and j denote different coordinates (x, y):

1. $B^t_{i,j}$ for storing angle compatibility between pairs of corners or pairs of curves, i. e., if they have similar orientations within a specified margin as shown in (2.15a) for corners and in (2.15b) for curves;

2. $A^t_{i,j}$ for storing angles between cell pairs, see (2.15c); and

3. $D^t_{i,j}$ for distances between cell pairs, see (2.15d).

Mathematically, they are assigned as follows:

$$B_{i,j}^{\text{Corner}} = \begin{cases} \text{ON} & \text{if } \left[\hat{\phi}(\lambda_i) \pm \frac{1}{2}\check{\phi}(\lambda_i)\right] \wedge \left[\hat{\phi}(\lambda_j) \pm \frac{1}{2}\check{\phi}(\lambda_j) \pm \Delta_e\right] \neq \varnothing \\ \text{OFF} & \text{otherwise.} \end{cases} \tag{2.15a}$$

$$B_{i,j}^{\text{Curve}} = \begin{cases} \text{ON} & \text{if connected by active } \Omega \text{ cells} \\ \text{OFF} & \text{otherwise.} \end{cases} \tag{2.15b}$$

$$A_{i,j}^{t} = \arctan\left(\frac{y_j - y_i}{x_j - x_i}\right) \cdot \left(\frac{180^\circ}{\pi}\right) \tag{2.15c}$$

$$D_{i,j}^{t} = \max\left(\left|x_j - x_i\right|, \left|y_j - y_i\right|\right). \tag{2.15d}$$

Finally, array $B_{i,j}^{\text{Corner}}$ is also checked for main-diagonal symmetry, such that only bidirectionally connected features will remain,

$$B_{i,j}^{\text{Corner}} = B_{i,j}^{\text{Corner}} \cap B_{j,i}^{\text{Corner}}. \tag{2.16}$$

2.3 FINAL SHAPE RETRIEVAL

The shape repertoire is $S = \{S_\square, S_{\text{rect}}, S_\triangle, S_\circ, S_{\text{ell}}\}$, denoting square, rectangle, triangle, circle and ellipse. All possible combinations coded in layer Ξ are processed, and candidate shapes are validated using the following rules: the correct number of features, their relative distances, connectivity and internal angles, and the centre of the shape. Specifically:

1. A candidate shape must possess a **correct number of features** of the type Corner or Curve which match the shape model. A Square or Rectangle has to include four condensed Corner cells a to d. In the case of a Triangle there are three, a, b and c. Mathematically,

$$\exists\, \xi_i = \text{Corner},\ \forall i \in \{a, b, c, d\} \tag{2.17a}$$

$$\forall\{S_{\square}, S_{rect}\},\ B_{i,j} = \text{ON},\ \forall i, j \in \{a, b, c, d\}, i \neq j \tag{2.17b}$$

$$\forall S_{\triangle},\ B_{i,j} = \text{ON},\ \forall i, j \in \{a, b, c\}, i \neq j. \tag{2.17c}$$

In the case of a Circle or Ellipse, there must be three Curve cells, e, f and g:

$$\exists\, \xi_i = \text{Curve},\ \forall i \in \{e, f, g\} \tag{2.18a}$$

$$\forall\{S_{\circ}, S_{ell}\},\ B_{i,j} = \text{ON},\ \forall i, j \in \{e, f, g\}, i \neq j. \tag{2.18b}$$

2. The **relative distances** between the shape's features must also match the shape model, i. e., a Square must have four pairs of Corners with about the same distances. Similar but different processes are applied to the other shapes. It should be stressed that, in the particular cases of Squares and Rectangles, the distances are tested over adjacent corners, such that possible diagonals inside the shapes are inhibited:

$$\begin{aligned} &\forall\{S_{\square}, S_{rect}\},\ \min(D_{i,j}) > 0.5 \times \max(D_{i,j}), \\ &\forall i, j \in \{a, b, c, d\}, i \neq j \wedge (i = a \wedge j \neq d) \wedge (i = b \wedge j \neq c) \end{aligned} \tag{2.19a}$$

$$\forall S_{\triangle},\ \min(D_{i,j}) > 0.6 \times \max(D_{i,j}),\ \forall i, j \in \{a, b, c\}, i \neq j \tag{2.19b}$$

$$\forall\{S_{\circ}, S_{ell}\},\ (D_{e,f} > 4) \wedge (D_{f,g} > 4) \wedge (D_{e,g} > 8),\ \forall i, j \in \{e, f, g\}, i \neq j. \tag{2.19c}$$

3. The candidate shape must exhibit **connectivity** between shape features, especially between (condensed) Corners, i. e., they must be linked by Bar cells, with **confirmatory evidence** $CE(\xi_i, \xi_j)$ of connectivity, by

analysing the number of Bar cells with perpendicular orientations between Corners or Curves. Mathematically,

$$\mathrm{CE}(\xi_i, \xi_j) = \begin{cases} \text{OFF} & \text{if } \#\,\xi_{(i\to j)} < 0.8 \cdot \sqrt{(x_j - x_i)^2 + (y_j - y_i)^2} \\ & \vee\ \#\,\xi_{(i\to j)} \left[= \mathsf{Bar} \wedge \hat{\phi}(\xi_{[i\to j]}) \perp m\right] < 0.8 \\ \text{ON} & \text{otherwise,} \end{cases} \tag{2.20}$$

with # the number of cells and $m = (y_j - y_i)/(x_j - x_i)$.

4. For polygons, the sum of the **internal angles** of the candidate shape must match the model shape, mathematically

$$\forall\{S_\square, S_{\text{rect}}\},\ \sum A_{i,j} \approx 360^\circ, \forall i, j \in \{a, b, c, d\}, i \neq j \tag{2.21a}$$

$$\forall S_\triangle,\ \sum A_{i,j} \approx 180^\circ,\ \forall i, j \in \{a, b, c\}, i \neq j. \tag{2.21b}$$

5. For Circles and Ellipses, a **centre of the shape** is estimated using the intersections of lines perpendicular to tangents of Curve cells. The intersection point of two perpendicular lines yields an estimate of the shape's centre (x_c, y_c),

$$(x_c, y_c) = \begin{cases} y_c = m_1(x_c - x_{m1}) + y_{m1} \\ x_c = (y_c - y_{m2} + m_2 \cdot x_{m2})/m_2, \end{cases} \tag{2.22}$$

with $m_1 = (x_e - x_f)(y_f - y_e)$, $m_2 = (x_f - x_g)(y_g - y_f)$, $x_{m1} = (x_e + x_f)/2$, $x_{m2} = (x_f + x_g)/2$, $y_{m1} = (y_e + y_f)/2$ and $y_{m2} = (y_f + y_g)/2$.

This process is applied to all triplets of Curve cells in 10×10 neighbourhoods of the Ξ layer. Resulting intersection points are then averaged to obtain a single centre estimate. Of course, this solution is less accurate in the case of ellipses.

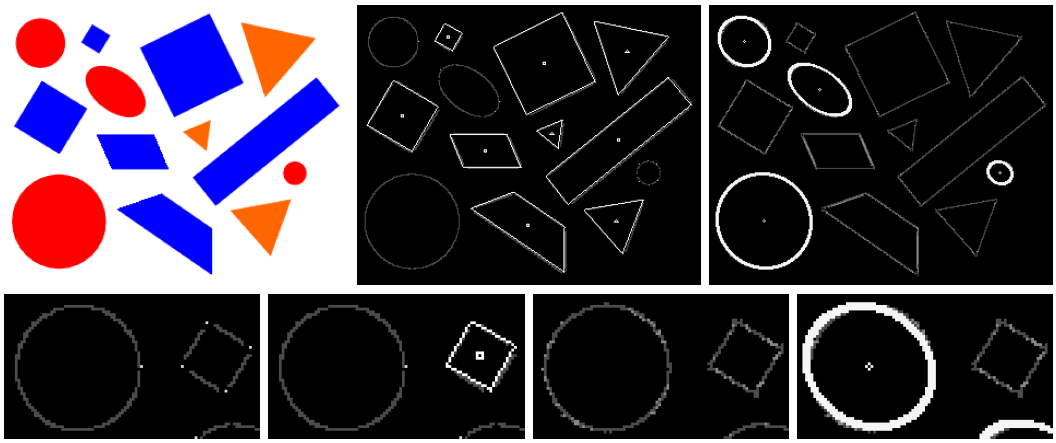


Figure 2.9: *Top row, from left:* artificial test image with different shapes, rotations and sizes, and detected shapes. *Bottom row:* magnification of the top-left corner of the test image, showing from left to right corner/bar detection with condensed corners, square detected, curve/bar detection, and circle detected.

In summary, specific shapes are detected by activating detection cells which apply the rules explained above: a Square and Rectangle have to obey the activation rules of Eqns 17a, b, 19a, 20 and 21a, all at the same time. For a Triangle these are Eqns 17a, c, 19b, 20 and 21b, and for a Circle and Ellipse Eqns 18a, b, 19c, 20 and 22 apply. Figure 2.5 (page 44) *(right)* shows an example of a detected Triangle and Square.

For a better comprehension of the results, Figure 2.9 *(top–left)* shows an artificial test image with different squares, rectangles, trapeziums, triangles, circles and ellipses, with different rotations and sizes. Detected shapes and their centres are shown to the *right*. The *bottom row* shows a magnification of the top–left corner of the test image with, *left to right:* corner/bar detection with condensed corners, the square detected, curve/bar detection, and the circle detected.

The *top* and *third row* in Figure 2.10 show two parts, of four frames each, of a sequence acquired from a moving car. The *second row* shows squares and triangles detected in the frames on the *top row*, whereas the *fourth row* shows circles and ellipses detected in the frames on the *third row*. The *bottom row* shows a frame with more traffic signs from another sequence and all detected shapes superimposed on the input image. All frames were resized

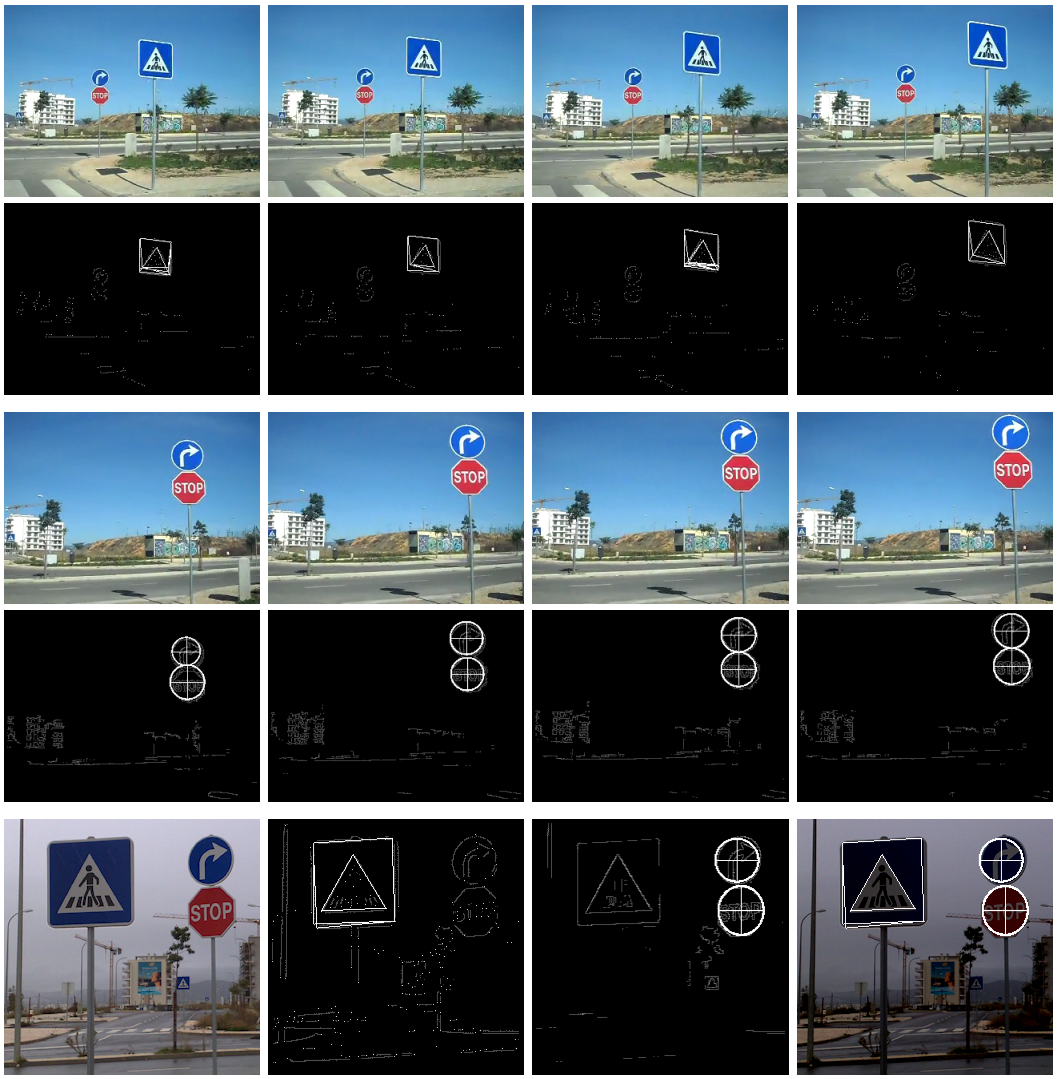


Figure 2.10: *Top four rows:* two series of frames of a sequence acquired from a moving car with detected squares and triangles *(2nd row)* and circles and ellipses *(4th row)*. *Bottom row, left to right:* another input image with traffic signs, detected triangles and squares, detected circles and ellipses, and all detected shapes superimposed on the input image.

to about 341×256 pixels, and the processing time of each was about 0.2 seconds on a normal PC. Figure 2.11 shows more results in the case of an office desk *(top row)*, a building, and a carpet with a golf theme, with successfully detected squares, trapeziums, triangles, circles and ellipses. In the golf carpet result, we can see that some balls have not been detected; these were too close to other objects such that their edges were not well separated.

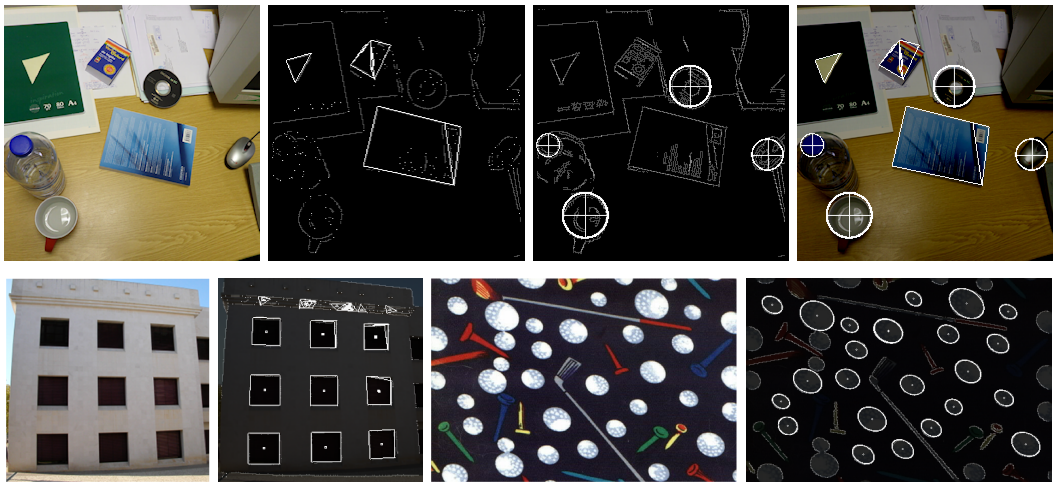


Figure 2.11: *Top, from left:* image of an office desk, detected triangles and squares, detected circles and ellipses, and all detected shapes superimposed on the input image. *Bottom:* results in the case of a building and a carpet with a golf theme.

As can be seen in Figs. 2.10 and 2.11, most important shapes as defined by colour contrast — colour conspicuity, the only information exploited here — have been detected, providing local gist of segregated objects in a spatial layout map, which can be used for subsequent object recognition by a sequential process steered by Focus-of-Attention (FoA).

2.4 DISCUSSION

Multi-scale representations of lines, edges and keypoints, extracted on the basis of simple, complex and end-stopped cells in cortical areas V1 and V2, can be used for invariant object categorisation and recognition (Rodrigues and du Buf, 2009a,b). These representations are complemented by saliency maps of colour, texture, disparity and motion information, which are thought to play an important role in FoA (Elazary and Itti, 2008; Martins et al., 2009). This processing is done in the normal pathway with ventral and dorsal (what and where) data streams, which proceed from the LGN via V1 etc. to the prefrontal cortex. These data streams are bottom-up but

with top-down attentional modulation from the prefrontal cortex down to the LGN (Saalmanna and Kastner, 2009).

As postulated by Rensink (2000), there may be two other subsystems for gist vision and spatial layout. These must be very fast, because they (1) serve to bias specific data paths related to specific objects in memory, i. e., the context serves to pre-select typical objects such that all objects held in memory which are out-of-context can be ignored, and (2) they prepare FoA for directing attention and our eyes to regions where important objects are expected. Global scene gist, for which computational models have already been developed (Bar, 2004; Siagian and Itti, 2007; Ross and Oliva, 2010; Rodrigues and du Buf, 2011b), cannot be directly linked to spatial layout, because the latter implies, by definition, a localised analysis: which types of objects are about where in a scene. The missing link can consist of local object gist, even with the possibility that this is extracted before global scene gist and, once local gist is available, global gist can be extracted completely at semantic level: detected objects determine the context and scene. In addition, object gist may contribute to solving the object segregation problem, i. e., if objects are complex in terms of coloured and textured regions. The class of general objects remains subject to further research, but here we have shown that at least *elementary* shapes indicating many man-made objects can be dealt with.

The model explored in this paper only exploits local colour contrast or colour conspicuity. As explained in the Introduction (Section 2.1), apart from the standard "retinal" ganglion cells there also are non-standard cells which code edges and their motion (Figure 2.1). Hence, colour information could be combined with texture and motion information for developing a retinal model of local object gist, retinal meaning the use of retinal information at some higher level, with the possibility that also disparity information could be integrated. The lowest level where this could happen is in the LGN, after the optic chiasm where information from the left and right eyes can be combined.

Figure 2.12: Partial occlusions. *Top:* because of colour contrast, only the inner circle and triangle of the traffic signs have been detected. *Bottom:* successful detection of square shapes on two partially occluded monitors in a laboratory scene.

The model as developed employs the normal processing strategy in the brain, with massively parallel processing at a low level and increasing complexity at a higher level. Using few cell layers it is possible to extract strictly local syntactical features and to combine them into "local-global" features like bars and corners, after which global semantics of elementary shapes can be extracted: man-made objects, often with square, rectangular, trapezoidal, triangular, circular and elliptical shapes. In addition, it is rather trivial to extract also the centres of these shapes. What is *not* trivial is to extract the shapes and their correct centres when they are partially occluded. In the current model, occlusions are already possible but only if the necessary corners of a shape are visible. Examples of detected shapes with such occlusions can be seen in Figure 2.12.

The shapes' rules as specified and applied in this paper must be complemented for dealing with other partial occlusions. This can be done with relaxation rules, such that a rectangle with one occluded corner can nevertheless be detected as a rectangle and not as a triangle. In such a case, more emphasis should be on the rectangle's edges, and their parts if some of the edges are also partially occluded. The latter is a direct application of Gestalt Theory's rules of proximity and good continuation. The rule concerning convexity has already been implemented, though implicitly, because all shapes in the repertoire are convex, whereas the symmetry rule can be applied to parallel edges in case of squares and rectangles.

The most interesting question concerns the way in which local object gist and the non-standard retinal cells can be integrated in the normal pathway for invariant object recognition. The trivial part of the answer is that the spatial layout map — the centres of shapes and their type — can be exploited in the prefrontal cortex for (a) biasing all objects in memory with the same shapes, and (b) updating the FoA map in order to prepare saccadic eye movements. Much less trivial are the non-standard cells and the new pathways as discussed by Masland and Martin (2007). It is possible that not only top-down attentional modulation from the prefrontal cortex influences processing down to the lower levels V4, V2, V1 and even the LGN, but that the same occurs bottom-up and at the same time. The difference may be that top-down modulation can be a serial process whereas bottom-up modulation can be a parallel one. Such questions are very speculative, and it may take some years before we know more about these issues, both the non-standard cells and good computational models of them, and their pathways to — and roles in — other visual areas.

3

SALIENCY AND FOCUS OF ATTENTION

FOCUS OF ATTENTION AND REGION SEGREGATION BY LOW-LEVEL GEOMETRY

ABSTRACT: Research has shown that regions with conspicuous colours are very effective in attracting attention, and that regions with different textures also play an important role. We present a biologically plausible model to obtain a saliency map for Focus-of-Attention (FoA), based on colour and texture boundaries. By applying grouping cells which are devoted to low-level geometry, boundary information can be completed such that segregated regions are obtained. Furthermore, we show that low-level geometry, in addition to rendering filled regions, provides important local cues like corners, bars and blobs for region categorisation. The integration of FoA, region segregation and categorisation is important for developing fast gist vision, i. e., which types of objects are about where in a scene.

KEYWORDS: Saliency, Focus-of-Attention (FoA), Region Segregation, Colour, Texture.

3.1 INTRODUCTION

Attention of animals, also primates and humans, is rapidly drawn towards conspicuous objects and regions in the visual environment. The ability to identify such objects and regions in complex and cluttered environments is key to survival, for locating possible prey, predators, mates or landmarks

for navigation (Elazary and Itti, 2008). But attention is only one aspect. We start to understand how our visual system works: (1) very fast extraction of global scene gist, (2) also fast local gist for important objects and a rough spatial layout map, (3) in parallel with (2) the construction of a saliency map for FoA, and only then (4) sequential screening of conspicuous regions for precise object recognition, using peaks and regions in the saliency map with inhibition-of-return in order not to fixate the same region twice, but with two strategies for FoA: first covert attention (automatic, data-driven) possibly followed by overt attention (consciously directed). In addition, our visual system is not analysing all information for constructing a complete and detailed map of our environment; it concentrates on essential information for the task at hand and it relies on the physical environment as external memory (Rensink, 2000).

In this chapter we concentrate on three aspects: (1) the construction of a saliency map for FoA on the basis of colour, which was shown to be very effective in attracting attention (van de Weijer et al., 2006a) and also texture (du Buf, 2007), (2) a first region segregation by employing low-level geometry in terms of blobs, bars and corners, and (3) using low-level geometry for allowing us to reduce significantly the dimensionality of texture features. We note that our approach is not based on the cortical multi-scale keypoint representation as recently proposed by Rodrigues and du Buf (2006), who built saliency maps which work very well for the detection of facial landmarks and for invariant object recognition on homogeneous backgrounds (Rodrigues and du Buf, 2008), but may lead to enormous amounts of local peaks in natural scenes.

3.2 COLOUR CONSPICUITY

Colour information in a saliency map was first used by Niebur and Koch (1996). Their model was later extended by Itti and Koch (2001), who integrated more features, for instance intensity, edge orientation and motion.

In our approach to create a saliency model which also contains cues for region and object segregation, we therefore start by using colour information, as this provides the most important input for attention (van de Weijer et al., 2006b), in order to build a colour conspicuity map which will later be combined with a texture map. But before using colour features the input images must be corrected because a same object will look different when illuminated by different light sources, i. e., the number, power and spectra of these.

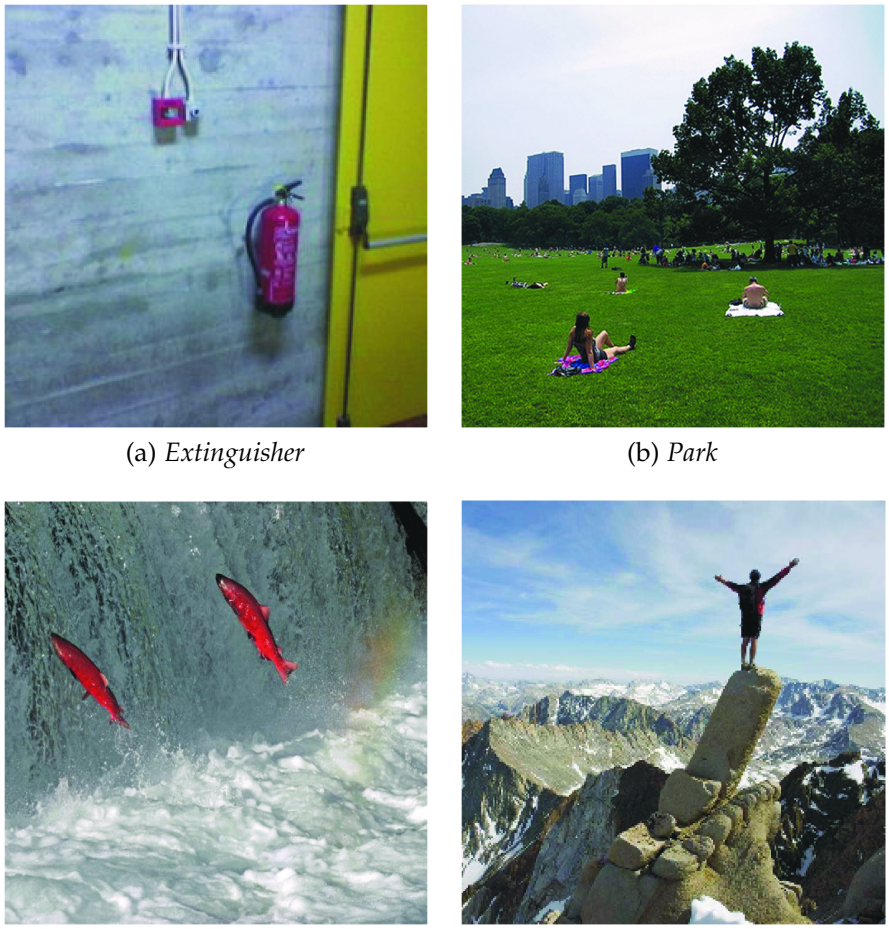
The processing consists of the following four steps: in the first, colour illuminant and geometry normalisation deals with correcting the image's colours. Let each pixel P_i of image $I(x, y)$ be defined as (R_i, G_i, B_i) and (L_i, a_i, b_i) in both RGB and Lab colour spaces, with $i = \{1...N\}$, N being the total number of pixels in the image. We first process the input image I_{in} using the two transformations described by Finlayson et al. (1998), as shown below, both in RGB colour space. Their method applies iteratively steps A and B, until colour convergence is achieved (4–5 iterations). Each individual pixel is first corrected in step A for illuminant geometry independency (i. e., *chromaticity*), by

$$P_i^A = \left(\frac{R_i}{R_i + G_i + B_i}, \frac{G_i}{R_i + G_i + B_i}, \frac{B_i}{R_i + G_i + B_i} \right), \tag{3.1}$$

followed in step B by global illuminant colour independency (i. e., *grey-world normalisation*),

$$P_i^B = \left(\frac{N \cdot R_i}{\sum_{j=1}^{N} R_j}, \frac{N \cdot G_i}{\sum_{j=1}^{N} G_j}, \frac{N \cdot B_i}{\sum_{j=1}^{N} B_j} \right). \tag{3.2}$$

After the process is completed, the resulting RGB image is converted to Lab colour space and the a_{cc} and b_{cc} components, where subscript cc stands for colour-corrected, are combined in I_{cc} together with the *unmodified* L_{in} channel from the input image I_{in}. The main idea for using the Lab space is that it is an almost linear colour space, i. e., it is more useful for



(a) *Extinguisher* (b) *Park*

(c) *Fish* (d) *Mountain*

Figure 3.1: Input images for saliency map processing.

determining the conspicuity of borders between regions. The reason for using the L_{in} component instead of the L_{cc} one is that, as observed by Finlayson et al. (1998), the simple and fast repetition of steps A and B does a remarkably good job. In fact, it does the job too well because all gray pixels (with values R=G=B from 0 to 255) end up having R=G=B=127. In other words, all information in grey image regions would be lost. Summarising, the initial I_{in} image in RGB is normalised to I_{cc} and then converted to the colour space $L_{in}a_{cc}b_{cc}$.

Figure 3.1 shows the four input images which will be used below, called *extinguisher*, *park*, *fish* and *mountain*, all of size 256×256 pixels with 8 bits for each colour component R, G and B. Figure 3.2 shows three results of colour correction applied to the *extinguisher* image, from *top–left to top–right:* the

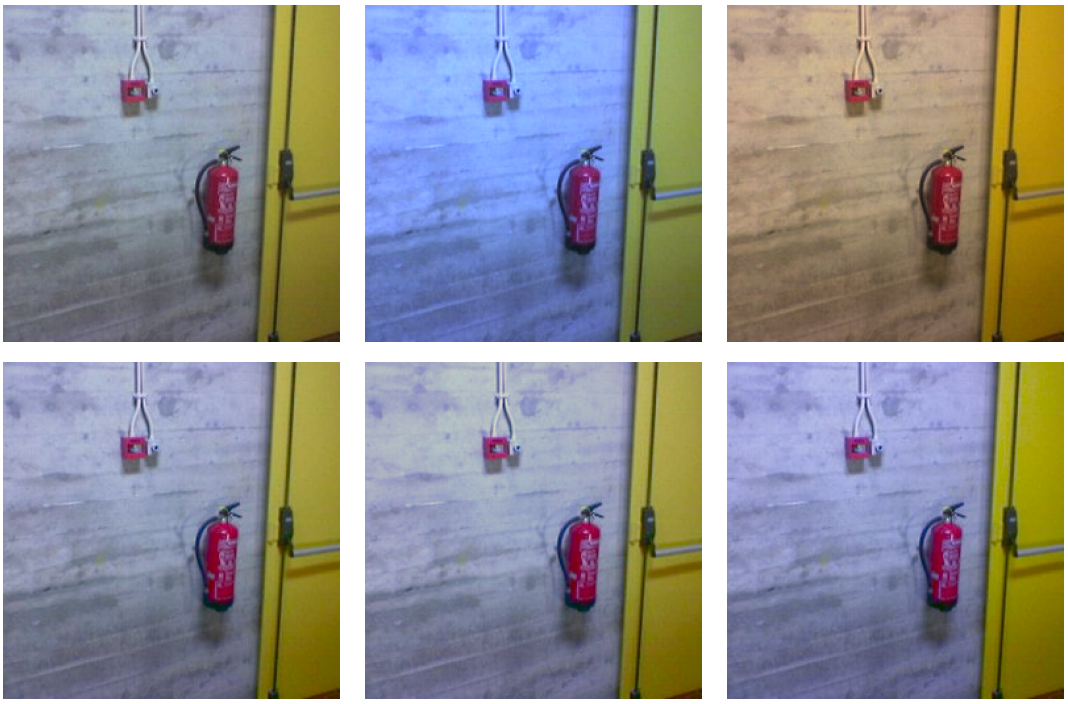


Figure 3.2: Colour illuminant and geometry normalisation. *Top row:* input *extinguisher* images under different illuminants. *Bottom row:* respective normalised results; see text.

original image, a modified image with a blue tint (R −12%, G +4% and B +50%), and a modified image with a warm white balance. The three results are shown below the input images. As can be seen, colour correction yields very similar images despite the rather large differences in the input images. Colour correction as explained above simulates colour constancy as employed in our visual system (Hubel, 1995).

The second step is to reduce colour inhomogeneities in the images by adaptive smoothing of the colour regions, while maintaining or even improving the boundaries between different regions. We propose a new, nonlinear, adaptive 1D filter, here explained in the horizontal direction but it can be rotated, which consists of a centred Difference-of-Gaussians (DoG)

$$F_{1,2}(x) = N_1 \left\{ \exp\left(\frac{-x^2}{2\sigma_1^2}\right) - \exp\left(\frac{-x^2}{2\sigma_2^2}\right) \right\}, \tag{3.3}$$

which is split into $F_1(x < 0)$ and $F_2(x > 0)$, and another centred Gaussian, which is *not* split,

$$F_3(x) = N_2 \cdot exp\left(\frac{-x^2}{2\sigma_2^2}\right), \tag{3.4}$$

taking $\sigma_1 \gg \sigma_2$. N_1 and N_2 are normalisation constants which make the integrals of all three functions equal to one. The three functions can be seen as a simulation of a group of three cells at the same position, but with different dendritic fields which are indirectly connected to cone receptors in the colour-opponent channels a and b of Lab, F_3 yielding the excitatory response of a receptive field of an *on-centre* cell, and $F_{1,2}$ yielding excitatory responses of two *off-centre* cells. From the three cell responses $R_{1,2,3}$ we first compute the contrast between the left (R_1) and right (R_2) responses; mathematically $C = |(R_1 - R_2)/(R_1 + R_2)|$. Then, based on the contrast C and the minimum difference between the centre response (R_3) and the left and right responses, the output R is determined by

$$R = \begin{cases} C\,R_1 + (1 - C)\,R_3 & \textit{if} \quad |R_1 - R_3| < |R_2 - R_3| \\ C\,R_2 + (1 - C)\,R_3 & \textit{otherwise}. \end{cases} \tag{3.5}$$

In words, if the contrast is low, as in an almost homogeneous region, the filter support is big, but if the contrast is high, at the boundary between two regions, the filter support is small. This adaptive filtering is applied to I_{cc} at each pixel position (x, y), first horizontally (R^H) and then vertically (R^V):

$$I_{ci}(x, y) = R^V\left[R^H\left[I_{cc}(x, y)\right]\right], \tag{3.6}$$

where subscript ci stands for colour-improved. In our experiments we obtained good results with $\sigma_1 = 7$ and $\sigma_2 = 3$, and adaptive filtering in horizontal and vertical directions was sufficient to sharpen blurred boundaries

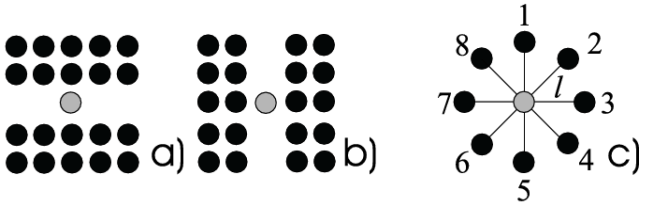


Figure 3.3: Gradient operators in vertical **(a)** and horizontal **(b)** orientations with summation areas of 2×5 pixels, and **(c)** the cluster of gating cells used for colour conspicuity.

even with oblique orientations. Furthermore, the processing is very fast because the three filter functions need only be computed once.

After colour correction and adaptive filtering, the third step serves to detect boundaries. In fact, for this purpose we could use the contrast function C described above, but in order to accelerate processing we apply a simple gradient operator, as shown in Figure 3.3(a–b), which requires only two convolutions, with mask sizes of 5×2 and 2×5, of the components of I_{ci}. These masks can be seen as dendritic fields of two cells, the two results being subtracted by a third cell which combines horizontal and vertical gradients:

$$\begin{aligned} \widehat{I}_{ed}(x,y) = &\sum_{left} I_{ci}(x,y) - \sum_{right} I_{ci}(x,y) \\ &+ \sum_{top} I_{ci}(x,y) - \sum_{bottom} I_{ci}(x,y). \end{aligned} \tag{3.7}$$

$\widehat{I}_{ed}(x,y)$ is then thresholded using

$$I_{ed}(x,y) = \begin{cases} 1 & \textit{if} \quad \widehat{I}_{ed}(x,y) > k \\ 0 & \textit{otherwise}, \end{cases} \tag{3.8}$$

where subscript ed stands for edge-detected. This yields a binary edge map by means of a cell layer in which cells are either active (response 1) or inactive (response 0). We apply a global threshold $k = \max(I_{ci}(x,y))$. Edge detection yields three distinct maps, one for each component of Lab colour space, I^L_{ed}, I^a_{ed} and I^b_{ed}, which can be combined.

In the last step, colour conspicuity at colour edges is calculated at each position in the edge map where there is an active cell $\left(I_{ed}^{L,a,b}(x,y) = 1\right)$. We define conspicuity Ψ at position (x,y) as the maximum difference between the colours in I_{ci} at four pairs of symmetric points at distance l from (x,y), i. e., on horizontal, vertical and two diagonal lines. Figure 3.3(c) shows a cluster of gating cells used for colour conspicuity. If the gating cells are called G_i, opposing pairs are (G_i, G_{i+4}), with $i = \{1,\dots,4\}$, for example (G_1, G_5). Partial conspicuity is then calculated independently for each of the colour components in Lab space, as defined by (3.9a), where $\vec{x}_i$ denotes the position of G_i relative to position (x,y). The final value is then calculated using the sum of all three colour components (3.9b).

$$\Psi_{L,a,b}(x,y) = \max_i\left(|I_{ci}^{L,a,b}(\vec{x}_i) - I_{ci}^{L,a,b}(\vec{x}_{i+4})|\right), \tag{3.9a}$$

$$\Psi_{Lab}(x,y) = \Psi_L(x,y) + \Psi_a(x,y) + \Psi_b(x,y). \tag{3.9b}$$

Results of colour conspicuity are shown in Figure 3.4 *(top)*, for the park and mountain images, using $l = 4$.

3.3 TEXTURE BOUNDARIES

Colour conspicuity Ψ_{Lab} includes the luminance component L and therefore luminance gradients, both in coloured image regions and in non-coloured or grey ones, but the processing as applied up to here is too local to capture texture as a region property. As different colours in surrounding or neighbouring regions attract attention, so do different textures because texture conveys complexity and therefore importance of regions to attend for screening.

Texture processing is in principle completely equal to colour processing, with adaptive filtering, gradient detection and the attribution of conspicuity to texture boundaries, but instead of using the three Lab components only the L one is used and texture features must be extracted from $L(x,y)$. Since

we are developing biologically plausible methods, it makes sense to apply Gabor wavelets as a model of cortical simple cells. Although very sophisticated texture models have been proposed on the basis of the Gabor model (du Buf, 2007), we will only use the spectral decomposition here because of speed. This frees CPU time for applying a reasonable number of frequency (scale) and orientation channels, which will be $8 \times 8 = 64$ in this chapter. Since Gabor filtering involves filter kernels which are relatively small (high-frequency textures because of viewing distance), all filtering can be done in the frequency domain (see e. g. Rodrigues and du Buf, 2004) and requires one forward Fast Fourier Transform (FFT) and 64 inverse FFTs, the latter parallelised on multi-core CPUs or even graphics boards (GPUs).

