UNIVERZA V LJUBLJANI

FAKULTETA ZA RAČUNALNIŠTVO IN INFORMATIKO

Matjaž Jogan

Postopno učenje razpršenih predstavitev za vizualno prepoznavo in kategorizacijo

Doktorska disertacija

mentor: prof. dr. Aleš Leonardis

Ljubljana, 2008

UNIVERSITY OF LJUBLJANA

FACULTY OF COMPUTER AND INFORMATION SCIENCE

Matjaž Jogan

Incremental Learning of Sparse Representations for Visual Recognition and Categorization

Dissertation

supervisor: prof. dr. Aleš Leonardis

Ljubljana, 2008

Povzetek

Predstavljamo model vizualne prepoznave in kategorizacije objektov na osnovi odkrivanja intenzitetnih in strukturnih ujemanj med interpretirano sliko in prototipno sliko objekta. Glavni prispevek te raziskave je predlog novega pristopa k vizualni kategorizaciji z uporabo sinhronega hierarhičnega iskanja ujemanj, pri čemer se visokonivojska ujemanja konstruirajo pragmatično, v več nivojih združevanja, izbire in inhibicije. Z iskanjem ujemanj lahko kategoriziramo objekte brez pretirane uporabe učenja, temveč le z iskanjem podobnosti med objekti. Kategorije objektov torej niso opredeljene z naborom značilnic, ki bi omogočale optimalno kategorizacijo, temveč kot mrežni sistem podobnosti med objekti.

Model uspešno združuje kategorizacijske metode, osnovane na lokalnih značilkah ter hierarhičen pristop združevanja, ki odpravlja siceršnjo odvisnost lokalnih metod od lokalnih intenzitetnih vzorcev, saj omogoča postopno gradnjo značilk, ki postajajo deskriptivne šele na stopnji, ko opisujejo večji delež objekta. Združevanje se prične na nivoju lokalnih značilk, ki opisujejo informativna in stabilna lokalna področja. Značilke v lokalnih soseščinah nato v hierarhičnem postopku združujemo v pare, ki preko nivojev hierarhije pridobivajo in prenašajo informacijo o strukturnih značilnosti lokalnih konstelacij. Lokalne soseščine se pri tem sistematično povečujejo, dokler ne pokrivajo celotne površine objekta. Združevanje značilk je uravnavano s sprotnim iskanjem ujemanj in inhibicijo značilk, ki ne vodijo k potencialnim ujemanjem na višjem nivoju. V nasprotju s hierarhičnimi modeli prepoznave, ki eksplicitno modelirajo hierarhično predstavitev kategorije objektov, temelji naš pristop na sprotni konstrukciji hierarhičnih značilnic.

Učenje kategorij je implementirano kot sprotno učenje lastnosti prototipa, in kot učenje in prilagajanje parametrov združevanja. Osnova za opis lokalnih področij je nabor filtrov, ki so pridobljeni z učenjem maksimalno razpršene predstavitve intenzitetnih zaplat. Postopek učenja je implementiran kot analiza

neodvisnih komponent (*Independent Component Analysis, ICA*) slikovne matrike zaplat enega ali večih objektov. Deskriptivnost nabora filtrov namenoma omejimo tako, da izračunamo le manjše število ICA komponent. Opis lokalne zaplate torej le grobo opredeljuje slikovno vsebino, nosi pa dodatno informacijo o generični skali in orientaciji regije. Da bi zmanjšali redundantnost take predstavitve, regije, ki ležijo na robovih, združimo z uporabo temeljnih pravil Gestalt. Z analizo združevanja značilk v prototipu se sistem nauči optimalnega strukturiranja prostora geometričnih konceptov, ki pogojuje parametre združevanja.

Hierarhično združevanje je paralelen proces, saj posamezni lokalni procesi združevanja, iskanja ujemanj in inhibicije ne vključujejo globalne informacije. Pari značilk se združujejo le v lokalnem območju, katerega oblika se sprotno prilagaja glede na ujemanja, in tako simulira nizkonivojsko pozornost. Simultana širitev lokalnih področij ter velikosti regij višjenivojskih značilk privede do nabora visokonivojskih značilk, ki pokrivajo večji del prototipa. Ker konstrukcija teh značilk poteka neodvisno, vsaka od značilk opisuje eno od ujemajočih kombinacij značilk na nižjih nivojih. Število visokonivojskih značilk odseva podobnost med prototipom in regijo ujemanj v interpretirani sliki.

Model preizkusimo na problemu kategorizacije segmentiranih objektov v zbirki *ETH80*, ter kategorizacije in lokalizacije objektov na nesegmentiranih slikah objektov različnih kategorij *Caltech-101*.

Rezultati dokazujejo uporabnost metode za kategorizacijo objektov v kanoničnem pogledu. Poleg same kategorizacije lahko metodo uporabimo za detekcijo potencialnih ujemanj med objekti na več nivojih podrobnosti. Zato zaključujemo z ugotovitvijo, da je metoda predvsem primerna kot sestavni del kognitivnega sistema, kjer so zahteve bistveno drugačne kot v primeru iskanja po slikovnih zbirkah.

Ključne besede: računalniški vid, vizualna kategorizacija, vizualno ujemanje, vizualno učenje, hierarhične metode

Abstract

We investigate a framework for visual object categorization in artificial cognitive systems that is based on a discovery of appearance and structural similarities between object exemplars and prototypes that represent a category. Our main contribution is a novel approach for visual categorization of objects by synchronous hierarchical matching to a prototype, where high level matches between an object and a prototype are gradually discovered through several steps of binding, selection and inhibition. We show that categorization can be achieved without an excessive collection of evidence or learning from examples. Instead, by discovering commonalities between an object and a prototype, categorization can be based on the notion of "family resemblances" that does not require an explicit modeling of features to define a category.

We combine the successful methods for categorization based on local features with a flexible and general prototype matching framework that balances the prevalent dependence of local methods on patch appearance by a gradual construction of high level features in a hierarchical feature binding process, where the descriptive power of features gradually increases from localized features with little or no structural information to complex composite features that span wider areas, until they encompass the whole object.

We start with a low level description, which defines relatively stable local regions of interest (ROI) based on their appearance. Features are then dynamically constructed, or activated, in order to search for possible matches between the view being interpreted and a prototype view of an object. The composite features describe geometrical and photometric properties of a local area that expands, until a high level focused response, with a *receptive field* that potentially covers the whole object, is constructed. The construction of high level features is not steered by matching to a predefined set of features, but is rather conditioned by a hierarchical matching and inhibition of ad-hoc features. The advantage of such an approach is that we do not have to define a category with respect to a set of

learning examples, since objects from the same category can be articulated using many different local geometrical and appearance matches.

In order to characterize the appearance variation within local regions, we learn a codebook of ICA filters which impose a maximally sparse response; we intentionally keep the number of filters in the codebook low. In that way, local regions are strongly informed in terms of their structural (orientation, scale) properties, but carry only a basic information on appearance. Clusters of local features that conform to a subset of Gestalt rules are then grouped. This reduces the redundancy and provides a raw description of linear structures and equi-curvature areas. Based on a prototype image, we also learn the structure of the geometric conceptual space.

The hierarchical binding is a massively parallel process that does not require global information. Twoplets of features are matched to twoplets of features, where the shape of the receptive field is conditioned to simulate attentive processing. The result is a set of matches of composite features at level the highest level of the hierarchy, where each of the features can be tracked to the constituent features at different levels. The decision whether the object categories match can be done based on the number of high level matches.

We evaluate our framework on two domains: the first set of experiments assesses the performance of the hierarchical categorization on segmented images of objects that come from multiple categories, and the task is to find the closest match to a given image. The second set of experiments assesses the performance of hierarchical categorization of objects in occluded images. We use the *ETH80* and the *Caltech-101* public databases.

The results show that our framework achieves a reliable categorization of objects in a canonical view. Besides categorization, it enables a reliable detection of an object as a whole, and of matching the object parts at different resolutions. We argue that the model could be integrated in a cognitive framework, where several other sources of information can be used to establish a stable interpretation of the scene. We also claim that categorization by visual matching to a prototype has biological plausibility, and that our architecture implements some ideas that originate in enactive cognitive science.

Keywords: computer vision, visual categorization, visual matching, visual learning, hierarchical methods

Contents

List of Figures

List of Algorithms

1

Introduction

The goal of the thesis is to investigate a possible implementation of visual object categorization of objects in an artificial cognitive system, that would be based on a discovery of appearance and structural similarities between object exemplars and prototypes that represent a category. The main contribution is a novel approach for visual categorization of objects by synchronous hierarchical matching, where high level matches between an object and a prototype are gradually discovered through several steps of binding, selection and inhibition. We show that categorization can be achieved without an excessive collection of evidence or learning from examples. Instead, by discovering commonalities between an object and a prototype, categorization can be based on the notion of "family resemblances" [166; 167] that does not require an explicit modeling of features to define a category.

1.1 Categorization as a challenge

Categorization has been for a long time the holy grail of cognitive science. In research on artificial cognitive systems, categorization is one of the key functionalities that enables intelligent acting and deliberating on the basis of data abstraction. When the data comes from devices that artificially sense the tangible world, we can name the process as *categorical perception*. Visual categorization is a subproblem of categorical perception, where scenes are interpreted in terms of categorical abstractions using vision.

For instance, a simple report like *"I see there is a chair in the office"* implies that a categorization was performed at multiple levels: the room, or the environment, was categorized as *"an office"*. One of the objects was recognized as *"a chair"*. Note that this label could be used for any chair, and not for the specific instantiation. Further, the spatial insertion of the chair was categorized as a situation where *"a ... is in the ..."*, therefore expressing a rather abstract spatial relation (as opposed to a more machine–like interpretation *"there is a chair*

at position (X,Y)", and implicitly reporting it as a predictable situation, which would not be the case for the sentence *"I see there is a levitating chair in the office!"*.

A key functionality of categorization is therefore to recognize a sensory phenomenon as a member of a category, or to attribute it a label that is consistent with its generic, and not only specific qualities. For visual categorization, the task is to categorically perceive a scene by using only its *visual* properties. As we however argue later in the thesis, it is not feasible to implement artificial models of perception that use only one sensory modality when, as it should be the case in cognitive systems, multiple modalities are available, together with functionalities such as contextual awareness, self-awareness, or embodied computation. Nevertheless, at least some aspects of visual categorization indeed use mainly the visual information, which makes them worthwhile to investigate and to discover.

The purpose of object categorization is to support categorical perception of objects. Given a representation of a number of object categories, the task is to infer the category membership of novel objects that are presented to the system. If the system knows about cats, cars and cups, it should label all objects that are cats, cars, and cups as members of the respective category. This task differs from object recognition where, given a representation of a specific object, the task is to recognize a novel instantiation of the same object.

If we compare the task of object categorization to object recognition, i.e. recognizing an object as a member of a category versus recognizing it as an individual instantiation, we would be tempted to say that categorization is easier than recognition. It is of course true that, in the case that all objects are labeled by their category, and that we can recognize all objects, we can also categorize them by looking up their respective labels. In this sense, categorization is a subordinate problem to recognition, and could be solved by defining a category *in extension*, e.g. by listing all existing objects that belong to a category. This is however certainly not a natural solution, although its technical implementation is viable, especially when considering the power of the Internet as a massive storage and retrieval machine [92].

The other way to define a category is *intentionally*, by defining, learning, or discovering properties that things must have to fall in a certain category. In recognition, the best features to search for are the ones that discriminate a certain object from all other objects. In categorization, however, features should discriminate between categories of objects, where objects that fall in the same category can be significantly different, or could be similar only

by specific cues, e.g. by shape, by texture, or by appearance. Furthermore, the importance of cues and features could be different for each specific exemplar in the category.

The set of defining features has to account for both *inter-class* and *intra-class* variability. For example, while the features should be general enough to cover all the different views of exemplars in a category, they should also be specific enough to discern all the objects in this category from objects in all other categories.

Other problems that are encountered in categorization are similar to those in exemplar recognition, and only add to the difficulty of the problem: we have to handle variation in illumination, scale, viewpoint, articulated motion of an object, and occlusion.

1.2 Problem formulation

In this work, we investigate how to perform visual categorization without an excessive modeling of the feature set that defines a category. We therefore narrow down the problem of visual object categorization to categorical matching between a prototype that represents a category and an unknown object. In contrast to the matching methods that are used for object recognition, or wide baseline matching, categorical matching has to establish correspondences between objects that are not inherently similar in appearance, nor in shape, however they still share some of the appearance or structure similarities which have to be discovered. We therefore construct ad-hoc high level features for each round of matching. The matching is conceived in a way that integrates local shape and appearance in a **general framework for hierarchical feature binding, matching and inhibition that constructs informative high level features by hierarchically binding low level features with weak information content**. We investigate **the role of learning in prototype based representations** and show how **learning a sparse representation of low level appearance** can steer and support the generation of plausible features at higher levels of the hierarchy. We also investigate how to learn, or adapt the low level feature detectors and the geometric properties to optimally encode the high level features. Finally, we show how the matching framework can be used for categorization by an evaluation of successfully constructed high level matches.

1.3 Motivation

This work is inspired by the optimistic results of many successful attempts to categorization in the computer vision community over the last few years. Researchers presented impressive results, both on *one-category* and *multi-category* categorization, and significant progress has

been achieved in our understanding of how to organize and use the appearance information using techniques from statistical learning and pattern recognition.

Categorization has been attacked with many different approaches: one possible, and today probably the most common approach, is to extract appearance based features, represent them in a feature space, and to apply a suitable pattern recognition technique. The other possibility is to carefully design representations and stages of categorization to reflect some specific aspects of visual categorization. For example, as objects are composed of parts, one might design a representation that reflects the property of compositionality. Further, volumetric objects can be modeled as compositions of 3-D generic shapes. An example of such approaches are e.g. the reconstruction methods, where an internal model of the world is reconstructed and compared to stored models. Another approach is to research methods that naturally learn regularities in the visual stimuli, and can devise appropriate building blocks that reflect the natural structure of categories, such as hierarchical compositionality, or feature sharing. These methods are particularly auspicious in terms of autonomy, robustness, and scalability.

There is however still a lack of insight on how to implement categorization abilities that would meet the requirements of advanced cognitive systems, i.e., systems with a significant level of autonomy, that operate over a large timespan and in an unpredictable environment, where novel information on categories has to be continuously integrated and apprehended.

The significant progress that has been made in the last decade in the area of object recognition using local features resulted in algorithms for repeatable detection of stable local regions that are invariant to variation in scale, rotation, and affine transformations [85; 87; 100], see Section 4.4. Our major motivation is to use the strong advantages of these methods, and to integrate them with a flexible and general prototype matching framework. Prototype based modeling and matching has received attention mostly in shape matching, but was only sparsely addressed in the visual learning community. We revise and combine these two approaches with the goal to alleviate the prevalent dependence of matching on local appearance or on global shape; we instead propose a hierarchical feature binding process, where the descriptive power of features gradually increases from low level, localized features, with little or no structural information, to composite features that span wider areas, until they encompass whole parts or objects. We also aim at a massively parallel and distributed computation model, that would not require global information to steer to a viable solution. We achieve this by modeling the binding at the level of local neighborhoods that gradually expand their receptive domain.

The gradual acquisition of the complexity of features through the hierarchy has an analogy in written language: while at the lowest level, isolated letters in a sentence carry almost no information about the encoded message, they gradually participate in the binding process. By binding immediate local context and letters across several positions, single words are inferred, which, at a higher level infers sentences and thoughts.

There are several interesting characteristics to note in this analogy. The first one is the indeterminacy of the partial sub-bindings: while words function as identifiers of the instantiated or abstract entities, their precise scope can only be inferred once they are put in context within the sentence, or, in some cases, even only when the whole message is understood. The second interesting property is the increased determinacy through different levels of interpretation. As it is well known, a text where some of the letters in words are interchanged or omitted, can be easily read without noticing its fallacies at the lowest level. This is mostly due to the fact that the text is read in synchrony at multiple levels, where upper levels impose the most favorable solutions at the lower levels.

The conjunction is that letters do not have to be fully recognized and that letters may bind not only to syllables, but also within and across words, settling for a plausible global explanation. The higher the hierarchy we go, the lesser is the invariance to such fallacies, as e.g. complete sentences have a much stronger tie with their potential meaning than words or letters (also meaning that the paradigmatic reading gradually takes over the syntactic interpretation).

In our framework, the intermediate bindings are not important in any other way than in the sense that they, under a set of constraints, lead to successful constructs at a higher level representation. This is in contradiction with the current trends in part–based categorization, where a search for constituent parts that can be shared among categories is seen as an essential basis for successful categorization. As we will show, the binding process is most efficient in binding partial structures, as they often exhibit a larger intra-category similarity in configuration, which is more prominent than the similarity of certain parts.

Another motivation is to implement a tentative framework as a platform to investigate the possibility of using a *prototype based* representation of categories that could replace the learned feature based representations that are commonly in use today. The prototype based representation is used to synchronize the binding at different levels of the hierarchy in order to search for a proper interpretation, and is tightly coupled to the overall architecture. While starting from a basic visual information, the binding hierarchy expands its implicit potential

for guiding the categorization process to a final result. The representation and the processing model therefore function in a *constructivist* fashion, where the majority of the information is derived in the process. Alternatively, such a representation could be easily encoded without any reference to appearance, for example as a code of perceptual actions that are needed to search for visual associations.

1.4 Our contribution

In this work, we perform visual categorization of objects using a prototype based representation. We investigate the role of learning in such representations, and show how a sparse representation of low level appearance can steer and support the generation of plausible features at higher levels of the hierarchy. We propose a framework that is based on hierarchical matching of increasingly complex sparse features. Features are locally binded, matched, and inhibited, until the bindings reach a high level of complexity and cover the whole object area. Our **original contributions to science** can be summarized as follows:

- We propose a **hierarchical matching** approach that is performed as a **distributed** and **synchronous** process, where local bindings are simultaneously constructed and then matched between the prototype image and the image being interpreted at different levels of the hierarchy. We therefore approach the problem with a brute-force, local to global search procedure that is designed to find *any possible match* that could support a similarity between the matched counterparts. We show how one can alleviate this otherwise intractable problem by dividing the task to separate computational subtasks (locality), by a simultaneous evaluation of partial results (synchronous inhibition), and by a gradual increase of complexity (hierarchical binding).
- We propose a **novel learning method** for the **estimation of a sparse codebook representation of local appearance**. The method is based on Independent Component Analysis of local regions of interest, and efficiently estimates an optimal codebook that exhibits a maximally sparse response pattern.
- We propose to adapt the parameters of the hierarchical binding framework by **learning the structure of the geometric conceptual space** based on the geometrical properties of the prototype. This step provides a conceptual representation of local geometric relations that adequately represents the variations with respect to a certain category of objects.
- In our framework, the **complex features** at higher levels of the hierarchy are **generated in an ad-hoc manner**, and are not part of the prototype based representation. We

show that objects can be efficiently categorized even without an explicit representation of complex structures and parts. We also argue that a search for similarities is a more natural approach to categorization that a match to a learned set of features.

- The major characteristics of our work (image in a canonical view as a prototype, the ad-hoc nature of high level features, synchronous distributed processing) partly support the **enactive theories of perception**, and offer a possible interpretation of the **role of imagination in cognitive systems**.
- We test our framework for categorization performance on two image databases. The experimental results show that our method can efficiently categorize objects in a canonical view. We also show that a high number of structural matches can be found automatically at different levels of detail.

We start with a low level description, which defines relatively stable local regions of interest (ROI) based on their appearance. Features are dynamically constructed, or activated, in order to search for possible matches between the view being interpreted and a model view of an object. The composite features describe geometrical and photometric properties of a local area that expands through the hierarchy, until a focused response, with a *receptive field* that covers the whole object, is constructed. The construction of high level features is not steered by matching to a predefined set of features, but is rather conditioned by a hierarchical matching and inhibition oh ad-hoc constructed features. The advantage of such an approach is that we do not have to define a category with respect to a set of learning examples, since objects from the same category can be articulated using many different local geometrical and appearance matches.

In order to characterize the appearance variation within local regions, the system learns a set of filters which impose a maximally sparse response; we intentionally keep the number of these filters low. In that way, local regions are strongly informed in terms of their structural (orientation, scale) properties, but carry only a basic information on appearance. The training stage therefore consists of an adaptation of the low level of the hierarchy to a certain prototype. Furthermore, we learn an optimal quantization of the geometric conceptual space, i.e. the angle, length and size attributes that steer the construction of novel features.

Clusters of local features that conform to a subset of Gestalt rules (being essentially co-centric or co-linear) are then grouped. This reduces the redundancy and provides a raw description of image elements, such as linear structures and equi-curvature areas.

The so derived features are then associated with the prototype: we bind features that fall within a receptive field, targeting an upper level where features are grouped in pairs and are determined by their low-level identity and by their local relative geometry. By simultaneous expansion of the receptive field and matching of binded features, only those pairs that match become potential candidates for a match at a higher level. We repeat the process until the level where the binding stabilizes in large receptive fields, typically outlining the object, or until the process ends because of a lack of successful matches.

The result is a set of matches of composite features at level N, where each of the features can be tracked to the constituent features at levels $N-1, N-2, \ldots 0$. The decision whether the object categories match can be done based on the number of high level matches. Since the number of features partially depends on the structural properties of the objects, the setting of fixed thresholds could lead to wrong decisions. The matches however provide a solid basis for other types of verification (using context, segmentation etc), which are not addressed in this thesis.

The problem of matching can be also seen as finding a subgraph within a noisy annotated graph, where annotations denote spatial relations and appearance characteristics between nodes, and is therefore extremely complex and time consuming. As we always grow the hypothesis based on independent processing of local neighborhoods, the method is inherently parallel, and can be distributed to any number of processors, achieving thus a constant time complexity with a fixed number of levels. If we want to check for n categories, we have to perform n sequential queries, resulting in a linear time complexity with respect to the number of categories, or n parallel queries, achieving a constant time complexity with respect to the number of categories.

1.5 Outline

The thesis is structured in three parts.

Part I is an introduction to the field of categorization, and a concise overview of the related work in the field of computer vision. We highlight some of the important aspects of categorization by an overlook of philosophical, cognitive neuroscience, psychology and cybernetics theories. A thorough treatise of the subject would require a significantly broader discussion;

we therefore focus on those issues that are important to contextualize our work and to highlight our major contributions in terms of their agreement or disagreement with the theoretical models. The reader with the main interest in the description of our approach may easily skip the initial chapters and proceed directly to Chapter 5.

Chapter 2 begins by outlining the key aspects of categorization as a problem sui generis. Starting with basic questions about the existence of categories, what categories are, and how are they grounded, we gradually narrow the discussion to problems that are important for categorization of objects. We focus in particular to different ways of determining the properties of a category, and introduce the Prototype theory, which we adopt as a basis of our framework.

Chapter 3 continues with a discussion on different models of representation in cognitive systems. We review the conceptual frameworks that were developed in the field of computational cognitive science, and introduce a taxonomy of representational models. In particular, we focus on the issues of *representation*, *symbolic* vs. *non-symbolic* paradigms, *object awareness*, and *feature binding*, which all touch upon questions that we tackle with the design of our architecture.

Chapter 4 reviews the approaches to object categorization that have been developed and tested within the field of computer vision. After a historical overview which starts in the early years of computer vision research, we focus to hierarchical methods and shape matching methods, which adopt some similar strategies as we do.

Chapter 5 briefly describes our approach and contextualizes it with respect to the releated work.

Part II is a detailed description of our framework for hierarchical categorization as a progression from low level processing through binding and learning to final categorization. In Chapter 6, Chapter 7, and Chapter 8 bring a detailed description of our method.

Chapter 6 describes how we process the raw image data and focus the processing to salient and stable local regions. Here we introduce the low level dictionary of Independent Component codes, which sparsely encode the variation in appearance within local regions. We also show how we create the discrete geometric parameters that encode the local geometry of feature twoplets.

Chapter 7 continues with a description of the hierarchical matching procedure, while Chapter 8 describes how the hierarchical binding is used for matching between a prototype image and test images.

Part III describes the experimental settings and results that were obtained using our framework. Chapter 9 describes the experimental results. We finish with a critical overview, description of failure modes, and of potential extensions of the method in Chapter 10.

Part I

Preliminaries

2

Categories

In this section we review some common views on categories, on what they are, and on how they can be defined. We focus on the difference between *categories by definition* and *prototype centered categories*, and on how objects can be categorized within a system of categories.

2.1 What is a category?

In general, categories stand for (systems of) generalizations across entities. In western philosophy, the debate on systems of categories can be tracked through most of its historical timeline. A very superficial observation of the various theories that emerged during the last three millennia could reveal two major paradigms: according to a more traditional viewpoint, categories stand for *things that are*, demarking thus different categories of things that constitute an objective reality. An opposite viewpoint would be less objectivist in the sense that it views categories as generalizations that are pertinent to a subjective conceptual system. In this view, categories are an abstraction of the sensory world according to a system of thoughts, past experiences, or language [149].

The question of categories is also related to the more metaphysical problem of existence of *universals* as universal characterizations of many objects, for example "greenness" that characterizes objects that are green, or "appleness" that characterizes all individual apples. The dichotomy between universal labels and particular entities that are attributed "names" had played a central role in ontological discussions already in ancient philosophy. Plato's answer constituted a *realist* view, which states that universals *exist independently* of particulars. Plato's *Forms*, or ideas, have a primacy over things, as they exist prior to instances,

and explain their occurrence. Things therefore merely partake of a Form by possessing its qualities.

A different form of realism was founded by Aristotle. His taxonomy of entities in the world consists of ten exclusive categories:

1. Substance	3. Quality	5. Place	7. Posture	9. Action
2. Quantity	4. Relation	6. Date	8. State	10. Passion

Objects fall within substances, which are divided to two sorts: *primary* substances denote individual instantiations, e.g. an individual man, animal, or object. *Secondary* substances (universals) are the species and genera of individuals. A more interesting aspect of the system is the way how primary and secondary substances are defined. Aristotle derived his classification system by a process of questioning on what can be asked about certain entities. In his view, secondary substances (e.g., man, animal) are those that can be SAID OF a subject, but can not be not PRESENT IN, e.g. "**Man** can be SAID OF Aristotle". Primary substances are those that both can not be SAID OF a subject, nor can they be PRESENT IN. Everything that can be PRESENT IN is a non-substance; if it can also be SAID OF, it stands for a secondary non-substance, e.g."**White** is SAID OF of this color" and "**White** is PRESENT IN a horse". Primary substances and non-substances are therefore particulars that are at the roots of a category tree ("this color", "Aristotle"). As it follows from Aristotle's method, universals are always SAID OF something particular OR of another universal. In contrast to Plato, Aristotle postulates a primacy of primary substances: universals can exist only by being instantiated [77], which reflects his empiricist beliefs.

Skeptic about the existence of an intrinsic division of reality, several philosophers established paradigms alternative to realist views. Rationalists, notably Spinoza and Descartes, already separated the subject from the reality by interfacing the direct access to the world by the agent of reason, which is also the fundamental aspect of cartesian dualism. Kantian *conceptualism* [149] treats categories as divisions governed by concepts that reside in the mind. The prerequisite for any (categorical) apprehension is that the subject experiences *himself* as existing in the act of apprehending [99]. The unity of consciousness is achieved simultaneously with the unity in the world (Kantian idealism), however it is always obtained as a subjective synthesis, via a subjective construction. Some aspects of conceptualism can be derived from empiricism, as both suggest that universals are thoughts or ideas constructed by mind [77].

Still in this line of thought, Husserl however shifted the focus from the subjective synthesis to a subjective reflection: categorization is a study of essences, based on essential insights about the types of meanings and correlative types of things [149]. In his *descriptivist* view of categories, he distinguishes between *meaning categories* and *formal-ontological* categories. While meaning categories relate judgments and include e.g. forms of conjunction and disjunction, formal categories relate objects and include notions such as part, whole, and relation. His method puts emphasis on *how* subjects *intentionally*, directly perceive, and, in a way, constitute objects by internalizing their objective nature through observation. In contrast to Kant, he gives the world a primacy over ideas, and states that the sole access to meaning is by a phenomenological study of the individual mental acts (of perception) that create meaning.

As an even more radical depart from realism, *nominalism* claims that there are no universals in the sense of entities, and recognizes only individuals.[1] As, in general, there are only words that refer to particular exemplars, the membership to a category is decided by *similarity* (see Section 2.3).

There are many important aspects of these theories that are directly related to the development of artificial systems that are capable of visual categorization of objects. For example, one could ask whether a cognitive system shall operate with objective categories that exist independently of a cognitive system, or whether it shall discover or learn its own categories. The first solution seems to require hard-wired knowledge to be programmed, or otherwise designed in some way by an external agent, e.g. a programmer. The second solution requires the system to be capable of an incremental and robust discovery of categories, and of a progressive acquisition of a subjective categorical system.

Another important question is how mechanisms for recognition of universals and those for recognition of instantiated objects are implemented – can this be achieved with a single mechanism or are there two dedicated subsystems for each of the tasks [78]? Is the implementation related to a taxonomy of categories, and, on how many levels can we name objects? And, once we decide on what categories relate to, how can we define them so that the system will be able to interpret scenes in terms of visual categories in sight?

Finally, categorical systems already imply a view on how cognition works, and how artificial cognition could be implemented. The realist views seem coherent with symbol systems,

[1] In the strict first-order logic sense, see [76].

where categorical perception and reasoning is governed by a fixed set of predicates. On the other hand, Husserl's system could be interpreted as advocating an embodied perception which functions as a "study" of elementary physical and chemical properties in the world.

2.2 A taxonomy of categories

In relation to how people spontaneously *name* objects, Jolicoeur, Gluck and Kosslyn [63] differentiate between *basic level* and *entry level* categories. Basic level categories were postulated already by Rosch [127], who provided psychological evidence that people name objects at an intermediate level of abstraction, which signals an appropriate degree of detail, considering a general context. For example, to name a chair as "a chair" is in agreement with the general naming of chairs, while calling it "furniture" or "office chair" is appropriate only in a specific context. However, objects that are atypical members of a category, for example "a penguin" are usually named at the entry level, which is subordinate to the basic level, therefore "a penguin" is called "a penguin", and not a "a bird". While at the basic level objects exhibit a level of similarity in shape, sensory-motor affordances, or other relevant cues (named *cue validity* by Rosch), this is not necessarily so at the entry level. In general, the entry point for a given object covaries with its typicality, which affects whether or not the object will be identified at the basic level [63]; atypical objects have their entry point at a level subordinate to the basic level.

Most of visual categorization systems consider the basic level, which is meant as an intermediate abstraction of visual concepts. Atypical objects should be therefore categorized at the entry level; for example, as there is not enough visual similarity between a penguin and a bird, penguins should form an entry level category. It is however an important question which is the acceptable level of visual similarity across the basic level to constitute a visually relevant category. For example, different species of dogs have only few common properties, maybe the only reliable property of them being that they are all four legged. Such categories can be defined on a basic level either by a property which is discriminative only in case that there are no other similar categories (e.g. if dogs are the only four legged furry animals), or if categorization is trained on so many exemplars that most of the features of subspecies are encompassed. It is important to note that visual perception alone is not efficient for a large number of even such common categories as "chairs", where a huge variety of visual appearances could be summarized by a simple sensory-motor affordance of "sitting".