In the spatial domain, Gabor filters consist of a real cosine and an imaginary sine component, both with a Gaussian envelope, which resemble receptive fields of simple cells with even $R^E_{s,i}$ and odd symmetry $R^O_{s,i}$, with i the orientation and s the scale. Responses of complex cells are modelled by taking the modulus $C_{s,i}(x,y) = \left[\left\{ R^E_{s,i}(x,y) \right\}^2 + \left\{ R^O_{s,i}(x,y) \right\}^2 \right]^{\frac{1}{2}}$.

Texture boundaries are obtained by applying three processing steps to the responses of complex cells, at each individual scale and orientation, after which results are combined: (a) the responses $C_{s,i}(x,y)$ are smoothed using the adaptive filter defined by eqns (3.3) to (3.5), obtaining $\widehat{C}_{s,i}$. The next step (b) consists of horizontal and vertical gradient detection $\overline{C}_{s,i}$, applying cells with dendritic fields of size 2×5 as shown in Figure 3.3(a–b) to $C_{s,i}(x,y)$. The final step (c) consists of summing the results at all scales and orientations

$$R(x,y) = \sum_{s,i} \overline{C}_{s,i}(x,y), \tag{3.10}$$

together with an inhibition of all responses below a threshold (we apply $0.1 \max\{R(x,y)\}$). Figure 3.4 *(bottom)* shows the results in the case of *park* and *mountain*. As can be seen, the information is more diffuse and complements that of colour processing *(top)*.

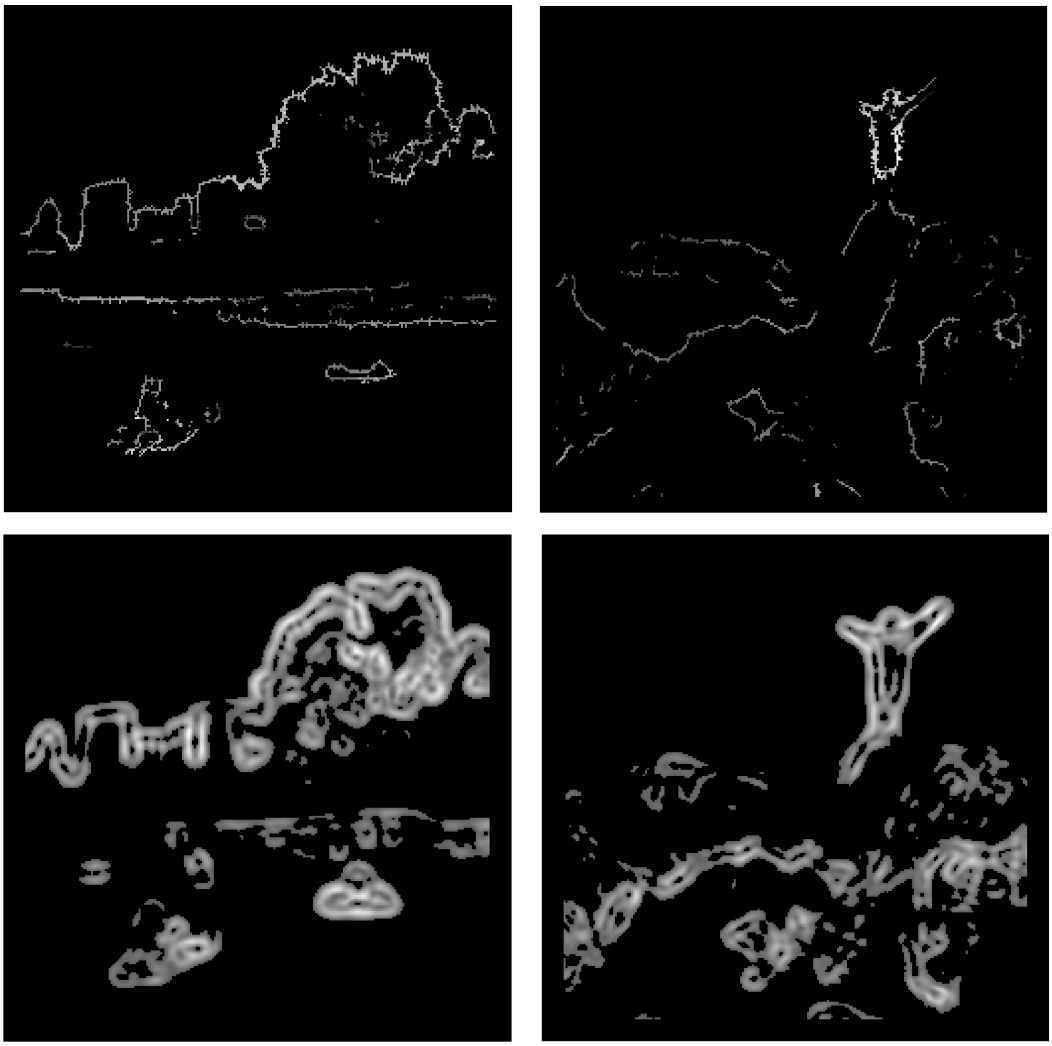


Figure 3.4: Colour boundary conspicuity *(top row)* for images *park (left)* and *mountain (right)*. *Bottom row:* texture boundaries.

3.4 SALIENCY MAP

A saliency map is built on top of colour conspicuity and texture boundary maps by using grouping cells which code local geometry. There are two levels of grouping cells. At the first, lower level, there are summation cells with a dendritic field size of $n \times m$, with the centres at a distance d; see Figure 3.5(a). In this chapter we use $m = n = 5$ and $d = 5$, such that the dendritic fields of the cells do not overlap. These cells sum activities in the colour or texture maps, hence boundary conspicuity at individual pixel

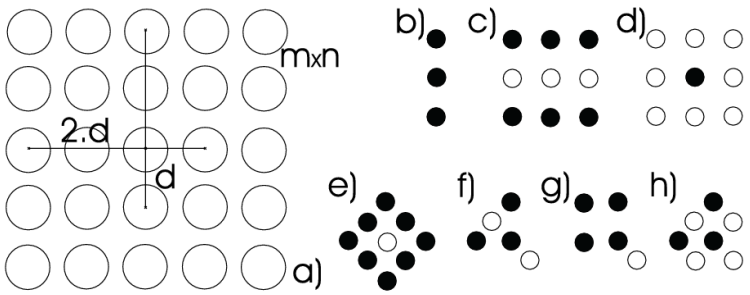


Figure 3.5: Grouping cells for low-level geometry: **(a)** cluster of cells with their dendritic fields (circles) on a 5×5 grid; **(b–h):** examples of spatial configurations.

positions is reinforced at this level, but also at the next level which deals with local geometry.

At the second level, there are many grouping cells, each one devoted to one geometric configuration on a 5×5 grid, but not all axons of the cells at the lower level are used. This allows a simple construction of spatial configurations, as shown in Figure 3.5(b–h), with up to four rotations, i. e., horizontal, vertical and two diagonal orientations. The *solid* and *open* circles refer to the use of the responses of the underlying summation cells: for a *solid* circle the sum S needs to be positive ($S > 0$), while for an *open* circle $S = 0$. The responses of all other unspecified summation cells in the grid are not used. Cells at this level take the maximum of the responses of the excited grouping cells, but only if the spatial configuration of the non-excited grouping cells is correct.

If the response R of configuration c is R^c, and if we call the configuration of cells which must be excited Ω_e^c and that of the cells which must not be excited Ω_{ne}^c, with $\Omega_e^c, \Omega_{ne}^c \in \Omega$, the 5×5 grid, and $\Omega_e^c \wedge \Omega_{ne}^c = 0$, then

$$R^c = \max_{i \in \Omega_e^c} S_i \Leftarrow \sum_{j \in \Omega_{ne}^c} S_j = 0. \tag{3.11}$$

The configurations shown in Figure 3.5(b–h) concern, respectively: a line (or an isolated contour), a bar (two parallel contours of a bar), two types of blobs and three types of corners. The (d) and (e) configurations are not rotated, but the other five are (horizontal, vertical and two diagonal orien-

tations), so in total there are 22 configurations in the total set C. Since there may be more than one configuration valid at the same position, the last cell layer determines the response of the maximum configuration, which yields the saliency map:

$$R(x, y) = \max_{c \in C} R^c(x, y). \tag{3.12}$$

The *top row* of Figure 3.6 shows results obtained when using only texture boundaries *(at left)*, and those when using only colour conspicuity *(at right)*. As can be seen, the maps are different but they complement each other, i. e., texture in general yields more diffuse areas (*park* and *mo*untain images) whereas colour conspicuity is more concentrated on contours. Combined results, using texture and colour, of all four images are shown in the *bigger images* of the figure. These final maps were created by taking the sum of the two values of the texture and colour saliency maps at each pixel position. It should be stressed that all images were normalised for visualisation purposes, with darker areas corresponding to less saliency and brighter ones to more saliency.

3.5 DISCUSSION

We introduced a simple, new, and biologically plausible model for obtaining saliency maps based on colour conspicuity and texture boundaries. The model yields very good results in the case of natural scenes. In contrast to the methods employed by Itti and Koch (2001), whose saliency maps are very diffuse versions of the entire input images, our method is able to highlight regions, a sort of pre-segregation of complex and conspicuous regions which is later required for precise object segregation in combination with object categorisation and recognition.

The saliency maps provide crucial information for sequential screening of image regions for object recognition and tracking: FoA by fixating con-

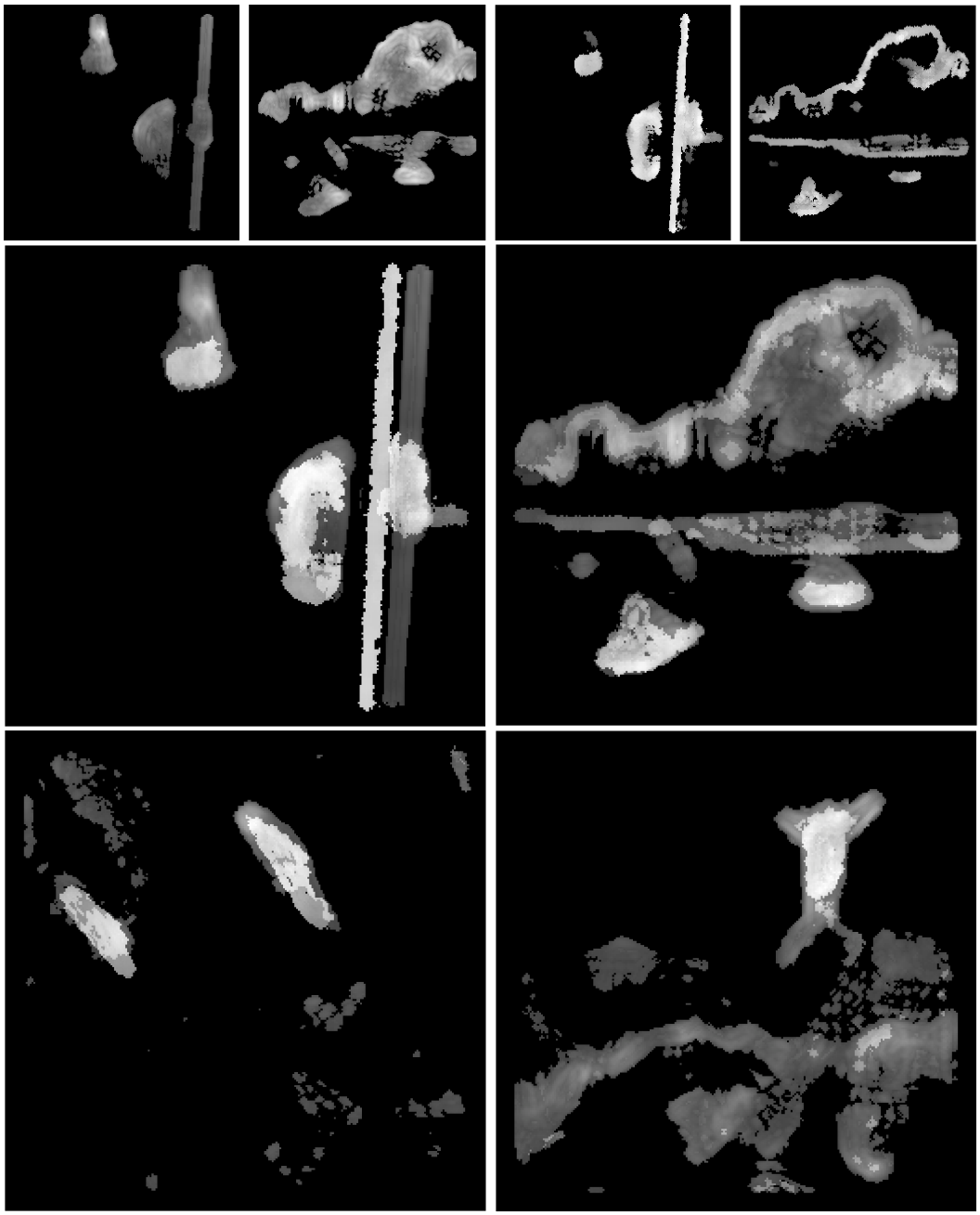


Figure 3.6: Saliency maps obtained by using only texture boundaries *(top row–left)*, only colour conspicuity *(top row–right)*, and by combining colour and texture *(bigger images)*.

spicuous regions, from the most important regions to the least important ones. Figure 3.7 shows an input image with toy cars *(left)*, the saliency map *(right)*, and the order, indicated by arrows, in which the regions will be processed by FoA. Fixation points were selected automatically by determining the highest response in the saliency map within each region, and regions

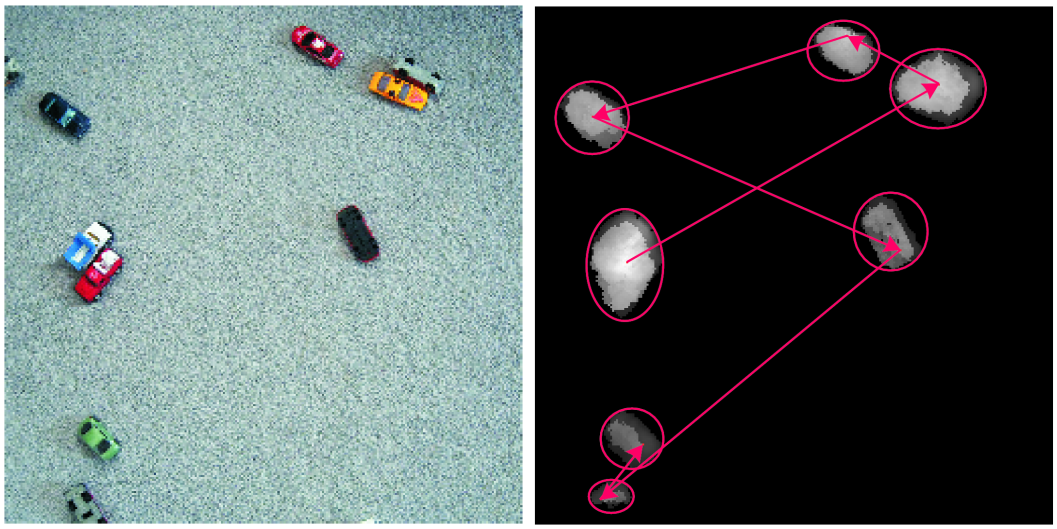


Figure 3.7: *Toy cars* image *(left)* and FoA-driven sequential screening of regions *(right)*.

are fixated using inhibition-of-return. Despite the fact that saliency based on texture boundaries is more diffuse than on colour conspicuity, car-region segregation is rather precise. The main reason for this precision is that low-level geometry processing mainly occurs at contours and inside objects, i. e., it does not lead to region growing.

The saliency model is now being extended by motion and disparity information, after which it can be integrated into a complete architecture for invariant object categorisation and recognition (Rodrigues and du Buf, 2006, 2008), which is based on multi-scale keypoints, lines and edges derived from responses of cortical simple, complex and end-stopped cells. This is beyond the scope of this chapter, but, as mentioned in the Introduction, very fast global and local gist vision are two basic building blocks of an integrated system. Until here, low-level geometry processing has only been used for producing saliency maps for FoA with segregated regions. But since low-level geometry information has already been extracted, it is therefore available for obtaining local object gist, for example providing cues which are used for a first and fast selection of possible object categories in memory (Bar et al., 2006). This is a purely bottom-up and data-parallel

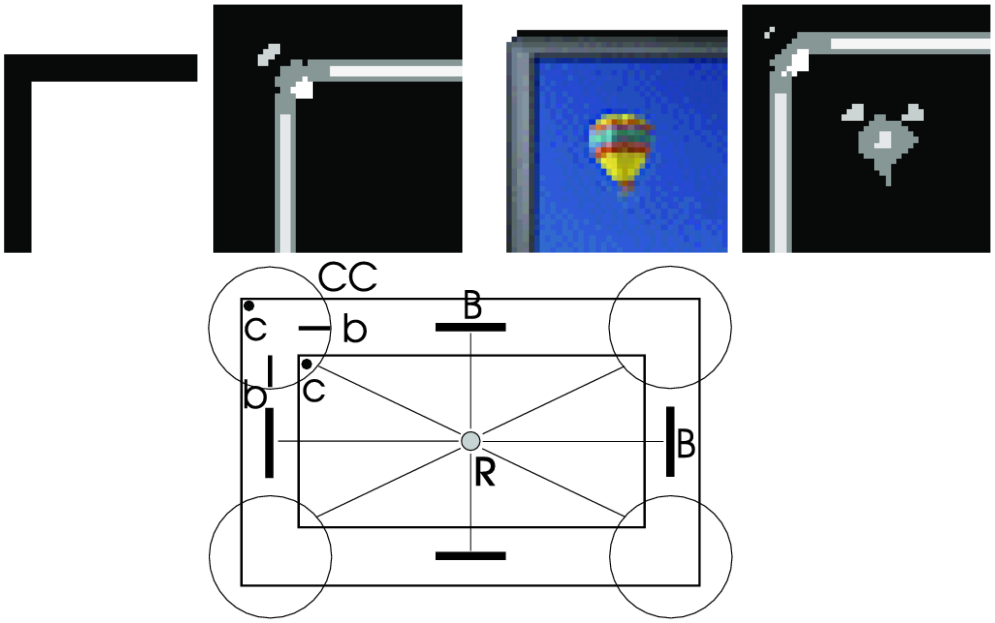


Figure 3.8: Low-level geometry *(top)* and example of mid- and high-level geometry groupings *(bottom)*; see text for details.

process for bootstrapping the serial object categorisation and recognition processes which are controlled by top-down attention.

Low-level geometry is difficult to visualise, because it consists of a large number of spatial maps, in this chapter limited to 22 but there could be more, one for each spatial configuration. Figure 3.8 *(top)*, shows detail images with a few configurations coded by different levels of grey, i. e., corners, bars and blobs. The input was an ideal rectangle with two sharp corners *(left)* and a computer monitor with a sharp inner corner but rounded outer one. Despite the outer corner being rounded, evidence for a corner has been detected at two pixel positions. These results were obtained on the basis of colour conspicuity, but later texture and other colour information can be combined, and low-level geometry should be used to construct mid- and high-level geometry. The latter idea is illustrated in Figure 3.8 *(bottom):* at low-level, corners (c) and bars (b) are detected. At mid-level, these can be grouped into a complex corner (CC), and at high-level the CCs, together with linking bars B, into a rectangle R. Such an R structure is typical for man-made objects, for example a computer monitor or a photo frame.

This example of high-level geometry is perhaps the last level below semantic processing: a computer monitor in combination with one or more photo frames is an indication for global scene gist: our office. In any case, the large number of features at the lowest level (64 Gabor channels) is reduced to the number of spatial configurations at low-level geometry, here 22. Groupings at mid-level (e. g., complex corner CC) may lead to less configurations, but at high-level (e. g., rectangle R) the number of configurations will increase again, because many elementary shapes must be represented. On the other hand, the precise localisation of configurations which is required at low-level is not necessary at higher levels; for example, grouping cells for complex corners CC may be located somewhere near the centres of the circles in Figure 3.8 *(bottom)*, as long as their dendritic fields are big enough to receive input from two corner and two bar cells. These aspects are subject to further research.

4

DISPARITY ENERGY MODEL

LUMINANCE, COLOR, VIEWPOINT AND BORDER ENHANCEMENT DISPARITY ENERGY MODEL

ABSTRACT: The visual cortex is able to extract disparity information through the use of special cells. This process is reflected by the Disparity Energy Model, that describes the role and functioning of simple and complex binocular neuron populations, and how they are able to extract disparity. This model explicitly involves cell parameters like their spatial frequency, orientation, binocular phase and position difference — however, it is a mathematical model. Our brain does not have access to such parameters, it can only exploit cell responses. Therefore, we introduce a new model for encoding disparity information implicitly by employing a trained binocular neuronal population. This model allows decoding of disparity information in a way similar to how our visual system could have developed this ability, during evolution, in order to accurately estimate disparity of entire scenes. At the same time, the monocular simple and complex cells can encode line and edge information useful to estimate disparity at borders. The brain should then be able, starting from a disparity draft to integrate all information including colour and perspective correction in the low-level disparity pathway to deliver better estimates to higher cortical areas.

KEYWORDS: disparity, visual cortex, population coding, colour, viewpoint, multiscale, lines, edges, conspicuity, gist.

4.1 INTRODUCTION

One of the intriguing functions of our visual cortex is to seamlessly extract disparity information from the surrounding environment. This is done after the Lateral Geniculate Nucleus (LGN), where information of the left and right retinae is relayed to the primary area V1, in the cortical hypercolumns. This is the first cortical processing stage, and disparity extracted there plays an important role in many other areas devoted to motor control, from walking about to precise eye-hand coordination, focus-of-attention and object segregation, even object recognition with partial occlusions. The development of better models is important to deepen our insights, but also for many practical applications, like in robotics where similar issues arise. In computer vision there are numerous approaches for stereo vision (Szeliski, 2011), but only few are biologically motivated. Of these, most have one common aspect: they are based on the widely accepted Disparity-Energy Model (DEM) (Ohzawa et al., 1997; Haefner and Cumming, 2008; Read, 2010; Martins et al., 2011b), which was first introduced from research into the cat's visual cortex and pathways.

The composition of disparity energy neurons (Haefner and Cumming, 2008) has led to different combinations of DEM subunits (with different weights and signs) into an energy complex cell. Other work has also tried to better explain the disparity-tuning curves of neurons in the rhesus monkey (Tanabe and Cumming, 2008). The use of windowed cross-correlation between the left and right eye's images to measure disparity could also explain the biological limits of stereopsis (Filippini and Banks, 2009).

In the case of uniform-disparity random-dot stereograms, the DEM model was able to explain that neurons tuned to horizontal disparities can also discriminate vertical disparities (Read, 2010). This ability comes as an emerging property from a neuronal system tuned to horizontal disparities, but with the ability of decoding vertical ones as a deviation from the expected neuronal response. It shows that the neuronal system is able to encode

much richer information than would be expected and, at the same time, concentrate neuronal resources on the most common cases while having the possibility of decoding rare ones.

Three exceptions of biological models that didn't use DEM information are the models by Pugeault et al. (2010), which combines geometric information with multi-modal constraints of local edge features, by Rodrigues et al. (2012) that combines multiscale lines and edges to retrieve a disparity wireframe model of the scene — the Line and Edge Disparity Model (LEDM) — and by du Buf et al. (2013), which employs the phase difference of the responses of complex Gabor filters to the left and right views. The latter model is often applied to real-world problems, although it has been shown to be very imprecise in terms of localisation of depth transitions.

Biological models applied to real-world scenes appeared only recently (Mutti and Gini, 2010; Pugeault et al., 2010; Martins et al., 2011b; Rodrigues et al., 2012; du Buf et al., 2013). The main reason for this lag is that models are usually only tested with very specific stimuli, such as random-dot stereograms or bar and grating patterns, in order to evaluate the model's theoretical performance (Read, 2010), or to prepare psychophysical experiments with minimal random noise (Tanabe and Cumming, 2008).

In this chapter, we propose that the disparity map is composed from different disparity cell maps built on top of each other, each refining the previously extracted disparity. We also propose that the first, rough disparity (disparity gist) is given by the DEM model, after which refinements based on colour, perspective correction (viewpoint) and border information are integrated to achieve the final disparity map.

In our improved DEM implementation we use two neuronal populations for obtaining disparities:

(1) An *encoding population* that consists of a set of neurons tuned to a wide range of parameters such as horizontal disparities, spatial frequencies and orientations. This population is trained on random-dot stereograms in order to learn the population codes for many different

disparities. It is also applied to real stereograms in order to obtain the local population codes. We use an encoding method similar to that of Read (2010), which is based on the DEM model of (Ohzawa et al., 1997), with proper normalisation to yield local correlations with neighbourhood weighting (Read and Cumming, 2006; Banks et al., 2004; Filippini and Banks, 2009). This is further explained in Section 4.3.1.

(2) A higher-level *decoding population* compares the local population code, at each image position, with all learned (trained) population codes, for estimating local disparity. This is further explained in Section 4.3.2. Basically, this second population implements a template-matching process similar to those of Tsai and Victor (2003) and Read (2010). This initial DEM model (disparity gist) is then integrated with colour, different viewpoints and border information, retrieved from the multiscale line and edge disparity model (LEDM) (Rodrigues et al., 2012) and low-level processes from object salience research (Martins et al., 2012) to achieve the final disparity map.

Our main contributions in this chapter are: (a) The adaptation of the biologically plausible DEM model to separate encoding and decoding populations, the extraction of disparity values in entire scenes, and the application to real-world images. (b) The integration of the DEM model with luminance, colour information and perspective correction (viewpoint). (c) The integration of two disparity models DEM and LEDM, to improve object boundary precision of the DEM. (d) The integration of different layers of disparity cell maps, with each layer improving the results from layer to layer. (e) The bio-inspired model has been tested with real-world scenes and can compete with state-of-the-art computer vision algorithms.

4.2 MONOCULAR AND BINOCULAR CELLS

In cortical area V1 there are simple, complex and end-stopped cells. Monocular receptive fields (RFs) of simple cells can be modelled by Gabor wavelets (Chen and Qian, 2004; Rodrigues and du Buf, 2009a; Martins et al., 2011b). Their parameters specify the orientation θ, spatial frequency f (or by $\lambda = 1/f$ the wavelength of the Gabor filters), receptive field size σ and spatial phase ϕ.

The responses of even and odd monocular simple cells, corresponding to the real and imaginary parts of a Gabor filter (Rodrigues and du Buf, 2009a), are denoted by $R^E_{s,i}(x,y)$ and $R^O_{s,i}(x,y)$, i being the orientation index according to θ. The scale s is given by λ. Responses of complex cells are modelled by the modulus $C_{s,i}(x,y) = \left[\{R^E_{s,i}(x,y)\}^2 + \{R^O_{s,i}(x,y)\}^2\right]^{\frac{1}{2}}$. There are two types of end-stopped cells, single and double, which are the first and second derivatives of $C_{s,i}$. These monocular cells will be used for the the LEDM model (see Section 4.5).

Cortical binocular cells can be based on pairs of monocular simple cells with different RFs, such that they can signal disparity if a same (but shifted) pattern is present in their RFs. However, binocular simple cells do not reliably signal disparity because they are also sensitive to the contrast and position of the pattern within their fields: disparity-tuning curves of simple cells measured with bright and dark bars, which have different phases, are very different (Ohzawa et al., 1997). The problem is that such tuning curves strongly depend on the phase of the pattern (Chen and Qian, 2004): any change to a pattern other than an amplitude scaling (average brightness and contrast) alters the cell phase response ϕ, which in turn affects disparity responses.

By contrast, binocular complex cells do not have separate excitatory and inhibitory subregions within their receptive fields, so they are not sensitive to local phase, but still to position, orientation and size of a pattern (Skottun and Freeman, 1984). They have also been found to be sensitive to fine

binocular disparity, and only complex cells respond to *dynamic* random-dot stereograms (Poggio et al., 1985). Complex cells also have a much finer disparity selectivity than what would be predicted by the size of their RFs (Ohzawa et al., 1997). An important advantage of binocular complex cells is that they ignore inverted local pattern polarities at their preferred disparity, in contrast to monocular complex cells (Ohzawa et al., 1997).

A phase-independent binocular complex cell can be made from two binocular simple cells S_1 and S_2 provided that their phase difference $\left|\phi_{S_1} - \phi_{S_2}\right| = \pi/2$, i. e., they are in quadrature. The response of a binocular complex cell is then obtained by summing the responses of these two binocular simple cells. Binocular simple cells are modelled by using monocular RFs with the same size, orientation and spatial frequency, but with different phases ϕ and positions on the retina $(\Delta x, \Delta y)$ (Read et al., 2009). The left $\left(\rho^L\right)$ and right $\left(\rho^R\right)$ RFs of monocular simple cells are defined by

$$\rho^{L,R}(x, y) = \exp\left(-\frac{x'^2_{L,R} + y'^2_{L,R}}{2\sigma^2}\right) \cos\left(2\pi f x'_{L,R} + \phi\right). \tag{4.1}$$

Since we will use phases in quadrature, i. e., $\phi = \{0, -\pi/2\}$, both ρ^L and ρ^R actually consist of two RFs: the sine and cosine components. In (4.1), x' and y' are the coordinates relative to the centre $(0, 0)$ and rotated according the cell's preferred orientation θ:

$$x'_{L,R} = x_{L,R} \cos\theta + y_{L,R} \sin\theta \tag{4.2a}$$

$$y'_{L,R} = -x_{L,R} \sin\theta + y_{L,R} \cos\theta. \tag{4.2b}$$

In case of binocular cells with left predominance (disparity of the left viewpoint), the offset coordinates Δx and Δy, which correspond to the cell's preferred horizontal and vertical disparities, are defined as follows: When the population code is trained (learned) with random-dot stereograms, the left RF is centred at $(0, 0)$ and the right one at $(-\Delta x, 0)$. When the cells are applied at all input stereogram positions, then $(x_L, y_L) = (x, y)$ and

$(x_R, y_R) = (x - \Delta x, y)$. We note that $\Delta y = 0$ is taken for all cells, as vertical disparity in the fovea is zero. Although it can be non-zero at other retinal positions Read et al. (2009), this effect is not applied here. For the detailed mathematical transformation from monocular simple to binocular cells see (Chen and Qian, 2004).

4.3 LUMINANCE DISPARITY-ENERGY MODEL

In this section we introduce the Luminance Disparity-Energy Model (L-DEM), and show how disparity maps can be extracted by exploiting binocular cell responses and comparing them with previously learned stimuli, by modelling cells sensitive only to luminance variations (Martins et al., 2011b). For this implementation we use two neuronal populations: (1) An encoding population and (2) a higher-level decoding population. For presenting our stereo results we use by default the reference viewpoint image of the left eye, unless stated otherwise.

4.3.1 *Disparity encoding population*

During the training phase, all binocular simple cells [see (4.1)] are located at the centre of the fovea, i. e., the centre of the RFs is at position $(0, 0)$. For the encoding population we selected a set of binocular simple cells, with parameters based on Read (2010):

(a) Orientation $\theta_i = (i \times \pi)/N_\theta$, with N_θ the number of orientations, here 8 (empirical tests showed that using more orientations yielded slightly better disparity estimates, but increases the total cell population; this value is a good compromise).

(b) Receptive field sizes $\sigma = \left\{2\sqrt{2}, 2, \sqrt{2}\right\}$. These are scaled by a factor of $\sqrt{2}$, as is the frequency (empirical results showed that bigger sizes

increase blur at objects' border regions and smaller sizes display too much error in disparity responses).

(c) RF spatial frequencies $f = \left\{\sqrt{2}/8,\ 1/4,\ 2\sqrt{2}/4\right\}$ cycles per pixel. These values are related to RF size by $\omega\sigma = \pi$ or $f = 1/2\sigma$ (the frequency bandwidth at all scales was 1.14 octaves).

(d) RF phases $\phi = \{0, \pi/2\}$, as only two values are needed to build a phase-invariant complex cell from two simple cells (Ohzawa et al., 1997).

(e) RF horizontal position disparity $\Delta x = \{0, ..., 59\}$ in steps of 1 pixel.

(f) RF phase disparity $\Delta\phi = 0$, implying no additional phase difference between the left and right RFs of each cell (equal ϕ phases for both). It is to be expected that in natural occurring images, the maximum response of a phase-shift disparity neuron is elicited when there is a different pattern of the same stimulus in the left and right RFs, something that never occurs in the real world (Haefner and Cumming, 2008; Tanabe and Cumming, 2008). Our empirical tests also showed that the use of phase differences — odd-symmetric disparity tuning curves — did not add significant information and sometimes even degraded the quality of disparity estimates. Possible roles for neurons tuned to phase disparities are explained further in Read and Cumming (2007).

In total, the above selection yields a population of $8 \times 3 \times 2 \times 60 \times 1 = 2880$ binocular simple cells (1440 complex cells); see below. The values were chosen to replicate physiological parameters of real cells, yielding precise disparity estimates in practice. For this, two steps are necessary: (1) *stereo energy coding*, used for the real images and in part for the training and (2) *model training*, that is only used in an initial step to train the model to discriminate disparities. Both are explained below.

4.3.1.1 *Stereo energy coding*

Responses of the left and right RFs of monocular simple cells (v^L and v^R) are obtained by convolving ($*$) the RFs with the corresponding left and right images $I^{L,R}(x, y)$:

$$v^{L,R}(x, y) = I^{L,R}(x, y) * \rho^{L,R}(x, y). \tag{4.3}$$

To simplify notation, below we skip (x, y). At each position, the response S of a binocular simple cell combines the responses of the left and right RF components (Ohzawa et al., 1997; Chen and Qian, 2004):

$$S = \left(v^L\right)^2 + \left(v^R\right)^2 + 2\,v^L v^R. \tag{4.4}$$

S can be split into the monocular term $M = \left(v^L\right)^2 + \left(v^R\right)^2$ and the binocular term $B = 2\,v^L v^R$. Biologically, this can be realised by combining the outputs of two energy neurons with phase disparities π apart. If such neurons are identical except for their phase disparities, then the first one computes $(M + B)$ and the second $(M - B)$. Both M and B are then available from the sum and difference of the two responses, i. e., $2M$ and $2B$ (Read, 2010).

For obtaining the local stereo energy E of a binocular complex cell which is invariant to the phases of local patterns in the input, one can either sum the responses of (a) many binocular simple cells with scattered phases ϕ in $[0, 2\pi]$, or (b) with just two phases $\pi/2$ apart. We could therefore apply the second case with $\phi = \{0, -\pi/2\}$: $E = \Sigma_{\phi=\{0,-\pi/2\}}\left(S_\phi\right)$. This stereo energy E, for each frequency, orientation and disparity, is similar to the cross-correlation between the filtered and windowed images (Filippini and Banks, 2009). However, E cannot be used as disparity estimate, since it not only reflects binocular energy (stimulus disparity between the left and right RFs), but also monocular energy (stimulus contrast inside each RF). This problem is solved by using M and B together with spatial pooling and effective binocular correlation.

4.3.1.2 *Spatial pooling*

Complex cells are normally modelled by taking the square root of the sum of the squared responses of the sine and cosine components of simple cells. This implies that the RF size of such complex cells is equal to that of the simple cells (the same Gaussian). However, RFs of real binocular complex cells are larger than those of simple cells (Chen and Qian, 2004). Therefore we apply this property by averaging M and B, using grouping cells with a Gaussian RF: $G^{sp}(x,y) = k\exp\left(-(x^2+y^2)/2\sigma^2\right)$. The normalisation factor $k = 1/2\pi\sigma^2$ and σ is the RF size of the corresponding simple cells: $\sigma \in \left\{2\sqrt{2},\, 2,\, \sqrt{2}\right\}$. This yields, for the two phases, $M^{sp}_{\phi} = G^{sp} * M_{\phi}$ and $B^{sp}_{\phi} = G^{sp} * B_{\phi}$. This pooling operation involves simple grouping cells with a dendritic field size defined by σ and is crucial to stabilise results in case of real-world images with noise and non-uniform disparity ranges.

4.3.1.3 *Effective binocular correlation*

In order to differentiate monocular energy from binocular energy one must use normalised correlation detectors (Read, 2010; Banks et al., 2004; Read and Cumming, 2006; Filippini and Banks, 2009). Such detectors have responses which range between $+1$, when the left and right images are identical, and -1, when the left image is an inverted-contrast version of the right one. This is achieved by dividing the pooled binocular term by the pooled monocular term, after which the result is pooled once more for increasing robustness:

$$\psi^{sp} = G^{sp} * \left(\frac{\sum_{\phi=\{0,-\pi/2\}} B^{sp}_{\phi}}{\sum_{\phi=\{0,-\pi/2\}} M^{sp}_{\phi}} \right). \tag{4.5}$$

The value ψ^{sp} relates to the correlation between local, filtered regions of the left and right views (Read and Cumming, 2007). The population of binocular correlation detectors ψ^{sp} is used for encoding disparity in the model. Normalising the stereo energy E to obtain the effective binocular

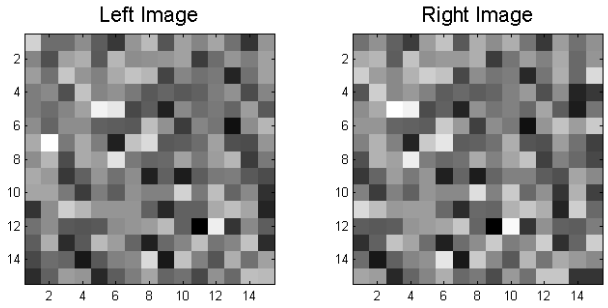


Figure 4.1: Example of a 15×15 random-dot stereogram used in the training phase, with a uniform 2-pixel shift and thus horizontal disparity (Δx_{stim}) of 2.

correlation removes the confounding effect of monocular contrast, and it allows us to extract the disparity from peaks in the population's activity code. ψ^{sp} has also the useful property that it exactly equals 1 when the actual disparity matches a cell's preferred disparity (Read, 2010). Please recall that ψ^{sp} is the short notation for $\psi^{sp}_{f,\theta,\Delta x}(x,y)$, i. e., there are three scales, eight orientations and 60 horizontal position disparities, hence 1440 binocular correlation cells which are applied at all image positions.