2.3 What defines a category?

Both the traditional (Aristotelian) view and in part also the conceptualist views postulate that categories have more or less clear boundaries. Properties of objects that are mediated by perception result in concepts[1] that are based on a discovery of a number of common features. As object naming and deliberating about their (categorical) properties, such as position, quality or quantity, is tied to language, the nature of category systems that we adopt seems tightly connected to the relation of language to the world and the relation of language to thought. Interestingly, most of the classical theories took language for granted, and it was only in the late 19th century that language entered the spotlight of philosophical, phenomenological and cognitive sciences.

An opposite take on categories is that they are constructed *ad-hoc*, and are task and situation related. For example, when looking for a certain tool to accomplish a task, objects can be on-the-fly categorized by their affordances. It is clear that this view on categories relies more on a subjective and active stance that a cognitive system is establishing via interaction, judgment, and comparison of similarities.

These two opposing views are partially reflected in the major paradigms that are currently used to explain how categories are defined: the definition based approach, the association based approach, and the prototype theory, which we also adopt in our work.

2.3.1 Categories by definition

The basic line of western thought was for long considering language as a phenomena that arises from naming essential *things* or *ideas*. For example, Plato claimed that for a word to have a meaning, there must be some essential trait that is associated to that word. Aristotle would use such meanings to *intentionally* determine the features that define a category. An Aristotelian system of categorization can be therefore regarded as being *definition based*, or based on a *feature based* representation. We can track down this notion of categories in many of today's frameworks for categorization in artificial cognition. In systems that use fixed, predetermined models, the important features that define a category are programmed, or modeled in advance, e.g., by specifying the characteristic shape of objects of interest, or by specifying the constituent parts and their spatial relationships. Other frameworks use perceptual learning to learn the set of features by, e.g. optimizing for the intra-category and inter-category discriminability, or by simply learning the most representative features that

[1] Often concepts and categories are considered to be "*... flip sides of the same coin ... a concept is an idea,that characterizes a set, or category of objects" [44; 139].* In Aristotelian theory, a concept is defined by a set of necessary and sufficient properties [44].

are common to a category of objects. Here, the intentionality is reflected in the choice of the algorithm, in the selection of category members, or in the pointing out relevant features, or relevant cues.

2.3.2 Categories by association

Wittgenstein formulated two different theories of thought that have become strongly influential among cognitivists. According to some readings of his first book, the Tractatus Logico-Philosophicus [165], language stands for the world by depicting it. Propositions function like pictures, and the way that propositions are formed reflects the structure of pictures. The *pictorial* theory of thought that he proposes postulates that pictures are vehicles of thought, and are therefore central to our thinking - we use pictures to *depict* the facts to ourselves.

In his later works [166; 167], the pictorial theory, or the notion of picture as a tool, was replaced by a notion of language as a tool, where the meaning of words is determined according to how the words (and propositions) are being used. The language, and therefore the world, is therefore structured by indulging in *language games*, which stand for the whole spectra of conversational and other activities. This makes language an unbounded and unstable phenomenon that represents a referential framework for everything we can deliberate on.

In picture theory, it is important that pictures refer to *facts*, not things, and facts are those who inhabit the logical space. It seems however that both theories propose that meaning depends on articulation, rather than on representation. This directly applies to categorical systematization: Wittgenstein's system was centered around the central notion of *family resemblances*, as the type of relationship that is primarily exhibited in language, and is related to the way our world is structured. In particular he rejects the idea that objects are named according a number of essential features, but states that objects are connected by a series of overlapping similarities, and that no feature needs to be common to all exemplars. He also rejected the notion of stable boundaries between "familiar things". As in his whole concept of language, boundaries are drawn only as part of the *language game*.

Familiarity of objects can be discovered by a process of association—either by a discovery of similarity between objects and memories of objects seen in the past, or by direct matching of objects in the field of view [92]. Naming of things is therefore not a process of answering the question "*What it is?*", but rather "*What is it like?*" [6; 92].

2.3.3 Prototype theory

Although the original idea of family resemblances did not include the notion of similarity with respect to a specific prototype, but postulates a rhizomatic network of similarity relations, it represented a departure from definition-based categories, and had an important influence on the *Prototype theory*, which was first formulated by Eleanor Rosch [126; 127]. Rosch claimed that categories can be only regarded to in terms of *degrees of membership*, expressed by distance to specific prototypes for which there exists a consensus.

In short, there are more and less representative exemplars that are connected to a category; the most representative can stand for a category as a prototype. Rosch proposed two different criteria for a prototype: in her first study, the prototype stimulus was the one that was the first one to be associated to that category [126]. In her later studies [127], the prototype is the most central member of a category, the one which is more representative for that category than other exemplars. As all the members of the category are not equally representative, the prototype theory leads to the notion of *graded* categories, and *graded* sets.

Related to the prototype theory is also the notion of *canonical perspective*, or *canonical view* [111]. Psychological experiments show that knowledge of objects, and the ability to identify and categorize an object, is maximally accessible from a specific perspective, which also embodies the maximum perceptual information content. If a prototype is therefore to be remembered, compared, or retrieved by means of imagery, the maximum perceptual grip is offered by canonical views of that object.

2.4 Similarity of exemplars

But how can we evaluate the *similarity* between an exemplar and a prototype? One of the interesting problems of similarity was posed by Watanabe in the form of an *Ugly Duckling Theorem* [26; 163]:

> Any two objects are as similar to each other as any two other objects, insofar as the degree of similarity is measured by the number of shared predicates.

According to Edelman [26], this problem directly leads to a weighing of the predicates according to their relevance. As he puts it, "... *a characterization of innate and acquired features of similarity may emerge if one considers how it is constrained (1) by the patterns of natural kinds prevailing in the world; (2) by the manner in which, in principle, distal objective similarities and dissimilarities can be mirrored in the proximal representations, and (3)*

by the architecture of the given perceptual system." However, the first criteria supposes that a subset of properties is (external or learned) is used to define a category, therefore resulting in a definition based category.

As an example, Edelman [27] proposes a similarity metrics that can be used whenever similarity is decided based on a set of classifiers x_i that evaluate the features of an object. If A and B are objects, and $\mathbf{x}(\mathsf{A})$ and $\mathbf{x}(\mathsf{B})$ their feature vectors, the simplest way to define a measure is by means of the Euclidean distance between $\mathbf{x}(\mathsf{A})$ and $\mathbf{x}(\mathsf{B})$, namely $||\overline{\mathbf{x}(\mathsf{A})\ \mathbf{x}(\mathsf{B})}||_2$. To account for scaling, one can replace the Euclidean distance with an angular measure, namely

$$S(\mathsf{A},\mathsf{B}) \approx \sum_i x_i(\mathsf{A}) x_i(\mathsf{B}) = \langle \mathbf{x}(\mathsf{A}), \mathbf{x}(\mathsf{B}) \rangle \tag{2.1}$$

To make that measure dependent on the context, additional weights can be used to weigh features according to contextually related entities; $W_i = W_i(\mathsf{A},\mathsf{B},\mathsf{C},\ldots)$. To weigh prototypes with respect to their distinction among other prototypes, an additional saliency measure $s(x_i,\mathsf{B})$ is introduced. The final measure is therefore

$$S(\mathsf{A},\mathsf{B}) \approx \sum_i W_i \frac{x_i(\mathsf{A}) x_i(\mathsf{B})}{s(x_i,\mathsf{B})} \tag{2.2}$$

Note that this metric is non-symmetrical, as the similarity between A and B depends both on the context, where $S(\mathsf{A},\mathsf{B}|\mathsf{C},\mathsf{D})$ will not be the same as $S(\mathsf{A},\mathsf{B}|\mathsf{E},\mathsf{F})$, and on the saliency of the prototypes.

A metric space embedding of prototype-based categories was most consistently elaborated in Gärdenfor's theory of conceptual spaces [44]. Here, similarity can be evaluated by evaluating the distance measures in a conceptual space, where concepts are modeled as partitionings in space delimited by convex manifolds.

As Gärdenfors claims, the *conceptual* nature of the representation (see Chapter 3) provides meaningful metrics, mostly because 1) information is sorted into domains, and 2) the relations between different properties are represented intrinsically, at a conceptual level that is more abstract than the raw data, and less detached than symbols. He however acknowledges that the definition of similarity based on distance in conceptual space is problematic: as concepts and properties are more primitive than similarity, we face the same problem of the Ugly Ducking; a careful selection of the properties that are included in a comparison has to be made.

Prototype based categorization has inspired mainly those models that use a shape based

representation, as shape of objects often exhibits a large "cue validity" in context of a basic level category. For example, Edelman's *Chorus of Prototypes* [28] represents novel shapes by similarity to familiar ones, which are represented as *landmarks* in a low dimensional *shape space* (Figure 2.1). Familiar shapes are learned form a number of exemplar objects of similar shapes (e.g. dogs). We discuss some implementations of shape matching techniques in Section 4.7.

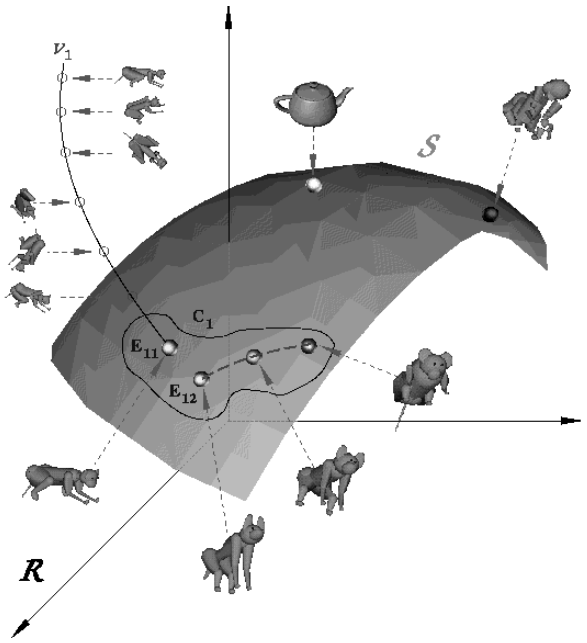


Figure 2.1: The Chorus of Prototypes space - Edelman's visualization of the space where similarity between prototypes and exemplars is measured. Exemplars, E11 and E12 belong to category C1; the shape space is depicted by the smooth manifold, while V1 denotes a subspace where different views of a particular exemplar reside. (From [27].)

♢

In our work, we also adopt a prototype based representation. An exemplar in a canonical view is chosen as a prototype, and serves to evaluate the similarity of other object, where similarity is measured as the number of successful synchronous bindings of high level features between the prototype and the target image. In that way, we implement exactly the notion of "family resemblance" as objects can be associated with the same basic level category, even if they do not exhibit a single feature that is common to all the exemplars in the category.

The representation is not designed to be a collection of the most relevant features, or by a representation in terms of relevant conceptualized properties. In contrast, we use a low level representation, and the relevant high level features are selected in the process of matching, in a fine-to-coarse hierarchical manner. We use a conceptual notion of spatial relations, however they are used only for the process of matching, not as part of the representation.

As our framework lacks a concise metric embedding, it is prone to same criticism that was directed toward the idea of family resemblance: for two arbitrarily selected objects, we can always find at least a small degree of similarity, and those two objects can be always linked by a series of intermediary family members.

♢

In this chapter, we examined the different aspects of what categories are, and how can categories be defined. In the following chapter, we proceed with a discussion on representations of objects that support categorization, and mechanisms for experiencing of categories.

3

Categorical representation and categorical perception

Perceptual categorization, or generic object recognition, is one of the seminal problems not only in the interdisciplinary field of cognitive science, but also in psychology, epistemology, phenomenology and neuroscience. Despite an increasing body of knowledge about specific mechanisms, laws and organization in perceptual processing, a consensual "big picture", that would explain how humans recognize and categorize objects they see, is still a far reaching goal. An important reason, and an agreeable excuse for this lack of understanding is the same nature of *categorization* and *recognition*.[1] As both processes indeed play an important part in high level cognition, they interfere with understanding, deliberating, taking action, and, through establishing a relation of one's inner cognitive process to external objectivity, of course under the assumption that such a relation exists, are part of the very core of self-conscious and subjective thought [55]. Theories on visual categorization therefore have to take into account not only the ways the brain processes the information that falls on the retina, but also in which form and at which stage this information enters the cognitive process and how do the two interfere.

Although there has been an enormous improvement in our understanding of the functional aspects of the visual processing pathway in human and primate brain [2], the evidence at hand is still not conclusive to a degree that would rule out, or entirely support any of the valid theoretical models that have coexisted in the last decades. While a lot is known about the

[1] Although categorization can be defined as recognition of object of same category, recognition can not be viewed in general as a subordinate process to categorization, as entirely different mechanisms could in principle be used for one or the other.

[2] For a review see [29]

functionalities of particular cells and areas, especially in LGN and V1, the center-surround property that is repeated through the visual hierarchy, and the dichotomy of the WHAT? and WHERE? pathways, the modes of interaction between the functional areas, the message passing, and the co-functioning of different sensory and motor areas are still to be explored. This is particularly the case for the functional properties of higher visual processing areas, where the actual event of visual recognition is supposed to happen.

If the perceptual pipeline is modeled as a stream of processing, then the visual information that is reflected by objects and transmitted by light rays enters the processing stream in a raw form, i.e. directly encoding the physical and chemical properties of the environment. The task of the perceptual system is then to process this raw information, in order to facilitate its use in whichever operation the cognitive system has to accomplish. The substrate of the cognitive system (the brains in biological systems, and the hardware and software in today's artificial systems) is therefore a vehicle for the visual content. As the vehicle has significantly different properties than the media in which light travels, the content has to be represented in way that can be efficiently passed and used by the cognitive modules. The visual information is therefore being *represented* in a way that significantly differs from its original form.[1]

After our discussion on what are categories and how category systems can be defined, we head on to the problem of how can (visual) categories be represented within a computational model of perception. This question is also one of the central problems in cognitive science, and refers to the more general problem of *representations* of the outside world that are used in cognition. We can, for example, pose the question: what sort of ideas are formed in the mind while perceiving? Are they icons, symbols, or something else? For example, a system's internal representation of the visual world could consist of only traits that reflect the properties of surfaces in terms of light emittance as captured by a certain portion of receptors in the visual field, therefore at any point encoding only appearance properties of the scene, centered in the viewer's coordinates. Other representations could use object-centered representations of 3D surfaces that form objects, being therefore centered on object's shape. In this section we address some of the key questions that arise—how can be visual data abstracted, how to acquire representations, how can representational architectures be implemented, and present some of the plausible representational models.

[1]Here we do not consider only the narrow notion of representation as a reconstruction of reality, but the extended notion of translation of perceptive data to a different substrate.

3.1 Reconstruction vs. recognition

We begin the discussion with an important distinction, namely the difference between *reconstructive* and *recognition* approaches, which present opposing strategies of using internal representations to match appearances of the 3D world that surrounds the perceiver. Approaches that are based on reconstruction typically use object or scene centered representations, and attempt to reconstruct the structure of the scene; this internal representation can be then used as an abstraction, based on which reasoning is performed. In contrast, recognition approaches use more or less invariant representations that can use momentary *views* to solve a subset of "visual problems", avoiding thus the cumbersome task of world-reconstruction. In this case, the representation is not a reconstruction that reflects the geometric, spatial and physical properties of the scene, but rather a pattern recognition machine that, by enacting appropriate behaviors performs *purposive* perceptual tasks [5; 146]. If we further confront these opposite views, where, according to reconstructive theories, a copy of the outside world is reconstructed within the cognitive system, and, on the other hand, according to non-reconstructive theories, the computational patterns within the system do not directly reflect the image of the world, we arrive at the possibility that it is the world itself that in a great part functions as a referential memory [12; 104; 110]. Apprehension can be therefore in a large part performed directly, in close interaction with the outside world, a view which complies both with Gibson's ecological view on perception [46], as with theories of enactive perception and emergent cognitive systems, discussed later in this chapter.

3.2 Centering, primitives and organization

Marr and Nishihara [93] postulated three dimensions that can serve to classify visual representations. According to their theoretical model, representations can be classified acording to the *centering of the coordinate system*, the set of basic *primitives* that they use, and according to the *organization* principles that govern the layout of primitives within the representation.

Centering of the coordinate system specifies the coordinates that are used as a reference for encoding the spatial configuration of things, which can be in general expressed in a *viewer centered*, *object centered*, or *scene centered* coordinate system. While the centering of coordinate system seems like a minor issue that influences only the implementation aspects of a cognitive system, it nevertheless significantly marks the perceptual experience. For example, a system capable of forming object centered descriptions may deliberate about

objects as separate entities, while a system that never extracts object-centered descriptions can think about objects only as entities that are part of an immediate environment.

The coordinate systems are used for a spatial insertion of primitives, from which descriptions of objects and scenes are formed. Primitives can in general vary in complexity, ranging from basic appearance patches, to complex 3D shapes. The major requirement for a system of primitives is to be *accessible*, *unique*, and *stable*. We discuss some of the choices that were used in computer vision systems in Chapter 4.

The organization criteria determines how are primitives organized into descriptions of objects and scenes. Part based organization can for example reflect the organization of physical entities, e.g. the compositionality of objects, which are composed of smaller, discrete parts, into various categorical spatial relationships. Organization can also optimize some of the computational factors in a cognitive system, e.g., shared parts representations use parts that can be shared across objects and categories, while a multiresolution hierarchical representation uses representation at different levels of detail. Non-hierarchical models treat primitives as if they reside on an unique level of abstraction, or at different levels of abstraction, but do not connect across levels in a hierarchical fashion.

This criteria becomes extremely important if we consider that humans do not merely categorize stimuli, but also apprehend their *structure*. If we postulate the requirements of a cognitive system in terms of *productivity* (the possibility to represent an unlimited number of objects), *systematicity*, and *independence* (separability of parts and relations), compositionality is a possible answer that satisfies all three requirements [28]. For Biederman, scenes can be efficiently described *exactly* because they can be represented in a compositional manner, therefore connecting perception and "the language of thought" [10].

3.3 Marr's computational model

Marr and Nishihara were the first who devised a complete computational model of visual perception [93], which already systematically defines and addresses the requirements that a representation of visual information, or more specifically, a representation of *object shape*, should fulfill. The five criteria that they used in the design of their representational model are *accessibility*, which refers to the feasibility of deriving the abstraction from the image, *uniqueness* with respect to an object or to a category, *stability* under different viewing conditions, *sensitivity* to subtle differences in stimuli, and *scope*.

Marr's model presumed several stages of processing. The first stage is a derivation of a primal sketch in the form of a description of primitives such as edges, light and dark areas, and contours. This rudimentary representation is based on a viewer-centered coordinate system. From this information, a $2^1/_2$-D sketch is calculated that represent surface primitives such as planes or convex and concave surfaces. The next step integrates these primitives to a number of separate representations of objects in the scene. These representations now reside in an object centered coordinate system, and are therefore invariant to viewpoint changes. The organization of primitives is a hierarchy, where primitives at different levels group to reflect different levels of detail, or *parts* in a structural decomposition of an object (Figure 3.1). The object centered descriptions are then inserted into a scene model, where objects are represented in a scene centered coordinate system.

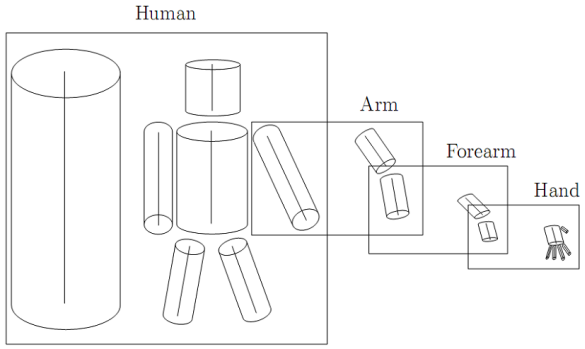


Figure 3.1: Organization of primitives in Marr's model - In Marr's model, the representation is organized as a hierarchy, where primitives at different levels group to reflect levels of detail and *parts* in a structural decomposition of an object. (From [93].)

3.4 Local and distributed representations

Another distinction is between *local* and *distributed*, or sparse, representations. In local representations, the entities being represented have an unique corresponding representative entity, forming thus a mapping that is essentially a one-to-one relation. A common example is the so-called *" grandmother cell"* theory [123], where a single computational unit represents a specific object, or, for example, a specific person,—i.e., someone's grandmother—and becomes active whenever the object is perceived. In contrast, distributed representations represent entities by sparse activation of several computing units that are distributed over the system. Both paradigms are well supported by respective bodies of experimental evidence, and share a same amount of criticism. Distributed representations raise the question

of integration, or binding of the sparse responses—the common objection is, who, if not a "grandmother cell", is observing the sparse activities? On the other hand, the major objection to local representation targets its non-robustness: a categorization process could not possibly stand or fall with the removal of single computing unit. This said, distributed representations have the obvious advantage of being more robust to failures within the system and to incomplete data. For example, if a part of the computing units, or modules, fails in computation, there would be, due to sparsity of processing, still enough evidence within the system that could stand for a certain phenomena. In contrast, local representations should in general fail with the failure of relevant local computing units, such as those that explicitly represent parts or objects. However, local representations have also some significant advantages. For example while superposition of activations is a problem for distributed representations, it is easily resolvable when units are responsible only for a certain type of percept.

It is however possible to mix and combine the two variations. For example, a hierarchical model that represents objects as composition of parts can be a distributed representation on the level of parts; on the other hand, if parts are connected to a central "grandmother" unit, it becomes a local representation, inasmuch as the system is incapable of recognizing the object if the "grandmother" cell is removed from computation.

3.5 Implementation

Plaut and Farah augmented the three criteria devised by Marr (centering, primitives and organization) by adding a dimension of *implementation* [119]. While, for Marr, implementation was strictly separated from the "algorithmic" aspects of the model, and was not of relevance to the representation itself, Farah argumentated the need for this dimension with the example of artificial neural network systems, where representation and implementation are practically indistinguishable [29].

The issue of *implementation* touches the very foundations of design of artificial and biological cognitive systems. The conventional view of *symbol systems*, which met a wide recognition in the wider field of artificial intelligence, advocates a *symbolic* type of representation, where visual processing leads to abstractions in form of symbols, which have to be interpreted and compared to a stored symbolic representation. The direct implication is that there have to exist bits of representations that reflect in a more or less detailed level some aspects of the outside world, and can be therefore traced, identified and interpreted by an external observer. These symbols can be either semantic - i.e. can directly participate in

high level cognition, as they represent entities that can be described by language or can be deliberated about, or pre-semantic, or perceptive, which have a "meaning" only when associated to a certain aspect of the perceived situation. Another implication is that there have to be two separate instances of information - one is perceptual, which is being derived form the stimuli, while the other is stored in memory and is searched for, or compared to the result of perceptual processing.

In contrast, systems that have a pure non-symbolic character can be classified as *emergent*, referring to the emergent nature of organization which is prevalent in today's theoretical models of non-symbolic computation [161]. The alternative label *dynamic systems* is used to express the fact that the computation in such systems can be characterized by the dynamics of the patterns of activity within a system. One example are the *connectionist systems*; there, the representation is embedded in the structure of the computational substrate, e.g. the memorized weights within the artificial neural network, and the computation is performed as a parallel, distributed computation of activities of simple, interconnected units. The "final" result (e.g. the active units at the highest level) does not require a search to a separate memory representation [29], or an explicit match, although it could be argued that the computation in neural networks is basically a similarity test of the signal with respect to the connection weights. While symbol representations are constructed by the accumulation of evidence and the organization of more or less invariant abstractions into memory, in emergent systems, learning is essentially an adaptive process of acquiring competences for executing a proper response to a stimuli.

3.5.1 Representations in symbol systems

Processing in symbol systems is based on manipulation of symbol tokens that is based on explicit rules; both symbols, rules, and manipulations are defined syntactically, i.e. independently of their semantics. [38; 103]. Concatenations of symbols form thoughts, which are connected via inferential relations that are not dependent on the semantic meaning of symbols. Cognitive processes are performed by means of *inference* (i.e., computation of logical consequences), and *parsing* (i.e., syntactical parsing and generation of streams of symbols). Symbols and predicates have to be stored in memory; symbols derived form perception are compared (searched for) to these symbols if they are to be recognized.

A central problem of symbolic representations is the *Frame problem*. The carriers of the

symbolic representation are *predicates*. As cognitive systems by definition actively operate in a dynamic environment, it becomes extremely hard to acquire new predicates, as the extending of the space of states, actions, and their consequences, leads to a combinatorial explosion of the logical inferences. Adequate learning and adaptation methods that would provide meaningful abstractions are not easily implemented by symbolic systems. For example, symbols, as representations of categories, have to adapt to new experiences, or they can not adequately develop their "cognitive potential". Consequently, much of the cognitive skills have to be preprogrammed.

Two other limitations that are frequently attributed to symbolic representations are the *Semantic gap*, and the related *Symbol grounding* problem, which also address issues that root in the "disconnected" nature of symbols. The semantic gap denotes the gap in the sequence of abstractions that are formed when connecting the raw perceptual stimuli to symbols. The more abstract is the memory representation, the larger is the semantic gap the system has to cross. The basic question therefore is whether a heavily abstract, symbolic representation, can adequately link to the perceptual stimuli that appear in a dynamic world.

The symbol grounding problem reflects the "parasitic" nature of symbols, as their semantic interpretation can not be made intrinsic to the system; at least not within a symbolist framework. If predicates, as syntactical bearers of computation, devoid of semantics, represent knowledge about the environment, there must exist *meaningful* sub-symbols that were used to derive the predicates in the first way. It is therefore not wise to implement cognitive functions by abstracting the connectedness of symbols to their origins.

Further, symbolic systems have also a problem with *similarity* (see section 2.4), as they can represent it only extrinsically, by rules or axioms [44]. Symbolic processing was also criticized for being an overly simplistic framework—often by using supportive evidence from neuroanatomy, which suggests that cognitive processing is performed using much densely interconnected units as simple feed–forward or feed–back streams of processing.[1]

3.5.2 Representations in non-symbolic systems

Non-symbolic processing can be traced back to a number of predecessor theories. For example, the general idea of processing in neural networks can be seen as computing by as-

[1] In higher visual areas, an estimate of 20% of the dendrites connect across the hierarchy levels; the other 80% of dendrite connections come from lateral interconnections or from other areas [107].

sociations, which was most notably promoted by the British empiricists. Inspired by Kantian conceptualism, where the act of organizing the chaotic percepts depends on schematic structures that organize contents, numerous structuralist and constructivist thinkers, such as Merleau-Ponty [99] or Piaget, developed theories where the objective symbolist viewpoint is substituted by subjective, dynamic and interactive view of cognition and perception. Notably, Piaget [114] proposed a concise theory of intelligence, where schematic structures are actively and subjectively constructed, or learned through interaction with the environment. Dynamical, self-organizing systems were particularly well adopted by constructivism, which promotes the idea that mental processes are but a construction of a particular cognitive system that reflects the external reality in a way that is viable for the system, rather than in an objective way.

Dreyfus' influential critique of symbolic AI [24] was, influenced by Heidegger's notions of *Dasien* (Being-there), and on the essential difference of *Vorhandenheit* (Presence-at-hand) to *Zuhandenheit* (Readiness-to-hand) [57], one of the many calls for models of intelligence that shift the focus from formal computations to the embodied agency and lived subjectivity, and towards a "referential totality", which supposedly alleviates the frame problem by extensive use of the relevant context. In robotics, Brooks called for embodied and situated solutions [12]. As part of the wave of *second order cybernetics* [39], Maturana and Varela's idea of *enactive systems* describes cognitive systems that are in a large part operationally autonomous, self-referential, and separated from the outside reality, but maintain an emergent equilibrium through structural couplings and a perception-action interaction at a low level [95; 160]. Both the embodied–embedded and the active, subjective paradigms led from passively observing (Presence-at-hand) to actively reaching (Readiness-to-hand) systems, therefore to systems that perceive the environment as a set of immediate action opportunities (*to use without theoretizing*). Instead of predicate based representation and computing, non–symbol theories defend a dynamic representation and an action-perception loop in close interaction with the immediate environment.

A widening support for enactive and non-symbolic perceptual models came also from biology and neuroscience; as an example, the milestone research of Freeman et al. [40] on rabbit's olfactory system reported an absence of any stimuli-specific representation; what they could model however were patterns of responses that were dependent on the behavioral response, and took a form of dynamic topographical maps of chaotic activities.[1]

[1] For a concise overview of enactive cognitive science see e.g. [97; 98] or [41]

Connectionist systems, or artificial neural networks, are the most researched instance of non-symbolic systems. They consist of interconnected units that process incoming information using a simple set of functions, exhibit some level of activity, and pass the output values to other units. Processing is performed in an parallel and distributed fashion. The memory is not a separate functional unit, but it is the network that stores the knowledge, which is learned by changing of the connection weights, that consequently modify the behavior of the network. Representations in artificial neural networks can be therefore seen as a high dimensional space of activities. If purely emergent, then representations are internal to the system, and represent only a point in the state space that reflects a state of the system with respect to a history of its perceptual inputs, context, and activity states. It is therefore not always possible for an external observer to provide an interpretation of a state space. Although self-organization often reflects the structure of the incoming data, this structure is always imminent to the system's architecture, physical instantiation, and operational modes.[1]

The traditional representationalist view adopted by symbol systems assumes in-the-world categories with more or less clear boundaries, which are represented by concepts that are based on a discovery of a number of shared features. The prototype theory [127] already argued against this view by claiming that categories can be only regarded in terms of *degrees of membership*, of distance to specific exemplars, for which there exists a consensus, and that categories are changing with experience. In an enactive framework, categories make sense only when coupled with a certain active stance. For example, the *sensimotor* theory of perception, most prominently advocated by O'Reagan and Noë [110], teaches that learning of categories is not a collecting of perceptual data, but rather an acquiring of a skill, for example to learn *how to see* a horse:

> ...visual experience does not arise because an internal representation of the world is activated in some brain area. On the contrary, visual experience is a mode of activity involving practical knowledge about currently possible behaviors and associated sensory consequences. Visual experience rests on know-how, the posession of skills. Indeed, there is no re-presentation in the brain, only the pictorial or 3D version required is the real outside version. What *is* required,

[1] Many connectionist systems are however not conceived to be purely emergent and dynamical, as they implement *by design* some computational preconceptions [45]. For example, a similarity measure, or a classification function, can be defined externally, e.g., by a labeling of can be trained by a learning machine on top of the outputs of the network.