4.3.1.4 *Learning the population code*

We trained the energy model to discriminate horizontal stimulus disparities (Δx_{stim}) ranging from 0 to 59 pixels with a stepsize of 1 pixel. To this purpose we generated the population codes to stimuli with known disparities. We used random-dot stereograms with uniform disparity, generated by random values with a Gaussian distribution with zero mean and unity standard deviation, with a Δx_{stim} horizontal offset between the left and right images. The gaps were filled by using randomly drawn pixels; see Figure 4.1. For each Δx_{stim} step we generated 1000 random-dot pairs. Hence, training involved 60,000 stereograms (for details see Martins et al., 2011b). For each stereogram, with $I^{L,R}$ the left and right views, we applied (4.3) and

(4.4), but only at the centre of the left and right images of each stereogram. The values of ψ were computed without spatial pooling, i. e.,

$$\psi = \frac{\sum_{\phi=\{0,-\pi/2\}} B_\phi}{\sum_{\phi=\{0,-\pi/2\}} M_\phi}, \tag{4.6}$$

because results are pooled over the 1000 random-dot stereograms for each disparity.

The effective binocular correlations ψ and ψ^{sp} are encoded by the neuronal population as a *mean spike count*,

$$\Psi = (1 + \psi)\, u, \tag{4.7}$$

where $u = 8$ is the average number of spikes elicited by a binocularly uncorrelated stimulus within the temporal discrimination window. Typical values of u are around 8 spikes, assuming a firing rate for the optimal disparity of 100 Hz and a temporal window of 160 ms (Read, 2010). Therefore, Ψ is in the range $[0, 2u]$, where $2u$ is the mean number of spikes a perfectly correlated binocular stimulus elicits from neurons tuned to its disparity.

Finally, Ψ was averaged ($A(\cdot)$) over the 1000 different stereograms for each Δx_{stim}. Averaging over random images serves to eliminate stimulus-dependent noise. This yields a population code for each Δx_{stim}:

$$W^{\Delta x_{stim}}_{f,\theta,\Delta x} = A\left(\Psi^{\Delta x_{stim}}_{f,\theta,\Delta x}\right). \tag{4.8}$$

In summary, W contains the number of spikes produced by neurons tuned to frequencies f, orientations θ and horizontal disparities Δx, averaged over all 1000 stimuli with the same disparity Δx_{stim}. The population code consists of 1440 binocular correlation cell responses (3 RF scales, 8 RF orientations and 60 RF horizontal position disparities) for *each* of the 60 different horizontal stimulus disparities (Δx_{stim}) of the random-dot stereograms — it is expected that the highest spike count will be when a cell's horizontal position disparity matches the stimulus disparity, i. e.,

$\Delta x = \Delta x_{stim}$. This learning process, which is the core of the method, can be seen as a replication of visual learning in early childhood, assuming that basic neural circuitry is the result of evolution, or, at least, needs adequate training to reach its full potential.

4.3.2 *Disparity decoding population*

As mentioned before, learning is done only once and in the centre of the random-dot stereograms. After training, the encoding population can then be applied at all positions (neighbourhoods) of real world input stereograms (excluding the border region). The disparity at each position is estimated by comparing the population code response with the learned codes. This is done by a second, higher-level *decoding* population. The disparity assigned to the position is the disparity of the best-matching code. Local disparity estimation is a simple matching process (Tsai and Victor, 2003): the input code of 1440 responses is matched or correlated with the 60 sets of 1440 trained codes. The final output is selected by the decoding population in a winner-takes-all strategy. Biologically, this probably involves associative memory which can also be based on a training process (Yang and Yao, 2008).

The matching process uses 60 correlation cells ("Corr") which compare $\Psi^{sp}_{f,\theta,\Delta x}$ and $W^{\Delta x_{stim}}_{f,\theta,\Delta x}$, i. e., the 1440 spike counts at the image position and the previous, learned 60 sets of 1440 spike counts:

$$r_{\Delta x_{stim}}(x,y) = \left[\mathrm{Corr}\left(\Psi^{sp}_{f,\theta,\Delta x}(x,y), W^{\Delta x_{stim}}_{f,\theta,\Delta x}\right)\right]^{+}, \tag{4.9}$$

where $[\cdot]^{+}$ is half-wave rectification. This avoids the problem of disparity in anti-correlated stereograms by setting any negative correlations to zero (Read and Eagle, 2000). Note that $r_{\Delta x_{stim}}$ is a vector of 60 correlation values, each related to a specific Δx_{stim} disparity that the population was trained to recognise, from 0 to 59. Disparity $D^{L}(x,y) = d$ is the one with maximum correlation: $r_{\Delta x_{stim}=d} = \max(r_{\Delta x_{stim}})$. Biologically, this corresponds to the

activation of a single disparity cell at each position, from the 60 possible. Mathematically, the implemented matching process (Corr) is the Pearson product-moment correlation coefficient with $A(\cdot)$ the average, σ_Ψ and σ_W the standard deviations of all 1440 responses:

$$r_{\Delta x_{stim}}(x,y) = \left[\frac{A\left(\Psi^{sp}(x,y)W\right) - A\left(\Psi^{sp}(x,y)\right) A\left(W\right)}{\sigma_{\Psi^{sp}(x,y)}\sigma_W} \right]^+ . \tag{4.10}$$

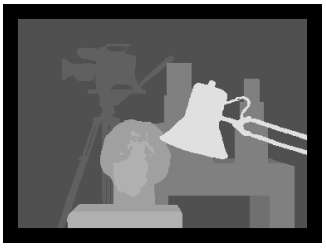
We emphasise that no further processing was applied until here, i. e., the pixels' disparity values were not corrected using any continuity constraints for homogeneous regions in combination with the detection of region boundaries, etc.

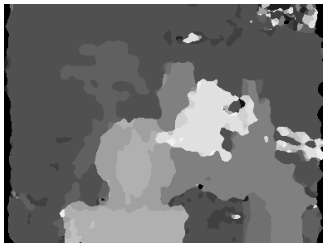
4.3.3 *Experimental results*

We tested L-DEM on various datasets, including the widely used stereograms *tsukuba*, *venus*, *teddy* and *cones* of the Middlebury stereo evaluation set (Scharstein and Szeliski, 2002; Scharstein, 2003), *aloe* and *cloth3* of the 2006 dataset, and *dolls*, *moebius* and *reindeer* of the 2005 dataset (Scharstein and Pal, 2007). The same datasets were used for all experimental results in this chapter.

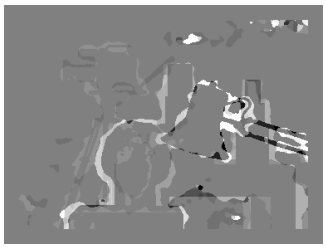
Figure 4.2 illustrates the different disparity cell layers for the *tsukuba* (Scharstein, 2003) stereo pair. Shown are: (a) *tsukuba*'s left image from the stereo pair, (b) ground-truth, (c) D_L layer (L-DEM), (d) bad pixels (in black) with an absolute disparity error $\geqslant 0.5$ and (e) signed disparity error retrieved from the Middlebury stereo site (Scharstein and Szeliski., 2012). The algorithm obtained good results in the Middlebury evaluation test (ranked there as "BioDEM [117]") (Martins et al., 2011b), also shown in Figure 4.7 (page 104). We will be comparing our further disparity algorithm improvements with these previous results as baseline.

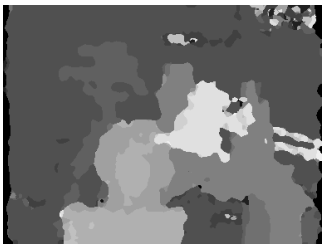

(a) *Tsukuba* left image


(b) Ground-truth


(c) D^{L} (L-DEM)

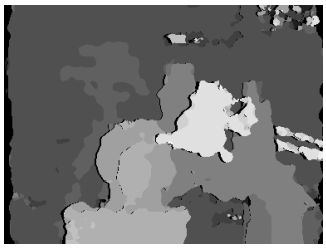

(d) $|\text{Error}| \geqslant 0.5$


(e) Signed D^{L} error

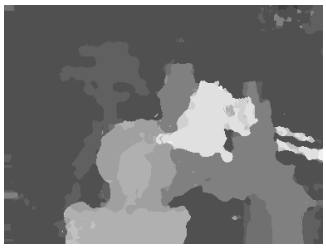

(f) D^{LC}_{Left}


(g) D^{LC}_{Centre}


(h) D^{LC}_{Right}

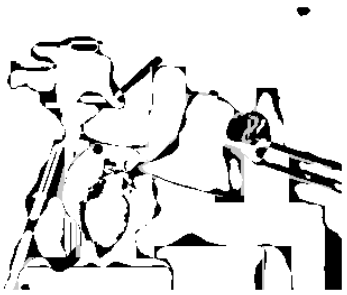

(i) $\widehat{D}^{LCV}$


(j) D^{LCV}


(k) $\widehat{D}^{LCVE}$


(l) D^{LCVE}


(m) D^{LCVB}


(n) D^{LCVB} $|\text{Error}| \geqslant 0.5$

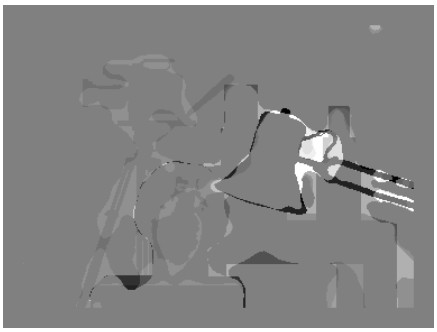

(o) Signed D^{LCVB} error

Figure 4.2: Disparity results for the cell map layers for the *tsukuba* (Scharstein, 2003) stereo pair. **(a):** *tsukuba* left image of the pair. **(b):** Ground-truth. **(c):** D^{L} (L-DEM) result. **(d):** Bad pixels (in black) with an absolute disparity error $\geqslant 0.5$. **(e):** Signed disparity error returned by (Scharstein and Szeliski., 2012). **(f):** Left–viewpoint D^{LC}_{Left}. **(g):** Centre–viewpoint D^{LC}_{Centre}. **(h):** Right–viewpoint D^{LC}_{Right}. **(i):** Viewpoint corrected $\widehat{D}^{LCV}$. **(j):** Background and occlusion corrected D^{LCV}. **(k):** Line and edge region enhanced $\widehat{D}^{LCVE}$. **(l):** Object border enhanced D^{LCVE}. **(m):** The final disparity map D^{LCVB}, after median smoothing. **(n):** Bad pixels with absolute disparity error $\geqslant 0.5$. **(o):** Signed disparity error of D^{LCVB}.

4.4 LUMINANCE, COLOUR AND VIEWPOINT DEM

This section addresses an improved disparity model, the Luminance, Colour and Viewpoint Disparity-Energy Model (LCV-DEM), which integrates colour and viewpoint (observer perspective) information to increase accuracy *vs.* the previous L-DEM implementation.

Studies involving chromatic representation in the primary visual cortex have evidenced that cone responses from the retina turn into three relatively independent spatio-chromatic colour channels after the LGN (Field and Chichilnisky, 2007), which are then transformed into several neural pathways, mixing colour responses with other cell parameters, like RF sizes and spatial frequencies (Wade et al., 2008). The majority of neurons in V1 seem to respond to pure isoluminant stimuli (i. e., colour sensitive), and around 50% of all neurons are sensitive to both luminance and isoluminant stimuli, being classified as either "colour-luminance" or "luminance-preferring" cells with some cone opponency (Johnson et al., 2004). Recent studies demonstrate that chromatic features help to solve the binocular matching problem in complex images and are consistent with the hypothesis of independent contributions of colour and luminance information (den Ouden et al., 2005; Krauskopf and Forte, 2002). It has also been reported that there exist neurons in V2 of macaques that are sensitive to both colour and disparity. This supports the notion that the primate visual system combines disparity and colour as early as in V2 (Ts'o et al., 2001).

For the LCV-DEM implementation we initially chose the LMS colour-space, which mimics the trichromatic neuronal encoding of cone responses after the LGN (Wade et al., 2008). However, the results with the LMS colour space were not significantly better than those with a simple variation of RGB (each channel is both luminance and colour coded), which is not surprising. Since the neuronal cells have so many different combinations of luminance or colour predominance, the system is able to be independent of the colour method used, as long as there is enough variety of weights

between different colour channels. We did, however, get better results when using retina cone weights (Stockman et al., 1993) for encoding luminance, suggesting that not only disparity is heavily luminance based, but also that it depends on luminance being highly discriminative of the scene being observed.

4.4.1 *Disparity encoding population*

This model uses similar cell population parameters as L-DEM, defined in Section 4.3.1, with additionally,

(g) RF shift μ: 3 values {Left, Centre, Right} of binocular RF composition, representing three possible combinations (Left, Centre and Right) for binocular receptive field dominance of horizontal position disparities, around a centre point, as further explained below.

We can improve disparity accuracy estimates by using two more RF dominances. As previously, the binocular simple cell RFs are defined by $\rho^{L,R}$ (4.1), where (x', y') are the offset coordinates relative to the centre $(0, 0)$ and rotated to the cell's preferred orientation, as shown in (4.2). For $\mu = \text{Left}$ we use $(x_{L,R}, y_{L,R})$ as shown in Section 4.2. For $\mu = \text{Centre}$ the RF coordinates are offset as $(x_L, y_L) = (x + \frac{\Delta x}{2}, y + \frac{\Delta y}{2})$ and $(x_R, y_R) = (x - \frac{\Delta x}{2}, y - \frac{\Delta y}{2})$. For $\mu = \text{Right}$ the RF coordinates are offset as $(x_L, y_L) = (x + \Delta x, y + \Delta y)$ and $(x_R, y_R) = (x, y)$.

4.4.1.1 *Stereo energy coding*

The LCV-DEM model also employs pairs of binocular simple cells in quadrature in order to construct phase-invariant complex cells. The responses of simple cells are obtained similarly to (4.3), but now with the previous DEM luminance-only channel (L) complemented by 3 new luminance/colour channels: $c = \{L, r, g, b\}$ with $r = R + \frac{G}{4} + \frac{B}{4}$, $g = \frac{R}{4} + G + \frac{B}{4}$, $b = \frac{R}{4} + \frac{G}{4} + B$ and L as in L-DEM (see Section 4.3). This represents the luminance-colour

sensitive cells to each of the RGB components, with an L channel representing luminance-predominant cells reflecting retina cone weights (Stockman et al., 1993). Responses of the left and right RFs of monocular simple cells ($v^{L}_{\mu,c}$ and $v^{R}_{\mu,c}$) are obtained by convolving ($*$) the RFs with the corresponding left and right images $I^{L,R}_{c}(x,y)$:

$$v^{L,R}_{\mu,c}(x,y) = I^{L,R}_{c}(x,y) * \rho^{L,R}_{\mu,c}(x,y). \tag{4.11}$$

In total this selection yields a population of $(8 \times 3 \times 2 \times 60 \times 1) \times (3 \times 4) = 34,560$ binocular simple cells (17,280 complex cells), twelve times larger than for L-DEM due to the 3 different viewpoints (μ) and 4 luminance and colour channels (c).

4.4.2 *Disparity decoding population*

The implementation uses the same decoding method as L-DEM, as specified in Section 4.3.2. However we are processing each of the four channels (c) as an "independent" image — this allows us to show the benefits of colour without having to train the population again.

For each (x,y), the correlation (Corr) coefficient is now calculated between $\Psi^{sp}_{\mu,c}$ and $W^{\Delta x_{stim}}_{f,\theta,\Delta x}$. The correlation vector $r_{\Delta x_{stim},\mu,c}$ now holds $60 \times 3 \times 4$ cell responses, one for each μ and c combination. At this step, 3 viewpoint–based D^{LC}_{μ} disparity maps are built independently. The disparity assigned to position $D^{LC}_{\mu}(x,y)$ will be the value d_{μ} of the maximum correlation, where $r_{\Delta x_{stim}=d_{\mu}} = \max\left(r_{\Delta x_{stim},c}\right)_{\mu}$ (winner-takes-all) between all Δx_{stim} and c values, keeping μ constant, yielding three different disparity maps $D^{LC}_{\mu}(x,y) = d_{\mu}$. Biologically this corresponds to an activation of a single disparity cell per pixel and per viewpoint (μ). Examples are shown in Figure 4.2, (f) to (h).

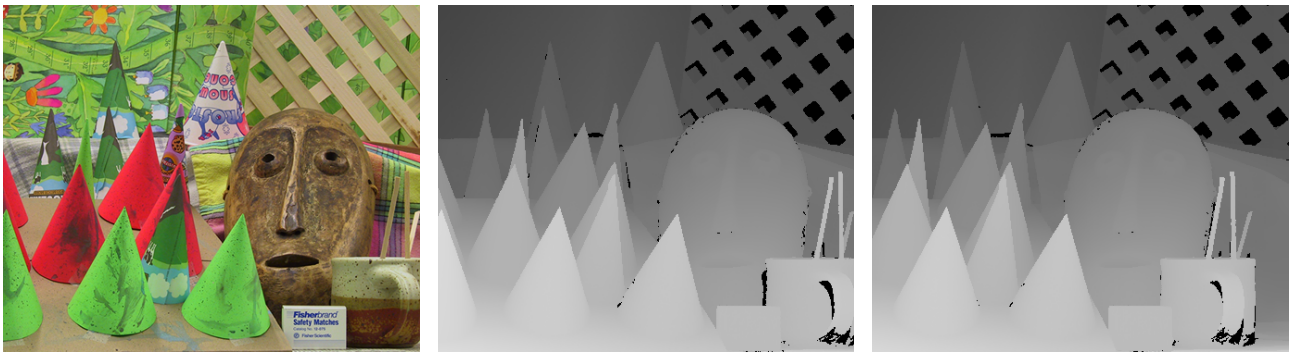
4.4.2.1 *Viewpoint correction layer*

Outputs from cell layers D_{μ}^{LC} are combined into a viewpoint correction layer, where the information from the three viewpoint disparity maps is used to select the most accurate information. This step can be seen as a fusion of the obtained stereo disparity maps relative to the perspective of an observer with a left-side viewpoint. This is done by shifting the maps to the right accordingly (each pixel's shift distance depends on its disparity value) and computing the median $M\{\cdot\}$ of the three maps:

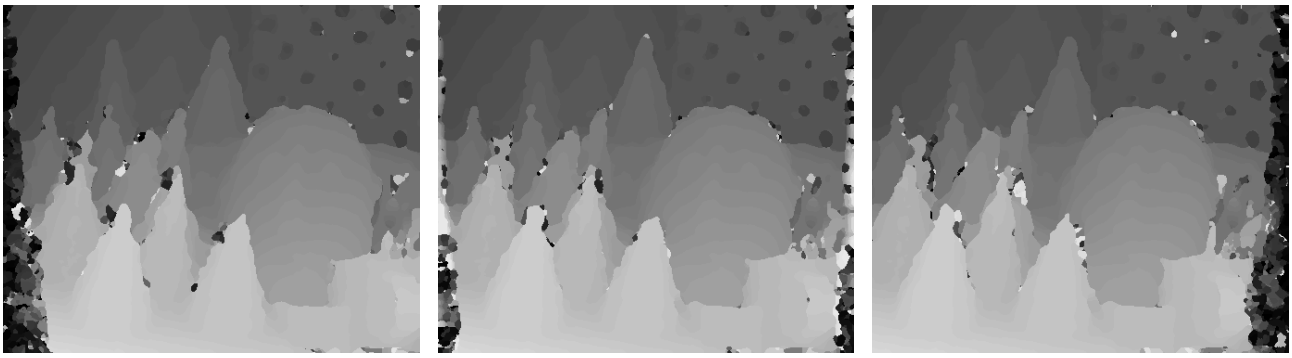
$$\widehat{D}^{LCV}(x,y) = M\Big\{ D^{LC}_{Left}(x,y), D^{LC}_{Centre}\big(x + 0.5\, D^{LC}_{Centre}(x,y), y\big), \\ D^{LC}_{Right}\big(x + D^{LC}_{Right}(x,y), y\big) \Big\}. \quad (4.12)$$

This can be seen in Figure 4.2(i). One of the benefits of this step — and combining viewpoints in general — is effectively increasing accuracy of disparity estimates at the left and right borders of objects, which are usually inaccurate due to viewpoint–occlusion (i. e., each eye will see some information that the other does not), eliciting a correspondence problem. This is more troublesome when the distance between left and right images of the pair is greater.

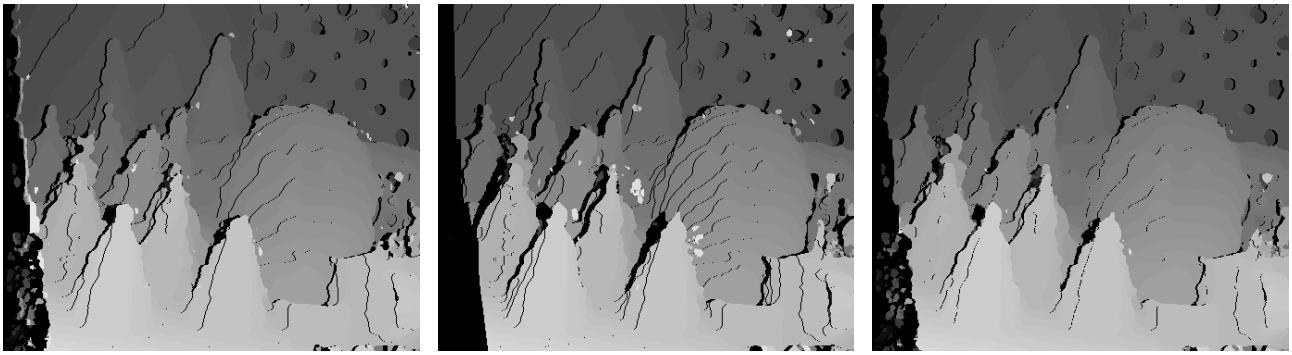
For illustration purposes, Figure 4.3 shows a better example of the benefits in combining viewpoints, for the *cones* stereo pair. Here, the left and right images are more separate, with a maximum disparity of 59px *vs.* only 15px for *tsukuba*. *Cones*' disparity maps show more differences between viewpoints. Respectively, (a) shows the left image of the pair, (b) the ground-truth from the left viewpoint and (c) from the right, with (d–f) displaying the different viewpoint disparity maps. We illustrate in (g) the effect of shifting the D^{LC}_{Centre} map to the Left viewpoint, corresponding to the middle term of (4.12): $D^{LC}_{Centre}\big(x + 0.5\, D^{LC}_{Centre}(x,y), y\big)$, and in (h) the even greater effect of shifting the D^{LC}_{Right} map to the Left viewpoint, corresponding to the



(a) *Cones* left image (b) Left ground-truth (c) Right ground-truth



(d) D^{LC}_{Left} (e) D^{LC}_{Centre} (f) D^{LC}_{Right}



(g) D^{LC}_{Centre} shifted to a Left view (h) D^{LC}_{Right} shifted to a Left view (i) $\widehat{D}^{LCV}$

Figure 4.3: Example of viewpoint correction results for the *cones* (Scharstein, 2003) stereo pair. **(a):** *cones* left image of the pair. **(b):** Left viewpoint ground-truth. **(c):** Right viewpoint ground-truth. **(d):** Left–viewpoint D^{LC}_{Left}. **(e):** Centre–viewpoint D^{LC}_{Centre}. **(f):** Right–viewpoint D^{LC}_{Right}. **(g):** Centre to Left viewpoint disparity shift. **(h):** Right to Left viewpoint disparity shift. **(i):** Fusion of the three Left (shifted) maps into $\widehat{D}^{LCV}$.

right term of (4.12):$D^{LC}_{Right}\left(x + D^{LC}_{Right}(x, y), y\right)$. The fusing of all three maps is showed in (i), where black pixels represent the uncertain disparity regions.

4.4.2.2 *Background and occlusion correction layer*

$\widehat{D}^{LCV}(x,y)$ will still show some uncertain/unknown disparities due to the occluded regions where the disparities were shifted from. To remove these, we determine which disparity is the possible background and assign to all smaller disparities the disparity of the background. Computationally, this process is done in four steps:

(a) Count how many cells ($\widehat{Nc}_{\Delta x_{stim}}$) are activated per disparity (Δx_{stim});

(b) Divide this by the square of the disparity $Nc_{\Delta x_{stim}} = \widehat{Nc}_{\Delta x_{stim}}/\Delta x^2_{stim}$ (this gives less priority to the "near" disparities, since it is expected that the background should be "far");

(c) Assign Δx_{stim} of the biggest value of $Nc_{\Delta x_{stim}}$ as the background (Δx_{bck});

(d) To values below Δx_{bck} are assigned the background disparity value Δx_{bck};

Afterwards, remaining inactive disparity cells receive the minimum median cell value of the closest active disparity cells in the epipolar plane, yielding $D^{LCV}(x,y)$. Results for *tsukuba* are shown in Figure 4.2(j).

4.4.3 *Experimental results*

We tested the LCV-DEM on the same Middlebury datasets as the L-DEM trials (Scharstein and Szeliski, 2002; Scharstein, 2003; Scharstein and Pal, 2007). Figure 4.2, (f) to (h) illustrate cell disparity maps for *tsukuba* — the D^{LC}_{Left}, D^{LC}_{Centre} and D^{LC}_{Right} images show results after luminance/colour grouping with three different viewpoints: Left, Centre and Right. The $\widehat{D}^{LCV}$ shows the integration of the three viewpoint maps into a single Left viewpoint (Figure 4.2i), and D^{LCV} shows the final LCV-DEM map after the background and occlusion correction layer (Figure 4.2j).

We can compare the results from D^{L} with D^{LCV} and verify that there are several improvements. Nevertheless, the edges and regions around objects

still lack a precise boundary definition. In the next section we will explain a model to assign disparity to line and edge features, and show how the integration of both disparity maps can be done. Section 4.6 (page 101) and Table 4.1 (page 103) show the quantitative results from the Middlebury stereo evaluation comparing L-DEM with LCV-DEM.

4.5 BOUNDARY ENHANCED LCV-DEM

In V1 there are monocular simple and complex cells which are thought to play an important role in coding the visual input: to extract multiscale lines and edges that are significant for object categorisation and recognition (Rodrigues and du Buf, 2009a). If lines and edges are extracted in area V1, where left and right retinal projections are closely together, one might even assume that depth is attributed to them. In other words, a "3D wire-frame representation" could be built in V1 for handling 3D objects and scenes. Although this idea is speculative, many V1 cells have been found to be tuned to different combinations of frequency (scale), orientation, colour and disparity. If not coded explicitly, disparity could be coded implicitly. In our Line and Edge Disparity Model (LEDM) we assume that lines, edges and disparity are coded explicitly. Disparity correction is based on enhancing disparity accuracy at object borders — Luminance, Colour, Viewpoint and Boundary enhanced Disparity-Energy Model (LCVB-DEM) — using low-level processes from an object salience model (Martins et al., 2012). This is done by using edge conspicuity and line/edge disparity information readily available in the V1/V2.

4.5.1 *Line and Edge Disparity Model*

The basic scheme for line and edge detection is based on responses of monocular simple cells modelled by Gabor filters (with the same orien-

tations as the binocular ones and scales $4 \leqslant \lambda \leqslant 24$ with a step size $\Delta\lambda = 2$: a positive/negative line is detected where R^E shows a local maximum/minimum and R^O shows a zero crossing. In the case of edges the even and odd responses are swapped. This gives four possibilities for positive and negative events. An improved scheme (Rodrigues and du Buf, 2009a) consists of combining responses of simple and complex cells, i. e., simple cells serve to detect positions and event types, whereas complex cells are used to increase confidence. Lateral and cross-orientation inhibition are used to suppress spurious cell responses beyond line and edge terminations, and assemblies of grouping cells serve to improve event continuity in the case of curved events. We denote the line and edge cell map by $\widehat{LE}_s(x, y)$.

Keypoint maps are also exploited in the LEDM model, as keypoints code line and edge crossings, singularities and points with large curvature. They are built from two types of end-stopped cells, single and double, which are modelled by the first and second derivatives of $C_{s,i}$. The latter are combined with tangential and radial inhibition schemes in order to obtain precise keypoint cell maps $KP_s(x, y)$. For a detailed explanation with illustrations see Rodrigues and du Buf (2006).

Figure 4.4 shows, *first row, left to right,* the multiscale line and edge coding at scales $\lambda = 4$ and $\lambda = 24$. Different grey levels, from white to black, show detected events: positive/negative lines and positive/negative edges, respectively. As can be seen, at fine scales many small events are detected, whereas at coarse scales only global structures remain.

The disparity to be assigned to each line/edge event is based on the left–right correspondence over scales. First we suppress events which may be due to noise. At each scale s of the left and right map of the stereo cell pair $\widehat{LE}_s^{L,R}(x, y)$, we compute the maximum response of the monocular complex cells $C_{s,i}$, but only at positions where events are detected. All events with small amplitude ($C_{s,i}$ below 5% of the maximum response) are inhibited. This yields $LE_s^{L,R}(x, y)$.

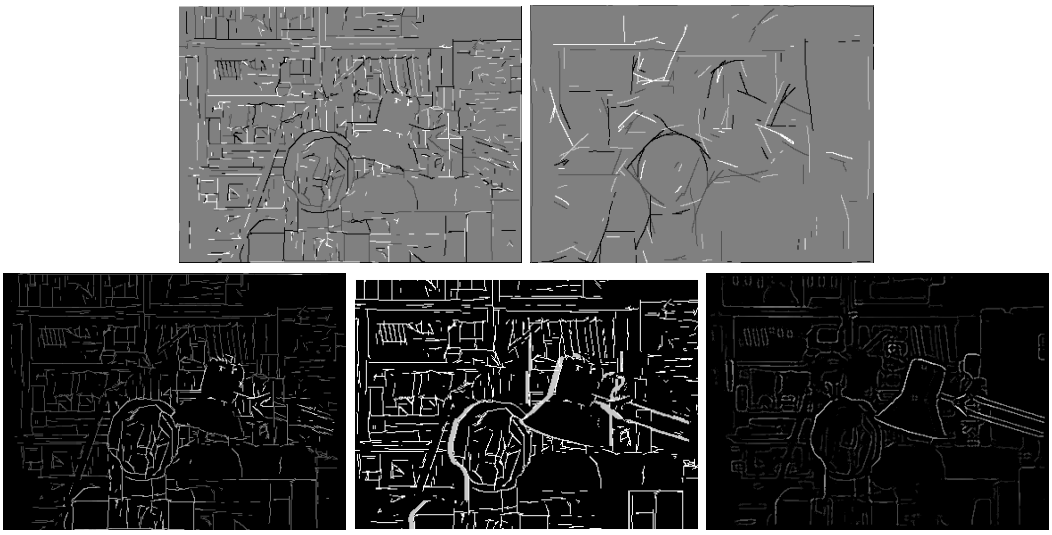


Figure 4.4: *First row, left to right:* multiscale line and edge coding at $\lambda = 4$ and $\lambda = 24$. *Second row:* D^{LE}, the correct disparities with an absolute mean error $\geqslant 0.5$ as returned by Middlebury evaluation site, and the conspicuity $\widehat{Co}$ after applying non-maximum suppression.

In the left map, at each event position (x_L, y_L) and at the finest scale ($s = 1$), LE_1^L is used to define regions of interest which are centred on each event position (x_L, y_L). These regions are grouping cells with circular receptive fields (RF). At the *same* position (x_L, y_L), other grouping cells are activated for all scales (still in the left cell map), with RF sizes depending on the scale: $2\lambda_s$. This *scale space* of the left cell map — or hierarchical set of grouping cells with RFs at all event positions from $s = 1$ — is used to accumulate displacement evidence of similar events at similar scales, but with relative (shifted) positions in the right cell map, LE_s^R; see below.

Basically, the RFs serve to "correlate" events in left and right cell maps as a function of the shift in the epipolar plane. This is done at all individual scales, after which the scales are combined. The right scale space LE_s^R shifts in Δx (epipolar line) with step size $\delta x = 1$, such that $1 \leqslant \Delta x \leqslant 60$, for a total of 59 shifts, at which the events in both scale spaces are binocularly RF "correlated" (similarly to L-DEM and LCV-DEM). The Δx with the maximum event *correspondence* (defined below) is then assigned to the disparity map $D^{LE}(x, y)$, where (x, y) still corresponds to event positions (x_L, y_L) of LE_1^L (Left viewpoint).

Computationally, at each scale and within each RF, four *correspondence* measures are combined with different weight factors:

(a) Counting all line/edge events with the same position, type (L/E) and polarity (+/−): $nLEtp_s$;

(b) As in (a) but only counting matching events irrespective of type and polarity: $nLEe_s$;

(c) Counting the number of complex cells with similar amplitudes at all event positions, i. e., $(C^L_{s,i} - 2) \leqslant C^R_{s,i} \leqslant (C^L_{s,i} + 2)$: $nLEa_s$; and

(d) Using $KP^{L/R}_s$, counting the number of keypoints with about the same coordinates, i. e., in small cell clusters of size 3×3: nKP_s.

Before combining the four measures, they are first normalised: $nLEtp_s$, $nLEe_s$ and $nLEa_s$ are divided by the number of events in LE^L_s, whereas nKP_s is divided by the number of keypoints in KP^L_s, each number being computed within each respective RF. The normalised numbers n are denoted by $\bar{n}$. The final *correspondence* is determined by combining the weighted and normalised measures over all scales:

$$\hat{C}_{\Delta x} = \sum_s (k_1 \cdot \bar{n}LEtp + k_2 \cdot \bar{n}LEe + k_3 \cdot \bar{n}LEa + k_4 \cdot \bar{n}KP), \tag{4.13}$$

with $k_1 = k_4 = 4$ and $k_2 = k_3 = 1$, empirically determined weights (small changes do not change significantly the final result). Finally, the Δx with the maximum value of $\hat{C}$ is stored in the depth map $D^{LE}(x, y)$. For more details see Rodrigues et al. (2012).

We applied LEDM to the Middlebury Datasets — Figure 4.4 shows in the *bottom row, left*: the disparity map D^{LE} coded by levels of grey, from dark (far) to white (near), the same as in L-DEM and LCV-DEM. The first image on the *bottom row* shows, in white, the correctly assigned disparities to the lines and edges as returned by the Middlebury evaluation site (absolute mean error $\geqslant 0.5$). The black regions were not evaluated. This error image shows that almost all lines and edges have a correctly assigned disparity.

4.5.2 *Edge conspicuity*

Employing results from a parallel V1 low-level process, *conspicuity* $\widetilde{Co}(x,y)$ is defined as the maximum difference between colours in $I_c(x,y)$, with $c = \{L, r, g, b\}$, at four pairs of symmetric positions at distance one from (x,y), i. e., on horizontal, vertical and two diagonal lines (Martins et al., 2012). Conspicuity $\widetilde{Co}$ is the maximum Euclidean distance between the colour pairs,

$$\widetilde{Co}(x,y) = \max_{i=1,\dots 4} \sqrt{\sum_c (I_c(\gamma_i)^2 - I_c(\widehat{\gamma}_i)^2)}, \tag{4.14}$$

with $\gamma = \left\{ (x-1, \{y, y-1, y+1\}), (x, y-1) \right\}$ and $\widehat{\gamma} = \left\{ (x+1, \{y, y+1, y-1\}), (x, y+1) \right\}$.

Only cells that respond higher than 10% of $\max(\widetilde{Co})$ (in order to remove low activity responses due to noise) are selected by non-maximum suppression, which yields conspicuity edge positions $\widehat{Co}$. The *tsukuba* $\widehat{Co}$ map is shown in Figure 4.4, *bottom row, at right*.

4.5.3 *Line and Edge region enhancement*

To enhance disparity accuracy in line and edge regions and remove small gaps we combine LCV-DEM with LEDM into an intermediate representation $\widehat{D}^{LCVE}$. For each event (x,y) in the D^{LE} map we check D^{LCV} in a small cell cluster, at the event position plus its left, right, top and bottom neighbours (N_4 neighbourhood). If it has nearly the same values, i. e.,

$$\left| M\left(N_4\left[D^{LE}(x,y)\right]\right) - M\left(N_4\left[D^{LCV}(x,y)\right]\right) \right| \leqslant T_i\,, \tag{4.15}$$

where T_i is an integer threshold value (1–5) and $M(\cdot)$ the median — the D^{LCV} cluster response at event position is propagated into $\widehat{D}^{LCVE}$. If not,

the same cluster region is filled with the values of D^{LE}. This way, we correct the LCV-DEM results using the LEDM responses. This process is applied in several cell layers (recursively) on top of the newly created $\widehat{D}^{LCVE}$ map, i. e., if it is not possible to fill it any more, but there are still gaps, we increment T_i by 1 and repeat the same procedure. In our experiments 5 was the maximum value. Biologically, this could correspond to 5 layers of $\widehat{D}^{LCVE}$ that activate neighbouring "idle" cells. The results can be seen in Figure 4.2(k).

4.5.4 *Object Boundary enhancement*

Despite the above process to correct ambiguous regions, some boundaries can still be improved. In real scenes, disparity borders are mostly linked to the contours of real objects, so we use a disparity sharpening process based on local contrast of disparity values, conspicuity information and line/edge boundaries to reach the final stage of this whole process, yielding D^{LCVB} — Luminance, Colour, Viewpoint and Boundary enhanced Disparity-Energy Model (LCVB-DEM).

With a special border-detection cell layer $B_d(x,y)$ which is active where $\widehat{Co}(x,y) > 0 \vee \left[D^{LE}(x,y) > 0 \wedge \left|D^{LE}(x,y) - D^{LCV}(x,y)\right| > 0\right]$, we devise two approaches to detect and correct bad disparity estimations *far* and *near* B_d active cells.