> however, are methods for probing the outside world – and visual perception constitutes one mode via which it can be probed.[1]

We met a similar philosophical standpoint already in Wittgenstein (Subsection 2.3.3), which addresses both perceptual and semantic aspects of conceptualization; in his philosophy, meaning is constructed thorough instantiation as a form of a game, concepts can be understood only in context of their usage and the whole act of perceiving, as perceiving *aspects*, not things:

> It shouldn't really be "Yes, I recognize this, it's a face" but "I recognize it, I see a face" [...] For the question is: "*What* do I recognize *as What*?" For "to recognize a thing as itself" is meaningless.[2]

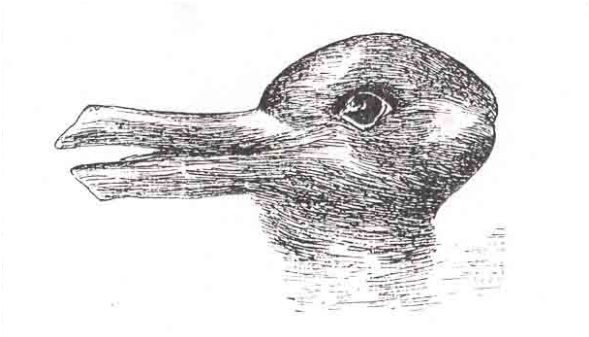


Figure 3.2: The duck-rabbit bistable figure - As appeared in Jastrow [60], adapted from *Harper's Weekly*, originally form *Fliegende Blätter*. The figure was later popularized through Wittgenstein [166], which used it to illustrate the difference between *seeing that* and *seeing as*.

A sketch based on Figure 3.2 that was used by Wittgenstein to illustrate the difference between *seeing that* and *seeing as* can be interpreted both in a constructivist manner, where expectations, attention and beliefs lead to a stable interpretation, and as an argument for dynamical systems, where the visual system's interpretation can only shift between two stable aspects, while making it impossible to perceive both objects simultaneously. The picture is not really interpreted (an interpretation would be "*I see a rabbit head which is simultaneously also a head of a duck*", but rather "seen as". In other words, the subject is not interpreting the image according to some inner representation, but is rather directly reporting what he sees by means of language [166].

[1] *O'Reagan and Noë, [110]*

[2] *Ludwig Wittgenstein, Philosophical grammar, 1.130 [167]*

3.5.3 Representations in hybrid systems

Several researchers share the opinion that the exclusivity of one type of representation and processing is unrealistic. For example, connectionist architectures can be either symbolic or non-symbolic[1], as the outcome of network processing can still be entities that are used in a symbolic computation at high level stages, for example in search or in discrimination, and such types of architectures can be referred to as *hybrid* [54; 161].

An even more important argument for hybrid models is the fact that perceptual information is in general used in many various ways that require significantly different levels of abstraction. For example, sensory-motor interaction with the immediate environment requires a rich information that describes the perceptual aspects of the physical manifestation of objects, but does not necessarily require symbols. Since emergent and connectionist systems model the data in a direct dependence of the outside world, they are a good candidate for representing such low level knowledge. On the other hand, efficient planning may be optimally performed only if the abstraction is detached from signal, while encompassing its semantic meaning.

For example, Stevan Harnad's proposed solution to symbol grounding [54] is based on bottom-up anchoring of symbols in non-symbolic representations of two kinds: *iconic representations* that represent sensory projections of objects and events, and are used for discrimination, and categorical representations which consist of learned or innate feature detectors that exhibit a selectivity for invariant features of object and event categories, and are used in *identification*. Stable object and event categories are not symbols per se, but represent just an inert taxonomy, however they can be assigned names, and higher-order symbolic representations are then grounded in these elementary symbols. Harnad therefore advocates a hybrid model: the symbolic functions *emerge* as a consequence of the bottom-up grounding of categories' names in their sensory representations [54].

In general, symbolic representations are explicitly positivist in the sense that they explain perception as the apprehension, or recovering of an *objective* world by means of perceiving an acting. Perceptual processes shall be therefore directly correlated with an objective realty; when they are influenced by subjective knowledge, this knowledge is in general obtained by experience, and is again measured against objective reality. This holds also for hybrid systems, as the symbolic and pre-symbolic symbols are usually treated as conditioned to

[1] It is still an open debate, under which circumstances connectionist systems can be classified as symbolic systems, see e.g. [38] for a discussion. However, there is a general agreement, that connectionist systems can operate with symbolic representations.

objective categories. As an example, according to Harnad, a higher order symbolic representations consists of symbol strings describing category membership relations, e.g. "An X is an Y that is Z". For example, in "A LEMON is an OVAL that is YELLOW", OVAL and YELLOW are concepts, or *basic categories* (see also Section 2.1), that directly encompass a category of percepts in the world, and should refer to external stimuli for which the subjects that participate in the exchange of propositions agree on their nature (e.g. YELLOW is a color such and such).

The *conceptual space representation* introduced by Gärdenfors (see Subsection 2.3.3) stands between the non-symbolic and the symbolic levels in terms of representational *granularity*. The representation consists of *dimensions* that represent perceptual qualities of objects. Quality dimensions span *domains*, which are needed for representing concepts. For example, spatial concepts and color concepts fall into spatial and color domain, respectively. Properties are a special case of concepts that are defined with the aid of a single dimension, or by a small number of integral dimensions from a single domain (for example the color space), while concepts may be based on several separable subspaces from multiple domains.

Other examples of hybrid systems can be found in representation of space. As an example, we developed a hybrid representation of space that discovers regularities in visual appearance to model a non-symbolic representation of an environment. This representation is then embedded in a parametric space spanned by measured ground truth coordinates, where the coordinates are given by an external observer [62]. We then developed a mapping method that does not require any external information, but instead reconstructs a global topological map using only partial, subjective observations. Although the global map reflects to some degree the environment it in reality represents only an action–perception subjective plan [140].

3.6 Perceptual learning of representations

If representations function as a "bridge" between the outside reality and the mind, an important question is how this representative knowledge is formed, acquired, and remembered. A particularly interesting aspect of representational theories is the question to what degree representations are organized in advance of experience, or *innate*, and to what degree they are *learned*, and what are the peculiarities that define the baseline between the two types of knowledge. Investigations in *perceptual learning*, which denotes the adaptability of perception to sensory experience that is largely independent from conscious forms of learning,

led to several theoretical models of how representations can be incrementally and adaptively constructed and modified in order to support a robust interpretation of natural scenes.

To achieve a better understanding of the processes involved in visual perceptual learning, several ways of learning under different laws, with different modes of guidance, and using different levels of innate knowledge are being investigated.[1] For example, a system that can autonomously discover the operationally relevant visual categories, and forms adequate representations that support robust scene interpretation in a variety of conditions, is said to have the capability of *unsupervised learning*. On the other side of the spectra, a system that can learn the categories only by an observation of labeled exemplars, is said to have a *supervised learning* ability. Between of those two extrema, a plethora of learning modes can be identified, e.g. *weakly supervised learning*, where only some of the exemplars are labeled by am external source, *tutor-guided learning*, where a tutor supervises the learning process and intervenes when necessary, or learning through interaction and communication with other systems.

In most computational frameworks, learning happens as learning of relevant features, or as learning of paterns in feature binding (Section 3.7). Researchers that advocate impenetrability of perceptual processes however state that the major part of the perceptual system can not be modified by experience, at least not in a way that is mediated by cognition (Figure 3.3). According to Phylyshyn [122], perceptual learning takes part only at the interface between vision and cognition, i.e. at the level of attention (see Section 3.7).

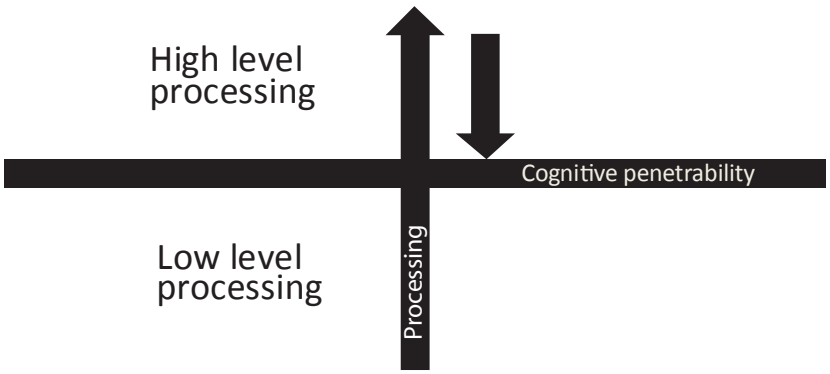


Figure 3.3: Cognitive penetrability of perception - To what extent do high level cognitive modules, e.g. memory, knowledge, experience, conscious attention, modify the processes that constitute visual perception?

[1] For a complete review on these issues, see e.g. [146]

As we already mentioned, symbol representations do not offer plausible ways of adaptation by learning. In contrast, non-symbolic representations, such as neural networks, easily adapt to changes in the environment, can efficiently learn to associate similar phenomena, and can discover the regularities and the structure of the perceptual data. However, in order to learn the relevant structures in real-life scenarios, a large amount of training data is needed. Furthermore, neural networks are in general not designed to autonomously learn and maintain constructs or concepts that structure the data in a more meaningful way. In short, they can not use even simple theories about the nature of the phenomena, and on the importance of concepts out of the basic functionality of the system [44].

In our example of a prototype based representation, the category might be formed by the process of finding similarities between the exemplars encountered, and by selecting the most central prototype that exhibits a well distributed distance to other objects. Where perceptual learning happens is in the adaptation of the low level filters that identify typical local structures, an in the adaptation of the concepts [44] such as relative size, length, and orientation that are used to construct high level matchings. We could therefore combine a supervised learning strategy (give the system a set of prototypes) combined with an unsupervised adaptation (the system autonomously expands and adapts the category representation). Further, As the synchronous binding discovers directed receptive fields that steer attention to relevant neighboring areas of a feature, one could associatively learn the feature–attention couplings and automatically learn to steer attention depending on category / feature pair.

3.7 Attention and feature binding

Perceiving an object as an autonomous entity is to simultaneously perceive its characteristics, or features, of different modalities, such as shape, structure, color, size, or pattern. The question on how are these features perceived—can they be perceived separately, or only as a whole—and, how are they integrated, can be summarized as the *feature binding problem*.

In biological vision, there is a strong evidence of modularity of representations in the visual system. Modalities such as local shape, hue and intensity, motion, or retinal disparity are supposedly represented in separate pathways, and give raise to a number of retinotopic *feature maps*. The binding problem therefore addresses the signal integration from different maps, therefore a spatial conjunction of information concerning visual attributes of the same item which gets spatially dispersed by a progression from the retina to higher visual

areas. More precisely, the higher levels of recognition (the WHAT? pathway) exhibit a significant degree of invariance to position. The spatial information is processed separately in the WHERE? pathway, and has to somehow be attributed to information from the feature maps. As Triesman points out [152], binding is especially important if we consider the modular processing of cluttered scenes—while perception of a single object does not require feature integration, modular signals could be easily misinterpreted in a crowded scene. One of the reasons for that is that neurons at higher levels of hierarchy have a wide receptive field, which could encompass stimuli that belong to more than one object [152]. A straightforward solution in the form of a direct binding of pre-learned conjunctions of features, involving units that consistently respond to complex shapes and patterns, as the ones discovered in IT by Tanaka [143; 144], is being extensively used in many hierarchical and non-hierarchical models (see Section 4.6), however the combinatorial aspects of learning every possible feature binding have not been properly evaluated yet.

Several other hypotheses on binding can be found in the literature, which typically adopt two opposing standpoints. One group of researchers acknowledges the problem as an important issue that has to be solved, and inquires into mechanisms such as synchronization, feature integration, etc. The second group of researchers either denies the role of features, or neglects the role of feature binding by giving primacy to the action centered nature of vision [145].

One of the classical models of feature integration is the one that states that consciously perceiving an object requires *attentional integration* of sensorial properties [153]. While feature maps that sparsely represent single attributes across the visual field are formed *preattentively*, conjunction of information in feature maps requires *attention*. Binding is therefore achieved by directing spatial attention serially to salient spatial locations.

According to Duncan [25], attention however does not require external selective mechanisms, but is rather achieved through a *global competition* between objects that settles the top-down and the bottom-up factors of experience and action. Attention is therefore an emergent state of balance, achieved by feedback and inhibition mechanisms. In Duncan's opinion, feature binding is supported by feature detectors that *directly* encode conjunctions of pairs of attributes.

Some other possible answers are given by models that use large-scale synchronization activity, for example special signals that temporally synchronize the neural activity across different modules [50]. According to an increasing body of evidence from neuroscience, synchrony is an important modality in brain signaling, and could be used to tags the units that

respond to the same object. It is not however clear which information is used to synchronize the units, especially if there are no temporal or spatial cues available for a prior segmentation of the signal. Synchronization has therefore to bridge similar problems as the original binding problem in spatial domain.

Taraborelli [145] however warns against a supposed fallacy in the studies of feature integration, which can be attributed to *internalization* and *externalization* of phenomenological and physical levels of perception. Namely, the phenomenological aspect (perceiving features in an object as a unity) is being internalized, in the sense that it is projected to the physical instantiation (features as signals in the neural substrate should be also unified). On the other hand, the segregation of feature processing in the physical instantiation is supposed to show in perceptual content (externalization).

3.8 *Gestalt* theory of perception

According to *Gestalt psychology*, the operational principles of cognition are primarily *holistic*, suggesting that the whole is not only a sum of its parts, but has properties that are imminent to its totality. In perception, gestalt theorist postulated a set of well known perceptual rules that govern the construction of particular elements to a perception of totality; namely *Closure*, *Similarity*, *Proximity*, *Symmetry*, *Good Continuation*, and *Common Fate* (Figure 3.4).

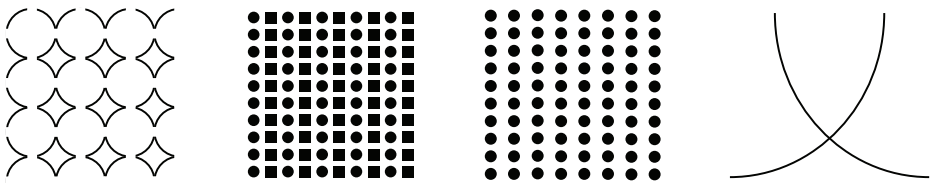


Figure 3.4: Gestalt rules of perceptual organization - (a) Closure: closed shapes are preferred over open shapes. (b) Similarity: similar features are grouped together. (c) Proximity: Close features are associated. (d) Good Continuation: contours that minimize change are preferred over abrupt changes.

As Gestalt rules govern binding of particulars, they could provide an alternative answer to feature integration; its main characteristic being that the mechanisms are considered to be in a large part given, or innate, and therefore not subject to change, learning and adaptability. In a connectionist framework, facilitatory connections between units across dimensions that

cover the gestalt laws (e.g. between units that respond to similar stimuli, or connections that enhance the contrasting responses between neighboring and distant stimuli) could account for a great deal of emergent binding, without requiring additional mechanisms as attention. From the innate nature of gestalt rules follows also the impenetrability of perception, and a clear division between perception and cognition. According to Köhler, *"Objects cannot exist for us before sensory experience has become imbued with meaning"*. Further, several gestaltists reject the idea of a "pure percept", as could arise for example as a result of passively observing e.g. a pure source of "redness"—in their opinion, the pure percept can have a particular significance only as part of the whole, which permeates it with meaning.

Several ideas that are related to gestalt psychology found their way to models of perception. However, while many frameworks borrow the gestalt ideas to explain subordinate processes, they are mainly limited to low level vision phenomena. In our framework, a subset of Gestalt rules is used to reduce the redundancy of low level feature responses.

3.9 Object awareness and object attention

According to feature integration theories, *object awareness* is achieved only after a unitary percept of features that occupy the same spatial location is constructed. Awareness therefore requires attention, attentional binding, or some other kind of binding mechanism. Although awareness can have several definitions and accounts, and is in itself, due to its relation to *consciousness*, subject of numerous debates, we will limit ourselves in our discussion to theories that relate awareness to the integration of all available information into "conscious" percepts.

In Kinsbourne's *integrated field theory* [29; 73], awareness emerges when various modality-specific perceptions, memories, actions, and plans are mutually consistent. Awareness (visual and in general) is therefore due to a consistent interpretation (here we can draw a parallel to constructivist theories), and is not subject to an interpretation in a special area (as was suggested by e.g. Descartes [20]. As we already mentioned in the previous section, the binding could emerge by a mechanism of synchronization, as advocated by Singer and Gray [137]. Damasio proposed a hybrid solution, where synchronization is mediated and integrated by specific "convergence zones" [18; 29].

In our architecture, feature binding is performed in a hierarchical fashion, and, by mechanisms of construction and by synchronous inhibition and matching, features get integrated

to reach a final solution, or a high level match, at a level of an object. The binding and matching is constrained by a prototype, and is therefore subject to an intentional comparison to a previously seen object. However, as the representation of the prototype can in principle incorporate features of different modalities, the process can be interpreted as a feature binding problem, where the synchronization happens at a local scale, in order to achieve a fully synchronized solution at the object scale.

At this point, we will not speculate how is the object prototype activated and stored. One of the tenets of enactive cognitive system is that the frame problem can be solved by contextual and situated processing; one could imagine that, once a stable reference frame that defines the spatial insertion of a system, a coarse contextual cue activates a contextual frame, that supports contextual priming of recognition [6]. In our case, the representation of a prototype is an icon, together with learned information on its perceptual categories, e.g the characteristic low level features, or the sampling of angles. Alternatively, we could store a number of preprocessed low level features prior to the hierarchical binding. In an extreme case, to imagine what an object looks like could trigger a comparison of the "mental image" to an object in the scene. In the next section we therefore review the phenomenon of mental imagery.

3.10 Mental imagery

By "mental imagery" we describe the quasi-perceptual phenomenon, where images of objects and scenes can be formed without an external stimuli. The same term is used also for other modalities of perception. According to Farah, this phenomenon is activated endogenously by high level cognitive modules, and it supposedly affects the low visual hierarchy, it therefore accounts for a deep cognitive penetrability (see Figure 3.3) [29]. While attention only modulates the activity of retinal signals, imagery is completely independent from any external signal, an therefore recreates the percept without any stimuli.

The true nature of of mental imagery is still controversial. The important issues can be summarized by two questions: 1) does mental imagery shares the same representations as normal visual perception, or does it operate only on more abstract, symbolic representations?, and 2) are mental images retinotopic (picture-like, or array-like), or are they propositional? Pylyshyn [122] strongly criticizes the picture-like nature of mental imagery, and advocates the idea that cognition, as well as thinking about images, uses only symbolic processing, distinct form the perceptual processes, which are, according to him, in a large part impenetrable.

As imaging something offers a chance to deliberate on its perceptual qualities, imaging can be classified as a kind of thought which has significantly different properties than verbal thought [75]. Images played a central role already in Aristotle's theory. Remembering involves the recall of imagery of past experiences, and thought is impossible without a *phantasma*; images are vehicles of thought and words take their meanings according to the images they represent. Phantasma may be experienced as a complex of sensory characteristic (sensory representation), or can only support thinking about the form of the perceived content (noetic representation). Wittgenstein's Tractatus [165] proposed a somewhat reversed pictorial theory of meaning, which states that propositions are meaningful insofar as they can be represented by an image, are therefore structurally similar to images, and bear imaginable facts.[1] For Kant, images serve to connect concepts to reality: imagination (*Einbildungskraft*) binds the incoherent percepts of the senses to a coherent image.

Starting with the seminal experiments on visual search performed by Kosslyn [74] the body of evidence on endogenous generation, and of an an *array-like* spatial characteristic of mental images is still growing. In a theoretical account that would support our synchronous inhibition model, where an image is compared to a prototype image, mental images would represent prototypes. By (consciously) imagining a prototype that represents a category, we can steer a binding process that discovers resemblances between the two,

The prototype theory of representation was conjoined with image similarity very sparsely. For example, Russell stated that *"...words represent because they are associated with mental images, and that the images themselves represent because they resemble their objects"* [128]. There is however a hidden discrepancy between "to resemble" and "to represent". Namely, an image of an object can not resemble an object, as two images (of the same object) would therefore resemble each other more than they would resemble the object. On the other hand, the object does not represent the image—resemblance is a symmetrical relationship, and representation is not. Before a cognitive system can recognize or use the relevant aspects of resemblance between a mental image and an object, it must already be able to *represent* the picture and its object, and their various features, to itself [148]. Representation is therefore a significantly different agent than an image that is compared to another image, as it represents the meaningful interpretation (see also Subsection 2.3.3).

In support of a role of imagery in enactive perception, Malafouris [91] proposes that images

[1] Wittgenstein however disbelieved a pictorial nature of mental imagery.

have a status of a "perceptual device", which is in large part enabled the human cognitive evolution by bringing forth new possibilities of engaging with and deliberating on our own perceptual apparatus. Images that are put in the world (by act of drawing and by technology) are therefore projections, or an "outward membrane" that makes visible the hidden inner concepts, or aspects of seeing. By being aware of ways of seeing, humans learned about the functional principles of vision that go far beyond simple processing, and, speculatively, led to the advanced categorization capabilities that are manifested in our species at this point of evolution.

The enactive cognitive framework accommodates imagery as an active phenomenon: ... *imagery is then experienced when someone persists in acting out the seeking of some particular information even though they cannot reasonably expect it to be there. We have imagery of, say, a cat, when we go through (some of) the motions of looking at something and determining that it is a cat, even though there is no cat (and perhaps nothing relevant at all) there to be seen. Visually imagining a cat is seeing nothing-in-particular as a cat.*" (cf. Thomas [148]). The enactive systems theory therefore considers imagery as manifestation of the more basic imaginative capacity of intentional perception (or seeing as, see Subsection 3.5.2).

In this chapter we reviewed some of the important aspects of categorization in terms of representation and generic object recognition. We specifically attempted to establish a theoretical basis that contextualizes our computational model. Some parallels that were already given in the text can be summarized as following:

We draw from the idea that categories are defined by similarity, and more specifically, we examine similarity to a certain prototype in a canonical view. The graded category membership is evaluated by a search for similarities that are not intentionally given in the model, but are rather discovered and constructed by the act of matching. The similarity is reflected by the number of different activations of distributed units, where a single activation represents a rather non-representative ad-hoc high level feature. Other than an image of the prototype in the form of image, we do not store any other high level description that would in any other way define the category. Low level binding is done by means of gestalt principles. Perceptual learning consists of low level perceptual tuning of receptors, and of conceptual calibration of geometric descriptors.

High level feature binding is performed as synchronous matching and selection that is guided by the image of a prototype. Feature binding is sparse, parallel, hierarchical, and ad-hoc, and binds appearance with local spatial information. As the high level binding incorporates all of the low level spatial relations, there is no need for additional mechanisms that bind location with other features. Furthermore, synchronous matching can in theory be performed with any modality—synchronization can be therefore seen as a gradual construction of a consensus on more and more complex descriptions. Attention in our framework is local and global: locally, each binding process activates a receptive field which depends on scale, orientation and coverage of the area where features should be searched for. Globally, attention emerges as the gradual vanishing of surrounding features, which focuses the processing resources on the object.

In the next chapter, we review the important contributions that were made in implementing systems for object categorization in computer vision research.

4

Categorization in computer vision

Early methods for visual recognition and categorization of objects can be tracked back to the pioneering work in computer vision. In spite of the general consensus that visual categorization is a much harder problem than visual recognition, and that appropriate techniques for categorization are yet to be discovered, one has to agree that many early methods for object recognition did represent a sound starting point for generic object recognition. The first attempts to scene interpretation were made under a strong influence of the research in artificial intelligence, which was primarily concerned with formal logic, predicates and symbolic processing. It is therefore natural that object recognition was seen as separate from the low level processing of images, and was based on discovering spatial arrangements of geometrical shapes.

4.1 Model based recognition

Roberts [125] was among the first to explicitly state the problem of recognition as the identification of well defined constituent parts. His recognition system grouped hypothetical vertices and edges of 3D polyhedral bodies, and evaluated these hypotheses using a projective model, where polyhedra stood for constituent parts of more complex arrangements. Later, research turned toward more complex constituent parts, or *primitives*, that were, according to Marr's postulates [93], designed or selected to be accessible, unique, stable, and at the same time flexible enough to provide a fine enough discrimination between shapes. A plethora of *"Recognition by Components"* methods that emerged used primitives such as generalized cylinders [120], Biederman's *geons* [10], or superquadrics [147] in 3D, or several curvature based representations in 2D. The part based models focused on a description of the essential *object-centered* geometric structure and on the interrelation of parts, in hope to encode properties that are usually shared between members of a category.

The performance of these methods was constrained by the challenging task of segmenting the relevant part descriptions from a variety of different possible views in cluttered scenes, often without using any prior knowledge about the contents in the image. Significant efforts were made in design of reconstructive methods that were able to recover 3-D or $2^1/_2$-D shape using stereo, multiple view geometry, motion, pattern or shading. According to Mundy [101], the majority of methods in this period performed perceptual grouping on boundary descriptions from single intensity views or range images, which caused difficulties with low contrast at object boundaries, high edge density in background, and occlusion by objects with complex texture. Today, partial models and 3D features can be reconstructed more reliably, mostly due to the advances in projective geometry, novel sensors, and due to the breakthrough of the dense range sensor technology. A retrieval of complete 3D models is however still unrealistic in the majority of situations.

4.2 View based recognition

The organizing of features into object descriptions can be also formulated as a bridging of the gap between low level and high level vision processing. One possible solution is to use a detailed object-centered model that constraints and steers feature organization and binding. Another solution is to change the representation in a way that would take into account the particular properties related to the acquisition of a specific view of the object, e.g., properties that depend on a specific viewpoint or illumination, and are usually directly reflected in the image that falls on the sensor. In *view-based* representations, the model is therefore a viewer-centered description, which supports the recognition under a variety of viewing parameters. This idea first came to fruition in the form of aspect graph models, where the representation consisted of a graph connecting multiple views; node connectivity reflected possible changes in the visibility and structure of elements within changing views. Notable contributions were made by Underwood and Coates [159] who base the description on visible object surfaces, or Dickinson et al. [23], who use geons as primitives.

Another way of dealing with this discrepancy is to use 2-D representations or 2-D projections of 3-D objects as models for the organization of features [86]. Grimson and Perez [51] attacked the problem of feature binding as a search of a consistent interpretation of a prior object model; by choosing sparse oriented edge segments as primitive features, the system exhibited a high level of robustness against occlusion and against fallacies of low level edge retrieval. Aspect graphs, feature binding and feature organization led to a line of research on

efficient matching of large and noisy structural representations, primarily graphs [16], or to a development of search oriented data structures such as *interpretation trees* [52].

4.3 Appearance based recognition

Nayar and Murase [102] introduced a view-based system which was inspired by previous work on *eigenfaces* by Turk and Pentland [155], and which could recognize a vast collection of objects from different views using an object-centered representation of appearance. Appearance, represented as a manifold in a low dimensional subspace, proved efficient for the retrieval of a wide class of weakly structured objects, and exhibited an inherent robustness to noise and occlusion [62; 81]. An important advantage of appearance based methods is that they do not depend on the structural complexity of the objects that are being represented, as is the case with geometric descriptions. However, the underlying representation is global, and the small local variations in structure, texture and shape often cause significant errors in recognition. One of the major arguments against appearance based methods are exactly the difficulties in using pure appearance for generic object recognition.

4.4 Local appearance methods

The drawbacks of global appearance representations motivated novel approaches that, rather than relying on global appearance, used *sparse* information on local *image patches*. The research on *scale space* by Lindeberg [85] led to methods for reliable detection of invariant region detection, while careful design of local descriptors, such as the ones used in Schmidt and Mohr [129], led to efficient and robust encodings of local patches. David Lowe [87] introduced an efficient feature transform which exhibited significant invariance to scale, illumination and rotation using a scale invariant detector centered at the maxima of the *Difference of Gaussians* in the scale space. Several other local region detectors use different criteria, such as 'surprise' over spatial scale [65], symmetry [88], stability of extremal region boundaries [94], principal curvature [19], or explicitly target invariance to affine planar transformations [100]. Although most of local region detectors focus on image patches, some of them are grounded on specific features, such as corners [100], or edges [156], and consequently cover areas with salient structure.

Based on local features, recognition of a previously seen object can be efficiently implemented by a matching step, followed or integrated with a verification step, e.g. by using planar geometry constraints [87]. Generic recognition however requires that information

encoded in local features represents a large number of members of the category. One way to achieve categorization using local features is by employing some variant of *bag of features* approach. Patches centered around keypoints, or centered on a regularly or randomly distributed positions can be for example used to calculate a visual dictionary [17; 138]; the members of the dictionary can be "learned" in a *generative* manner, e.g. by clustering, by multivariate modeling, or can be selected by e.g. Probabilistic Latent Semantic Analysis (pLSA) [138]. Grauman and Darrell [49] implemented an efficient *discriminative* method, where scattered features are matched using a pyramid match kernel: at each pyramid level, features that fall in the same bin add to a weighted number of matches, which results in an estimation of an optimal partial matching between sets of features. Viola and Jones [162] use Boosting to learn a discriminative cascade of features, where weak classifiers immediately reject most of the patches that enter the detection algorithm. The cascade can be realized with filters that are simple and fast to compute.

4.4.1 Spatial information

Bag of features approaches typically do not incorporate any type of *spatial information*; objects, or topics, could therefore in practice be detected in arbitrarily scattered images of the same scene. Several researchers therefore augmented bag of feature models with spatial information. Notably, Sudderth et al. [142] model the spatial locations of features by encoding the information on their reference position in a graphical model.

Spatial information can be used to explicitly model the distribution of features with respect to the object, and to represent structured sets (or singletons) of features as *parts*, which reflect the compositionality of objects. A number of successful *part-based models* that emerged aimed at exploiting the reduced complexity of such representations (as parts reside on a higher level of abstraction, there are fewer parts that define an object than features) for the recognition of an increasingly larger number of categories. Part-based models differ mainly in the connectivity of features that represent parts, in modeling of spatial relations, in how they model local appearance, and in how the part-based representation is used for recognition. For example, Fei-Fei and Perona [30] use a constellation model which can be used in a Bayesian learning framework to achieve efficient *one shot* learning with as little as 1-5 training images. Crandall et al. [15] use *k-fans*, which are defined with connectivity of neighboring parts to k reference parts. Notably, k regulates what type of geometric invariants can be represented by the grouping. E.g. 1-fans define translation invariant models, a 2-fan defines scale and translation invariant model, and a 3-fan defines an affine invariant

model. In general, the part variability within a category is represented *explicitly* by a joint probability density function on the shape of constellation and the visual appearance of the parts [31; 164]. Leibe et al. [80] and Opelt et al. [109] model the distribution of parts implicitly: in the process of recognition, each of the features votes for the location of the center of the object.

4.4.2 Methods based on local contour

Many object categories can be robustly recognized by their contours. As it is evident form psychophysical studies, even sparse contour segments represent a strong cue that supports recognition even when the whole contour of an object is not seen, or can not be retrieved. Local contours therefore represent a good choice for a representation of local shape; in particular when it is not obscured by textured patterns, and describes a clear outline of an object or an object part's shape. As local contours have a well defined spatial extension, orientation, and various choices of descriptors that are invariant to rotation and affine transformations, they can be used in frameworks similar to those with appearance based local regions, or even in combination with appearance based local regions.