For the *far* case (i. e., regions that should be homogeneous), we analyse relationships between small disparity spikes or bumps and their surrounding areas. If the inside median disparity of a small cell cluster (less then 20 px) M_{in} is different from that of its border (outside perimeter) M_{out}, and if $M_{in} > M_{out}$ then the cell cluster disparity is reduced to the value M_{out}. Otherwise if $M_{in} < m_{out}$, with m_{out} the minimum value of the border region (perimeter), then the cell cluster disparity is increased to m_{out}. This eliminates spurious zones.

For the *near* case (i. e., object borders), every active border in B_d is subjected to a sharpening process, based on six median M detection cell clusters (each of size 3×3), orthogonal to the B_d edge orientation and sliding along an arbitrary distance range (d, from 1 to 25px). Three clusters are applied on both sides of the border, with a distance between the centres of one pixel. A block of three clusters slides away from the border until a continuity criterion is met: if M1 (closest to border), M2 and M3 (farthest from border) are the three median disparity values from the three cell clusters, when $M1 \neq M2$ by 2 (rise/drop) and $M3 = M2$, then the disparity M2 is propagated to the border position. This process is applied to both sides (180°) of the border, independently. The completion of this process returns the intermediate disparity map D^{LCVE}; see Figure 4.2(l).

The final step consists in inhibiting all inconsistent disparity cells and assigning to each position the most probable disparity within a small RF; also the borders are smoothed. This process is similar to a median smoothing filter and is achieved by applying a circular cell cluster to $D^{LCVE}(x, y)$ (radius of 6; slightly bigger or smaller sizes do not affect the ranking in the Middlebury test, despite slightly improving/degrading individual images), and assigning to the centre (x, y) position the median disparity value within the cluster. This yields the final disparity map LCVB-DEM denoted by D^{LCVB}. It is shown in Figure 4.2(m).

4.5.5 *LCVB-DEM Experimental Results*

Figure 4.2 shows the D^{LCVE} map in (l) and the final disparity map D^{LCVB} in (m). By subjecting the last result to the Middlebury evaluation test we obtain the *"bad pixels absolute disparity error* $\geqslant 0.5$*"* and *"signed disparity error"* of D^{LCVB}, respectively shown in (n) and (o). When comparing (m) with the results obtained from L-DEM in (c) we can observe significant improvements.

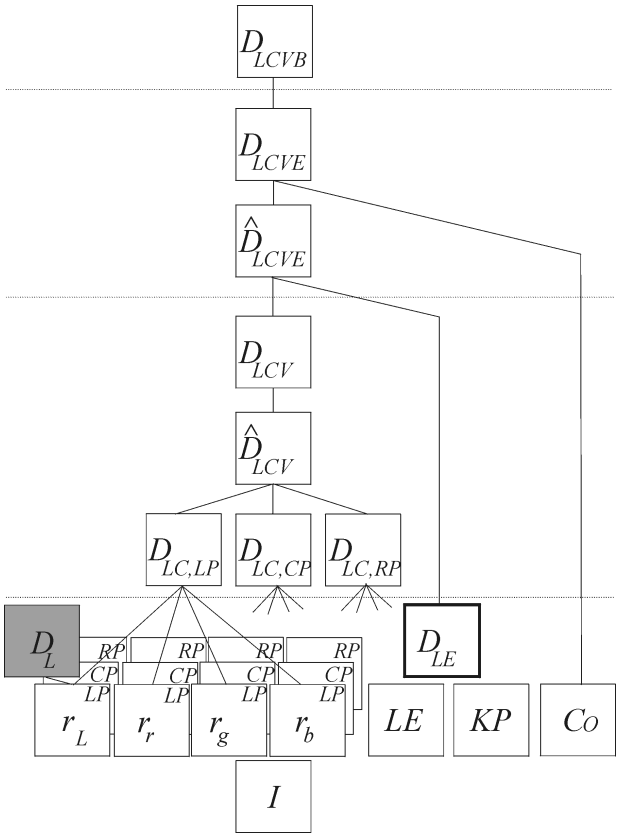


Figure 4.5: Summary of features and disparity maps for building the LCVB-DEM.

Figure 4.5 shows in detail all intermediate disparity maps needed to create the LCVB-DEM model, divided in three big layers. In grey, the L-DEM is similar to the classical DEM. Table 4.1 shows quantitative results from the Middlebury stereo evaluation, comparing the final maps of each layer. This table is explained below. In the next section we will show results from other images and discuss the different disparity models qualitatively and quantitatively.

4.6 RESULTS

As mentioned in Section 4.3.3, we tested the model at the different implementation steps on various stereograms, including *tsukuba*, *venus*, *teddy*, *cones*, *aloe*, *cloth3*, *dolls*, *moebius* and *reindeer* (Scharstein and Szeliski, 2002;

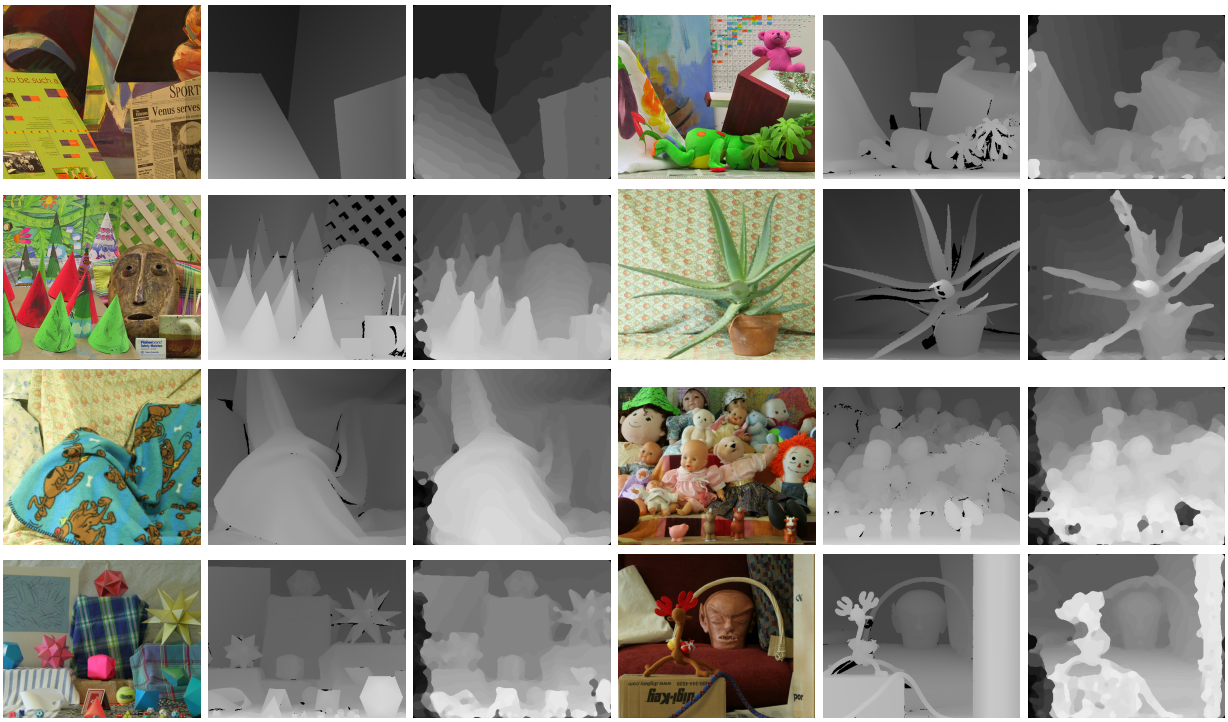


Figure 4.6: LCVB-DEM Middlebury dataset results. Each row triplet shows the left image of each input stereogram, the ground-truth, and disparity result.

Scharstein, 2003; Scharstein and Pal, 2007). Figure 4.6 shows the left image of the pair along with the groundtruth and our result.

Table 4.1 shows the evolution of error results, using the Middlebury evaluation page (Scharstein and Szeliski., 2012), for *tsukuba*, *venus*, *teddy* and *cones*, at different steps of the model which correspond to the final disparity maps in each layer (see Figure 4.5). *Top to bottom* is shown: (1) L-DEM; (2) LCV-DEM with integration of colour and viewpoint; (3) LCVE-DEM with integration of colour, viewpoint and LEDM; and finally (4) LCVB-DEM integrating all above steps with object border enhancement. The results show improvements in all layers of the model, with the number of error pixels decreasing.

We can see that our model performs best in non-occluded regions (*nonocc* columns) but is not as good near depth discontinuities (*disc*). This was expected, because L-DEM and LCV-DEM struggle at border transitions, which is why the LEDM model is used to improve the LCV-DEM; it improves results but without yet achieving outstanding results — still, the *disc* error de-

Table 4.1: Comparison of disparity error values in different model layers

	Tsukuba			Venus			Teddy			Cones			Avg %
Algorithm	*nonocc*	*all*	*disc*	*nonocc*	*all*	*disc*	*nonocc*	*all*	*disc*	*nonocc*	*all*	*disc*	bad pxls
L-DEM	16.5	18.2	39.8	8.64	9.83	36.0	19.3	27.2	44.0	11.7	20.8	29.7	23.5
LCV-DEM	13.1	14.6	35.9	8.53	9.69	35.2	19.0	26.7	43.5	11.5	20.5	28.9	22.3
LCVE-DEM	11.5	12.4	30.7	6.73	7.13	15.2	16.9	23.9	37.2	11.5	18.9	28.0	18.3
LCVB-DEM	11.1	12.0	32.0	5.94	6.31	14.9	16.6	23.6	36.9	11.5	18.9	27.1	18.1

creases more than a factor of two in the *venus* case. The *all* columns include regions and border regions (even those half-occluded). *"Avg % bad pxls"* gives a general indication of how well the methods perform, as it shows the average percentage of bad pixels over all twelve columns; for details see Scharstein and Szeliski. (2012).

Overall, best results were obtained for images without many small details. This is related to the size of the RFs in the cell population; smaller RFs are required to resolve the smallest details but unfortunately also increase binocular correspondence errors. Figure 4.7 shows our result when compared to the ranked results of other methods (which can include more sophisticated post-processing and top-down methodologies like image segmentation for yielding massively improved pixel-to-pixel correspondences). This table was copied from the Middlebury online evaluation webpage, applying the smallest available error threshold ($\leqslant 0.5$) to emphasise that a biologically-inspired algorithm can achieve very competitive results.

We can also see that the LCVB-DEM method improves the results achieved with the L-DEM (*BioDEM [117]*) method. Overall, we achieved a good position in the average ranking table: rank 71.2 between 3.9 (best) and 130.0 (worst), on a total of 135 evaluated methods. With LCVB-DEM we rise 64 positions relative to *BioDEM*, from place 96.6 to place 71.2. Finally, to the best of our knowledge, our method is ranked highest when compared with other biologically inspired methods.

Error Threshold = 0.5 Error Threshold...		Sort by nonocc			Sort by all			Sort by disc				
Algorithm	Avg. Rank	Tsukuba ground truth nonocc	Tsukuba ground truth all	Tsukuba ground truth disc	Venus ground truth nonocc	Venus ground truth all	Venus ground truth disc	Teddy ground truth nonocc	Teddy ground truth all	Teddy ground truth disc	Cones ground truth nonocc	Cones ground truth all	Cones ground truth disc	Average Percent Bad Pixels
(...)														
InfoPermeable [109]	70.5	25.7 120	26.2 117	21.2 68	8.64 88	9.34 88	15.0 71	15.0 62	22.1 65	29.2 55	7.68 34	15.1 48	15.1 30	17.5
VSW [108]	70.6	19.2 74	19.5 65	18.5 37	8.17 81	8.65 78	13.2 51	17.4 82	23.2 77	31.4 72	13.1 85	18.3 74	20.4 71	17.6
YOUR METHOD	71.2	11.1 23	12.0 25	32.0 123	5.94 45	6.31 46	14.9 88	16.6 78	23.6 81	36.9 109	11.5 70	18.9 78	27.1 108	18.1
PlaneFitBP [32]	71.4	12.7 38	13.6 38	16.2 22	8.58 86	9.12 84	12.7 45	18.4 91	24.3 86	31.4 71	15.3 101	21.9 99	23.9 96	17.3
AdaptDispCalib [36]	71.8	24.6 109	24.7 100	21.3 70	7.14 66	7.56 65	15.0 70	18.8 94	25.2 92	29.7 58	9.21 47	15.1 47	16.7 43	17.9
RTAdaptWgt [114]	72.8	19.2 73	19.9 72	21.4 74	9.61 99	10.2 99	16.0 79	15.8 71	22.2 66	28.0 44	11.5 71	17.5 64	19.1 62	17.5
(...)														
OptimizedDP [70]	96.0	24.0 99	25.5 108	22.7 82	12.4 112	13.7 113	23.7 108	17.1 81	25.0 90	30.3 63	14.1 92	22.2 104	24.6 100	21.3
ESAW [86]	96.1	19.2 71	19.7 70	22.8 84	11.0 107	11.7 105	18.4 90	19.4 98	25.9 98	33.6 93	18.5 116	24.0 110	28.8 111	21.1
DOUS-Refine [87]	96.3	21.6 96	22.8 87	21.1 63	10.8 104	11.8 107	22.8 104	18.2 89	25.2 93	33.8 94	16.3 108	23.0 105	30.0 115	21.5
BioDEM [117]	96.6	16.5 56	18.2 58	39.8 132	8.64 88	9.83 94	36.0 127	19.3 97	27.2 105	44.0 121	11.7 75	20.8 92	29.7 114	23.5

Figure 4.7: Ranked comparison of the LCVB-DEM method (listed as "YOUR METHOD") and the original L-DEM method (listed as "BioDEM") on the Middlebury evaluation webpage (Scharstein and Szeliski., 2012), with the strictest error threshold ($\leqslant 0.5$), ordered by average ranking.

4.7 DISCUSSION AND CONCLUSIONS

We presented an algorithm for disparity estimation based on the Disparity Energy Model that yields good results even when compared with state-of-the-art computer vision algorithms. Unlike other models, it does not rely on specific filter parameters to estimate disparities, it only employs filter responses. Therefore, the model must be trained, but only once. First, a population of binocular complex cells based on simple cells is defined, which is trained by using random-dot stereograms. It is then applied at all image positions in order to compare the local activation code with the learned activation codes, which yields local disparity estimates.

The number of filters involved is rather large: for the L-DEM model, 2880 simple cells on the basis of which 1440 complex cells are constructed, all at every retinotopic (image) position, and 17,280 binocular cells for the LCVB-DEM model. In reality, our visual cortex counts many more cells in V1 — the latest estimates point to 190 million (DiCarlo et al., 2012), but that number could well be more closely around 243 million (average volume

of V1 of 5,405mm^3×45,000 cells/mm^3). However, our model does not introduce any new filters or many other cells. It only exploits cells which are already available: pairs of simple cells at different positions are only wired together, and they also serve other purposes, like multi-scale line and edge coding, which was used for the LEDM model, but also object recognition and brightness perception (Rodrigues and du Buf, 2009a). In addition, disparity, as for optical flow, is very important for object segregation, supplementing surface features like colour and texture. Also, as shown by Pugeault et al. (2010), spatial structures can be linked both in 2D and 3D by using constraints like good continuity.

Interestingly, the fact that disparity is extracted in the hypercolumns of V1, where left and right projections are close together and also lines and edges are coded, suggests that our visual system may attribute depth to detected lines and edges already at that level. Hence, our brain could use a sort of "wireframe" representation (as used in computer graphics) to model solid objects and employ this for 3D object recognition. Furthermore, post-processing of local disparity estimates can be based on edge information: edges between homogeneous regions are often caused by occlusions, exactly where disparity is not continuous and detail is visible in one projection but not in the other. Therefore, disparity estimation astride edges can be steered by detected edges, using phase tuning, and smoothed in homogeneous regions.

Disparity training is applied to the encoding population in order to prepare the matching process, but the decoding population is a fixed neural network. It involves subtractive and divisive normalisation in combination with half-wave rectification, for which plausible neuronal mechanisms have been proposed (Read, 2010). However, the decoding population could also be trained, even dynamically adapting itself to local image content by neural plasticity.

The role of colour in biological disparity models is still rather speculative (den Ouden et al., 2005), with little research into biological disparity models

that employ colour, even with the already existing evidence that disparity-sensitive neurons are also isoluminant-sensitive. Meanwhile, our empirical evidence suggests that colour may definitively play a significant role in improving the luminance discrimination of cells, and that it significantly improves disparity estimations. Also, we combined colour conspicuity around border regions (Martins et al., 2012) to better define disparity transitions.

The role of *perspective correction*, to shift the viewpoint of disparity maps to yield better estimates, is also biologically plausible: as even V1 cells display the ability to shift their RFs, see e. g., Fu et al. (2004). This could be particularly useful for multi-view stereo.

Finally, we propose and illustrate that the classical DEM (L-DEM), and LEDM can be used as a disparity "gist," i. e., they can give us a fast draft of the environment either from binocular energy complex cells or from object contours (the bottom layer of Figure 4.5). These maps are appropriate for person (or robot) navigation, as they are only based on quickly extracted visual features in the same layer. In a second layer, the DEM is combined with colour and perspective correction, giving a more accurate disparity map, but still with a lack of border definition. However, this map can be enough for object categorisation. In the 3rd layer, information about edges is integrated into the final LCVB-DEM disparity map. This map can be appropriate for object recognition. In summary, we have two disparity gist-like maps, one with edge information (LEDM) and one with inaccurate, but precise generic region information (L-DEM) — they combine colour, viewpoint and in the future also texture and other features, to achieve an even more precise map (LCVB-DEM).

Highlighting other possibilities for further research, the composition of disparity energy neurons suggests the possibility of using a different combination of DEM subunits in an energy complex cell, with different weights and signs (Tanabe and Cumming, 2008; Haefner and Cumming, 2008). The role of phase-tuned cells is also an interesting subject for debate (Read and Cumming, 2007; Rodrigues and du Buf, 2009a), as their use can be seam-

lessly integrated into our model, signalling false disparity matches that can be corrected.

5

OBJECT CATEGORISATION

PROTO-OBJECT CATEGORISATION AND LOCAL-GIST USING IMPLICIT CODING OF LOW-LEVEL SPATIAL LAYOUT FEATURES

ABSTRACT: Object categorisation is a research area with significant challenges, especially in conditions with bad lighting, occlusions, different poses, and similar objects. This makes systems that rely on precise information unable to perform efficiently, like a robotic arm that needs to know the exact grasp pressure for different kinds of objects. We propose a biologically-inspired object categorisation framework that relies on robust low-level object shape, using edge conspicuity and disparity features for categorisation, with a trained Neural Network classifier, that can consequently bootstrap a scene gist system.

KEYWORDS: Disparity, 3D, stereo vision, colour, population coding, learning, biological model, object, categorisation, verification, neural network, LDA, PCA, visual cortex.

5.1 INTRODUCTION

Object recognition and categorisation is one of the primary research areas in computer and biological vision, since objects represent the most important visual stimuli available to humans and are key to survival. There are many visual pathways related to object recognition and categorisation, es-

pecially focusing on quick shape categorisation, which is essential for scene gist. In this chapter we focus on the transition between low-level syntax and low-level semantics, using elementary information such as surface lighting, colour and stereo disparity. The goal is to develop an integrated system for fast local gist vision: *which* types of objects are about *where* in a scene. This is necessary to bootstrap and guide, even alleviate, the processing in the ventral and dorsal data streams — which are known to serve two goals: the dorsal stream, also called the *where* or *vision-for-action* stream, is mostly devoted to optical flow and stereo disparity, whereas the ventral stream, also called the *what* or *vision-for-perception* stream, is devoted to object categorisation and recognition (Konen and Kastner, 2008; Farivar, 2009). However, the dorsal stream can also play a very important role in object categorisation (Gottlieb, 2007; Janssen et al., 2008; Konen and Kastner, 2008), which is the main focus of this chapter.

This integrated system must first solve two hard problems: (a) The first one is of paradoxical nature, as precise object categorisation and recognition in the ventral stream requires object segregation, but object segregation has been usually regarded as only possible if the system already knows what the object is (assuming of course that objects are seen against complex backgrounds). Consequently, we explore the possibility for a system to segregate an image region or an object even before knowing what it is, based on robust low-level and local-shape features like edge conspicuity and disparity, as we will show later. (b) The second problem is that object categorisation and recognition is a sequential process: while fixating one object, its features must be routed to normalised object templates held in memory. This routing blocks the system until categorisation and recognition has been achieved, after which the system is released for dealing with another object. Therefore Rensink (2000) proposed a non-attentional "scene schema" consisting of concurrent spatial-layout and gist subsystems which both drive attentional object categorisation and recognition, all employing "proto-objects" resulting from low-level vision. Gist vision addressed so

far mostly concerns global gist of entire scenes (Bar, 2004; Siagian and Itti, 2007; Ross and Oliva, 2010; Rodrigues and du Buf, 2011a). However, there is already some research into local parts of scenes that allow a quick pre-categorisation of geometrically-shaped objects (Martins et al., 2012). Here we extend the use of fixed geometric shapes and aim at representing any kind of proto-object shape.

The use of depth information has shown good results in the context of general object detection (Quigley et al., 2009), as depth information is resilient to lighting and tonal variations, provides geometrical hints and is efficient for separating foreground from background. Most recent methods for object recognition with RGB-D images use hand-designed features such as SIFT for 2D images (Lai et al., 2011), Spin Images for 3D point clouds (Johnson and Hebert, 1998), specific colour, shape and geometry features (Lu and Rasmussen, 2012; Bo et al., 2011) or Neural Networks with Deep-Learning (Socher et al., 2012).

Succinctly, we will address the possible construction of "proto-objects" using normalised shapes, resulting from low-level attentional edge conspicuity and disparity processes. Edge conspicuity is a measure for object salience that highlights the transitions in colour/lighting at the borders of objects and relies heavily on the simultaneous colour and luminance contrast of an object with its background, so it is able to represent both *salience* and *shape* information. This is especially useful for constrained processing systems that have limited resources and must first prioritise image regions or shapes to process, like it is common in robotics. Besides conspicuity, disparity information is also of paramount importance for shape categorisation and recognition, since it is only mildly affected by variations in pose and illumination.

Previous work on low-level shape feature extraction (Martins et al., 2012) in Chapter 2 used adaptive feature detectors for extracting *corners*, *edges* (bars) and *curvature* information from objects and defined simple but efficient rules for classification based on geometric relationships between those

features. It was shown that many man-made objects with geometric properties obey those rules, yielding good results. We now propose a more elaborate and generic shape extraction method that can define a shape feature vector without explicitly constraining the feature-search space, generalising the process of object categorisation, also without needing to define geometric relationships between features.

We also use disparity and conspicuity information for object detection, for inhibiting the background of scenes and for highlighting conspicuous objects in the foreground, so that a robust shape feature vector can be obtained.

Our main contributions are a quick object *detection* and *categorisation* framework that relies on salient scene information (both conspicuity– and disparity–based) to simultaneously detect foreground objects, retrieve object shapes and disregard superfluous information. These object shapes, represented by a shape vector, can be used in a feed-forward processing scheme, like a Neural Network, to quickly assess their shape category, effectively being a type of "proto-object" representation that only needs very few data. We also aim to prove that structural object information that is available from biologically-inspired salience methods can successfully recognise objects with good accuracy, and is also capable of obtaining information not available in luminance-based methods.

Section 5.2 explains the used databases, Section 5.3 details the steps necessary for the detection of objects and encoding of their shapes, Section 5.4 details the experimental conditions and categorisation processes, Section 5.5 presents the evaluation trials and Section 5.6 the discussion and conclusions based on the data.

5.2 OBJECT CATEGORISATION DATABASES

An ideal database for the present research would need two requirements: (1) RGB camera-rectified stereograms of objects with full object revolutions

and (2) a large collection of objects with different categories. Unfortunately, at present time, there is no object database that complies to both, so we needed to use two databases, each for a different purpose:

(a) The CSCLAB Image database (CSCLAB ID) (Murphy-Chutorian and Triesch, 2005) has RGB camera-rectified stereograms of objects in frontal views. It is used to illustrate the first part of our work, dealing with object detection, segmentation and shape extraction, on different kinds of complex scenes/backgrounds (Section 5.3). This database contains different kinds of mundane objects in ten different backgrounds.

(b) The RGB-D Object Dataset (RGB-D OD) (Lai et al., 2011) has a large collection of objects with 360° revolutions (with 3 different camera heights), but has only single RGB images (no stereo pairs) with matching depth images for each object. It is used to compare our object categorisation performances with those obtained by other authors. Since RGB-D OD was created by using a high-resolution RGB camera and an IR light pattern to measure disparity (a prototype RGB-D PrimeSense camera, similar to Microsoft's Kinect), it allows us to establish a baseline for expected performance of a system which employs precise depth information.

Examples of used object data are shown in Figure 5.1. The CSCLAB ID (Figure 5.1:a–j) contains stereo-rectified images of objects under various backgrounds (only left viewpoint images are shown). We use the stereo image pairs for exemplifying our whole object processing algorithm, from detection to shape extraction, using disparity and conspicuity mechanisms. RGB-D OD contains cropped object images with corresponding range maps. In Figure 5.1:(k–t) we show 5 example objects of categories "apple" and "cellphone." We use both range maps and RGB-D data to detect and extract object shape information and then obtain categorisation performance estimates.

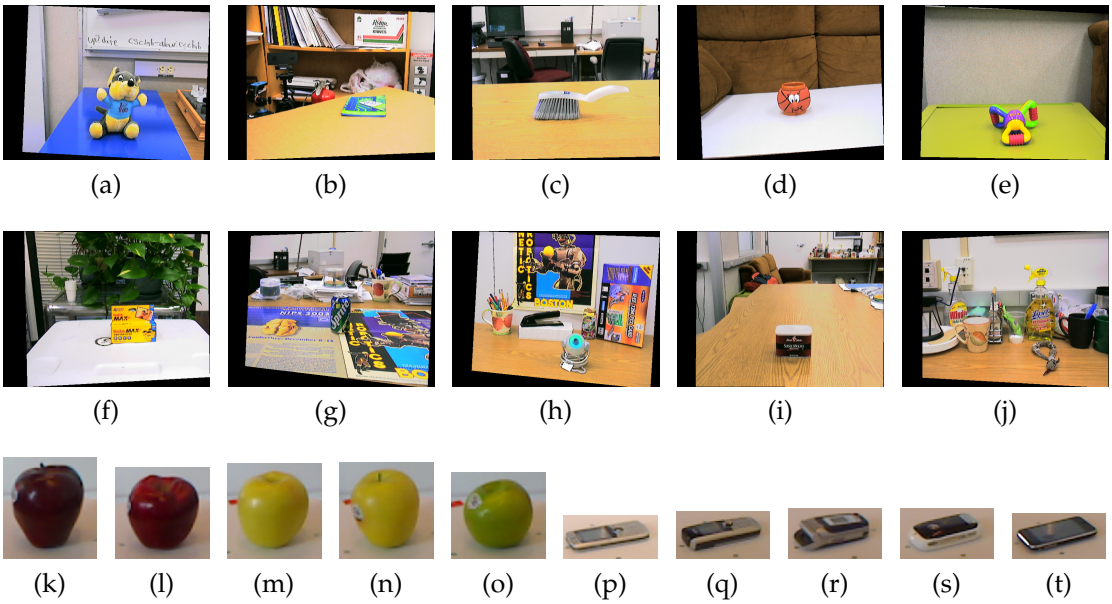


Figure 5.1: Examples of objects. **(a–j):** CSCLAB image database, left stereogram images of ten example objects for each of the ten different backgrounds. **(k–t):** RGB-D object database, five apples and five cellphones at 0° turntable position and 30° height.

5.2.1 *CSCLAB Image Database*

The CSCLAB Image Database (Murphy-Chutorian and Triesch, 2005) was created at the Complex Systems and Cognition Laboratory at the University of California, San Diego. It consists of a single view of 50 objects for training and 498 heterogeneous scenes for testing, each containing from 3 to 7 objects, with similar poses in 10 different backgrounds. Objects can be significantly occluded and display subtle differences in scale, viewpoint and illumination conditions.

5.2.2 *RGB-D Object Dataset*

The RGB-D Object Dataset (Lai et al., 2011) contains visual and depth images of 300 distinct objects with multiple views. The chosen objects are commonly found in home and office environments, where personal robots are expected to operate. Objects are organised into a hierarchy taken from

WordNet hypernym/hyponym relations, which is a subset of the categories in ImageNet. Data was recorded with the cameras mounted at three different heights relative to a turntable where the object was located, at approximately 30°, 45° and 60° above the horizontal plane. One revolution of each object was recorded at each height. Each video sequence was recorded at 20 frames per second and contains around 250 frames, giving a total of 250,000 RGB + Depth frames.

5.3 OBJECT DETECTION AND SHAPE CODING FRAMEWORK

This section describes how low-level features can be combined to detect an object in a complex scene, yielding a binary segmentation mask of the object's outline. This mask is then used to extract a shape feature vector that describes the object's contour and is independent of object size, but dependent on object perspective. We will illustrate this process using images from the CSCLAB Object Database (Murphy-Chutorian and Triesch, 2005), that has each object in 10 different backgrounds. Below we introduce our biological disparity model, edge conspicuity model (*border saliency*), classifiers used and classification rules in case of identification and verification experiments. For having a completely biological framework we will also consider — apart from the two disparity cell populations used (one for encoding and another for decoding) — a third population: a trained neural-network classifier.

5.3.1 *Disparity-Energy Model*

Here we applied the Luminance, Colour and Viewpoint Disparity-Energy Model (LCV-DEM) implementation for extracting disparity maps from all stereogram image pairs in the CSCLAB database. For a detailed explanation, we refer the reader to Section 4.4, where it is explained in de-

tail. Obtained disparities (D) for the CSCLAB object #107 (a *beer bottle* — Figure 5.3a) are exemplified in Figure 5.3(b).

5.3.2 *Disparity-based background inhibition*

LCV-DEM disparities are then used for an initial foreground/background segregation. This is done by first inhibiting the background disparity responses that are below an empiric threshold of the normalised disparity values. This results in a *foreground*–only disparity image, defined $\forall(x, y)$ as:

$$D^{+}(x, y) = D(x, y) > 0 \tag{5.1a}$$

$$D^{n}(x, y) = \frac{D - \overline{D^{+}}}{\sigma_{D^{+}}} \tag{5.1b}$$

$$D_{fg}(x, y) = \begin{cases} D(x, y) \text{ if } D(x, y)^{n} > 0.1 \\ \text{OFF otherwise.} \end{cases} \tag{5.1c}$$

The notation $\overline{(\cdot)}$ reflects the mean and $\sigma_{(\cdot)}$ the standard deviation. Threshold parameter 0.1 was empirically chosen as a good compromise. This step is effectively a disparity–based global scene pre-segmentation process that allows for posterior processing of only foreground regions. The result of this step can be seen in Figure 5.3(c).

5.3.3 *Edge conspicuity model*

Edge conspicuity has yielded good results in object shape discrimination, using luminance and colour differences to differentiate object shapes (Martins et al., 2012). It is also a process similar to the one explained in Section 2.2, up to Section 2.2.1.4.

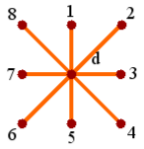


Figure 5.2: Conspicuity orientations mask. The four triplets of conspicuity orientations (1,5), (2,6), (3,7) and (4,8).

Succinctly, it starts with a step of colour smoothing inside image regions, removing redundant information which is not necessary for shape detection, while preserving each region's boundaries. This also helps to stabilise regional differences astride boundaries, between the inside and outside of objects. This is done using a cell layer that outputs a result similar to an adaptive Difference-of-Gaussians (DoG) filter with edge preservation (for an in-depth discussion, please refer to Martins et al., 2012). The resulting colour image from this step is defined as $I_c(x, y)$.

Gathering the smoothed results as a V1 low-level process, *conspicuity* ($\widetilde{C}$) is now defined as an edge salience measure that represents the maximum difference between smoothed colours in $I_c(x, y)$ (with $c = \{L, u, v\}$, explained below) at four triplets of symmetric positions at distance $l = 1$ from (x, y), i. e., on horizontal, vertical and two diagonal lines (Martins et al., 2012):

$$\widetilde{C}_{Luv}(x, y) = \max_{i=1}^{4} \left(\sqrt{\sum_{c} \left(I_c^{L,u,v}(\vec{x}_i)^2 - I_c^{L,u,v}(\vec{x}_{i+4})^2 \right)} \right). \tag{5.2}$$

Figure 5.2, shows the positions of the clusters of gating cells. If these are called G_i, opposite pairs are (G_i, G_{i+4}), with $i = \{1, \ldots, 4\}$, for example (G_1, G_5) and (G_4, G_8). The main idea for using the CIE *Luv* colour space is that it mimics double-opponent colour cells found in human vision, making it very useful for estimating the conspicuity of borders between image regions. We also found out empirically that the CIE *Luv* colour space generally performed better for conspicuity discrimination than the CIE *Lab* colour space. Only cells that respond higher than 10% of $\max(\widetilde{C})$ are kept (in or-

der to remove low activity responses due to noise), which yields conspicuity edge positions $\widehat{C}$. Results of $\widehat{C}$ are shown in Figure 5.3(d).

5.3.4 *Foreground object detection*

At this stage, we use a combination of conspicuity and disparity information to estimate possible foreground object locations. For the present work, we are only interested in processing the most conspicuous foreground object in each scene (the most salient), so the rest of possible object locations are discarded (they could also be processed in queue, using a FoA top-down approach).

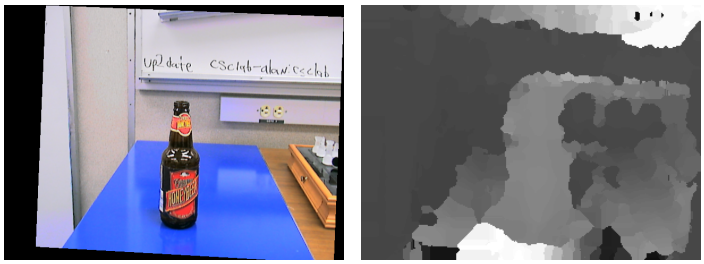
5.3.4.1 *Background conspicuity inhibition*

Since disparity and conspicuity extraction is done in parallel by two different cell populations, we also envision a quick V1/V2 low-level process that combines them for signalling important regions for low-level attention (Rensink, 2000) and FoA gaze. Since we already have foreground disparity regions from Section 5.3.2 (D_{fg}), at this stage the *active* conspicuity cells in $\widehat{C}$ are compared against the *foreground* disparity areas (D_{fg}), inhibiting any conspicuity responses outside these areas. This is exemplified in Figure 5.3(e).

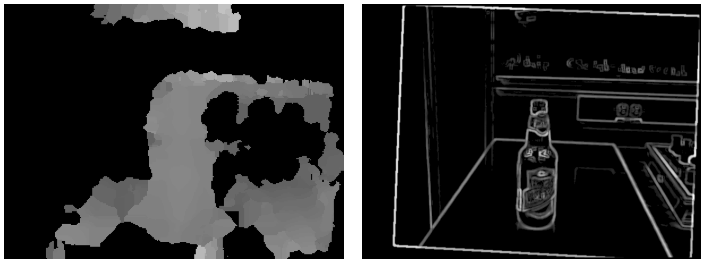
5.3.4.2 *Disparity-based object detection*

From the foreground disparity and conspicuity maps, object detection is done using a weighted combination of both. First, we count the number of active foreground conspicuity cells in each possible disparity 3-plane slice $(d-1,\ d,\ d+1)$, with $1 < d < D_{fg}^{max} - 1$. This can be defined as

$$A_d^{\widehat{C}} = \sum_{d-1}^{d+1} \left[\widehat{C}(x,y) > 0\right]_d .$$



(a) Left stereogram (b) LCV-DEM



(c) D_{fg} (d) Conspicuity $(\widehat{C})$

(e) $\widehat{C} \wedge D_{fg}$ (f) D_{ObjReg}

(g) $\widehat{C}^{A}_{ObjReg}$ (h) $\widehat{C}^{A}_{Obj}$

(i) Object mask (j) Object segregation

Figure 5.3: "*Beer bottle*" shape extraction. **(a)** left image of the stereo pair; **(b)** LCV-DEM disparity map; **(c)** foreground disparity map; **(d)** scene edge-conspicuity; **(e)** foreground scene conspicuity map; **(f)** foreground object region disparity; **(g)** foreground object region conspicuity; **(h)** object conspicuity; **(i)** object shape mask; **(j)** object segregation using the shape mask applied to the left stereogram image.

Next, we select a single disparity plane slice δ using two conditions: it is the closest possible disparity plane (highest d) that also has the highest cell count (highest $A_d^{\widehat{C}}$ value when $\delta = d$). For this we use an empirically formulated criterion, such that δ satisfies:

$$A_\delta^{\widehat{C}} \cdot \delta^4 \geqslant \left[A_d^{\widehat{C}} \cdot d^4 \right]_{d_{min}}^{d_{max}}.$$

The disparity value δ therefore represents the closest disparity plane slice that simultaneously has the most active conspicuity cells, and is also the best candidate for the closest and most conspicuous foreground object. Since we wish to expand the disparity slice to encompass all possible disparity values of the object in question (as objects rarely occupy only a single disparity plane), we define an empirical *range* parameter (r) such that the final disparity object slice will range from $[\delta - r, \delta + r]$, with $r = \sigma_{D^+}/2.5$ (the 2.5 value was empirically chosen). The *probable region* of foreground disparity ranges where the object is located is then defined as

$$D_{ObjReg}(x, y) = \begin{cases} D_{fg}(x, y) & \text{if } \delta - r \leqslant D_{fg}(x, y) \leqslant \delta + r \\ OFF & \text{otherwise.} \end{cases}$$

This is illustrated in Figure 5.3(f). We can see that only the disparity values present in the bottle region were preserved, while the rest was discarded.