Contour segments have been successfully used by Selinger and Nelson [132] in their seminal work, where contours are encapsulated in context patches, which represent the local appearance in a layered framework (see also section 4.6). A combination of contour segments and appearance local features can be found in Fergus et al. [32], where feature constellations from [31] are extended to accommodate the contour information. Ferrari et al. [33] combined k-connected groups of adjacent "roughly straight" contours, represented them with histograms, and used sliding windows for detection. The aforementioned work by Opelt [109] uses a "boundary fragment model", where local geometry is also represented by a constellation. A similar model was recently developed by Shotton [59].

4.5 Object manifolds

The global and local appearance, and the spatial and geometric properties can be in general represented in feature space. Under the condition that the relation of similarity of exemplars is preserved in feature space, feature projections form smooth manifolds, which can form a solid basis for intuitive low dimensional representations. This property led to what became the paradigm of continuous categorical representations [28; 157], where category members lie on separable smooth manifolds that represent the intra-category variations. Within this

framework, the acquisition of a representation that encompasses the variation of a specified category of objects, or integrates a certain level of invariance to viewing parameters, can be formulated as *generative* (visual) learning, as it is essentially a modeling of the prior joint probability of data x representing object instances y as the probability density function $p(x,y)$. On the other hand, *discriminative* learning is essentially a training of a classifier as a direct mapping $c(x,y) \mapsto \{0,1\}$, which directly models the posterior probability $p(y|x)$. The specific properties of the two types of learning have been extensively discussed in the pattern recognition and data mining communities, and are beyond the scope of this work. With respect to representation, while generative learning is appropriate for unsupervised training on large, noisy, unlabeled or weakly labeled data, the discriminative methods can efficiently discover the manifolds that divide different categories of data, especially if combined with nonlinear *kernel* projections, or projections to a high dimensional space. Furthermore, while the discriminative models provide only a decision function, generative models give a probabilistic model of the original data, and therefore support reconstruction of the data. Generative models therefore *preserve* the intrinsic properties of visual signals (appearance), which results in a number of favorable consequences.

4.6 Hierarchical models

Most of the part-based models mentioned in the previous section have a flat structure, and therefore are not suitable to mimic the hierarchical compositionality of objects. Hierarchical models that emerged already in the geometric era [10; 93], and in the neural network era [43], gained a renewed attention in the last decade — in the appearance era [3; 11; 34; 36; 61; 124; 142; 158]. Hierarchical representations can represent scenes at several levels of abstraction, where relations between entities at different levels of hierarchy define the hierarchical relations of parts of the scene. A clear advantage is that such models can be efficiently processed by solving sub-problems, e.g. with a divide and conquer strategy. Further, the compositionality of the scene can be naturally represented as a subsequence of entities and their constituent parts. Parts represented at the lower levels of the hierarchy are usually simple, and can be efficiently shared across categories, while parts at higher layers are more specific.

Amit and Geman [3] proposed a bottom-up processing model of visual selection by local groupings of edge fragments based on their loose geometrical relationships. The groupings have no a-priori geometrical or semantical content, but are selected purely based on their statistical properties. The relative pose is determined using the three point local coordinate

system. Local groupings are basic triplets of edge segments. Global groupings are then triplets of local triplets. The representation is trained over a number of images, and a representative feature set is selected to support the visual selection. After training, visual selection is performed by a search of candidate bases. The final categorization is then performed using classification trees on partition registered and standardized edge data. Similarly to our approach, this method uses pragmatical feature bindings. In contrast to our method, features are binded in two fixed levels, and the derived couplings are less informative - they do not grow until reaching a consensual area of coverage.

Selinger and Nelson [132] argue for a hierarchical model of perceptual grouping: they devise a four level perceptual grouping hierarchy. The first two levels extensively use the *gestalt* principles of good continuation and proximity (see Section 6.4), while the third level uses the concept of *familiarity*. At the lowest level, contour segments are extracted form images. At the second level, *context patches* are designated, that anchor on key contours form the first level, and carry information on all the contours that intersect them. Key contours are selected among the longest contour sections between points of high curvature. At the third level, similar context patches can be associated by similarity into groups. The model groups that are used to drive the grouping process are extracted from clean views of objects and stored in memory. During recognition, the context patches are matched to stored groups, and the results used to construct clusters that are consistent with a particular model view. The database constructed for recognition is generated from views of the object taken around the whole viewing sphere. The fourth level topologically organizes these views and therefore offers support for deliberating on object's affordances.

Bouchard and Triggs [11] proposed a generative hierarchical model, where local features cluster in higher level parts based on the consistency of estimated structural parameters. The lowest level contains elementary parts, which model the appearance classes of features within a certain neighborhood. Parts are then modeled by their location and scale distributions, and by their parent assignment probabilities. The model is learned using the Expectation-Maximization method, where the number of parts has to be specified in advance by the user.

Riesenhuber and Poggio's H-MAX architecture [124] is designed to model the wide body of evidence in neuroscience, that explains the visual hierarchy in terms of a gradual increase in specificity of neurons (and of an increase in complexity of the stimuli they respond to) and of a gradual acquirement of invariance to scale and translation. In practice, they implement two basic functions: *simple* computing units are designed to integrate the responses from

lower level units in order to respond to complex stimuli that are composed from simpler parts. *Complex* computing units integrate responses of an identity class of simple units over position and scale using a MAX-like function, and therefore generate responses that are invariant to these two transformations. An unsupervised generative learning step is used to tune the simple cell weights, acquiring thus a basic vocabulary of parts; a second step of discriminative learning is used to differentiate between categories of stimuli based on responses at lower levels. The starting level of processing is basically a convolution with a set of Gabor filters in a number of orientations and at different scales.

In our architecture, the basic level is calculated on regions of interest, which account for the initial rotation and scale invariance. We learn the set of filters in order to represent the appearance, which is related to the feature detector, with maximally sparse responses. The feature units that fire in the process of categorization are not pre-learned, nor do they exist as representing a possible subpart of an object that might appear - rather, all imaginable combinations of features have an equal probability of being activated, as the only criteria for successful categorization is the successful construction of commonalities between the scene and a prototype.

Hierarchical multi-scale models are widely used in image segmentation and image modeling. *Tree-structured belief networks*, based on structures like *Quadtree models*, and *Dynamic trees* [141] are examples of data structures that encode the increasing abstraction from pixels to segmented parts of the image, and which support probabilistic inference such as belief propagation. Because of their hierarchical structure, image parsing can be done more efficiently than with flat models, for example Markov random fields.

Todorovic and Ahuja [150] generate a canonical categorical representation which is based on *segmentation trees*. A tree matching algorithm that finds maximum subtree isomorphisms between segmentation trees of two images is used to unsupervisely match images containing objects of different categories. After matching, the set of matched subtrees is used to derive a canonical representation by a tree-union operation. Besides the structure of the subtrees, the evaluation criteria considers also geometric and photometric properties of segmentation regions and their neighborhood. Consistencies in category trees can be used to extract *subcategories*, which are shared among exemplars [151].

Jin and Geman [61] developed a *"composition machine"*, which consists of a dense hierarchical set of units with sparse activations. They start with a *"Markov backbone"* and a

Bayesian framework. However, as compositional systems are non-Markovian, a non-Markov perturbation is used to assess the model; its function is to capture regularities of the arrangements in the *part-whole* relationship. Furthermore, as the model integrates evidence from a wide array of so called *bricks*, contextual information inherently contributes to the results of interpretation.

Fidler and Leonardis [34; 35; 36] show how the building blocks of the hierarchy can be learned. In their approach, parts are learned sequentially, layer after layer. In learning the compositions, local neighborhoods are defined in part-centered coordinate systems, and group into *s*-compositions, *s s* being limited to form compositions of a maximum of five subparts. The learning is statistics driven, where a selection process selects the stable compositions to pass to the next layer. The lower layers are learned to be category-independent, while upper layers are constructed by using specific categories; this is later extended to a learning of categories using multiple levels of the hierarchy [35].

Similarly, Piater et.al [115; 116; 117] devise a framework for unsupervised learning of visual feature hierarchies, which however starts with a sparse layer of local features based on interest point locations. Features at intermediate layers of the hierarchy are learned as compositions of subordinate features that tend to occur at stable positions relative to each other. The hierarchical structure is formulated as a graphical model, which provides a statistical representation of the variability of shape and appearance. Spatial relations are defined in terms of distances [116], distances and angle [115], or spatial probability density [117] of the relative configuration of a pair of features. High level features are inferred using Belief Propagation. In [117], a bottom-up and top-down feature learning is presented, which is inspired by Lee and Mumfords' computational model of visual cortex [79]. According to Lee and Mumford, the visual stream (LGT to V4) performs Bayesian inference within an undirected Markov chain, using both bottom-up input and top-down expectations. In [117], a Markov network is used to model and infer feature posterior probabilities based on pairwise compatibility potentials of features, which are directly related to pairwise spatial relations. The models result in about 10 layers of 10 to 100 vertices per layer, and can be trained to recognize several objects, while the authors do not report any experimental validation on categories.

Another related hierarchical model is the one proposed by Granlund and Moe [48], which is a concise proposal of an invariant representation, where invariant triplets of features are

hierarchically binded, learned, and used for recognition. Considerable invariance and compactness of this model yielded optimistic results , however the model was not extended to categorization.

4.7 Structural Matching

Several methods attempt at matching (hierarchical) structural descriptions of objects, i.e., descriptions that concentrate on object's shape. Advanced and efficient methods have been developed both for 2D and 3D matching, however their efficiency is often limited, when shape representations are difficult to obtain, e.g. in noisy images, or in occluded or partial 3D reconstructions. Structural representations can be viewed as shape signatures, which significantly reduce and abstract the rich appearance and/or 3D shape information to a simple low dimensional description. For example, Makadia et al. [90] use 3D shape signatures based on spherical harmonics, and Ohbuchi et al. [105] propose the use of distributions that are sampled from one of many shape functions. Chen et al. [13] proposed to represent 3D objects by their omnidirectional appearance, by constructing a *light-field*, 2D projections from a dense sampled viewpoint distribution of on a sphere, encoded with Fourier coefficients and Zernike moments.

A popular representation of shape is skeletonization, or the extraction of a 1D skeleton as a shape of curves that capture the essential topology of the underlying object. In 2D, the skeleton is defined as the medial axis, which is the locus of the centers of maximal inscribed discs, or, in other words, the connected set of points which are equidistant from at least two points on the boundary of the object. In 3D, the skeleton is derived using the *medial surface*, which requires more elaborated techniques [89]. *Shock graphs* [72; 136] are skeletons, represented as labeled, directed adjacency graphs: nodes represent protrusions, necks, bends and seeds, and each of these skeleton features has an order - typically, they are described as first to fourth order shocks. Adjacent shocks of the same order are combined to a single node, and each node is labeled by the distance form the curve, and with first order curves with orientation and length. The shock graph conforms to a non-context-free grammar. *Shock trees* that can be derived by cutting the loops in the shock graphs, offer several possibilities for robust matching algorithms that draw from graph theory.

If we define matching as a *bipartite graph matching* problem, relaxation methods can be used to find the solution [71]. Siddiqi et al. [136] use an eigenvalue characterization of a shock tree and a bipartite matching framework to match shock graphs. In general, spectral

properties of graph adjacency matrices are often used for matching graph and tree structures [96; 133; 134]: in this case, subtrees are characterized by the sums of eigenvalues of their adjacency matrices, which are structural signatures that are invariant to similarity transformations; the largest similar subgraph can be found by matching subtrees with similar sums of eigenvalues. The bipartite matching algorithm then recursively finds an optimal one-to-one match between a model graph and the graph of the image being interpreted.

Pellilo et al.[112] represented trees as association graphs (TAG) that retain the original hierarchical properties. Subtree isomorphisms can be then found by a quadratic programming framework that uses a maximal clique formulation. The methods in [112] and [136] focus on a one-to-one matching. As Dickinson argues, matching for object recognition *requires* a many-to-many shape matching solution, in particular because *"... one-to-one feature correspondence between exemplars in a given category may not exist at the level of extracted features, but may exist at the level of groups of features"* [22]. However the many-to-many matching problem is intractable if unconstrained, as any subset of features could match any other subset from the other exemplar, and requires specific heuristic solutions that translate the graph matching problem to a more tractable domain. As an example, Keselman and Dickinson [69] formulate the problem of finding a *Lowest Common Abstraction* (LCA), where objects match only at a certain level of abstraction, requiring thus a search for potential groupings of matches that lead to a plausible solution. Their method finds a LCA of two exemplars through a recursive decomposition of their silhouettes, which significantly constrain the otherwise intractable problem. Many-to-many matching has been attacked with the Levenshtein distance [131], i.e. by finding a minimal set of renaming, adding, deleting, merging, and splitting operations that transform one feature graph into another.

In [68], Keselman and Dickinson propose a skeleton matching algorithm that embeds the graphs with low distortion in a low dimensional Euclidean space. In the metric space, the many-to-many matching is solved using *Earth Mover's Distance* algorithm. Figure 4.1 shows a results of matching skeleton graphs, where corresponding groups of nodes are colored with same color. Graph embedding in an Euclidean space is only one of the many embeddings that are can represent the graph labeling and structure. Notably, Granlund [47] introduced a monopolar channel representation, which implies a mapping of signals into a higher-dimensional space in a way that introduces locality with respect to the geometric and property space.

The framework proposed in [134] extends the spectral representation from [136] to accom-

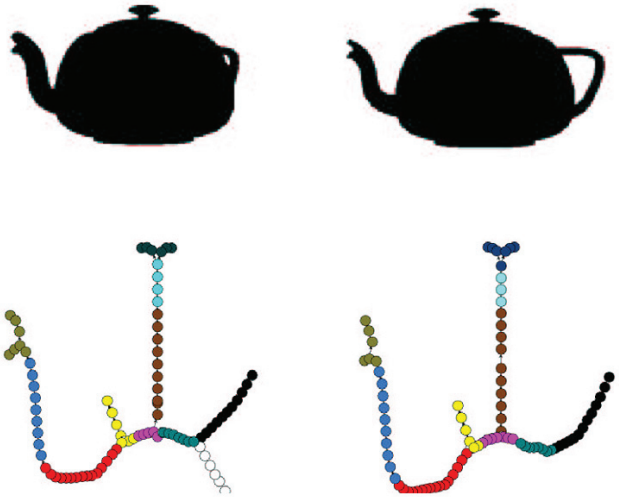


Figure 4.1: Many-to-many structural matching - The results of matching skeleton graphs with many-to-many correspondence, as presented in [68].

modate the shape of a tree in a low dimensional vector, where each of the nodes in the hierarchy is represented by a *topological signature vector* (TSV), composed form sorted sums of subgraph eigenvalues. In [135], the authors propose a framework for categorical shape recognition where a directed acyclic graph is used to represent hierarchical composition of parts, which are at the low level detected as multiscale blobs. In this work, the spectral graph matching from [134] by augmenting it with a geometric signature that describes the geometric properties of blobs. Blob and ridge extraction is performed using automatic scale selection [85]. For blob detection, they use the square of the normalized Laplacian operator, while for ridges, a multiscale ridge detector is used that detects elongated structures. To further characterize the shape features, *orientation* and *anisotropy* are calculated from eigenvalues of a windowed second moment matrix.

Multiple feature responses that originate form the same structures are removed by estimating the disjunct volume of their associated Gaussian kernel that reflects the spatial coverage of features. Ridges that lie on a longer ridge structure are linked using criteria of similar orientation and overlap. The resulting set of features is then assembled in a directed acyclic graph: the feature at the coarsest scale is chosen as the root, and features that overlap with the root are designated its children. This process is then iterated, until there are no more overlapping features - at this point, the feature at the coarsest scale of those that are unassigned is selected as the new root. A scene super-root note is then inserted to bind all root nodes. Every pair of features (every graph vertex), and sibling nodes that share the same parent, are associated with geometric attributes:

- *distance*
- *relative orientation*
- *bearing*, and
- *scale ratio*.