The next object detection process selects only the conspicuity values of the foreground object. First, conspicuity values are normalised by

$$\widehat{C}^+(x, y) = \widehat{C}(x, y) > 0$$
$$\widehat{C}^n(x, y) = \frac{\widehat{C} - \overline{\widehat{C}^+}}{\sigma_{\widehat{C}^+}},$$

after which we determine $\widehat{C}^{A}_{ObjReg}$, which represents the *significantly active* conspicuity cells of the object region (i. e., active–above–average in the region),

$$\widehat{C}^{A}_{ObjReg}(x,y) = \begin{cases} \widehat{C}^{n}(x,y) & \text{if } \widehat{C}^{n} > 0 \wedge D_{ObjReg} > 0 \\ OFF & \text{otherwise.} \end{cases}$$

The result of this step is shown in Figure 5.3(g) — pixels darker than the background represent less–than–average conspicuity values ($\widehat{C}^{n} < 0$) and are discarded, while pixels brighter than the background represent the most active conspicuity cells in the region ($\widehat{C}^{n} > 0$) and are kept in the $\widehat{C}^{A}_{ObjReg}$ map.

There is now a further refinement step to see if the active conspicuity cells in the object's region effectively belong to the object in question (border ownership). This connectivity test is done mathematically using four dilation iterations on $\widehat{C}^{+}$ to close small gaps between active conspicuity cells, and afterwards keeping only the biggest connected area. This area is then intersected with $\widehat{C}^{A}_{ObjReg}$, yielding the final conspicuity image with just the object's active cells ($\widehat{C}^{A}_{Obj}$), as shown in Figure 5.3(h). This process can also be explained biologically using an equivalent higher-level grouping cell population with big RFs, that activates when the lower $\widehat{C}^{+}$ layer has enough cells active within each higher-level RF region.

5.3.5 *Object mask*

The next step uses the $\widehat{C}^{A}_{Obj}$'s cells — that are now correctly segregated from everything else, to extract an *object mask*, which is useful for both segmentation and shape categorisation. This is done in two steps:

5.3.5.1 *Non-maximum suppression*

Cell layer Ω, built on top of the $\widehat{C}^{A}_{Obj}$ layer, applies non-maximum suppression in order to extract the positions where $\widehat{C}^{A}_{Obj}$ has a local maximum in horizontal, vertical and diagonal directions, in 3×3 neighbourhoods. This is achieved by four oriented cell clusters plus one grouping cell at the output. Mathematically,

$$\Omega_{(x,y)} = \begin{cases} \widehat{C}^{A}_{Obj(x,y)} & \text{if } \left[\widehat{C}^{A}_{Obj(x,y)} > \widehat{C}^{A}_{Obj(x-1,y-1)} \wedge \widehat{C}^{A}_{Obj(x,y)} > \widehat{C}^{A}_{Obj(x+1,y+1)}\right] \\ & \vee \left[\widehat{C}^{A}_{Obj(x,y)} > \widehat{C}^{A}_{Obj(x-1,y+1)} \wedge \widehat{C}^{A}_{Obj(x,y)} > \widehat{C}^{A}_{Obj(x+1,y-1)}\right] \\ & \vee \left[\widehat{C}^{A}_{Obj(x,y)} > \widehat{C}^{A}_{Obj(x-1,y)} \wedge \widehat{C}^{A}_{Obj(x,y)} > \widehat{C}^{A}_{Obj(x+1,y)}\right] \\ & \vee \left[\widehat{C}^{A}_{Obj(x,y)} > \widehat{C}^{A}_{Obj(x,y+1)} \wedge \widehat{C}^{A}_{Obj(x,y)} > \widehat{C}^{A}_{Obj(x,y-1)}\right] \\ \text{OFF} & \text{otherwise.} \end{cases} \tag{5.3}$$

5.3.5.2 *Contour continuity filling*

The contour gaps of Ω are closed using a process similar to morphological closing (four dilations followed by four erosions). This fills gaps in the contours and results in Ω^{f}. The aim is to get fully active Ω^{f} cells for the whole contour of the object, so that these can be used for segregation. After the contours are connected, the inside of the shape is filled for yielding a binary segmentation mask. The resulting binary object mask is shown in Figure 5.3(i) and the object mask applied to the left stereogram input image (segregated object) is shown in Figure 5.3(j).

Figure 5.4 shows more results for the *beer bottle* shape in nine different backgrounds. In only one case (*poster table*) the conspicuity of the pattern on the table close to the bottle is so complex that segregation using only disparity and conspicuity information is not sufficient.

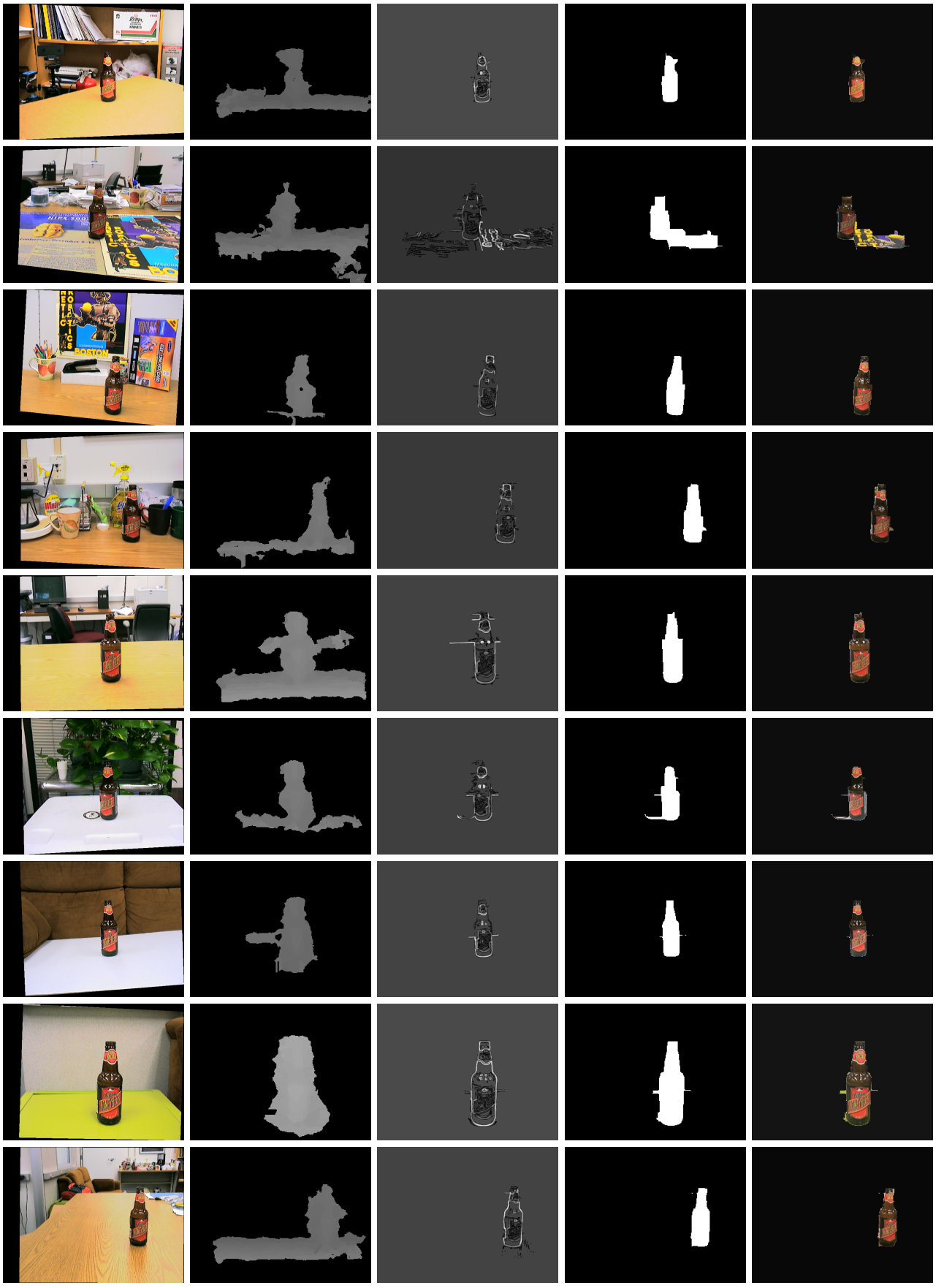


Figure 5.4: "*Beer bottle*" shape extraction under different backgrounds. *Left to right columns:* (1) left image of the stereo pair; (2) foreground object region disparity; (3) object conspicuity; (4) shape mask; (5) object segregation.

5.3.6 *Shape feature vectors*

Humans can recognise objects using several vision cues, even when presented with different views of the same object. We know that shapes of objects depend on the observer perspective — the shape of a particular object can be rather stable when the observer's viewpoint rotates around it (e. g., an orange) or it can change significantly (e. g., a statue). Also, objects can appear with any arbitrary rotation (e. g., upside-down, or a bottle lying flat on a table), which is also an obstacle to solve.

For the scope of this chapter, we chose to address the problem of observer perspective using the most common object poses, since it is the most relevant case for local gist processing and probably even hard-coded in a very quick proto-object neural pathway (Yanulevskaya et al., 2013; Martin and von der Heydt, 2013). It also makes sense from an evolutionary, survival perspective — predators in upright or running poses are much more dangerous than when they are lying down or even upside-down. Our approach combines the information to solve the shape–encoding problem using a hypothetical proto-object *shape feature vector* (S) that represents shape from a common-oriented perspective viewpoint.

5.3.6.1 *Shape encoding*

For implementation, our shape representation builds upon the *centroidal profiles* methodology (Davies, 2004), which represents shape boundary distances in a polar coordinate system. For retrieving a shape vector S we first calculate the object's centroid coordinates. This is then assigned position $(0, 0)$. The perimeter of the object is then followed counter-clockwise, from $-180°$ to $180°$. The retrieved perimeter positions are then converted to polar coordinates, yielding a rotation angle $\theta \in [-180°, 180°]$ and a corresponding distance ρ. The retrieved values are then sampled from $[\theta, \theta + 1°]$ in intervals of $1°$, storing, for each interval, the ρ_i^{max} and ρ_i^{min} values inside it. The first describes the outer–shape of the object and the second the inner–

shape. Both are needed to characterise objects and make shapes immune to outliers. We then define S as:

$$S = \left\{ \left[\rho^{max}_{[\theta,\theta+1^\circ]_i} \right]_{\theta=-180^\circ,i=0}^{\theta=180^\circ,i=358} , \left[\rho^{min}_{[\theta,\theta+1^\circ]_i} \right]_{\theta=-180^\circ,i=359}^{\theta=180^\circ,i=718} \right\} , \tag{5.4}$$

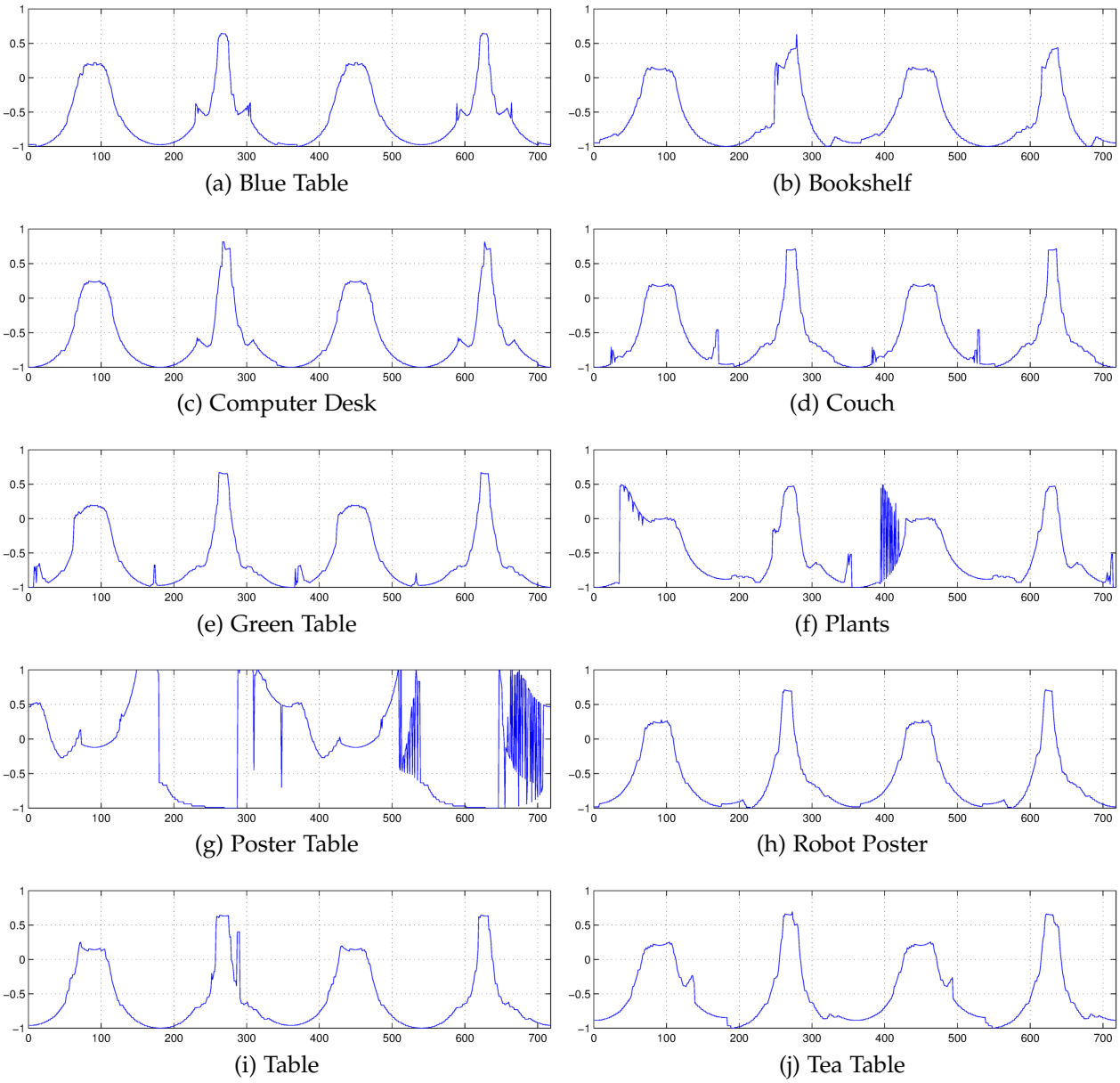
which is a vector of 718 elements (359 from ρ^{max} and 359 from ρ^{min}) describing the contour following of the object's shape. Before yielding the final feature vector values, we normalise S to be invariant to specific object sizes, as

$$S^n = \frac{S - \bar{S}}{\sigma_S} . \tag{5.5}$$

Examples of the shape vectors for the object *beer bottle* retrieved in each background can be seen in Figure 5.5. Notice that the shapes are very consistent, with the exception of the *poster table* background (g) where the object was not properly detected and segregated. The *implicit* encoding of local shape features can be seen if we imagine the derivative of these curves: positive or negative slopes represent *bars/edges* or *curves*, (with a zero slope being a perfect circular curvature) and spikes/signal changes (zero-crossings of the second derivative) represent *corners*.

5.4 OBJECT SHAPE CATEGORISATION FRAMEWORK

This section explains the experimental setup for object categorisation, using the RGB-D Object Dataset (Lai et al., 2011). We chose to use this dataset for its sheer volume of object data — 300 objects with full revolution data, taken at approximately 30°, 45° and 60° above the horizon (a total of 250,000 RGB plus Depth frames).



(a) Blue Table (b) Bookshelf

(c) Computer Desk (d) Couch

(e) Green Table (f) Plants

(g) Poster Table (h) Robot Poster

(i) Table (j) Tea Table

Figure 5.5: "*Beer bottle*" 718-element shape vector, by backgrounds. The first 359 elements of each vector correspond to the outer shape distances ρ^{max} (359 intervals of 1°) and next 359 elements to the inner shape distances ρ^{min}.

5.4.1 *Experimental setup*

The RGB-D Object Dataset has 300 objects in 51 categories. We sampled the turntable data similarly to Lai et al. (2011), using only every 5th video frame, for a total of 41,942 RGB-D images. From these, we extract an equal

number of conspicuity maps and object shape representations, using the methods described before.

For every one of the 51 categories, we left the first object out for testing and used all the remaining for training. Thus, our *training* set, per experimental condition, consists of 34,921 images, while the *test* set counts 7,021 images.

We applied seven experimental conditions on the training and test sets, each condition using a different data type (luminance, colour, conspicuity, disparity, or shape), and combinations of these. Thus, for each condition we measured categorisation performance using: (1) luminance images of the cropped objects; (2) colour images; (3) conspicuty images; (4) disparity images; (5) shape representation; (6) shape representation and disparity images; (7) shape plus grayscale, conspicuity and disparity images.

5.4.2 *Data normalisation*

Prior to classification, there are some pre-processing steps. (a) As RGB-D images are already cropped to the object area, they only need a rescaling step (as each object crop has a different size), so they are all rescaled to 60×60 pixels. (b) For I_i^L, the original RGB image (I_i^C) is reduced to grayscale format. (c) We also normalise all images by subtracting the mean and dividing by the standard deviation, which yields luminance–only I_i^{Ln}, colour I_i^{Cn}, conspicuity $\widehat{C}_i^n$ and disparity D_i^n. Shape vectors S_i^n are already normalised, so they are used with all their 718 components. Data examples are illustrated in Figure 5.6.

5.4.3 *Classifiers*

We applied artificial neural networks (NNs) as classifiers, which have a biological background. By using NNs we can: (a) prove that the framework can

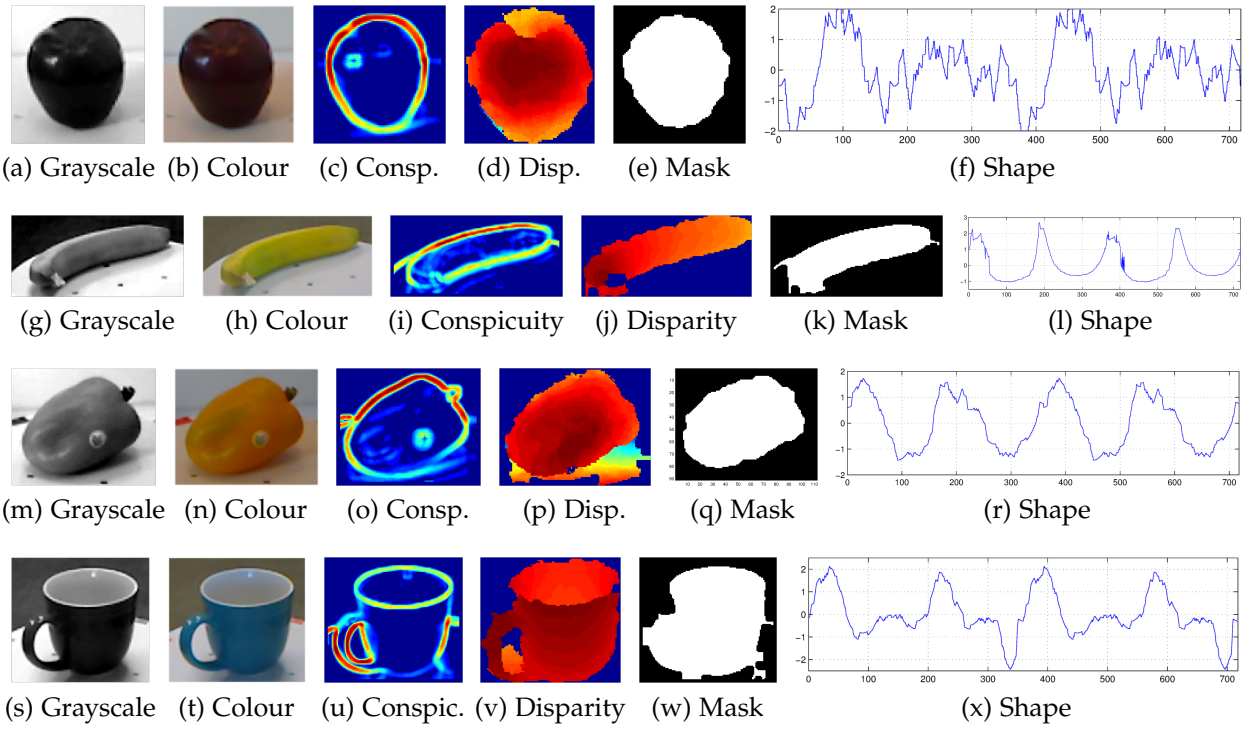


Figure 5.6: Examples of RGB-D OD objects used for classification trials, with respective data types. *Top to bottom rows:* an apple, banana, bellpepper and cup.

be completely implemented by applying biological principles, still obtaining very good performance and (b) show that shape information provides robust features for object categorisation.

5.4.3.1 *Neural Network*

We used a two-layer, feed-forward neural network with log-sigmoid hidden and output neurons (Moller, 1993). The hidden layer was composed of 2000 neurons, chosen after empirical testing. The network was trained to classify the 51 object types of the training set in the RGB-D OD (see Figure 5.7). Training was done iteratively by *resilient backpropagation*, with as stopping criterion a maximum of 1000 epochs. The performance criterion was the *mean-squared normalised error* with a regularisation factor of 20% to avoid over-fitting. This means that the NNs converged gradually towards the final solution, with the outputs of the networks not being binary (the output of

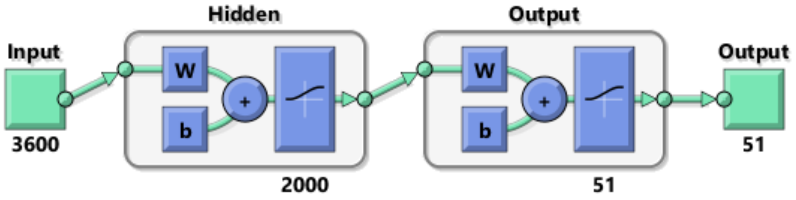


Figure 5.7: Neural Network topology used for RGB-D OD classification. Two-layer, feed-forward neural network with 2000–neuron log-sigmoid hidden layer and 51–neuron log-sigmoid output layer. The input vector is either of size 718 for shape vectors, 3600 (60×60 px) for image vectors or the sum of the sizes for multi-modalities.

a given object category is maximum and those of all other categories are smaller but not necessarily zero).

5.4.4 *Classification rules*

Generally, an object classification system operates in either *verification* mode or in *identification* mode. In verification, the goal is to accept or reject an identity claim that is being presented to the system. In identification there is no identity claim; the system must find the object class which best matches the input object. In both cases there will be true and false positives and negatives, and the goal is to minimise the false ones.

5.4.4.1 *Verification mode*

As mentioned above, this mode serves to verify the identity of an object whose representation is being presented to the system. The output feature vector $\mathbf{v}$ is a (trained) function of the actual input $\mathbf{u}$ — the latter consists of either I^{Ln}, I^{Cn}, $\widehat{C}^n$ or D^n, each with 60×60 pixels (3600 elements), S^n with 718 elements, $\{S^n, D^n\}$ with 4318 elements or $\{S^n, D^n, \widehat{C}^n, I^{Ln}\}$ with 11,518 elements. In the NN case, $\mathbf{v}$ is the output of the network with dimension N, the number of known object classes (so $\mathbf{v}$ has 51 elements for RGB-D OD).

Feature vector $\mathbf{v}$ must be compared with the stored feature vector $\mathbf{v}_c$ if the claimed identity is U_c, with $c \in \{1, 2, \ldots N\}$. Vectors $\mathbf{v}_c$ store the NN classification responses for each of the template input vector types $\mathbf{u}_c$ $\left(I_c^{Ln},\right.$

I_c^{Cn}, $\widehat{C}_c^n$, D_c^n, S_c^n, $\{S_c^n, D_c^n\}$ or $\{S_c^n, D_c^n, \widehat{C}_c^n, I_c^{Ln}\}$) in the database. The result R of the comparison is binary: either accept (1) or reject (0) the identity claim, considering a predefined acceptance threshold Δ. This is formulated as (Štruc and Pavešić, 2010):

$$R = \begin{cases} 1, \text{ if } \xi(\mathbf{v}, \mathbf{v}_c) \geqslant \Delta \\ 0, \text{ otherwise,} \end{cases} \tag{5.6}$$

where $\xi(\cdot, \cdot)$ is a similarity function, for which we use the cosine similarity measure (Štruc and Pavešić, 2010),

$$\xi_{cos}(\mathbf{v}, \mathbf{v}_c) = -\frac{\mathbf{v}^T \mathbf{v}_c}{\|\mathbf{v}\| \|\mathbf{v}_c\|}. \tag{5.7}$$

5.4.4.2 *Identification mode*

A system running in this mode tries to identify an object by finding the feature template (the object class) in the database which best matches the input feature vector (Štruc and Pavešić, 2010):

$$U = \begin{cases} U_c, \text{ if } \xi(\mathbf{v}, \mathbf{v}_c) = \max_{k=1}^{N} \xi(\mathbf{v}, \mathbf{v}_k) \geqslant \Delta \\ U_{N+1}, \text{ otherwise,} \end{cases} \tag{5.8}$$

where $\xi(\cdot, \cdot)$ is ξ_{cos}. Now U_{N+1} is the case when the input vector $\mathbf{v}$ cannot be matched with any of all N objects in the database. In all cases considered, the template feature vector $\mathbf{v}_c$ of an object class is the mean of the output feature vectors resulting from the training data of the respective class, i. e., obtained after training using the I_c^{Ln}, I_c^{Cn}, $\widehat{C}_c^n$, D_c^n, S_c^n, $\{S_c^n, D_c^n\}$ or $\{S_c^n, D_c^n, \widehat{C}_c^n, I_c^{Ln}\}$ maps.

5.5 EXPERIMENTAL RESULTS

5.5.1 *Performance measures*

For presenting performances of the different experimental setups, we measured standard error and recognition rates which are commonly used in object categorisation/recognition (Štruc and Pavešić, 2009, 2010) — remember that *categorisation* is in fact *recognising* an object *class* or *category*.

5.5.1.1 *Cumulative Match Characteristic (CMC)*

In case of class identification experiments we show results in the form of recognition rates. We first computed the Rank One Recognition (ROR) rate on the test set:

$$\text{ROR} = \frac{n_{ca}}{n_{ni}} 100\%, \tag{5.9}$$

where n_{ca} is the number of images correctly assigned to the right object and n_{ni} is the total number of test images. ROR rates were complemented by the ranking beyond the first position, i. e., from rank 1 to rank τ. This yields a CMC curve that plots recognition rate by rank. CMC results are particularly useful for a local-gist system, where the top τ matches can be used for scene categorisation bias. In calculating the recognition rate for the τ-th rank, identification is considered successful if the correct identity is among the top τ results.

5.5.1.2 *Detection Error Trade-off (DET)*

In case of class verification experiments we measured the False Acceptance Rate (FAR) and the False Rejection Rate (FRR), as well as the Half Total Error Rate (HTER). FAR and FRR are defined by

$$\text{FAR} = \frac{n_{ar}}{n_r} 100\%; \ \text{FRR} = \frac{n_{ra}}{n_a} 100\%, \tag{5.10}$$

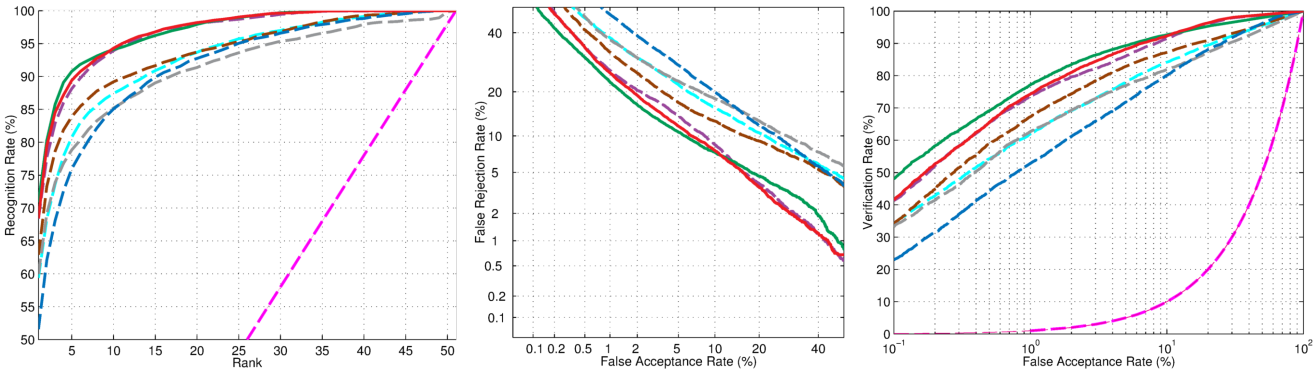
with n_{ar} the number of accepted illegitimate identity claims, n_r the number of all illegitimate identity claims, n_{ra} the number of rejected legitimate identity claims, and n_a the number of all legitimate identity claims. HTER is the average

$$\text{HTER} = (\text{FAR} + \text{FRR})/2. \tag{5.11}$$

Both FAR and FRR depend on the value of the acceptance threshold Δ. When the one decreases, the other increases. To show the effect of Δ on FAR and FRR, the two error rates must be plotted against each other, for all possible values of the acceptance threshold, in the form of detection error trade-off (DET) curves. These relate FAR and FRR for different values of Δ on a scale defined by the inverse of a cumulative Gaussian density function (Štruc and Pavešić, 2010). A DET curve can be summarised by the Equal Error Rate (EER), the point where FAR = FRR, with a lower value representing a more accurate result. Instead of the normal HTER, we will list HTER_{min}, i. e., the point of the DET curve which is closest to the origin, with $\text{HTER}_{\text{min}} \leqslant \text{EER}$. We will also include verification results at two FAR rates: from a moderate $\text{FAR}_{1\%}$ to a more stringent $\text{FAR}_{0.1\%}$.

5.5.1.3 *Receiver Operating Characteristic (ROC)*

In case of verification experiments, we also present results in terms of ROC curves that show the true acceptance rate, also known as verification rate or sensitivity (%), for a range of increasing FAR values (meaning decreasing specificity). Random–guess results will be plotted as dashed magenta lines. ROC curves are often summarised by the area under the curve: Area Under ROC (AUC). A larger AUC implies a better result. For example, even with a stringent false acceptance rate, the verification rate (the number of true positives) should still be high.



(a) CMC (b) DET (c) ROC

Figure 5.8: CMC, DET and ROC performance curves for the 51 object classes of RGB-D OD. Single modality results are given by dashed lines and joint results by solid lines. Shape results in blue, Disparity results in purple, Conspicuity results in brown, Luminance results in grey, Colour results in cyan, Shape + Disparity results in red and Shape + Disparity + Conspicuity + Luminance results in green. CMC and ROC random–guess rates are dashed in magenta.

Table 5.1: NN RGB-D OD Performance Results (51 Object Classes)

	Identification	**Verification**				
Data type	$ROR_{\%}$	$EER_{\%}$	$AUC_{\%}$	$HTER_{min\%}$	$FAR_{1\%}$	$FAR_{0.1\%}$
Shape	51.57	15.12	92.41	14.82	52.74	23.33
Disparity	68.38	9.15	96.71	9.04	73.49	41.19
Conspicuity	62.91	11.76	93.74	10.83	67.26	34.00
Luminance	59.99	15.14	91.63	13.78	62.71	33.16
Colour	59.38	13.56	92.68	12.79	61.93	34.14
Shape + Disparity	68.45	8.60	**96.98**	8.22	74.29	41.50
Shape + Disp + Consp + Lum	**70.89**	**8.11**	96.21	**7.83**	**77.13**	**48.27**

Note: best results in each column are printed **bold**.

5.5.2 *Performance assessments*

Performance curves and data for the experimental setups are shown in Figure 5.8 and in Table 5.1. Different curves are used for the seven experimental conditions: using only S^n (Shape) vectors (as blue dashed lines), using only D^n (Disparity) maps (as purple dashed lines), using only $\widehat{C}^n$ (Conspicuity) maps (as brown dashed lines), using only I^{Ln} (Luminance) images (as grey dashed lines), using only I^{Cn} (Colour) images (as cyan dashed lines) and

using a combination of either $\{S^n, D^n\}$ (Shape + Disparity, as red solid lines) or a combination of $\{S^n, D^n, \widehat{C}^n, I^{Ln}\}$ (Shape + Disparity + Conspicuity + Luminance, as green solid lines). The CMC and ROC random–guess rates are represented by dashed magenta lines.

In summary, identification results of the object class (categorisation) based on Shape alone are able to achieve a ROR of 51.57%, quickly rising to a rank–5 recognition rate around 77% which is indeed promising for a proto-object categorisation system. However, verification rates are lower than in the other trials, showing that shape alone is not discriminative for class verification purposes at low error rates, but it still achieves around 80% at $FAR_{10\%}$ error. Overall, shape results are impressive considering the shape vector is less than 20% smaller than the other feature spaces.

Interestingly, Disparity–only results are in the top 3 results in both class identification and verification (see Table 5.1). The overall best performance is obtained when combining the 4 modalities, although the shape plus disparity results are very close.

Colour achieved slightly better CMC results than Luminance on class identification (except for the ROR rate) and both achieved similar rates on verification. Conspicuity was overall able to achieve better results than both, with a bigger margin on verification trials.

We compare our categorisation results with other authors in Table 5.2. Lai et al. (2011) used a state-of-the-art computer vision algorithm for *Shape* categorisation, based on features extracted from the 3D location of each depth pixel, relying on spin–images that capture the spatial distribution of a randomly sampled set of 3D points, expressing it into a 16×16 histogram. These are used to compute efficient match kernel (EMK) features using random Fourier sets and principal component analysis (PCA) to arrive at a 2703-dimensional shape descriptor (3.76 times larger than ours). For *Vision* categorisation they used SIFT descriptors on a dense grid of 8×8 cells with EMK features at two scales, followed by PCA, achieving a 1500-dimensional vector to be used along with texton histograms (from oriented Gaussian

Table 5.2: RGB-D OD Categorisation ROR% Rate Comparison

Classifier	Number of Features for Shape/Vision	Shape	Vision	All
Linear SVM (Lai et al., 2011)	Spin Images, efficient match kernel (EMK), random Fourier sets, width, depth, height / SIFT, EMK, texton histogram, colour histogram.	53.1	74.3	81.9
kSVM (Lai et al., 2011)	(same)	64.7	74.5	83.8
Random Forest (Lai et al., 2011)	(same)	66.8	74.7	79.6
SVM (Bo et al., 2011)	3D shape, physical size of the object, depth edges, gradients, kernel PCA, local binary patterns, multiple depth kernels.	78.8	77.7	86.2
CKM (Blum et al., 2012)	SURF interest points	—	—	86.4
SP+HMP (Bo et al., 2013)	Surface normals	81.2	82.4	87.5
CNN-RNN (Socher et al., 2012)	ZCA whitening, softmax classifier	78.9	80.8	86.8
Proto-NN	Shape vector only, Disparity only, or Shape + Disparity respectively.	51.57	68.38	68.45

filter responses), a colour histogram and a mean/standard deviation of each colour channel, as visual features.

Lai et al. results can be seen in Table 5.2, along with our results from the 718-dimensional shape vector, 3600-dimensional disparity vector and 4318-dimensional shape+disparity vector. Since Lai et al. only provided ROR data, we can not compare the evolution in classifier performance (neither CMC nor DET/ROC rates). We also compare our categorisation results in Table 5.2 with Bo et al. (2011); Blum et al. (2012); Bo et al. (2013); Socher et al. (2012).

5.6 DISCUSSION AND CONCLUSIONS

According to the attentional *Coherence Theory* (Rensink, 2000), low-level proto-object shapes are continually formed, rapidly and in parallel across the visual field — they are volatile, lacking strong coherence until being stabilised by FoA-gaze, afterwards dissolving when FoA is released. In our

model, we postulate that available disparity and conspicuity information, when combined, is able to quickly highlight all important objects and to resolve border ownership of the outlines of objects. Higher level, oriented, grouping cells can encode the distance from the centre of the object to its border, for very specific orientations. In our case, we implemented two populations of 359 of these cells, with orientations separated by 1 degree (a total of 718 cells). The reason for using two populations for a seemingly similar task is that borders are seldom unique: complex shapes can have several border transitions at certain orientations (i. e., they are not star-shaped). To resolve this problem, one of the (359 cell) populations encodes the distance to the first, closest border near the centre of the object, whereas the second population encodes the farthest border. This allowed for significant resilience in categorising complex shapes. Both populations served as inputs to a simple, feed-forward *proto-shape classifier*, which can reside in the LIP, dorsal pathway (Konen and Kastner, 2008; Janssen et al., 2008). LIP shows shape activation times of 62 ms (Lehky and Sereno, 2007), well within global gist recognition times and also has access (via areas MT and MST), to the Superior Colliculus for eye and head control (Gottlieb, 2007), crucial for FoA, making it a prime candidate area for integrating low-level attention with a proto-object categorisation role.

The implicit shape coding used can be seen as a natural evolution of the explicit coding done in Martins et al. (2012) (Chapter 2) — which is done at a much lower level, possibly using non-standard retinal ganglion cells (Masland and Martin, 2007) — while the current encoding is expected in V1/V2 with decoding done in the dorsal pathway towards the LIP.

Obtained results clearly emphasise the role of shape and 3D information in object categorisation, more than the obvious benefit of both being invariant to lighting conditions. Both represent a strong structural representation, enough to classify objects with good accuracy, quickly surpassing a rank–5 recognition rate of 76% using shape and 88% when using LCV-DEM data, perhaps more than enough for bootstrapping gist vision, since this cate-

gorisation can be done by the visual system in parallel streams — quickly classifying just a few familiar objects in each scene is enough for hugely biasing scene recognition.

Our results also indicate that the proposed proto-object shape categorisation method is almost able to compete with much more computationally complex methods, e. g., based on state-of-the-art spin images with EMK features (Lai et al., 2011). Perhaps even more important is that a system which employs cortical neuronal processes, i. e., which can be thought of as mimicking part of our visual system, can be applied to a real-world problem, and it can compete with advanced methods in computer vision.

Finally, we note that categorisation performance of conspicuity features was also able to beat both luminance and colour data, highlighting its discriminative capabilities. In further research it makes sense to expand categorisation with other kinds of low-level features, such as lines/edges and keypoints that are readily available from simple, complex and end-stopped cells in V1/V2 (Rodrigues and du Buf, 2009a).

6

FACE RECOGNITION

EXPRESSION-INVARIANT FACE RECOGNITION USING A BIOLOGICAL DISPARITY ENERGY MODEL

ABSTRACT: Face recognition is a research area with significant challenges, especially in case of varying lighting, occlusions, different facial expressions, ageing, and even identity spoofing. Such factors make normal images less reliable for biometric identity recognition, requiring to complement or substitute them by other biometric information. In this chapter we present and explore a biological model of stereo vision: a Disparity Energy Model. We show that it provides precise 3D disparity maps which are suitable for identity recognition and verification. We test disparity information, both alone and in combination with image data, yielding state-of-the-art results. We also compare results with those obtained by precise laser range maps.

KEYWORDS: Stereo vision, visual cortex, disparity, population coding, face recognition, verification, neural network, LDA, PCA.

6.1 INTRODUCTION

Human faces are among our most important visual stimuli, as they are paramount to socialisation. Extensive neuropsychological research has shown that we can obtain a lot of information from faces: gender, age, race, emotional state, direction of gaze and even physical health (Bruce, 1990). Neural systems involved in face recognition become active very early in life. In in-

fancy, faces provide non-verbal information which is essential for communication and survival (Hess and Thibault, 2009). During the first six months, infants quickly develop the capacity for detecting and recognising faces. Newborns already show a visual preference for faces and the capacity of prototyping them very rapidly (Goren et al., 1975; Walton and Bower, 1993). At about four months, infants are able to discriminate upright faces from upside-down ones, and at six months they show different brain potentials in case of familiar *vs.* unfamiliar faces (Fagan, 1972; Haan and Nelson, 1997). Hence, faces provide the most important biometric cues (Ives et al., 2005).

In computer vision, face processing consists of two tasks: (1) detecting faces in all types of scenes, and (2) recognising the persons associated with the detected faces. Significant problems are involved in both tasks, ranging from partial occlusions to dealing with different facial expressions, even extreme ones. These are huge hurdles for face recognition technology (Franco and Nanni, 2009). Therefore, a robust face recognition system should employ techniques which give reliable results, regardless of any differences in acquired images. Most face recognition methods rely heavily on image processing techniques which are normally not related to models of cortical processing. Siagian and Itti (2004) proposed a biologically inspired face *detection* model based on saliency, gist and gaze data, to sequentially detect image regions containing faces. For a general survey on face recognition approaches we refer to Li and Jain (2011).

One of the more difficult problems of face recognition is to deal with facial expressions, where 3D structural information can help immensely. In addition, 3D face recognition is attracting more research effort as 3D data is less affected by pose, illumination, scaling and even person ageing (Scheenstra et al., 2005). Disparity information from stereo is mainly used for face *detection*, since it has several advantages for that. In case of *recognition*, several studies reported significant improvements when combining luminance/colour data with disparity.

Kosov et al. (2009) applied stereo information to suppress false negatives in a real-time face detection system. Later they applied eigenface recognition on a custom dataset of 34 persons, using both luminance and disparity information (Kosov et al., 2010). They tested robustness to additive Gaussian noise, at different noise levels, with leave-one-out cross validation. Disparity data improved recognition rates by an average of 7.7% over all noise levels.

Tsalakanidou et al. (2003) combined colour with disparity information to increase recognition rates by 10%, from 87.5 to 97.5%, on the XM2VTS database (Messer et al., 1999). Sun et al. (2007) also improved the performance of a 2D eigenface recognition system by employing disparity data. They trained the system on neutral expressions, and then tested it on smiling expressions (case A), on other expressions (case B), and on smiling *and* other expressions (case C). On a custom 100-person dataset, the Rank One Recognition (ROR) rate in case A improved from 94 to 97%. In cases B and C, ROR rates increased from 89 to 91% and from 91.5 to 94%, respectively. When only employing disparity data, they achieved ROR rates of 73% (case A), 61% (case B) and 67% (case C).

Many 3D methods exploit Gaussian curvature or apply morphing approaches. For a survey on 3D face recognition we refer to Scheenstra et al. (2005). Hayasaka et al. (2009) proposed a 3D face recognition framework, based on stereo vision, which employs iterative closest point analysis. It was tested on a dataset of 15 persons with five expressions (neutral, smiling, angry, surprised and sad). Measured Receiver Operating Characteristic (ROC) curves showed different EERs (equal error rates, where the false acceptance rate equals the false rejection rate; definitions are given in Section 6.4.1) for different facial regions: only nose 8%, eyes and nose 10%, only mouth 34%, eyes plus nose plus mouth 16%, and entire faces 20%; a lower EER means a better performance. Although the nose region may be least affected by different expressions, it has been found that the regions around the eyes and nose are the most informative ones for person recognition (Tistarelli

et al., 2007). More quantitative results are given in Section 6.4 (*Experimental Results*), where we compare our own recognition rates with those obtained by others who used the same data sets as we did.

In this chapter, our main contribution is to show that structural face information from a biological disparity model can be employed to recognise persons with different facial expressions with good accuracy, in fact yielding state-of-the-art recognition rates. Our disparity method, which is based on trained, binocular Disparity-Energy Model (DEM) cells (Martins et al., 2011b), is paired with an also-trained neural-network classifier. Hence, our framework can be seen as being completely biologically plausible, providing a baseline for future research into biological methods. To the best of our knowledge, the biological DEM model has never been applied to real faces. We also compare disparity-based recognition rates with results obtained by employing a 3D range scanner, i. e., "classical" stereo vision *vs.* more advanced technology.

One of the biggest problems when comparing results is the lack of completeness in existing literature. For example, using the Binghamton University 3D Facial Expression database, which is described below, some studies applied reduced subsets of all represented persons (Mpiperis et al., 2008) or considered only specific expressions (Venkatesh et al., 2012). Most studies presented identity recognition results — often only ROR rates instead of cumulative matching curves — without verification results (Kaushik et al., 2009). Because of this, we also complement existing research by presenting standard performance curves for different combinations of data (luminance, disparity and both) and three different classifiers, all on both identity *recognition* and *verification*, using two common and large databases.

The rest of this chapter is organised as follows. Section 6.2 deals with the face recognition setup. In Section 6.3 we explain the face recognition framework, the DEM implementation, the classifiers and classification rules. In Section 6.4 we present experimental results, and in Section 6.5 the discussion and conclusions.

6.2 FACE RECOGNITION SETUP

We will consider two databases: (a) The Binghamton University 3D Facial Expression database (BU-3DFE) (Yin et al., 2006) is used to show all steps of the biological approach. This database contains dense triangle meshes constructed by using six cameras and structured light projection. We use the 3D meshes to create stereo views for disparity estimation. (b) The Texas 3D Face Recognition database (Texas-3DFR) (Gupta et al., 2010) is used to compare disparity-based performances with those obtained by using other depth maps. Since Texas-3DFR was created by using a laser range finder, it allows us to establish a baseline for expected performance of a system which employs more precise depth information.

Examples of face data are shown in Figure 6.1. BU-3DFE (Figure 6.1:a–g) contains textured face meshes with different expressions (happiness, disgust, fear, anger, surprise, sadness and neutral) of each person. These meshes are converted to stereo image pairs from which disparity maps are created, and these maps can be compared with the input meshes to check the quality of the disparity algorithm. Texas-3DFR (Figure 6.1:h–o) contains images with corresponding range maps, of each person with different expressions and in different lighting conditions.

6.2.1 *Binghamton University 3D Facial Expression database*

BU-3DFE (Yin et al., 2006) comprises 3D mesh and image data of 100 persons, 56 female and 44 male, from 18 to 70 years old, and from seven ethnicities. Each person is represented with six expressions (happiness, disgust, fear, anger, surprise and sadness) with four levels of intensity, from mild to prominent, plus a neutral one. Hence, there are 25 3D models of each person in VRML format, a total of 2500 models. An example of a mesh

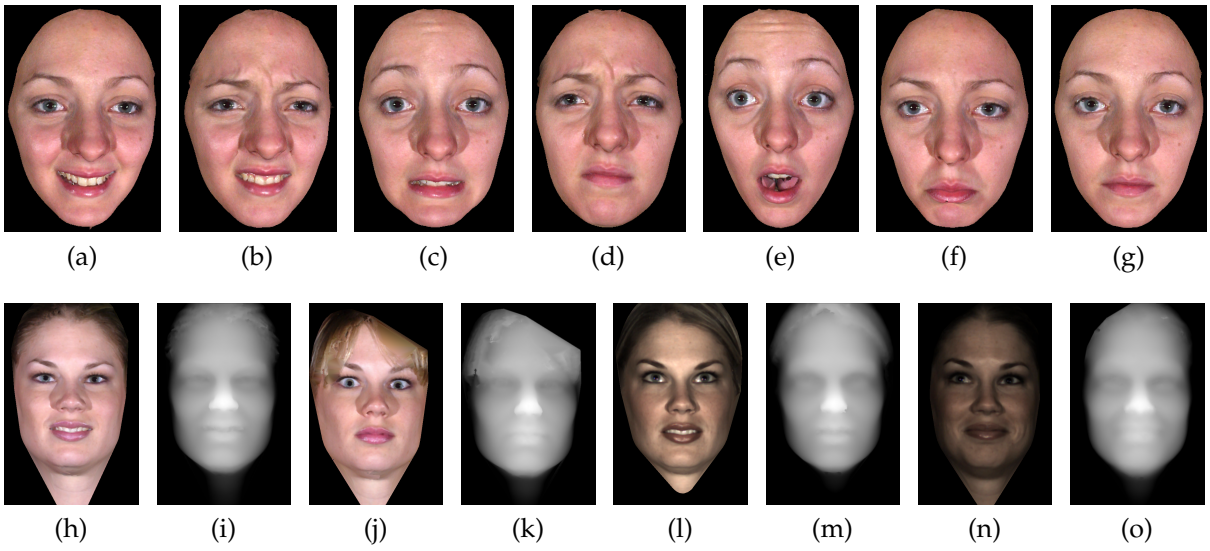


Figure 6.1: Examples of faces. **(a–g):** BU-3DFE database, frontal views of textured face meshes with different expressions (happiness, disgust, fear, angry, surprise, sadness and neutral) for the same person. **(h–o):** Texas-3DFR database, four images with corresponding range maps of the same person, with different expressions and lighting conditions.

can be seen in Figure 6.2(a) with the corresponding texture map (image) in Figure 6.2(b).

For each of the 25 mesh models of the 100 persons ($i \in \{1, ..., 2500\}$) we created a stereogram with two views, separated by 0.2° relative to the main vertical axis, $(I_i^L, I_i^R) = (-0.1^\circ, +0.1^\circ)$, simulating the perspective views of the left and right cameras; see Figure 6.2:(c–d). The views were created by using OpenGL, combining the 3D mesh with texture (image) mapping, with perspective projection, oblique diffuse lighting and Gouraud shading. The two images serve as input for the stereo disparity model. Resulting disparity maps D_i (see Section 6.3) are aligned with I_i^L (Figure 6.2f) and used for face recognition after normalisation, as explained in Section 6.3.2. We also stored OpenGL's Z-buffer data (Figure 6.2e) in order to compare the input depth maps with estimated disparity maps.

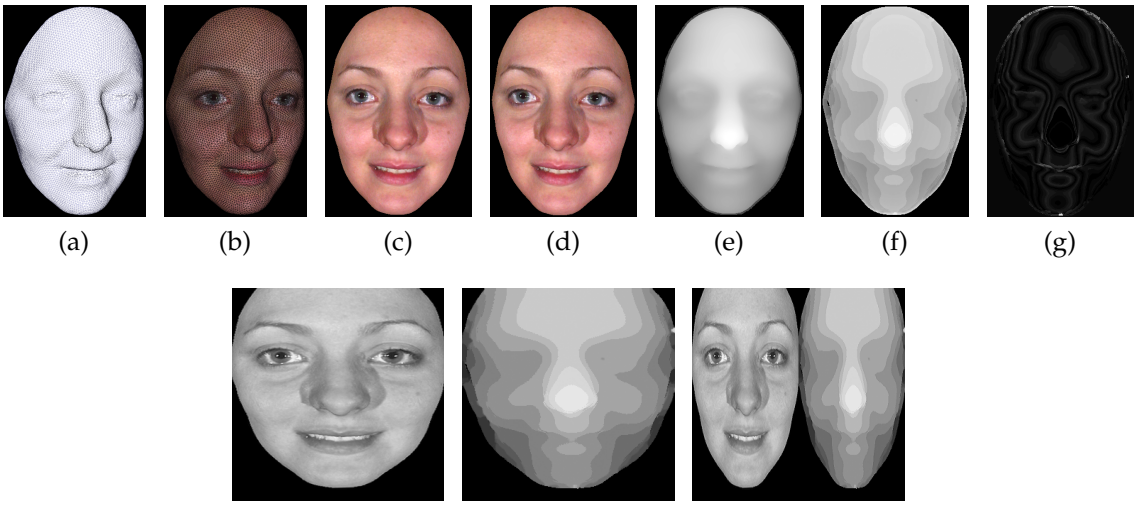
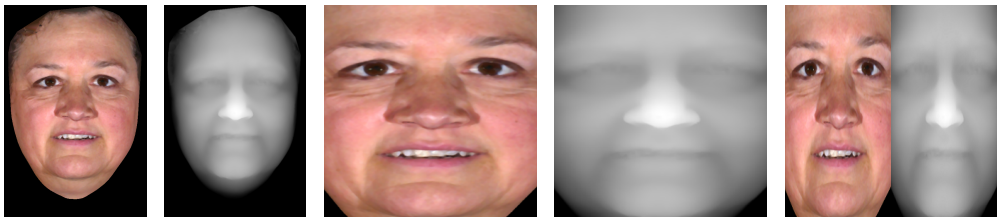


Figure 6.2: One BU-3DFE face. **(a–b):** 3D wireframe (mesh) and texture map. **(c–d):** Left and right images of the stereogram extracted from the mesh. **(e):** Z-buffer depth map. **(f):** L-DEM disparity map D. **(g):** Difference between Z-buffer and L-DEM maps. **(h–i):** Normalised image and disparity map (left I^n and D^n; 300×300 pixels). **(j):** Joint image and disparity map (J^n; 300×300 pixels).

6.2.2 *Texas 3D Face Recognition database*

Texas-3DFR (Gupta et al., 2010) consists of 1149 sets of colour images and grayscale depth maps, pixelwise matched, of 118 persons, from 22 to 75 years old ($i \in \{1, ..., 1149\}$). It contains both men and women from five ethnic groups. Each person is represented by a number of sets which varies from 1 to 89 (if 1, the person can be included in a training or test set for checking false positives or negatives). Each set has 25 manually-annotated anthropometric facial fiducial points. The faces are neutral (emotionless) or expressive: smiling or talking, with open or closed mouth and/or eyes. As already mentioned, we use the precise range data (R_i) to establish a baseline for person recognition when more precise disparity data becomes available, either from a better diparity model or from another source. Figure 6.3 shows a Texas-3DFR example, i. e., the image (a) and corresponding range map (b) of person 001.



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

Figure 6.3: One Texas-3DFR face. **(a–b):** Person 001's image and corresponding range map R. **(c–d):** Normalised image and range map (I^n and R^n; 300×300 pixels). **(e):** Joint luminance and range map (J^n; 300×300 pixels).

6.2.3 *Experimental setup*

We applied three experimental conditions, each one using a different data type (luminance, disparity, or both) for the training and test sets. Thus, for each condition we measured recognition performance (a) using only grayscale images of the left stereogram (I^L), (b) using only disparity maps (D^L), and (c) the joint data (J^L), where luminance and disparity are horizontally reduced by 50% to keep the same number of pixels as in I and D; see Figure 6.2(j) and Figure 6.3(e).

The BU-3DFE database has 25 expressions per person, which are labelled from lowest to highest intensity per expression as #1-4 anger, #5-8 disgust, #9-12 fear, #13-16 happiness, #17 neutral, #18–21 sadness, and #22–25 surprise. For the *training* set we selected two images per person: the neutral expression (#17) and the happy one with the lowest intensity (#13), which was generally a faint smile. These were selected to reflect normal conditions when building a recognition system, i. e., persons are typically photographed with a neutral or slightly expressive face.

For *testing* purposes, we defined three different test sets: (1) a NEUTRAL set composed of all remaining 11 expressions with intensity values one and two (#1, #2, #5, #6, #9, #10, #14, #18, #19, #22, #23), (2) an EXPRESSIVE set with the 12 highest intensities per expression, i. e., intensity values three

and four (#3, #4, #7, #8, #11, #12, #15, #16, #20, #21, #24, #25), and (3) the ALL set which combines all 23 expressions of the first two sets.

In the case of Texas-3DFR, and for comparing results, we used the same sets as Gupta et al. (2010). The training set (called *gallery* set in Gupta et al., 2010) contains one image and/or range map of 105 persons with neutral expressions. The test set (called *probe* set in Gupta et al., 2010) contains 663 images and/or range maps of 95 persons with neutral and arbitrary expressions. The number of images (and maps) of each person varies from 1 to 55. As in Gupta et al. (2010), the entire ALL test set is also subdivided into NEUTRAL and EXPRESSIVE sets.

6.3 FACE RECOGNITION FRAMEWORK

In this section we describe our biological disparity model, the classifiers, and the classification rules in case of identification and verification experiments. We focus on one biological disparity model, the energy model, although there exist alternative models. One alternative model, which employs the phase difference of the responses of complex Gabor filters to the left and right views, is often applied to real-world problems, although it has been shown to be very imprecise in terms of localisation of depth transitions (du Buf et al., 2013). The energy model has, to the best of our knowledge, only been applied in theoretical studies concerning visual perception, but never to a real-world problem like face recognition.

Cortical area V1 comprises simple, complex and end-stopped cells, but the latter are not employed here. Receptive fields (RFs) of monocular simple cells can be modelled by Gabor filter kernels (Chen and Qian, 2004; Rodrigues and du Buf, 2009a; Martins et al., 2011b). Their parameters specify the preferred orientation θ, spatial frequency f, receptive field size σ and spatial phase ϕ. Binocular cells are based on pairs of simple cells with RFs at different positions, such that they can signal disparity if a same but shifted pattern is present in the RFs. However, binocular simple cells do not

reliably signal disparity because they are also sensitive to the contrast of the pattern in their fields: disparity-tuning curves of simple cells as measured with bright and dark bars, which have different Fourier phases, are very different (Ohzawa et al., 1997). Any change of a pattern other than an amplitude scaling (average brightness and contrast) alters the Fourier phase, which in turn affects disparity tuning (see Chen and Qian, 2004; Martins et al., 2011b).

We use a Luminance Disparity-Energy Model (L-DEM) to estimate local disparities. We apply two neuronal populations: (1) An *encoding population* that consists of a set of neurons tuned to a wide range of parameters such as horizontal disparities, spatial frequencies and orientations. This population is trained on random-dot stereograms in order to learn the population codes for many different disparities. It is also applied to real face stereograms in order to obtain the local population codes. We use an encoding method similar to that of Read (2010), which is based on the DEM model of Ohzawa et al. (1997), with proper normalisation to yield local correlations with neighbourhood weighting (Read and Cumming, 2006; Banks et al., 2004; Filippini and Banks, 2009). (2) A higher-level *decoding population* compares the local population code, at each image position, with all learned (trained) population codes, for estimating local disparity. Basically, this second population implements a template-matching process similar to those of Tsai and Victor (2003); Read (2010). The estimated local disparities are then used to evaluate face recognition on the BU-3DFE database. For having a completely biological framework we will, apart from other classifiers, also consider a third population: a trained neural-network classifier.

6.3.1 *Disparity Energy Model*

We used the Luminance Disparity-Energy Model (L-DEM) implementation for extracting disparity maps for all stereogram face pairs in the BU-3DFE

database. For a detailed explanation, we refer the reader to Section 4.3, where it is explained in detail.

For validation purposes, we also extracted 500 depth maps (Z-buffer images) from the 3D meshes of the BU-3DFE dataset: the first 10 women and the first 10 men. These were compared with the corresponding L-DEM disparity maps. The average correlation was 98.7%. An example Z-buffer image is shown in Figure 6.2(e), with the corresponding L-DEM map and difference image in Figure 6.2:(f–g). Taking into account that the dense 3D meshes are accurate and that the simulated stereo views are therefore also accurate (Gouraud shading of dense meshes), we may conclude that the accuracy of the L-DEM model is very good. Of course, the L-DEM model is less accurate at depth discontinuities, but this mainly occurs at face contours; see Figure 6.2(g).

6.3.2 *Data normalisation*

Prior to classification, face images I_i^L and disparity maps D_i^L must be normalised. Since we are focusing on face recognition and not on face detection, all images and maps must be properly aligned and localised (for details see Viola and Jones, 2004).

In case of BU-3DFE, we used the centre eye coordinates of I_i^L. These were hand-picked for all 2500 stereograms. Then I_i^L was rotated and scaled such that the centres of the eyes are located at predefined pixel positions: the interocular distance d_i is computed and the midpoint between the eyes is determined. After rotation, the line connecting the eyes is horizontal. After scaling, the top of the face region is $0.8d_i$ higher and the bottom is $1.75d_i$ lower than the midpoint; the left and right boundaries are at $0.9d_i$ (Štruc and Pavešić, 2010). Colour images are converted to grayscale format. All images and disparity maps were initially cropped to a size of 300×300 pixels (but see below), and normalised by subtracting the mean and dividing

by the standard deviation. This yields I_i^n and D_i^n. Examples are illustrated in Figure 6.2:(h–i).

In case of Texas-3DFR, we derived the centre eye coordinates from the fiducial points, see Gupta et al. (2010). The faces were then rotated and scaled as was done in case of BU-3DFE. This yields I_i^n and R_i^n. For an example see Figure 6.3:(c–d). Below, for simplicity the range map R is also denoted by D.

The normalised images and disparity/range maps are still feature vectors with a large dimension (300×300 pixels). This makes matching costly, so we further reduced the size. Empirical results showed that we can average 6×6 pixel blocks to reduce the size to 50×50 pixels without losing classification performance.

6.3.3 *Classifiers*

We applied and compared three classifiers. Artificial neural networks (NNs) have a biological background, whereas the others, LDA and PCA, apply linear sub-space projections. All are common in computer vision.

By using these classifiers we can (a) prove that the framework can be completely implemented by applying biological principles, still obtaining state-of-the-art performance, (b) show that stereo information provides robust features for face recognition, and (c) demonstrate that the proposed framework does not depend much on the type of classifier.

6.3.3.1 *Neural Network*

We used a two-layer, feed-forward neural network (NN) with sigmoid hidden and output neurons (Moller, 1993). The hidden layer was composed of 150 neurons, chosen after empirical testing. The network was trained to classify the 100 persons of the training set in case of BU-3DFR (see Figure 6.4), and the 105 persons in case of Texas-3DFR. Training was done iteratively by

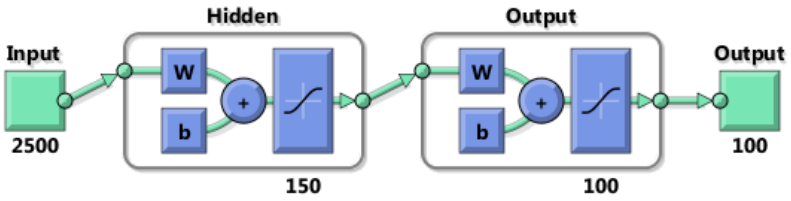


Figure 6.4: Neural Network topology used for face recognition. Two-layer, feed-forward neural network with 150-neuron sigmoid hidden layer and 100-neuron sigmoid output layer, used in case of BU-3DFE. In case of Texas-3DFR the output layer counts 105 sigmoid neurons.

scaled conjugate gradient-descent backpropagation, with as stopping criterion a minimum gradient of 10^{-7}. The performance criterion was the *mean-squared error* with a regularisation factor of 20% to avoid over-fitting. This means that the NNs converged gradually towards the final solution, with the outputs of the networks not being binary (the output of a given person is maximum and those of all other persons are smaller but not necessarily zero).

6.3.3.2 *Sub-space projections*

We applied two linear sub-space projection methods: linear discriminant analysis (LDA) and principal component analysis (PCA) (Delac et al., 2005). LDA creates a subspace where the between-class variations are maximised and the within-class variations are minimised; see the implementation in Štruc and Pavešić (2010). PCA applies an orthogonal transformation to convert a set of possibly correlated variables into a set of linearly uncorrelated variables called principal components. The first principal component has the largest variance, and each successive component has the next largest variance with the constraint that they must be orthogonal to the preceding components. For the implementation see Wechsler (2007). Since we did not experiment with feature reductions, and in order to apply the same classification rules as in the case of neural networks with almost the same number of features (see below), we simply take the $N-1$ most significant features after training PCA and LDA, with N being the number of persons (100 for BU-3DFE and 105 for Texas-3DFR).

6.3.4 *Classification rules*

In verification, the goal is to accept or reject the identity claim that is presented to the system. In identification there is no identity claim; the system must find the person template which best matches the input face. In both cases there will be true and false positives and negatives, and the goal is to minimise the false ones.

6.3.4.1 *Verification mode*

As mentioned above, this mode serves to verify the identity of a person whose face is being presented to the system. The output feature vector $\mathbf{v}$ is a (trained) function of the actual input $\mathbf{u}$ which consists of either I^n, D^n or J^n, each with 50×50 pixels. In the NN case, $\mathbf{v}$ is the output of the network with dimension N, the number of known persons. In case of LDA and PCA, $\mathbf{v}$ is the subspace projection of $\mathbf{u}$ with dimension $N-1$ (for simplicity we took the $N-1$ most discriminative dimensions). Feature vector $\mathbf{v}$ must be compared with feature vector $\mathbf{v}_c$ if the claimed identity is U_c, with $c \in \{1, 2, \dots N\}$. Vector $\mathbf{v}_c$ is also a (trained) function of the template input vector $\mathbf{u}_c$ (I_c^n, D_c^n or J_c^n, each with 50×50 pixels) in the database. The result R of the comparison is binary: either accept (1) or reject (0) the identity claim, considering a predefined acceptance threshold Δ. This is formulated as (Štruc and Pavešić, 2010):

$$R = \begin{cases} 1, \text{ if } \xi(\mathbf{v}, \mathbf{v}_c) \geqslant \Delta \\ 0, \text{ otherwise,} \end{cases} \tag{6.1}$$

where $\xi(\cdot, \cdot)$ is a similarity function. We apply the cosine similarity measure (Štruc and Pavešić, 2010) in case of NN and LDA,

$$\xi_{\cos}(\mathbf{v}, \mathbf{v}_c) = -\frac{\mathbf{v}^T \mathbf{v}_c}{\|\mathbf{v}\| \|\mathbf{v}_c\|}, \tag{6.2}$$

and the Mahalanobis cosine similarity measure in case of PCA,

$$\xi_{m-cos}(\mathbf{v}, \mathbf{v}_c) = -\frac{\mathbf{v}^T C^{-1} \mathbf{v}_c}{\|\mathbf{v}\| \|\mathbf{v}_c\|}, \tag{6.3}$$

with C the covariance matrix of $\mathbf{v}_c$. C is diagonal and its elements are the (eigenvalue) variances of the corresponding components (Wechsler, 2007).

6.3.4.2 *Identification mode*

A system running in this mode tries to identify the person by finding the feature template in the database which best matches the input feature vector (Štruc and Pavešić, 2010):

$$\Omega = \begin{cases} \Omega_c, \text{ if } \xi(\mathbf{v}, \mathbf{v}_c) = \max_{\kappa=1}^{N} \xi(\mathbf{v}, \mathbf{v}_\kappa) \geqslant \Delta \\ \Omega_{N+1}, \text{ otherwise,} \end{cases} \tag{6.4}$$

where $\xi(\cdot, \cdot)$ is ξ_{cos} for NN and LDA, and ξ_{m-cos} for PCA. Now Ω_{N+1} is the case when the input vector $\mathbf{v}$ cannot be matched with any of all N persons in the database. In all cases considered, the template feature vector $\mathbf{v}_c$ of a person is the mean of the feature vectors of the training data of that person, i. e., of the I, D or J maps.

6.4 EXPERIMENTAL RESULTS

6.4.1 *Performance measures*

For presenting performances for the different experimental setups, we measured standard error and recognition rates which are commonly used in face recognition (Štruc and Pavešić, 2009, 2010). These are (see Section 5.5.1):

(a) Cumulative Match Characteristic (CMC) — In case of identification experiments, the Rank One Recognition (ROR) rate and the CMC, i. e., from rank 1 to rank τ with $\tau \leqslant 25$. CMC results are particularly useful

in law enforcement applications, where a trained specialist can closely inspect the top τ matches.

(b) Detection Error Trade-off (DET) — In case of verification experiments we measured the False Acceptance Rate (FAR) and the False Rejection Rate (FRR), as well as the Equal Error Rate (EER) (where FAR =FRR) and Half Total Error Rate (HTER), with HTER being the average of FAR and FRR. Instead of the normal HTER, we will list HTER_{min}, i. e., the point of the DET curve which is closest to the origin, with $\text{HTER}_{\text{min}} \leqslant \text{EER}$. We will also include verification results at different FAR rates: from a moderate $\text{FAR}_{1\%}$ to $\text{FAR}_{0.1\%}$ and to a very stringent $\text{FAR}_{0.01\%}$.

(c) Receiver Operating Characteristic (ROC) — In case of verification experiments, we also present results in terms of ROC curves. These are often summarised by the area under the curve: Area Under ROC (AUC). A larger AUC implies a better result. For example, even with a stringent false acceptance rate, the verification rate (the number of true positives) should still be high.

6.4.2 *Performance assessments*

6.4.2.1 *BU-3DFE results*

Performance curves and data for the experimental setups are shown in Figure 6.5 and in Table 6.1. Different curves are used for the three experimental conditions: using only I^{n} grayscale luminance images of the left stereogram (dash-dot lines), using only D^{n} disparity maps (dotted lines), and using the linear combination J^{n} of both (solid lines). Figure 6.5:(a–c) show NN performance curves, Figure 6.5:(d–f) show LDA results, and Figure 6.5:(g–i) show PCA results. The colours of the curves indicate facial expression sets: NEUTRAL in blue, EXPRESSIVE in red, and ALL in black.

As expected, disparity-only results are still worse than luminance-only results, but best performances were obtained by combining luminance with disparity (Lum+L-DEM). In identification, highest ROR rates (see Table 6.1) were obtained by PCA (NEUTRAL set) and LDA (EXPRESSIVE and ALL sets). However, Lum+L-DEM differences with NN results are very small: less than 0.3% (NEUTRAL), about 2% (EXPRESSIVE), and about 1% (ALL set). In contrast, NN clearly outperformed PCA and LDA in all verification experiments; see Table 6.1 and Figure 6.5:(b–c and e–f). Hence, overall NN classification can compete with the other classifiers.

An important conclusion is that disparity significantly contributes to increase recognition rates of the EXPRESSIVE set (NN Lum *vs.* NN Lum+L-DEM). In identification, the difference between ROR rates on the NEUTRAL set is 2.4%. The ROR difference on the EXPRESSIVE set is 3.5%. However, in verification, the corresponding $FAR_{0.01\%}$ differences are much bigger: about 4% on the NEUTRAL set and a very significant improvement of about 11% on the EXPRESSIVE set. This implies that disparity information is crucial when verifying very expressive faces.

Worst results were consistently achieved by PCA in the L-DEM-only case and on all three test sets, both in identification and in verification. Nevertheless, PCA performed better than LDA at very low FAR rates ($FAR_{0.01\%}$), especially on the NEUTRAL and ALL sets. Although disparity-only results are worst, they are still very promising because of quickly increasing CMC curves in case of NN and LDA; see Figure 6.5:(a and d). At rank–5 all identification rates were better than 92%, even on the EXPRESSIVE test set.

Figure 6.6 shows our CMC curve in case of NN LUM+L-DEM applied to the ALL dataset together with CMC curves of three state-of-the-art methods. Hajati et al. (2012), who used patch geodesic moments to interpolate 2.5D data, achieved a ROR rate of 84.8%. We achieved a ROR rate of 94.13%, and already at rank–6 the 99% line is crossed. Hence, our results are much better than those obtained with patch geodesic moments (Hajati et al., 2012),

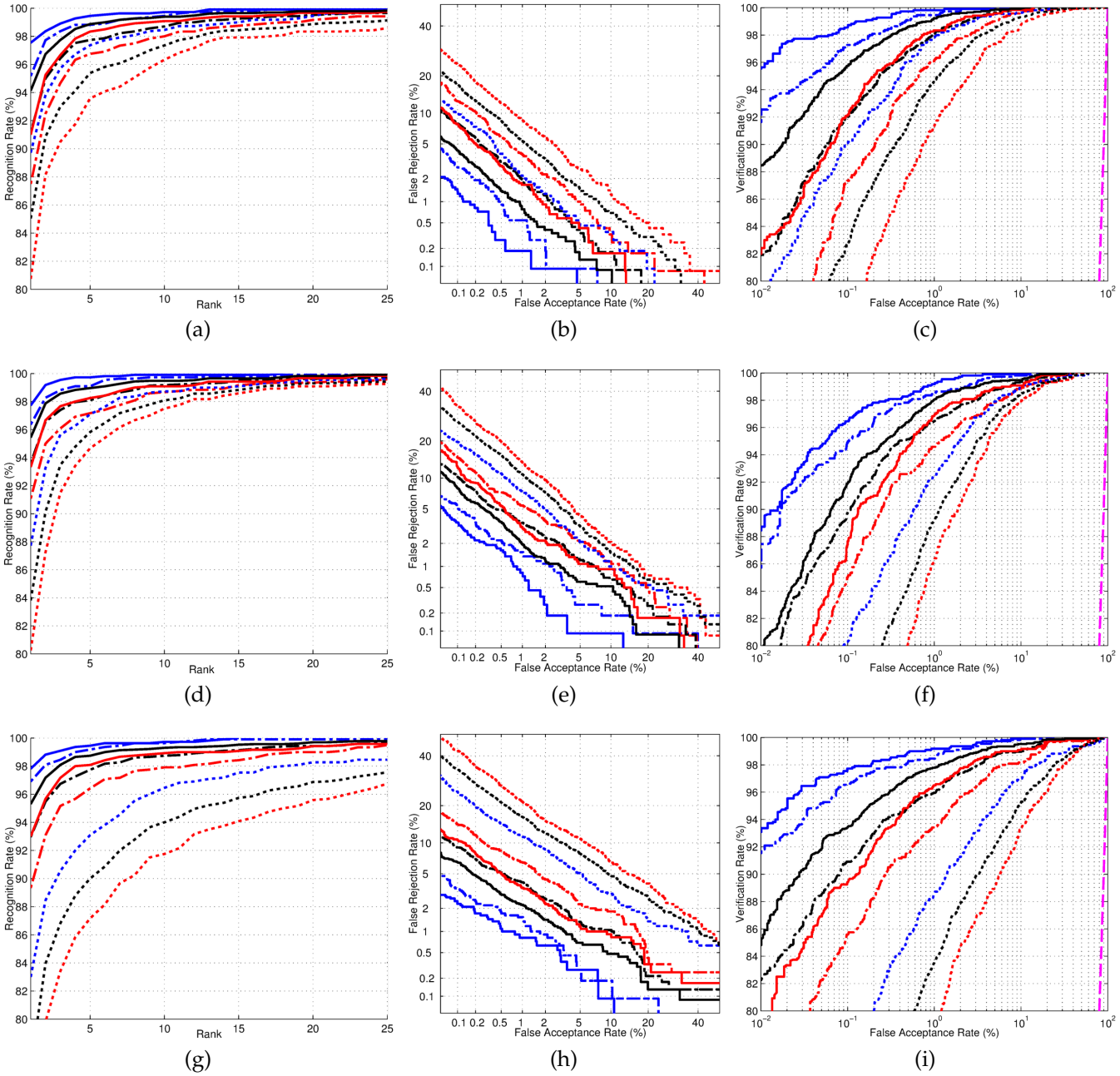


Figure 6.5: CMC (left), DET (middle) and ROC (right) performance curves in case of BU-3DFE. **(a–c):** Neural Network results; **(d–f):** LDA results; **(g–i):** PCA results. Luminance-only results are given by dash-dot lines, disparity-only results by dotted lines, and joint results by solid lines. NEUTRAL set in blue, EXPRESSIVE set in red, and ALL set in black. ROC random guess rates are dashed in magenta.

Table 6.1: BU-3DFE performance results

			Identification	Verification					
Test set	Classifier	Data	$ROR_{\%}$	$EER_{\%}$	$AUC_{\%}$	$HTER_{min\%}$	$FAR_{1\%}$	$FAR_{0.1\%}$	$FAR_{0.01\%}$
Neutral	NN	Lum	95.18	0.63	99.98	0.59	99.45	97.27	91.64
Neutral	NN	L-DEM	89.73	1.54	99.86	1.48	97.82	90.09	77.45
Neutral	NN	Lum+L-DEM	97.55	**0.36**	**99.99**	**0.34**	**99.82**	**98.55**	**95.55**
Neutral	LDA	Lum	96.36	1.36	99.90	1.20	98.45	94.91	85.73
Neutral	LDA	L-DEM	87.73	3.18	99.48	3.08	92.55	80.27	63.45
Neutral	LDA	Lum+L-DEM	97.73	0.91	99.96	0.84	99.18	96.45	88.45
Neutral	PCA	Lum	96.91	1.18	99.93	1.10	98.45	96.55	91.55
Neutral	PCA	L-DEM	83.00	5.01	98.71	4.80	88.64	75.00	54.91
Neutral	PCA	Lum+L-DEM	**97.82**	0.91	99.95	0.76	99.18	97.73	93.09
Expressive	NN	Lum	87.50	2.10	99.77	2.04	96.17	87.33	70.83
Expressive	NN	L-DEM	80.75	3.50	99.41	3.41	91.33	76.00	55.08
Expressive	NN	Lum+L-DEM	91.00	**1.41**	**99.91**	**1.32**	**98.33**	**92.08**	**82.00**
Expressive	LDA	Lum	91.08	3.08	99.58	2.77	94.58	85.00	68.33
Expressive	LDA	L-DEM	80.25	4.50	99.07	4.48	86.25	66.58	47.50
Expressive	LDA	Lum+L-DEM	**93.33**	2.07	99.74	1.87	96.92	87.42	69.58
Expressive	PCA	Lum	89.33	3.34	99.40	3.14	93.42	85.58	71.67
Expressive	PCA	L-DEM	70.75	8.17	97.54	7.81	78.67	57.58	37.92
Expressive	PCA	Lum+L-DEM	93.00	2.18	99.66	2.12	96.50	89.42	76.42
All	NN	Lum	91.17	1.44	99.89	1.36	98.04	91.96	81.91
All	NN	L-DEM	85.04	2.62	99.65	2.58	94.65	82.78	65.48
All	NN	Lum+L-DEM	94.13	**0.92**	**99.96**	**0.88**	**99.22**	**95.74**	**88.30**
All	LDA	Lum	93.61	2.26	99.74	2.16	96.48	89.70	76.74
All	LDA	L-DEM	83.83	4.08	99.26	3.94	89.17	73.00	55.35
All	LDA	Lum+L-DEM	**95.43**	1.43	99.85	1.41	98.04	91.87	79.00
All	PCA	Lum	92.96	2.44	99.66	2.20	95.96	90.91	82.26
All	PCA	L-DEM	76.61	6.65	98.12	6.38	83.57	66.09	46.13
All	PCA	Lum+L-DEM	95.30	1.66	99.81	1.57	97.78	93.43	84.83

Note: best results for each test set are printed **bold**.

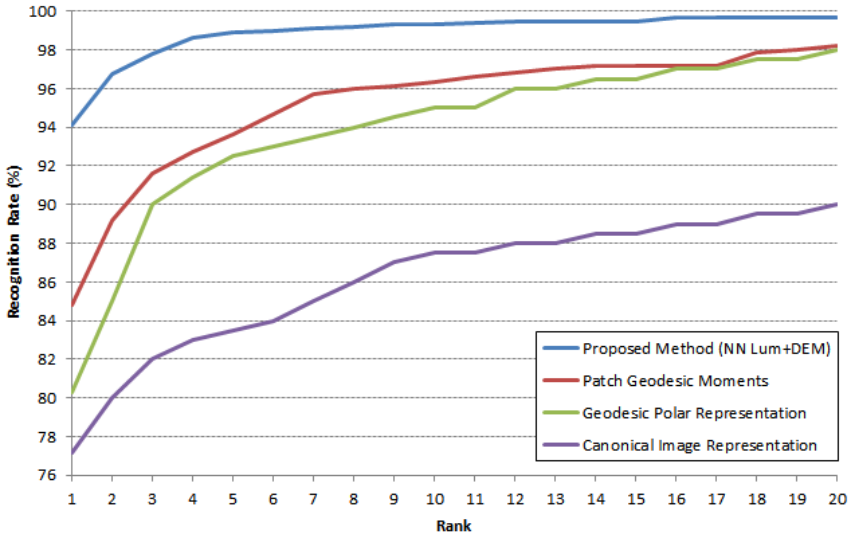


Figure 6.6: Comparison with other authors of CMC curves obtained with the BU-3DFE dataset. Adapted from Hajati et al. (2012).

the geodesic polar representation (Mpiperis et al., 2007) and the canonical image representation (Bronstein et al., 2007).

It is less easy to compare our results with those of most other studies, because we used all 100 persons of the dataset (in most studies only a subset was used) with only two training images per person (most studies used more training images). Kaushik et al. (2009) obtained an ROR rate of 95.33% and an EER of 4.67% when using a small sample of 695 test images. We used 200 training images and 2300 test images, achieving a similar ROR rate of 94.13% (NN Lum+L-DEM; ALL set), but a much better EER of 0.92%. Venkatesh et al. (2012) also used grayscale images, depth maps and a combination of both, but tested only 81 persons and a subset of images with only expression intensity levels 3 and 4 (in total 320 images, of which 48 were expressive) after training on a much larger set of 81 neutral images plus 191 expressive ones. Their best result in identity recognition was obtained when combining image and disparity data (*"stacked G–D, intensities 3,4"*) in Venkatesh et al. (2012): an ROR rate of 98.96% using NN classification. When only using disparity data, the ROR rate was 79.17%, which is comparable to our 80.75% in case of NN+L-DEM and the EXPRESSIVE dataset. However, we did not use expressive images for training. Mpiperis et al. (2008) used

50 persons for bootstrapping a face surface model and the remaining 50 were divided into a training and test set, attaining an ROR rate of about 86% and a rank–5 rate of 97%. We used a much larger test set and obtained an ROR rate of 94.1% and a rank–5 rate which is very close to 99% (NN Lum+L-DEM; ALL set; Table 6.1; Figure 6.6).

Finally, we can compare our results with those obtained by using another database. Sun et al. (2007) trained their method on neutral expressions and tested it on smiling and other expressions. By employing disparity data, they could improve ROR rate from 91.5 to 94%. This is comparable to our test on the EXPRESSIVE set, where we could improve LDA-based ROR rate from 91.1 to 93.3% (to be fair: the average over NN, LDA and PCA improved from 89.3 to 92.4%). However, their disparity-only ROR rate was 67%, whereas we obtained an average of 77.3% over NN, LDA and PCA (if we exclude PCA, even 80.5%). In summary, we may claim that our results are at state-of-the-art level.

6.4.2.2 *Texas-3DFR results*

Performance curves and data are shown in Figure 6.7 and in Table 6.2. Overall, in case of Texas-3DFR the diffences between the classifiers, the test sets, and the data used (luminance, range and both) are smaller than in case of BU-3DFE. NN classification showed the best performance in both identification and verification tests. In verification, PCA performed best on combined luminance and range data, but in identification the ROR rate was only 90.71% when using only range data of expressive faces. In general, LDA results are comparable with the other results. Combining luminance with range data seems to consistently produce better results for the hardest $FAR_{0.01\%}$, especially in case of expressive faces.

CMC curves in Figure 6.7 show that rank–5 recognition rates are well over 98% in all the tests performed. ROR rates (Table 6.2) show that the difference between range-only and range-plus-luminance results is about only 1% for neutral faces, about 5% for expressive faces, and then a mere 2% in case

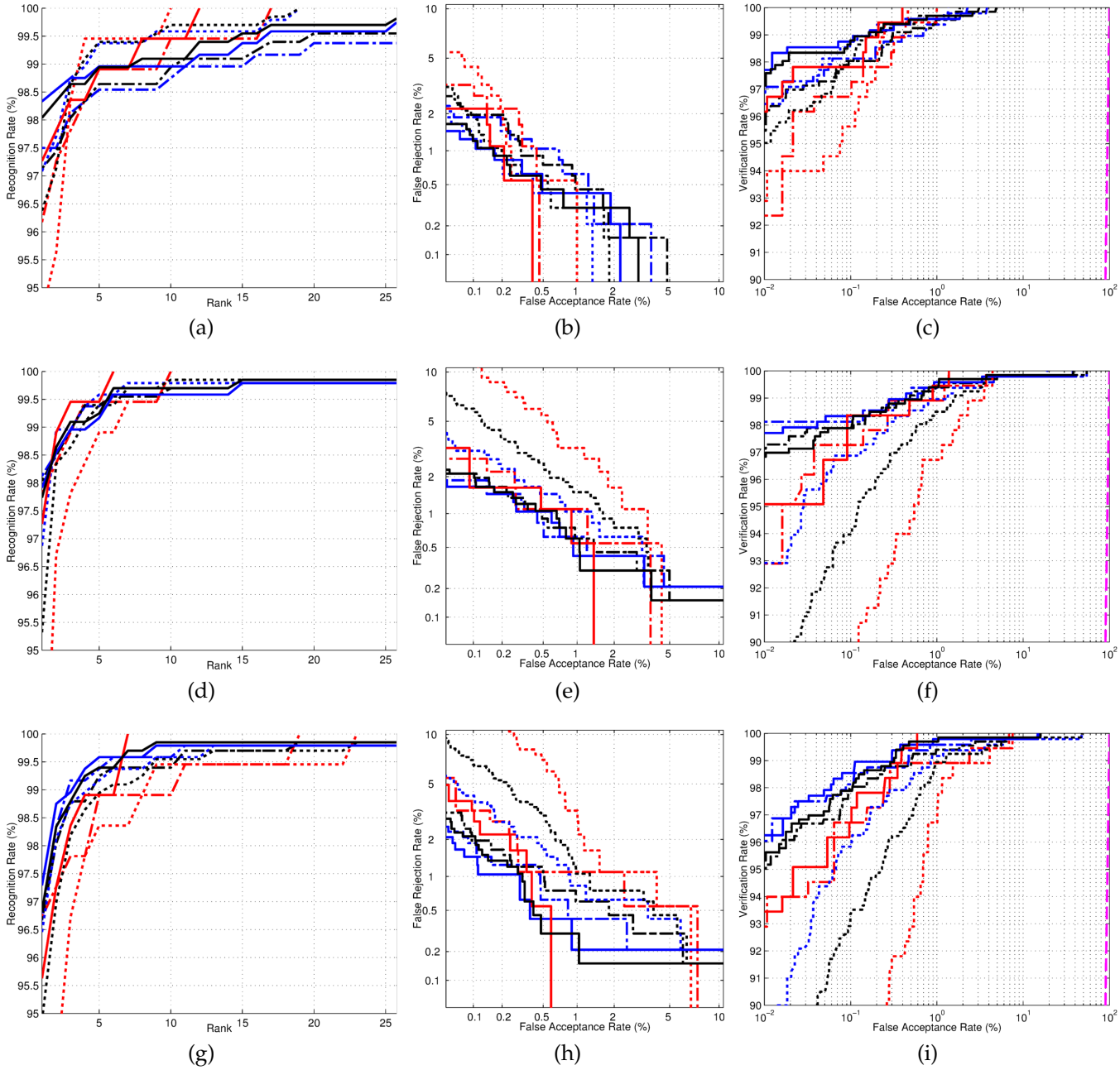


Figure 6.7: CMC (left), DET (middle) and ROC (right) performance curves in case of Texas-3DFR. **(a–c):** Neural Network results; **(d–f):** LDA results; **(g–i):** PCA results. Luminance-only results are given by dash-dot lines, disparity-only results by dotted lines, and joint results by solid lines. NEUTRAL set in blue, EXPRESSIVE set in red, and ALL set in black. ROC random guess rates are dashed in magenta.

Table 6.2: Texas-3DFR Performance Results

			Identification		Verification				
Test set	Classifier	Data	$ROR_{\%}$	$EER_{\%}$	$AUC_{\%}$	$HTER_{min\%}$	$FAR_{1\%}$	$FAR_{0.1\%}$	$FAR_{0.01\%}$
Neutral	NN	Lum	97.50	0.79	99.98	0.69	99.38	98.13	96.88
Neutral	NN	Range	97.08	**0.42**	**99.99**	0.41	99.58	**98.75**	96.46
Neutral	NN	Lum+Range	**98.33**	0.45	**99.99**	0.45	99.58	**98.75**	97.71
Neutral	LDA	Lum	98.13	0.62	99.94	0.56	99.38	98.13	**98.13**
Neutral	LDA	Range	96.88	1.04	99.86	0.92	98.96	96.88	92.92
Neutral	LDA	Lum+Range	97.92	0.67	99.91	0.66	99.58	98.33	97.71
Neutral	PCA	Lum	96.67	0.62	99.96	0.55	99.58	98.13	96.04
Neutral	PCA	Range	96.46	0.86	99.86	0.86	99.17	95.83	88.96
Neutral	PCA	Lum+Range	97.29	**0.42**	99.96	**0.40**	**99.79**	98.54	96.25
Expressive	NN	Lum	96.17	0.50	**99.99**	0.23	**100.00**	96.72	92.35
Expressive	NN	Range	94.54	0.55	**99.99**	0.40	**100.00**	95.63	92.90
Expressive	NN	Lum+Range	97.27	**0.47**	**99.99**	**0.20**	**100.00**	97.81	**96.17**
Expressive	LDA	Lum	**97.81**	1.09	99.96	0.68	98.91	97.27	92.90
Expressive	LDA	Range	91.26	1.72	99.89	1.70	96.72	86.34	75.96
Expressive	LDA	Lum+Range	97.27	0.99	99.98	0.69	99.45	**98.36**	95.08
Expressive	PCA	Lum	96.72	1.09	99.94	0.69	98.91	96.72	92.90
Expressive	PCA	Range	90.71	1.58	99.86	1.31	96.72	86.34	70.49
Expressive	PCA	Lum+Range	95.63	0.55	99.98	0.29	**100.00**	96.72	93.44
All	NN	Lum	97.13	0.75	99.98	0.61	99.55	98.04	95.48
All	NN	Range	96.38	0.58	**99.99**	0.44	**99.70**	98.04	95.02
All	NN	Lum+Range	**98.04**	0.46	**99.99**	0.42	**99.70**	**98.79**	**97.13**
All	LDA	Lum	**98.04**	0.75	99.95	0.65	99.40	97.89	**97.13**
All	LDA	Range	95.32	1.36	99.87	1.22	98.49	93.97	87.93
All	LDA	Lum+Range	97.74	0.75	99.93	0.68	99.40	97.89	96.83
All	PCA	Lum	96.68	0.75	99.95	0.64	99.40	98.04	95.02
All	PCA	Range	94.87	1.05	99.87	1.02	98.94	93.36	84.31
All	PCA	Lum+Range	96.83	**0.45**	99.97	**0.39**	**99.70**	97.89	95.32

Note: best results for each test set are printed in **bold**.

of all faces. Hence, precise range data can almost compete with luminance data. In the case of disparity (L-DEM) data, see Table 6.1, the corresponding differences are 11, 15 and 13%, respectively. This implies that disparity data cannot yet compete with range data. In addition, overall ROR rates based on luminance-only data were lower in case of BU-3DFE than in case of Texas-3DFR. We need to keep in mind that these two databases are similar but not equal in terms of facial expressions and lighting.

Focusing again on Texas-3DFR, our results can be directly compared with those of Gupta et al. (2010), because we used the same training and test sets. However, Gupta et al. applied three LDA steps: (1) of all 300 Euclidean plus 300 geodesic distances, the 106 plus 117 most discriminative distances were determined; (2) of those 223, the best 123 were selected; and (3) a Fisher LDA classifier was trained, using Euclidean distances, to retain only 11 features. In contrast, we employed rougly ten times more LDA and PCA components, together with the cosine distance (LDA) and the Mahalanobis cosine distance (PCA), because these were reported to be the most appropriate distances which also provided the best results (Wechsler, 2007). Nevertheless, our results are at least comparable with those of the anthropofaces method (Gupta et al., 2010), as can be seen in Tables 6.3, 6.4 and 6.5. In these tables, manual and automatic refer to manually annotated and automatically detected fiducial points, the number of points being 10 or 25. Hence, these tables show that it is not really necessary to search for and apply fiducial face points in order to achieve very high recognition rates and very low error rates. Hasan et al. (2012) used a 3D Radon transform on the Texas-3DFR and achieved a moderate ROR rate of 86% with a dataset partitioning which was similar to the one we used.

6.4.2.3 *BU-3DFE vs. Texas-3DFR results*

Above we compared individual results in Table 6.1 and Table 6.2. Here we summarise and compare them by averaging data over the NEUTRAL, EXPRESSIVE and ALL sets. In each set we took the NN Lum+L-DEM result (BU-

Table 6.3: Texas-3DFR ROR$_{\%}$ Rate

Algorithms	Neutral	Expressive	All
Anthroface 3D (25 manual)	**98.8**	95.6	97.9
Anthroface 3D (10 manual)	**98.8**	96.2	**98.0**
Anthroface 3D (10 automatic)	97.3	95.6	96.8
NN Lum+Range	98.3	**97.3**	**98.0**

Table 6.4: Texas-3DFR EER$_{\%}$

Algorithms	Neutral	Expressive	All
Anthroface 3D (25 manual)	0.84	1.58	1.00
Anthroface 3D (10 manual)	1.10	2.34	1.68
Anthroface 3D (10 automatic)	1.65	2.81	1.98
NN Lum+Range	**0.45**	**0.47**	**0.46**

Table 6.5: Texas-3DFR Area Above ROC$_{\%}$

Algorithms	Neutral	Expressive	All
Anthroface 3D (25 manual)	0.07	0.08	0.08
Anthroface 3D (10 manual)	0.12	0.18	0.14
Anthroface 3D (10 automatic)	0.14	0.25	0.18
NN Lum+Range	**0.01**	**0.01**	**0.01**

3DFE) and the NN Lum+Range result (Texas-3DFR). We also averaged the best results in each set. Hence, we can compare NN with best results but also the (summarised) results obtained with the two databases. In case of BU-3DFE, NN classification yielded the best performance in all sets (verification) except for ROR rates (identification). As we can see in Table 6.6, the differences between NN and best results in case of Texas-3DFR are also very small. All performance measures indicate that BU-3DFE results are worse than Texas-3DFR results. The differences in Area Under ROC (AUC) are probably insignificant, perhaps similarly to those in EER and HTER. Significant are the differences in ROR and FAR rates, especially at the more stringent FAR$_{0.1\%}$ and FAR$_{0.01\%}$ levels. However, in Table 6.1 we can see that BU3D-FE FAR rates obtained with the EXPRESSIVE set are lower than those

Table 6.6: Comparison of Texas-3DFR and BU-3DFE results

		ROR%	EER%	AUC%	$HTER_{min\%}$	$FAR_{1\%}$	$FAR_{0.1\%}$	$FAR_{0.01\%}$
Texas-3DFR	NN Lum+Range	98.21	0.47	99.99	0.36	99.76	98.45	97.00
BU-3DFE	NN Lum+L-DEM	94.23	0.90	99.95	0.85	99.12	95.46	88.62
	Difference	3.98	0.43	0.04	0.49	0.64	2.99	8.38
Texas-3DFR	Best in set	98.06	0.45	99.99	0.33	99.83	98.63	97.14
BU-3DFE	Best in set	95.53	0.90	99.95	0.85	99.12	95.46	88.62
	Difference	2.53	0.45	0.04	0.52	0.71	3.17	8.52

Note: results averaged over NEUTRAL, EXPRESSIVE and ALL sets.

with the NEUTRAL and ALL sets. In contrast, Texas-3DFR does not include very expressive faces, which means that FAR rates of the EXPRESSIVE set are not much lower than those of the NEUTRAL and ALL sets. If we take the averages over only the NEUTRAL and ALL sets — to be fair, the EXPRESSIVE set is still represented in the ALL set of BU-3DFE — the NN Lum+(R/D) differences between Texas-3DFR and BU-3DFE become smaller: $FAR_{0.1\%}$ goes from 2.99 to 1.31%, and $FAR_{0.01\%}$ goes from 8.38 to 5.08%. The FAR differences in case of "Best in set" also become smaller. Apart from expressions, the lighting conditions and the difference between disparity and range data may provide an explanation of the remaining differences, but this requires further investigation.

6.5 DISCUSSION AND CONCLUSIONS

Obtained results clearly emphasise the role of 3D information in face recognition, not only because of the obvious benefit of being invariant to lighting conditions. It represents significant structural information, enough to classify faces with good accuracy, quickly surpassing a rank–5 recognition rate of 95% when using L-DEM disparity data and even 99% when using more accurate range data. Moreover, 3D structural information complements luminance data for increasing the robustness of a recognition system. This is especially useful in identity verification where very low false acceptance and rejection rates are required.

Our results also indicate that the proposed method can compete with many computationally more complex ones, e.g., patch geodesic moments (Hajati et al., 2012), geodesic polar representation (Mpiperis et al., 2007), canonical image representation (Bronstein et al., 2007), surface-based control-points (Kaushik et al., 2009), 3D face matrices (Venkatesh et al., 2012), geodesically aligned bilinear models (Mpiperis et al., 2008), anthropometric 3D face recognition (Gupta et al., 2010) and the 3D Radon transform (Hasan et al., 2012). Perhaps even more important is that a system which employs cortical neural processes, i. e., which can be thought of as mimicking part of our visual system, can be applied to a real-world problem, and it can compete with very advanced methods in computer vision.

Of all classifiers, neural networks consistently gave best overall results, especially when applied to disparity and range data. In the case of BU-3DFE, they performed better than LDA and PCA in the more stringent verification trials with very low false acceptance rates. Interestingly, the biological stereo model, which cannot yet provide high-resolution disparity maps, appeared to contribute crucial information for successful recognition. Although BU-3DFE is not strictly a stereo database, results obtained with it are similar to those obtained with the range data of Texas-3DFR. This suggests that at least similar performances can be obtained with real stereo databases.

In further research it makes sense to check recognition and verification rates when using non-frontal facial views, also because stereo information can be used to retrieve face normals. In addition, recognition rates in the case of specific facial expressions can be explored, in conjunction with the detection of emotion intensities. The use of complementary information available in the Gabor-filter disparity model, especially phase information, has recently been shown to produce very good face recognition results (Štruc and Pavešić, 2010; Günther et al., 2012).

7

CONCLUDING REMARKS

REACHING AN INTEGRATED FRAMEWORK FOR LOCAL-GIST WITH PROTO-OBJECT CATEGORISATION AND FACE RECOGNITION

ABSTRACT: This chapter summarises the previous chapters, main achievements and presents directions for further research.

7.1 SUMMARY

Chapter 1 introduced the scope of this thesis along with the biggest differences between Computer Vision (CV) and Human Vision (HV)-based methods. It elaborated a bit on the main research topics, offering a brief overview of each one and of the cortical aspects of HV and how they serve different purposes.

In Chapter 2 we discussed the building blocks of a low-level non-attentional gist system based on semantic representations of low-level scene and object properties. The proposed framework uses bottom-up data streams, from the retina, via LGN, to V1 and further, which are modulated by attention and short-time memory in the ventral and dorsal regions of the prefrontal cortex. A specific aspect was how these processes are bootstrapped, and probably continuously guided, by an extremely fast analysis devoted to scene gist and spatial layout using *local object gist* for simultaneous object segregation, attention, and spatial layout, instead of a more common *global*

scene gist approach. We focused on man-made objects which are dominated by a simple geometric shape repertoire: squares, rectangles, trapeziums, triangles, circles and ellipses. It was shown how these shapes can be detected by a hierarchy of a few cell layers, with strictly bottom-up or data-driven processing. We argued that this processing may occur in very early vision, possibly only employing signals from non-standard retinal ganglion cells. Although proposed to play a role in the fast dorsal stream, similar processing may occur in the slower ventral stream.

In Chapter 3 the effect of cell-based object boundary conspicuity and texture was studied for saliency maps and Focus-of-Attention (FoA). We showed that these maps can be constructed based on colour and texture boundaries. Furthermore, we showed that low-level geometry, especially the one extracted through conspicuity, in addition to rendering filled regions, provides important local cues like corners, bars and blobs for region categorisation. The integration of FoA, region segregation and categorisation is important for developing fast gist vision, i.e., which types of objects are about where in a scene.

In Chapter 4 we described a cortical stereo disparity model (the Disparity-Energy Model) that involves the role and functioning of binocular simple and complex neuron populations and how they are able to extract disparity. We employed a trained neuronal population to encode disparity data implicitly and applied it to several real-world scenes from a ranked evaluation test set and compare our results with those of other authors. The key aspect is that this implementation allows decoding of disparity information in a way similar to how our visual system could have developed this ability, during evolution, in order to accurately estimate disparity of entire scenes. Also, the same simple and complex cells can be used to encode line and edge information, useful to further refine disparity values, especially at object borders. We also proposed that the brain is then able to integrate both information in the low-level disparity pathway, delivering better estimates to higher cortical areas.

In Chapter 5 we integrated our Disparity-Energy Model and Local Object Gist implementations into a unified framework, supporting object detection and categorisation. Our proposition relies on using a parallel low-level bottom-up driven attention scheme for detecting possible object shapes in complex scenes, followed by robust low-level proto-object shape retrieval and classification. The process uses local conspicuity and disparity features for both detection and categorisation, followed by a feed-forward neural network classifier (probably in MT/MST/LIP pathways) that can consequently bootstrap a bottom-up scene gist system.

In Chapter 6 we applied the Disparity-Energy Model implementation to face recognition, comparing its performance with other state-of-the-art face recognition systems. We showed that our implementation provides precise-"enough" disparity maps which are suitable for facial identity recognition and verification. We also tested performance using disparity information, both alone and in combination with image data, yielding state-of-the-art results. We also compared our results with those obtained by precise laser range maps. Overall, our HV-based Disparity-Energy Model framework can compete with and even surpass many CV computationally intensive face recognition methods, highlighting the fact that a system which employs cortical neural processes, i. e., which can be thought of as mimicking part of our visual system, can be applied to a real-world problem.

7.2 INTEGRATED ARCHITECTURE

The proposed integrated architecture is summarised in Figure 7.1. The *solid arrows* represent procedures implemented in this thesis, with *dashed arrows* representing expected model links for future research. *Green arrows* represent the low-level *global* gist architecture developed by Rodrigues and du Buf (2011a).

The current architecture relies on both non-standard retinal ganglion cells and standard ones. The non-standard pathway can serve to quickly boot-

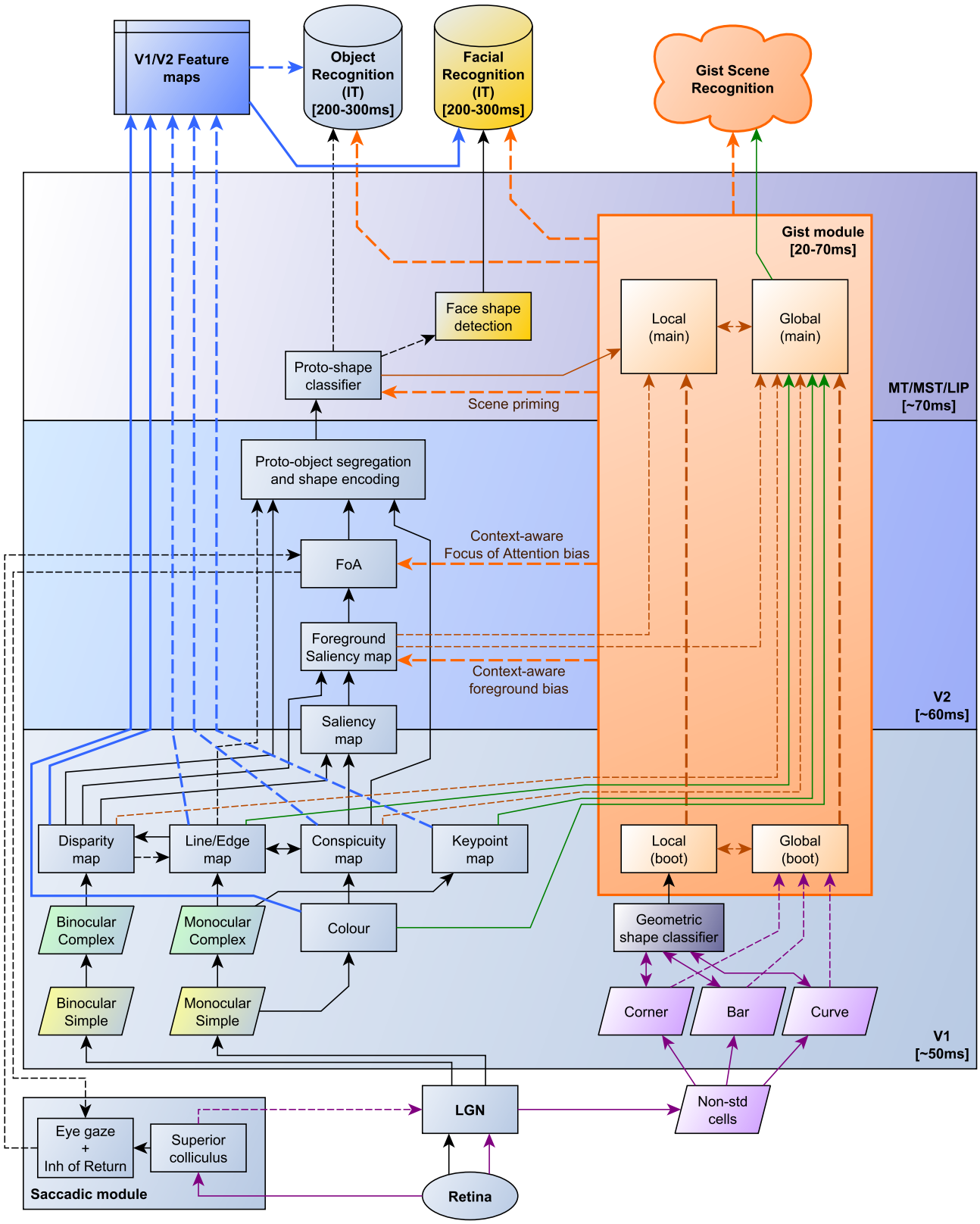


Figure 7.1: Proposed integrated cortical architecture.

strap the low-level gist module, by using specific, hard-coded, shape descriptors (*corners*, *bars* and *curves*) to feed a *geometric shape classifier*, which can be constructed using higher-level grouping cells (Martins et al., 2012). Since low-level geometry information has already been extracted, it is therefore available for obtaining local object gist, e. g., providing cues which are used for a first and fast selection of possible object categories in memory (Bar et al., 2006). This is a purely bottom-up and data-parallel process for bootstrapping the serial object categorisation and recognition processes, which are controlled by top-down attention. Recent research also suggests that we actually categorise objects before we have segregated them, or that both processes occur in parallel. This means that by the time we realise that we are looking at something, our brain already knows what that thing is (Oliva and Torralba, 2006). Therefore, Rensink (2000) proposed a non-attentional "scene schema" consisting of concurrent spatial-layout and gist subsystems which both drive attentional object recognition, all employing "proto-objects" resulting from low-level vision. Similarly, Yanulevskaya et al. (2013) also focused on salient proto-object detection within an object-based attention theory. However, gist vision addressed so far only concerns global gist of entire scenes (Bar, 2004; Siagian and Itti, 2007; Ross and Oliva, 2010; Rodrigues and du Buf, 2011b). Global scene gist can be used to bias — select or exclude — object templates in memory in the matching process: when in a classroom it is not very likely that we see a horse. But global gist lacks localisation. On the other hand, when seeing a horse it is not very likely that we are in a classroom. Local object gist has the advantage of solving, or at least contributing, to the spatial-layout subsystem proposed by Rensink (2000). Although both global and local gist can determine context, probably with a straight relation between them, local gist can solve important problems like a first and fast object categorisation, localisation and segregation, the latter being related to figure-ground organisation (Craft et al., 2007).

A similar low-level attentional view is described by Martin and von der Heydt (2013), who measured spike time correlations in a monkey's visual

cortex. They concluded that specific grouping cells in V1/V2 were able to specifically enhance the activity of neurons whose receptive fields fit their grouping templates, linking neurons to "proto-object structures." In this context, we expect that global gist features will suffice for initial discrimination between very different scene types (in regards to spatial layout, like a "forest" *vs.* a "city") but will severely lack detail in more difficult scenes (like an "office" *vs.* a "classroom"), where local gist can be of great benefit — since most man-made objects tend to follow well-defined geometric shapes — when employed in conjunction with global scene properties (Rodrigues and du Buf, 2011a). This view is also reinforced by Groen et al. (2013), who concluded that gist seems to depend on two stages: an early, automatic stage, where local-contrast responses present in the LGN or V1 seem to play a very important role, followed by a later, task-dependent stage.

Along the standard path, binocular simple and complex cells are used to create a *disparity map* (Martins et al., 2011b), while the monocular versions can be used for the *line/edge map* (Rodrigues et al., 2012), *conspicuity map* (Martins et al., 2012) and *keypoint map* (Rodrigues and du Buf, 2011a). According to the attentional *Coherence Theory* (Rensink, 2000), low-level proto-object shapes are continuously and rapidly formed, and in parallel across the visual field — they are volatile, lacking strong coherence until being stabilised by FoA–gaze, and afterwards dissolve when FoA is released. In our model, we postulate that available disparity and conspicuity information, when combined, is able to quickly highlight all important objects and to resolve border ownership of the outlines of objects. Higher level, oriented, grouping cell populations can encode the distance from the centre of the object to its border, for very specific orientations. These population responses can serve as inputs to a simple, feed-forward *proto-shape classifier*, probably residing in area LIP in the dorsal pathway (Konen and Kastner, 2008; Janssen et al., 2008). LIP also shows shape activation times of 62 ms (Lehky and Sereno, 2007), well within global gist recognition times and has access

(via areas MT and MST), to the Superior Colliculus for eye and head control (Gottlieb, 2007), crucial for FoA, making it a prime candidate area for integrating low-level attention with a proto-object categorisation role.

Salient foreground regions can also serve as inputs for FoA, which can use available proto-object shapes. The ventral stream (V4 and PIT) can further refine shape information and bootstrap either object recognition or face recognition, since we know the cortex employs a dedicated pathway for face processing (Biederman and Kalocsai, 1997).

Gist is expected to have a significant biasing effect throughout this whole process, which is an area of interest for further research. When the brain establishes a background/foreground split in a scene, gist can bias the split based on the scene context (i. e., a forest *vs.* an office space) helping to choose, for each, the best close-to-far range for foreground depth. Later it can influence FoA by increasing the priority of salient shapes depending on context (e. g., a very close "bear"-shape in a "forest" gets top priority!). Gist can also employ scene context to influence the categorisation of objects or faces, prioritising some in detriment of others.

7.3 ACHIEVEMENTS

There are three major achievement categories. These are summarised in this section and they are the basis for future research that will be presented in the next section.

LOW-LEVEL GIST VISION WITH OBJECT CATEGORISATION. The developed cortical architecture framework proposes a new bottom-up concept of "local-gist" vision, based on exploring the quick identification of low-level proto-object shapes. We explored two different mechanisms for this, (1) using *corner*, *bar* and *curve* cells, which yield good results for geometric-shaped objects (Chapter 2) and (2) using a radial space distance-based encoding, where cells can encode shape vectors as pop-

ulation codes (Chapter 5). We proposed possible biological mechanisms of how the brain is able to use low-level pathways propagating upwards from non-standard ganglion cells to V1/V2/LIP, quickly focusing on segregating salient proto-objects from scenes and simultaneously using their shapes for categorisation. Key aspects are:

- The development of a FoA-based saliency measure based on conspicuity propagated from colour-opponent cells (conspicuity can reflect changes in lightness and colour between patches);
- Using both HV-inspired saliency and disparity to achieve foreground/background separation;
- Saliency map used for object segregation and for queueing further processing, focusing on the most salient objects first (for top-down FoA control);
- Quick extraction of proto-object shapes, either from:
  - Object borders obtained from active conspicuity cells;
  - Disparity surfaces obtained from Disparity-Energy Model (DEM) cells.

DISPARITY-ENERGY MODEL. We developed a DEM implementation that relies on learned neuronal population coding, targeted for real-world images, benchmarking it with other state-of-the-art CV stereo models. To our knowledge, this is the first DEM implementation which uses population coding that is capable of delivering very good results for various real-world cases. We further developed our implementation by using conspicuity data that is readily available in the V1/V2 pipeline (from a parallel process) to refine disparity discontinuities at borders of objects. Key aspects are:

- The development of a DEM that relies on a neuronal population coding firing scheme, resulting from exposure of the neuronal population (with real-like cell parameters and characteristics) to

multiple stereo stimuli and consequent learning how to discriminate individual disparity values;

- Integrating colour, different viewpoints, lines/edges, and edge conspicuity to create a robust Luminance, Colour, Viewpoint and Boundary enhanced Disparity-Energy Model (LCVB-DEM);
- Getting good results from a DEM model that approaches results of CV models;
- Ranking our disparity models on the Middlebury dataset, which allows for objective quality comparison with other authors.

FACE RECOGNITION. We evaluated our DEM performance on two state-of-the-art stereo databases, testing both identity and verification recognition with very different facial expressions, using different kinds of classifiers, and compared results with precise laser-acquired disparity data. The proposed DEM framework worked extremely well, even achieving results that topped CV models:

- Grayscale and DEM images classified with Neuronal Networks were compared to Linear Discriminant Analysis and Principal Component Analysis;
- Obtained results show that our biological framework can achieve state-of-the art recognition performance even when persons are displaying very different facial expressions or under different lighting conditions;
- Extensive performance data was elaborated, easing future comparisons with methods from other authors.

7.4 FINAL CONSIDERATIONS AND FUTURE RESEARCH

The work on this Thesis took about 4 years, yielding significant achievements in exploring both the local gist pathway and disparity models based on HV.

In general, our feeling is that there are still many uncharted waters when dealing with low-level local gist vision — object recognition (or categorisation) is still a relatively new research area, where we only inspected the "tip of the iceberg." Still, the architecture's foundations are good, with Chapter 5 showing very good results for a proto-object framework that aims to simultaneously extract from the same data *detection* and *categorisation*. This is rather uncommon in CV methods but makes perfect sense in HV methods, where the brain always tries to minimise energy spent, untangling information as it propagates through the visual cortex areas. Our ROR rate for proto-object categorisation (51 categories, Lai et al., 2011) starts at 50% and quickly rises to 76% at rank–5 and 85% at rank–10 — which means that there is a huge potential for shape data to bias possible scene gist and object templates in (associative) memory for precise recognition of both *scenes* ("where am I?") and *objects* ("what is this?").

We are particularly satisfied with our HV–modified LCVB-DEM — especially due to face-recognition results that can rival current state-of-the-art CV systems. Disparity decoding fits perfectly in a low-level pipeline that can run in parallel with the local gist pipeline, happening mostly inside V1/V2 (Tanabe and Cumming, 2008). Although we did not focus on face detection, the same process used in object detection can be extrapolated for face detection, sending "oval" shapes to a face-recognition pipeline. Our recognition results are indeed very encouraging, showing the remarkable ability for disparity-based structural face data to be almost sufficient to recognise persons from faces, while maintaining a good degree of robustness to interfering facial expressions that usually hamper recognition by classical approaches. We are also interested in pursuing this subject further, by comple-

menting our DEMs with Gabor phase disparity data (for increased disparity resolution), opening the door to facial *expression* recognition.

The aspects mentioned above are selections of many possibilities or doors which have been opened by the research reported in this thesis. The study and modelling of a cortical architecture remains a continuing goal, even after the state-of-the-art has advanced: more cells with additional functionalities can be integrated, feature extractions can be further refined, small and big building blocks can be moved to other positions with other functional and/or structural roles and timings. This thesis is not a final period, neither in basic research concerning the cortical architecture nor in all applications that can be envisioned.

We are continually faced with a series of great opportunities,
brilliantly disguised as insoluble problems.

— John W. Gardner (1912 – 2002).

BIBLIOGRAPHY

Banks, M. S., S. Gepshtein, and M. S. Landy (2004, March). Why is Spatial Stereoresolution so low? *J. Neurosci. 24*(9), 2077–89. (Cited on pages 76, 82, and 148.)

Bar, M. (2004). Visual objects in context. *Nat. Rev. Neurosci. 5*, 619–629. (Cited on pages 22, 54, 111, and 171.)

Bar, M., K. S. Kassam, A. S. Ghuman, J. Boshyan, A. M. Schmid, A. M. Schmidt, A. M. Dale, M. S. Hämäläinen, K. Marinkovic, D. L. Schacter, B. R. Rosen, and E. Halgren (2006, January). Top-down facilitation of visual recognition. In *Proc. Natl. Acad. Sci. U. S. A.*, Volume 103, pp. 449–54. (Cited on pages 22, 70, and 171.)

Baran, B., B. Dogusoy, and K. Cagiltay (2007, July). How do adults solve digital tangram problems? Analyzing cognitive strategies through eye tracking approach. In *Proc. 12th Int. Conf. Human–Computer Interact.*, Volume 4552 of *Springer LNCS*, pp. 555–563. (Cited on page 31.)

Baumann, F., A. Ehlers, K. Vogt, and B. Rosenhahn (2013). Cascaded Random Forest for Fast Object Detection. In *Image Anal.*, Volume 7944 of *Springer L.N.C.S.*, pp. 131–142. (Cited on page 9.)

Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. *Psychol. Rev. 94*(2), 115–147. (Cited on page 1.)

Biederman, I. and P. Kalocsai (1997). Neurocomputational bases of object and face recognition. *Philos. Trans. R. Soc. Biol. Sci. 352*, 1203–1219. (Cited on pages 14, 28, 31, and 173.)

Blum, M., J. Wulfing, and M. Riedmiller (2012, May). A learned feature descriptor for object recognition in RGB-D data. In *2012 IEEE Int. Conf. Robot. Autom.*, pp. 1298–1303. (Cited on page 135.)

Bo, L., X. Ren, and D. Fox (2011, September). Depth kernel descriptors for object recognition. In *2011 IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, pp. 821–826. (Cited on pages 111 and 135.)

Bo, L., X. Ren, and D. Fox (2013). *Unsupervised Feature Learning for RGB-D Based Object Recognition*, Volume 88 of *Springer Tracts in Advanced Robotics*. (Cited on page 135.)

Bomberger, N. A. and E. L. Schwartz (2005). The structure of cortical hypercolumns: Receptive field scatter may enhance rather than degrade boundary contour representation in V1. *J. Vis. 5*(8), 891. (Cited on page 35.)

Bronstein, A. M., M. M. Bronstein, and R. Kimmel (2007). Expression-Invariant Representations of Faces. *IEEE Trans. Image Proc. 16*(1), 188–197. (Cited on pages 158 and 165.)

Bruce, V. (1990, December). Perceiving and Recognising Faces. *Mind Lang. 5*(4), 342–364. (Cited on page 139.)

Chen, Y. and N. Qian (2004). A Coarse-to-Fine Disparity Energy Model with both Phase-Shift and Position-Shift Receptive Field Mechanisms. *Neural Comput. 16*(8), 1545–1577. (Cited on pages 77, 79, 81, 82, 147, and 148.)

Craft, E., H. Schütze, E. Niebur, and R. von der Heydt (2007, June). A neural model of figure-ground organization. *J. Neurophysiol. 97*(6), 4310–26. (Cited on pages 22 and 171.)

Creem-Regehr, S. H. (2009). Sensory-motor and cognitive functions of the human posterior parietal cortex involved in manual actions. *Neurobiol. Learn. Mem. 91*, 166–171. (Cited on page 24.)

Crick, F. and C. Koch (2003). A framework for consciousness. *Nat. Neurosci. 6*, 119–126. (Cited on page 21.)

Davies, E. R. (2004). *Machine Vision: Theory, Algorithms, Practicalities*. Elsevier. (Cited on page 124.)

Deco, G. and E. T. Rolls (2005). Attention, short-term memory, and action selection: a unifying theory. *Prog. Neurobiol. 76*, 236–256. (Cited on page 24.)

Delac, K., M. Grgic, and S. Grgic (2005). Independent Comparative Study of PCA, ICA, and LDA on the FERET Data Set. *Int. J. Imaging Syst. Tech. 15*(5), 252–260. (Cited on page 151.)

den Ouden, H. E. M., R. van Ee, and E. H. F. de Haan (2005). Colour helps to solve the binocular matching problem. *J. Physiol. 567*(2), 665–671. (Cited on pages 88 and 105.)

DiCarlo, J. J., D. Zoccolan, and N. C. Rust (2012, February). How does the brain solve visual object recognition? *Neuron 73*(3), 415–34. (Cited on pages 1, 5, 12, and 104.)

du Buf, J., K. Terzic, and J. Rodrigues (2013). Phase-differencing in stereo vision: solving the localisation problem. In *Proc. 6th Int. Conf. Bio-inspired Syst. Signal Process.*, Barcelona, Spain, pp. 254–263. (Cited on pages 75 and 147.)

du Buf, J. M. H. (2007). Improved grating and bar cell models in cortical area V1 and texture coding. *Image Vis. Comput. 25*(6), 873–882. (Cited on pages 58 and 65.)

Elazary, L. and L. Itti (2008). Interesting objects are visually salient. *J. Vis. 8*(3), 1–15. (Cited on pages 53 and 58.)

Fagan, J. F. (1972, December). Infants' Recognition Memory for Faces. *J. Experim. Child Psychol. 14*(3), 453–476. (Cited on page 140.)

Farivar, R. (2009). Dorsal-ventral integration in object recognition. *Brain Res. Rev. 61*(2), 144–153. (Cited on pages 21, 26, and 110.)

Farrajota, M., J. Martins, J. Rodrigues, and J. du Buf (2011). Disparity energy model with keypoint disparity validation. In *Proc. 17th Port. Conf. Pattern Recognit.*, Porto, Portugal, pp. 70–71. (Cited on pages xi and 15.)

Faubel, C. and G. Schoner (2009). A neuro-dynamic architecture for one shot learning of objects that uses both bottom-up recognition and top-down prediction. In *Proc. 2009 IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, pp. 3162–3169. (Cited on pages 32 and 34.)

Field, G. D. and E. J. Chichilnisky (2007). Information processing in the primate retina: circuitry and coding. *Annu. Rev. Neurosci. 30*, 1–30. (Cited on page 88.)

Filippini, H. and M. Banks (2009). Limits of Stereopsis explained by Local Cross-Correlation. *J. Vis. 9*(8), 1–18. (Cited on pages 74, 76, 81, 82, and 148.)

Finlayson, G. D., B. Schiele, and J. L. Crowley (1998). Comprehensive Colour Image Normalization. *Proc. 5th Eur. Conf. Comput. Vis. 1406*, 475–490. (Cited on pages 36, 37, 59, and 60.)

Franco, A. and L. Nanni (2009, July). Fusion of Classifiers for Illumination Robust Face Recognition. *Expert Syst. Appl. 36*(5), 8946–8954. (Cited on pages 13 and 140.)

Fu, Y.-X., Y. Shen, H. Gao, and Y. Dan (2004). Asymmetry in Visual Cortical Circuits Underlying Motion-induced Perceptual Mislocalization. *J. Neurosci. 24*(9), 2165–2171. (Cited on page 106.)

Gall, J., N. Razavi, and L. V. Gool (2012). An introduction to random forests for multi-class object detection. In *Proc. 15th Int. Conf. Theor. Found. Comput. Vis.*, Dagstuhl Castle, Germany, pp. 243–263. Springer. (Cited on page 9.)

Gao, D. (2005). Integrated Learning of Saliency, Complex Features, and Object Detectors from Cluttered Scenes. In *2005 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Volume 2, pp. 282–287. IEEE. (Cited on page 9.)

Goren, C. C., M. Sarty, and P. Y. K. Wu (1975, October). Visual Following and Pattern Discrimination of Face-like Stimuli by Newborn Infants. *Pediatrics 56*(4), 544–549. (Cited on page 140.)

Gottlieb, J. (2007, January). From thought to action: the parietal cortex as a bridge between perception, action, and cognition. *Neuron 53*(1), 9–16. (Cited on pages 6, 110, 136, and 173.)

Greene, M. and A. Oliva (2009a). The briefest of glances: the time course of natural scene understanding. *Cogn. Psychol. 20*(4), 137–179. (Cited on page 8.)

Greene, M. R. and A. Oliva (2009b). Recognition of natural scenes from global properties: Seeing the forest without representing the trees. *Cogn. Psychol. 58*, 137–176. (Cited on pages 8 and 22.)

Grill-Spector, K. and N. Kanwisher (2005). Visual recognition: as soon as you know it is there, you know what it is. *Psychol. Sci. 16*, 152–160. (Cited on page 11.)

Groen, I., S. Ghebreab, V. Lamme, and H. S. Scholte (2013, July). Two stages in scene gist processing revealed by evaluating summary statistics with single-image ERPs. *J. Vis. 13*(9), 1059. (Cited on pages 8 and 172.)

Günther, M., D. Haufe, and R. P. Würtz (2012). Face Recognition with Disparity Corrected Gabor Phase Differences. In *Artif. Neural Networks Mach. Learn.*, Volume 7552 of *Springer LNCS*, pp. 411–418. (Cited on page 165.)

Gupta, S., K. R. Castleman, M. K. Markey, and A. C. Bovik (2010). Texas 3D Face Recognition Database. In *Proc. IEEE Southwest Symp. Image Anal. Interpret.*, pp. 97–100. (Cited on pages 143, 145, 150, and 165.)

Gupta, S., M. K. Markey, and A. C. Bovik (2010, June). Anthropometric 3D Face Recognition. *Int. J. Comp. Vis. 90*(3), 331–349. (Cited on pages 147 and 162.)

Haan, M. and C. A. Nelson (1997, April). Recognition of the Mother's Face by Six-Month-Old Infants: A Neurobehavioral Study. *Child Dev. 68*(2), 187–210. (Cited on page 140.)

Haefner, R. M. and B. G. Cumming (2008, January). Adaptation to Natural Binocular Disparities in Primate V1 explained by a Generalized Energy Model. *Neuron 57*(1), 147–58. (Cited on pages 7, 74, 80, and 106.)

Hajati, F., A. A. Raie, and Y. Gao (2012, March). 2.5D Face Recognition using Patch Geodesic Moments. *Pattern Recognit. 45*(3), 969–982. (Cited on pages 155, 158, and 165.)

Hasan, M., W. Jouhar, and M. Alwan (2012). 3-D Face Recognition Using Improved 3D Mixed Transform. *Int. J. Biometrics Bioinforma. 6*(1), 278–290. (Cited on pages 162 and 165.)

Hayasaka, A., K. Ito, T. Aoki, H. Nakajima, and K. Kobayashi (2009). A Robust 3D Face Recognition Algorithm Using Passive Stereo Vision. *IEICE Trans. Fundam. Electron. Comm. Comput. Sci. E92–A*(4), 1047–1055. (Cited on page 141.)

Hess, U. and P. Thibault (2009, January). Darwin and Emotion Expression. *Am. Psychol. 64*(2), 120–128. (Cited on page 140.)

Horwitz, G. D. and C. A. Hass (2012, June). Nonlinear analysis of macaque V1 color tuning reveals cardinal directions for cortical color processing. *Nat. Neurosci. 15*(6), 913–9. (Cited on page 6.)

Houtkamp, R. and P. R. Roelfsema (2010, December). Parallel and serial grouping of image elements in visual perception. *J. Exp. Psychol. Hum. Percept. Perform. 36*(6), 1443–59. (Cited on page 21.)

Hubel, D. H. (1995). *Eye, Brain and Vision*, Volume 22 of *Scientific American Library series*. New York. (Cited on pages 5, 6, 37, and 61.)

Itti, L. and C. Koch (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Res. 40*(10-12), 1489–1506. (Cited on page 9.)

Itti, L. and C. Koch (2001). Computational modeling of visual attention. *Nat. Rev. Neurosci. 2*(3), 194–203. (Cited on pages 58 and 68.)

Ives, R., Y. Du, D. Etter, and T. Welch (2005, August). A Multidisciplinary Approach to Biometrics. *IEEE Trans. Educ. 48*(3), 462–471. (Cited on page 140.)

Janssen, P., S. Srivastava, S. Ombelet, and G. A. Orban (2008, June). Coding of shape and position in macaque lateral intraparietal area. *J. Neurosci. 28*(26), 6679–90. (Cited on pages 110, 136, and 172.)

Johnson, A. and M. Hebert (1998). Efficient multiple model recognition in cluttered 3-D scenes. In *Proc. 1998 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 671–677. (Cited on page 111.)

Johnson, E. N., M. J. Hawken, and R. Shapley (2004). Cone inputs in macaque primary visual cortex. *J. Neurophysiol. 91*(6), 2501–2514. (Cited on page 88.)

Karttunen, R. and S. Häkkinen (1982). Discrimination of traffic signs in the peripheral areas of the field of vision. *Reports from Liikenneturva 25/1982, Central Organization for Traffic Safety, Helsinki (Finland)*. (Cited on page 30.)

Kaushik, V. D., A. Budhwar, A. Dubey, R. Agrawal, S. Gupta, V. K. Pathak, and P. Gupta (2009, December). An Efficient 3D Face Recognition Algorithm. In *Proc. 2009 IEEE 3rd Int. Conf. New Technol. Mobil. Secur.*, pp. 1–5. (Cited on pages 142, 158, and 165.)

Kiani, R., H. Esteki, K. Mirpour, and K. Tanaka (2007). Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. *J. Neurophysiology 97*, 4296–4309. (Cited on pages 25 and 26.)

Kohn, A. and J. Movshon (2003). Neuronal adaptation to visual motion in area MT of the macaque. *Neuron 39*, 681–691. (Cited on page 24.)

Konen, C. S. and S. Kastner (2008). Two hierarchically organized neural systems for object information in human visual cortex. *Nat. Neurosci. 11*(2), 224–231. (Cited on pages 21, 26, 110, 136, and 172.)

Kosov, S., K. Scherbaum, K. Faber, T. Thormahlen, and H.-P. Seidel (2009, November). Rapid Stereo-Vision Enhanced Face Detection. In *Proc. 16th IEEE Int. Conf. Image Process.*, pp. 1221–1224. (Cited on page 140.)

Kosov, S., T. Thormahlen, and H.-P. Seidel (2010, September). Rapid Stereo-Vision Enhanced Face Recognition. In *Proc. 17th IEEE Int. Conf. Image Process.*, Hong Kong, pp. 2437–2440. (Cited on page 141.)

Krauskopf, J. and J. D. Forte (2002). Influence of chromaticity on vernier and stereo acuity. *J. Vis. 2*(9), 6. (Cited on page 88.)

Krüger, N., M. Lappe, and F. Wörgötter (2003). Biological Motivated Multimodal Processing of Visual Primitives. *Interdiscip. J. Artif. Intell. Simul. Behav. 1*(5), 53–59. (Cited on page 46.)

Lai, K., L. Bo, X. Ren, and D. Fox (2011). A Large-Scale Hierarchical Multi-View RGB-D Object Dataset. In *Proc. IEEE Int. Conf. Robot. Autom.*, pp. 1817–1824. (Cited on pages 111, 113, 114, 125, 126, 134, 135, 137, and 176.)

Lehky, S. R. and A. B. Sereno (2007). Comparison of shape encoding in primate dorsal and ventral visual pathways. *J. Neurophysiol. 97*(1), 307–319. (Cited on pages 23, 24, 27, 29, 30, 34, 136, and 172.)

Li, S. Z. and A. K. Jain (Eds.) (2011). *Handbook of Face Recognition* (2nd ed.). London: Springer. (Cited on pages 14 and 140.)

Logothetis, N. K. and D. L. Sheinberg (1996, January). Visual object recognition. *Annu. Rev. Neurosci. 19*, 577–621. (Cited on page 2.)

Lowe, D. G. (2004). Distinctive Image Features from Scale-Invariant Keypoints. *Int. J. Comput. Vis. 60*, 91–110. (Cited on page 9.)

Lu, Y. and C. Rasmussen (2012). Simplified Markov Random Fields for Efficient Semantic Labeling of 3D Point Clouds. In *IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, pp. 2690–2697. (Cited on page 111.)

Luoma, J. (1992, January). Immediate responses to road signs of alerted and unaltered drivers: An evaluation of the validity of eye movement method. Technical report, Transportation Research Board — Annual Meeting, Washington DC. (Cited on page 29.)

Manor, B. R. and E. Gordon (2003). Defining the temporal threshold for ocular fixation in free-viewing visuocognitive tasks. *J. Neurosci. Methods 128*, 85–93. (Cited on page 31.)

Manor, B. R., E. Gordon, and S. W. Touyz (1995). Consistency of the first fixation when viewing a standard geometric stimulus. *Int. J. Psychophysiol. 20*, 1–9. (Cited on page 31.)

Martens, M. H. and M. Fox (2007). Does road familiarity change eye fixations? A comparison between watching a video and real driving. *Transp. Res. Part F – Traffic Psychol. Behav. 10*, 33–47. (Cited on pages 29 and 30.)

Martin, A. and R. von der Heydt (2013, July). Firing synchrony between neurons reveals proto-object representation in monkey visual cortex. *J. Vis. 13*(9), 289. (Cited on pages 124 and 171.)

Martins, J., M. Farrajota, R. Lam, J. Rodrigues, K. Terzic, and J. du Buf (2012). A disparity energy model improved by line, edge and keypoint correspondences. In *Proc. 35th Eur. Conf. Vis. Percept.*, Volume 41, Alghero, Italy, pp. 76. (Cited on pages xi and 15.)

Martins, J., J. Rodrigues, and J. M. H. Buf (2008). Region segregation and saliency using colour information. In *Proc. 14th Port. Conf. Pattern Recogn.*, Volume D, Coimbra, Portugal. (Cited on pages xi and 15.)

Martins, J., J. Rodrigues, and J. M. H. Buf (2009). Object segregation and local gist vision using low-level geometry. In *Proc. 32nd Eur. Conf. Vis. Percept.*, Regensburg, Germany, pp. 41–42. (Cited on pages xi and 15.)

Martins, J., J. Rodrigues, and J. du Buf (2011a). Local Gist Vision of Man-made Objects. In *Web Proc. IV Rovereto Atten. Work.*, Volume 4, Rovereto, Italy. (Cited on pages xi and 15.)

Martins, J., J. Rodrigues, and J. M. H. du Buf (2009). Focus of Attention and Region Segregation by Low-level Geometry. *Proc. Fourth Int. Conf. Comput. Vis. Theory Appl. 2*, 267–272. (Cited on pages xi, 15, 36, 40, and 53.)

Martins, J., J. Rodrigues, and J. M. H. du Buf (2012). Local object gist: meaningful shapes and spatial layout at a very early stage of visual processing. *Gestalt Theory 34*(3/4), 349–380. (Cited on pages xii, 9, 15, 76, 94, 98, 106, 111, 116, 117, 136, 171, and 172.)

Martins, J. A., J. Rodrigues, and J. du Buf (2011b, December). Disparity Energy Model using a Trained Neuronal Population. In *Proc. IEEE Int. Symp. Signal Proc. Info. Tech.*, Bilbao, Spain, pp. 287–292. (Cited on pages xi, 7, 15, 74, 75, 77, 79, 83, 86, 142, 147, 148, and 172.)

Martins, J. A., J. Rodrigues, and J. du Buf (2014a). Expression-Invariant Face Recognition using a Biological Disparity Energy Model. *Submitt. to IEEE Trans. Image Process.*. (Cited on page xii.)

Martins, J. A., J. Rodrigues, and J. du Buf (2014b). Luminance, Colour, Viewpoint and Border Enhancement Disparity-Energy Model. *Prep. Submiss.*. (Cited on page xii.)

Martins, J. A., J. M. F. Rodrigues, and J. M. H. du Buf (2014c). Proto-Object Categorisation and Local-Gist using Implicit Coding of Low-Level Spatial Layout Features. *Prep. Submiss.*. (Cited on page xii.)

Masland, R. H. and P. R. Martin (2007). The unsolved mystery of vision. *Curr. Biol. 17*(15), R577–582. (Cited on pages 27, 28, 56, and 136.)

Messer, K., J. Matas, J. Kittler, J. Luettin, and G. Maitre (1999). XM2VTSDB: The Extended M2VTS Database. In *Proc. 2nd Int. Conf. Audio Videobased Biometric Pers. Authentic.*, Volume 964, pp. 965–966. (Cited on page 141.)

Moller, M. F. (1993, January). A Scaled Conjugate Gradient Algorithm for Fast Supervised Learning. *Neur. Net. 6*(4), 525–533. (Cited on pages 128 and 150.)

Mpiperis, I., S. Malassiotis, and M. G. Strintzis (2007). 3-D Face Recognition with the Geodesic Polar Representation. In *Proc. IEEE Trans. Info. Forensics Secur.*, Volume 2, pp. 537–547. (Cited on pages 158 and 165.)

Mpiperis, I., S. Malassiotis, and M. G. Strintzis (2008). Expression-Compensated 3D Face Recognition with Geodesically Aligned Bilinear Models. In *Proc. 2nd IEEE Int. Conf. Biometrics Theory, Appl. Syst.*, pp. 1–6. (Cited on pages 142, 158, and 165.)

Murphy-Chutorian, E. and J. Triesch (2005). Shared Features for Scalable Appearance-Based Object Recognition. In *Proc. 7th IEEE Work. Appl. Comput. Vis.*, Volume 1, pp. 16–21. (Cited on pages 113, 114, and 115.)

Mutti, F. and G. Gini (2010). Bio-inspired disparity estimation system from energy neurons. *Proc. IEEE Int. Conf. Appl. Bionics Biomech.*, 1–6. (Cited on page 75.)

Niebur, E. and C. Koch (1996). Control of Selective Visual Attention: Modeling the 'Where' Pathway. *Neural Inf. Process. Syst. 8*, 802–808. (Cited on page 58.)

Ohzawa, I., G. C. DeAngelis, and R. D. Freeman (1990). Stereoscopic depth discrimination in the visual cortex. *Science (80-.). 249*, 1037–1041. (Cited on page 7.)

Ohzawa, I., G. C. DeAngelis, and R. D. Freeman (1997, June). Encoding of binocular disparity by complex cells in the cat's visual cortex. *J. Neurophysiol. 77*(6), 2879–909. (Cited on pages 7, 74, 76, 77, 78, 80, 81, and 148.)

Oliva, A. and A. Torralba (2006). Building the gist of a scene: the role of global image features in recognition. *Prog. Brain Res. Vis. Percept. 155*, 23–26. (Cited on pages 11, 22, and 171.)

Peng, Q. and B. E. Shi (2010). The changing disparity energy model. *Vision Res. 50*(2), 181–192. (Cited on page 23.)

Pinna, B. (2010). New gestalt principles of perceptual organization: an extension from grouping to shape and meaning. *Gestalt Theory 32*(1), 11–78. (Cited on pages 18, 19, 20, and 31.)

Pinna, B. and A. Reeves (2006). Lighting, backlighting and watercolor illusions and the laws of figurality. *Spatial Vision 19*, 341–373. (Cited on page 18.)

Poggio, G. F., B. C. Motter, and S. T. Y. Squatrito (1985). Responses of neurons in visual cortex (V1 and V2) of the alert macaque to dynamic random-dot stereograms. *Vis. Res. 25*, 397–406. (Cited on page 78.)

Pugeault, N., F. Woergoetter, and N. Krueger (2010). Disambiguating multimodal scene representations using perceptual grouping constraints. *PLoS One 5*(6), e10663. (Cited on pages 32, 75, and 105.)

Quigley, M., S. Batra, S. Gould, E. Klingbeil, Q. Le, A. Wellman, and A. Y. Ng (2009). High-accuracy 3D sensing for mobile manipulation: Improving object detection and door opening. In *Proc. IEEE Int. Conf. Robot. Autom.*, pp. 2816–2822. (Cited on page 111.)

Read, J. C. and R. A. Eagle (2000, January). Reversed Stereo Depth and Motion Direction with Anti-correlated Stimuli. *Vis. Res. 40*(24), 3345–58. (Cited on page 85.)

Read, J. C. A. (2010, April). Vertical Binocular Disparity is encoded implicitly within a Model Neuronal Population tuned to Horizontal Disparity and Orientation. *PLoS Comp. Bio. 6*(4), e1000754. (Cited on pages 7, 74, 75, 76, 79, 81, 82, 83, 84, 105, and 148.)

Read, J. C. A. and B. G. Cumming (2006, January). Does Depth Perception require Vertical-Disparity Detectors? *J. Vis. 6*(12), 1323–55. (Cited on pages 76, 82, and 148.)

Read, J. C. A. and B. G. Cumming (2007, October). Sensors for Impossible Stimuli may solve the Stereo Correspondence Problem. *Nat. Rev. Neurosci. 10*(10), 1322–8. (Cited on pages 80, 82, and 106.)

Read, J. C. A., G. P. Phillipson, and A. Glennerster (2009, January). Latitude and Longitude Vertical Disparities. *J. Vis. 9*(13), 11.1–37. (Cited on pages 78 and 79.)

Rensink, R. (2000). The dynamic representation of scenes. *Vis. cogn. 7*(1-3), 17–42. (Cited on pages 9, 21, 54, 58, 110, 118, 135, 171, and 172.)

Rodrigues, J., D. Almeida, J. Martins, R. Lam, and J. du Buf (2008). An integrated framework for combining gist vision with object segregation, categorisation and recognition. In *Proc. 31th Eur. Conf. Vis. Percept.*, Volume 37 suppl., Utrecht, The Netherlands, pp. 117–118. (Cited on page xi.)

Rodrigues, J. and J. du Buf (2006). Multi-scale keypoints in V1 and beyond: object segregation, scale selection, saliency maps and face detection. *BioSystems 2*, 75–90. (Cited on pages 58, 70, and 95.)

Rodrigues, J. and J. du Buf (2009a). Multi-scale Lines and Edges in V1 and beyond: Brightness, Object Categorization and Recognition, and Consciousness. *BioSystems 95*, 206–226. (Cited on pages 53, 77, 94, 95, 105, 106, 137, and 147.)

Rodrigues, J. and J. du Buf (2011a, December). A cortical framework for scene categorization. In *Proc. Int. Conf. Comput. Vis. Theory Appl. (VISAPP 2011)*, Vilamoura, Portugal, pp. 364–371. (Cited on pages 8, 111, 169, and 172.)

Rodrigues, J. and J. du Buf (2011b). A cortical framework for scene categorization. *Proc. Int. Conf. on Computer Vision-Theory and Applications, Vilamoura, Portugal, 5-7 March*, 364–371. (Cited on pages 22, 54, and 171.)

Rodrigues, J. and J. M. H. du Buf (2004). Visual cortex frontend: integrating lines, edges, keypoints and disparity. *Proc. Int. Conf. Image Anal. Recogn. Springer L*(1), 664–671. (Cited on page 65.)

Rodrigues, J. and J. M. H. du Buf (2008). Invariant multi-scale object categorisation and recognition. *Proc. 3nd Iber. Conf. Patt. Recogn. Image Anal. Springer L*, 459–466. (Cited on pages 58 and 70.)

Rodrigues, J. and J. M. H. du Buf (2009b). A cortical framework for invariant object categorization and recognition. *Cogn. Process. 10*(3), 243–261. (Cited on pages 23, 26, and 53.)

Rodrigues, J., J. Martins, R. Lam, and J. du Buf (2012). Cortical multiscale line-edge disparity model. In *Proc. Int. Conf. Image Anal. Recognit.*, Aveiro, Portugal. Springer LNCS 7324. (Cited on pages xi, 15, 75, 76, 97, and 172.)

Ross, M. G. and A. Oliva (2010). Estimating perception of scene layout properties from global image features. *J. Vis. 10*(1), 1–25. (Cited on pages 8, 22, 54, 111, and 171.)

Rubin, E. (1921). Visuell wahrgenommene Figuren. *Kobenhavn: Gyldendalske*. (Cited on page 18.)

Saalmanna, Y. B. and S. Kastner (2009). Gain control in the visual thalamus during perception and cognition. *Curr. Opin. Neurobiol. 19*(4), 408–414. (Cited on page 54.)

Scharstein, D. (2003). High-accuracy stereo depth maps using structured light. *Proc. IEEE Comp. Soc. Conf. Comp. Vis. Pattern Recogn.*, 195–202. (Cited on pages 86, 87, 92, 93, and 102.)

Scharstein, D. and C. Pal (2007). Learning Conditional Random Fields for Stereo. *Proc. IEEE Conf. Comp. Vis. Pattern Recogn.*, 1–8. (Cited on pages 86, 93, and 102.)

Scharstein, D. and R. Szeliski (2002). A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Int. J. Comput. Vis. 47*, 7–42. (Cited on pages 86, 93, and 101.)

Scharstein, D. and R. Szeliski. (2012). Middlebury Stereo Vision – Evaluation Webpage, http://vision.middlebury.edu/stereo/eval. (Cited on pages 86, 87, 102, 103, and 104.)

Scheenstra, A., A. Ruifrok, and R. C. Veltkamp (2005). A Survey of 3D Face Recognition Methods. In *Proc. 5th Int. Conf. Audio Video-Based Biometric Pers. Auth.*, Volume 3546, Chapter 93, pp. 891–899. Springer. (Cited on pages 140 and 141.)

Serre, T., M. Kouh, C. Cadieu, U. Knoblich, G. Kreiman, and T. Poggio (2005). A Theory of Object Recognition : Computations and Circuits in the Feedforward Path of the Ventral Stream in Primate Visual Cortex. Technical report, MIT-CSAIL-TR-2005-082, Cambridge, MA, USA. (Cited on pages 10 and 11.)

Siagian, C. and L. Itti (2004). Biologically-Inspired Face Detection: Non-Brute-Force Search Approach. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Work.*, Volume C, pp. 62–70. (Cited on pages 14 and 140.)

Siagian, C. and L. Itti (2007). Rapid Biologically-Inspired Scene Classification Using Features Shared with Visual Attention. *IEEE Tr. Robot. 29*(2), 300–312. (Cited on pages 22, 54, 111, and 171.)

Skottun, B. C. and R. D. Freeman (1984). Stimulus specificity of binocular cells in the cat's visual cortex: ocular dominance and the matching of left and right eyes. *Exp. Brain Res. 56*, 206–216. (Cited on page 77.)

Smith, A. T., M. B. Wall, A. L. Williams, and K. D. Singh (2006). Sensitivity to optic flow in human cortical areas MT and MST. *Eur. J. Neurosci. 23*(2), 561–569. (Cited on page 23.)

Socher, R., B. Huval, B. Bath, C. D. Manning, and A. Y. Ng (2012). Convolutional-Recursive Deep Learning for 3D Object Classification. In *Adv. Neural Inf. Process. Syst.*, pp. 665–673. (Cited on pages 111 and 135.)

Stockman, A., D. I. A. MacLeod, and N. E. Johnson (1993, December). Spectral sensitivities of the human cones. *J. Opt. Soc. Am. – A 10*(12), 2491. (Cited on pages 89 and 90.)

Stoerig, P. and A. Cowey (1997). Blind sight in man and monkey. *Brain 120*, 535–559. (Cited on page 28.)

Sun, T.-H., M. Chen, S. Lo, and F.-C. Tien (2007, August). Face Recognition using 2D and Disparity Eigenface. *Expert Syst. Appl. 33*(2), 265–273. (Cited on pages 141 and 159.)

Szeliski, R. (2011). Stereo correspondence. *Comput. Vis.*, 467–503. (Cited on pages 7 and 74.)

Tanabe, S. and B. G. Cumming (2008, October). Mechanisms Underlying the Transformation of Disparity Signals from V1 to V2 in the Macaque. *J. Neurosci. 28*(44), 11304–14. (Cited on pages 74, 75, 80, 106, and 176.)

Terzić, K., D. Lobato, S. Saleiro, J. Martins, F. Farrajota, J. M. F. Rodrigues, and J. M. H. du Buf (2013). Biological Models for Active Vision: Towards a Unified Architecture. In *Comput. Vis. Syst. – Spec. Issue 12*, Volume 7963 of *Springer LNCS*, pp. 113–122. (Cited on pages xii and 16.)

Thorpe, S. J. (2009, April). The speed of categorization in the human visual system. *Neuron 62*(2), 168–70. (Cited on page 1.)

Tistarelli, M., L. Brodo, A. Lagorio, and M. Bicego (2007). Recognition of Human Faces: From Biological to Artificial Vision. In *Adv. Brain, Vision, Artif. Intell.*, Volume 4729 of *Springer LNCS*, pp. 191–213. (Cited on page 141.)

Troncoso, X. G., S. L. Macknik, and S. Martinez-Conde (2011). Vision's first Steps: Anatomy, physiology, and perception in the retina, lateral geniculate nucleus, and early visual cortical areas. In G. Dagnelie (Ed.), *Vis. Prosthetics Physiol. Bioeng. Rehabil.*, Springer Science+Business Media, pp. 23–57. (Cited on page 23.)

Tsai, J. J. and J. D. Victor (2003, February). Reading a Population Code: a Multi-Scale Neural Model for representing Binocular Disparity. *Vis. Res. 43*(4), 445–66. (Cited on pages 76, 85, and 148.)

Tsalakanidou, F., D. Tzovaras, and M. Strintzis (2003, June). Use of Depth and Colour Eigenfaces for Face Recognition. *Patt. Recog. Lett. 24*(9-10), 1427–1435. (Cited on page 141.)

Ts'o, D. Y., A. W. Roe, and C. D. Gilbert (2001, May). A hierarchy of the functional organization for color, form and disparity in primate visual area V2. *Vision Res. 41*(10-11), 1333–1349. (Cited on page 88.)

van de Weijer, J., T. Gevers, and A. D. Bagdanov (2006a). Boosting color saliency in image feature detection. *IEEE Trans. Pattern Anal. Mach. Intell. 28*(1), 150–156. (Cited on page 58.)

van de Weijer, J., T. Gevers, and A. D. Bagdanov (2006b). Boosting Color Saliency in Image Feature Detection. *IEEE Tr. PAMI 28*(1), 150–156c. (Cited on page 59.)

Venkatesh, Y. V., A. A. Kassim, J. Yuan, and T. D. Nguyen (2012, October). On the Simultaneous Recognition of Identity and Expression from BU-3DFE Datasets. *Patt. Recog. Lett. 33*(13), 1785–1793. (Cited on pages 142, 158, and 165.)

Viola, P. and M. J. Jones (2004, May). Robust Real-Time Face Detection. *Int. J. Comp. Vis. 57*(2), 137–154. (Cited on page 149.)

Vogel, J., A. Schwaninger, C. Wallraven, and H. H. Bülthoff (2007). Categorization of natural scenes: Local versus global information and the role of color. *ACM Trans. Appl. Percept. 4*(3), 19. (Cited on page 8.)

Vondrick, C., A. Khosla, T. Malisiewicz, and A. Torralba (2012, December). Inverting and Visualizing Features for Object Detection. Technical report, M.I.T. (Cited on page 9.)

Štruc, V. and N. Pavešić (2009). Gabor-Based Kernel Partial-Least-Squares Discrimination Features for Face Recognition. *Inform. 20*(1), 115–138. (Cited on pages 131 and 153.)

Štruc, V. and N. Pavešić (2010). The Complete Gabor-Fisher Classifier for Robust Face Recognition. *EURASIP J. Adv. Signal Process. - Spec. Issue Adv. Image Process. Def. Secur. Appl. 1*, 31. (Cited on pages 130, 131, 132, 149, 151, 152, 153, and 165.)

Wade, A., M. Augath, and N. Logothetis (2008). fMRI measurements of color in macaque and human. *J. Vis. 8*(10), 1–19. (Cited on page 88.)

Wall, M. and A. Smith (2008). The representation of egomotion in the human brain. *Curr. Biol. 18*, 191–194. (Cited on page 24.)

Walton, G. E. and T. Bower (1993, May). Newborns Form "Prototypes" in less than 1 Minute. *Psychol. Sci. 4*(3), 203–205. (Cited on page 140.)

Wechsler, H. (2007). *Reliable Face Recognition Methods: System Design, Implementation and Evaluation*. Springer. (Cited on pages 151, 153, and 162.)

Wertheimer, M. (1923). Untersuchungen zur Lehre von der Gestalt. II. *Psychologische Forschung 4*, 301–350. (Cited on page 18.)

Yang, S. and X. Yao (2008, October). Population-Based Incremental Learning With Associative Memory for Dynamic Environments. *IEEE Trans. Evol. Comp. 12*(5), 542–561. (Cited on page 85.)

Yanulevskaya, V., J. Uijlings, and J.-M. Geusebroek (2013, January). Salient object detection: From pixels to segments. *Image Vis. Comput. 31*(1), 31–42. (Cited on pages 11, 124, and 171.)

Yin, L., X. Wei, Y. Sun, J. Wang, and M. J. Rosato (2006, April). A 3D Facial Expression Database For Facial Behavior Research. In *Proc. 7th IEEE Int. Conf. Autom. Face Gesture Recognit.*, pp. 211–216. (Cited on page 143.)

Zhang, X., L. Zhaoping, T. Zhou, and F. Fang (2012). Neural Activities in V1 Create a Bottom-Up Saliency Map. *Neuron 73*(1), 183–192. (Cited on page 6.)

COLOPHON

Final Version as of November 21, 2013 (`classicthesis` version 1.0).