The authors formulate the matching problem as computing similarity between two graphs $G1$ and $G2$, that represent two object examples. As they state, the problem can not be considered as a label-consistent isomorphism problem, as *there may not exist significant subgraphs common to* $G1$ *and* $G2$. They therefore define two separate measures:

Structural similarity accounts for the similarity in the shapes of two graphs in therms of nodes, and branching distributions. Basically, this description is obtained with a procedure similar as the one in [134].

Geometrical similarity accounts for consistency in relative positions, orientations, and scales of nodes in two graphs. For every sibling pair of nodes, a distribution vector is computed for each of the geometric attributes. Given two graphs G and G', similarity between nodes can be computed in terms of the Earth Mover's Distance of their respective distributions (Figure 4.2).

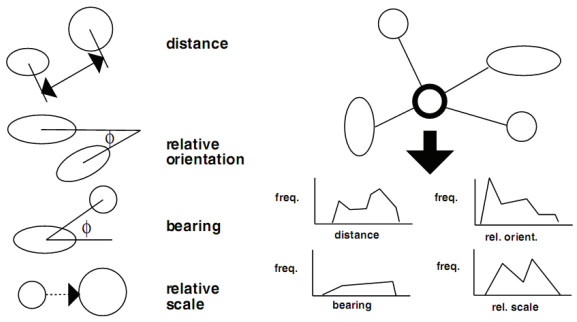


Figure 4.2: Geometric signature - Forming the geometric signature in [135]. Each node is represented by distribution vectors of geometric attributes such as distance, relative orientation, bearing, and relative scale.

This representation is then used in a graph matching framework, inspired by Reiner's algorithm for finding the largest common subtree. Nodes as first treated in a bipartite matching framework, annotated with the geometric similarity. The best pairwise node correspondence obtained after a maximum cardinality maximum weight bipartite matching is extracted, while the algorithm recursively matches the subtrees, resulting from splitting the graph at the matched nodes. The authors report excellent results on COIL-20 and ETH-80 databases.

However, as with most of the graph matching approaches, it is not clear how far could these solutions reach in heavily occluded scenes.

Belongie et al. [8] introduce an efficient shape context descriptor, which captures, for each point, the spatial distribution of other local points in the image. Correspoding points on two similar shapes will have similar shape contexts, which greatly improves the assignment of local matches. In [9], Berg, Berg and Malik reformulate the problem of recognizing object categories as a problem of *deformable shape matching*. In their interpretation, shapes are represented by a sampling of points from contours of shapes, which are then treated as keypoints, therefore described using local appearance and a similarity function. Deformable shape matching can be therefore seen as a search for an optimal matching between two sets of keypoints, taking into the account the cost of deformation, associated with the geometric distortion, and the smoothness of the deformable transformation. The rationale behind this approach is similar to that of our method. However, rather than to explicitly design the matching process, we discover the low level matchings through a gradual hierarchical construction of high level geometric consistencies. A similar undertaking could be achieved by spectral matching, as for example in the work by Hebert et al. [82], which however show only results obtained on DoG regions. In their latest work [70], the spectral matching technique is used for unsupervised discovery of category related features, where "soft" matches and spectral matching is used to generate a large number of links between training images, in which significant link groupings are found using link analysis.

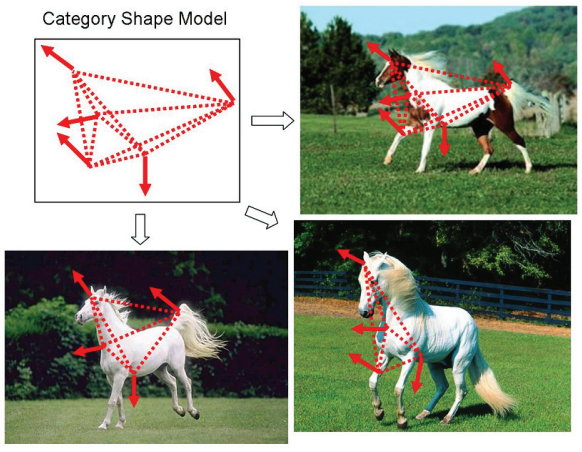


Figure 4.3: Category shape model from [83] - The category shape model in [83] is generated by integrating many example nonsegmented images of a category. The model is a graph whose edges are abstract pairwise geometric relationships.

Hebert et al. [83] propose a shape matching framework that uses a representation composed of cliques of fully-interconnected parts and encoding the pairwise relationships between them (Figure 4.3). The search for consistent correspondences is formulated as an energy optimization. They use an efficient algorithm to find optimal discrete sets of correspondences, and a learning algorithm to learn the parameters used in representing the geometric constraints. The category shape model in [83] is generated by integrating many example non-segmented images of a category. The model is a graph whose edges are abstract pairwise geometric relationships. The excellent results they report are obtained by using purely shape features and no appearance cues.

4.7.1 Categorization by association and similarity of exemplars

Most of the approaches to structural or appearance matching match images to a representation that is learned from a large number of exemplars. Few attempts try to minimize the learning set, and only some of them achieve an acceptable performance using a representation of only a few, or possibly only one exemplar. Methods that attempt to reformulate the categorization problem as recognition by matching, or recognition by association, for example by deformable shape matching [9], or by shock tree matching [135] can successfully match exemplars that exhibit some similarity (shape, structure), however, when used on a large number of objects that have to be "sorted" to categories, they all face the problem of defining an adequate similarity metrics.

In a framework for similarity based categorization, one can therefore choose to learn not the representation, but rather the distance metrics. Frome et al. [42] propose a framework where an image-to-image distance function is learned for all images. By matching triplets of images (two images from the same category and one outlier), a local weighing vector is learned that reflects the discrimination potential of each of the local features in an image and that maximizes the similarity between same category exemplars, and minimizes the similarity between the exemplar and the outlier image.

In Chum et al. [14], a similarity function is learned comprising a weighted sum of pairwise distance between visual words, pairwise distance between edge directions, and a cost factor that accounts for the deviation in aspect ratio over a set of training images. In an iterative process, rectangular regions of interest are constructed to encompass the features that (globally) maximize the similarity between exemplars, and those exemplars that generate a stable ROI are selected as representatives of a category. In detection, individual features are used

to generate ROI hypotheses (each of the features is spatially associated to the ROI rectangle). Multiple hypotheses are aggregate using mean-shift clustering, and the final detection is achieved by iterative cost function minimization.

In a similar fashion, Efros et al. [92] propose a framework to learn object similarity where individual distance metrics are learned for each of the exemplars. The learned metrics is based on linear combinations of elementary distances, such as color, shape, or texture; the elementary distances are calculated as distances between region-based features, where regions are obtained form multiple, bottom-up segmentations of the image.

♢

In this chapter we reviewed the computer vision approaches to categorization. In the next chapter, we give a more concise description of our architecture, and draw parallels to the existing work.

5

Our contribution

Our framework for object categorization is based on a number of the paradigms that were described in the previous chapters. In this chapter we therefore summarize the architecture and connect each of the key aspects with the related paradigms.

5.1 Architecture in brief

Our architecture implements categorization as matching to a *prototype icon based representation*. The prototype is an *image*, or an *icon*[1] of a representative exemplar of a category. The appearance and the structural properties of the prototype are matched to the image that is being interpreted. Matching is performed as a parallel and *hierarchical* process of *binding*, *matching* and *inhibition* of progressively complex features. Although the matching is hierarchical, and progressively operates on a coarse to fine progression of features, categories are not represented hierarchically, as features at higher levels of the hierarchy are always constructed in an ad-hoc manner. We therefore avoid any kind of high level representation, but design the matching as a local-to-global, bottom-up *ad-hoc construction* of high level features, where the construction of features depends on their relevance for the current matching process.

Hierarchical matching with the prototype is preceded by a learning of a sparse codebook representation of local appearance, and a learning of the structure of the geometric conceptual space based on the geometrical and appearance properties of the prototype. The learning step tunes the matching hierarchy to a specific prototype. Matching is a brute-force process in the sense that it is designed to find *any possible match* that could support a similarity between

[1] In general, image is an iconic representation when it stands for a 'motivated' sign that operates through some sort of visual resemblance, cf. [91].

the matched counterparts. Local bindings extend their receptive fields until they discover a satisfactory number of 'good' features, or until they exhaust the available information. Consequently, matching is largely indeterministic, as it can not be easily modeled for a certain category, and is specific for each of the prototype-exemplar pair. Intermediate features do not have an explanatory value which would be related to a representation of a category, nor they are compared to any type of stored representation, but are rather transient, dynamic, action-specific emergent constructs that explain only the similarities between the prototype and the exemplar.

An outline of the architecture is given in Figure 5.1. It can be shorty summarized by the following processing paradigms:

Prototype based representation A category is represented by a prototype image.

Discovery of scale and orientation invariant local features We develop a local feature detector which detects salient regions with intrinsic scale and orientation that responds both to edges and non-edges.

Learning a sparse codebook of local appearance The appearance content in local regions is encoded using a small codebook. The codebook is derived by learning the independent components of the signal encompassed by local regions.

***Gestalt* grouping of low level features** The spatial redundancy of features is reduced by a *Gestalt* grouping step.

Learning the structure of the geometric conceptual space By learning the structure of the geometric conceptual space based on the geometrical properties of the prototype, a conceptual representation of local geometric relations is provided that adequately represents the variations with respect to a certain category of objects.

Hierarchical synchronous feature binding and inhibition Neighboring features in a receptive field are grouped to n-plets with an explicit encoding of between-feature geometry. The prototype image is used to match the features at each level of the hierarchy and to inhibit the features that do not lead to a potential high level match.

Local to global matching Features are synchronously constructed, matched and inhibited at k levels of the hierarchy. Matching starts at a local level; at each level the receptive field expands until it encompasses the whole object.

Massive parallelization Each of the bind-match-inhibit operation is performed autonomously and independently of the global process. The procedure is therefore highly parallelizable.

Categorization as successful construction of high level features High level matches are constructed by a hierarchical matching of object (or scene) to a prototype. The number of successfully constructed matches reflects the similarity of the objects.

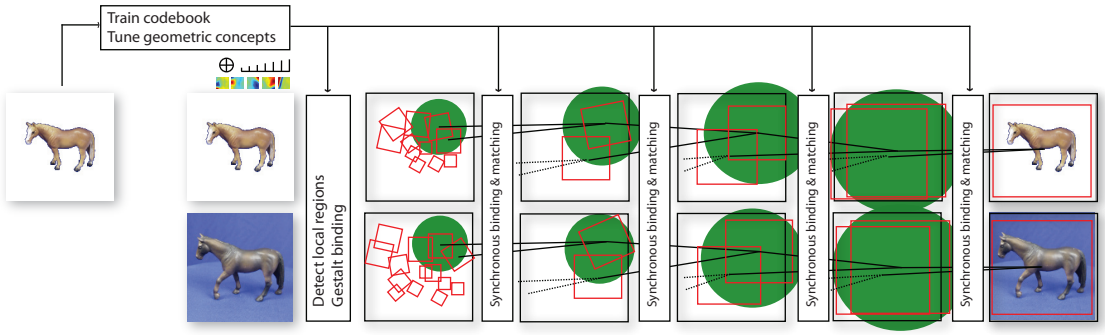


Figure 5.1: Framework for hierarchical categorization and matching - The training stage is followed by a number of hierarchical processing steps of matching between a prototype and an exemplar image. For clarity, only one binding branch that links two features at the lowest level is depicted in detail. The red squares stand for the area encompassed by a feature, while the green area represents the receptive field of the feature. The illustration does not represent true data.

We start with a low level description, which defines relatively stable local regions of interest based on their appearance. Features are dynamically constructed in order to search for possible matches between the view being interpreted and a prototype view. The composite features describe geometrical and photometric properties of a local area that expands through the hierarchy, until a focused response with a *receptive field* that covers the whole object is found. As we construct high level features for each categorization round, we do not match the high level features to a predefined set of features, but rather evaluate the similarity between the prootype and the image that is reflected in the number of successfully constructed features at each level of the hierarchy. The advantage of such an approach is that we do not have to define a category with respect to a set of learning examples, and similarity between objects from the same category can be articulated using many different local geometrical and appearance features.

In the training stage, we first learn a sparse codebook as a set of filters which impose a maximally sparse pattern of activation. The codebook is learned on appearance patches of local regions of interest. This codebook can be learned based on local regions of the prototype image alone; additional (correctly categorized) images can be used to refine the codebook in a continuous manner, as there is no higher level of the representation that would

depend directly on the codebook.[1] We intentionally keep the number of these filters low—in that way, local regions are informative in terms of their structural properties (orientation and scale), while the information on their appearance is only rudimentary. Groupings of local features that conform with a subset of Gestalt rules (that group essentially co-centric or co-linear concatenations) are then formed to provide a basic description of image elements such as lines and parts of curves. In the second stage of training, a quantization of the geometric conceptual space, i.e., the angle, length and size parameters, is learned based on the prototype image. Both the sparse codebook and the quantization of the geometric concepts are then used in hierarchical binding.

The low level features are then associated between the prototype and an exemplar image: in each image we bind features that fall within a receptive field, targeting an upper level were features are grouped in pairs and are determined by their low-level identity and by their local relative geometry. By simultaneous region growing and selection, we can steer the binding process by selecting only those pairs that match in both images, as only those are potential candidates for a match at a higher level. Following that procedure, we can repeat the process until on some level the binding process stabilizes in a set of large receptive fields, typically enclosing the object we search for, or until the process terminates because a lack of successful matches at lower levels.

As a result we therefore obtain a set of matches at level N, where each of the responses can be tracked down the tree to the constituent features at levels N-1, N-2,... 0. The number of high level matches reflects the similarity of the prototype to the part of the scene that is encompassed by the constructed features.

The problem of matching could be interpreted as a search for a sub-graph within a noisy annotated graph, where annotations denote the between–nodes spatial relations and node appearance characteristics, and is therefore extremely complex and time consuming. Our heuristic approach independently processes of local neighborhoods to find a large number of partial matches. The method is inherently parallel, and can be in in theory distributed to any number of processors, achieving thus a constant time complexity with a fixed number of levels. If we want to check for n categories, we have to either perform n sequential queries, which results in a linear time complexity with respect to the number of categories, or perform n queries in parallel, achieving thus a constant time complexity with respect to the number of categories.

[1] Incremental updating of hierarchical representations with multiple learnable layers is a non-trivial task.

5.2 Prototype based representation

A category is represented by a prototype (see Subsection 2.3.3) in a canonical view. As the prototype based representation introduces the graded notion of categories [127], it naturally leads to categorization by association [92]. We measure the similarity by the number of constructed activations that the stimuli triggers. Due to its simplicity, the prototype representation can be thought of as an *envelope*, which is used to unfold the *familiarity* of an object seen in the past, e.g. a schemata that informs how a certain category of objects is perceived. The prototype image could be evocated by means of mental imagery (see section 3.10), or could be composed of a series of basic atoms, such as gestalts, and their inter-relations, which can be expressed in terms of an "*operational knowledge*" (how to construct, where to look).

5.3 Scale and orientation invariant local features

We base our binding process on local regions that are scale and orientation invariant. Invariant regions have been widely used in the last decade, and we already gave a brief overview of local recognition methods that use invariant regions in Section 4.4. Regions are designed to encompass a scale invariant image region with some minimum of appearance information. In order to enhance the structural informativeness of the local regions, we promote edge centered regions. Several other methods used edges or edge contours [9; 109; 132; 156]; however our architecture is largely agnostic to the type of features, as long as they are abundant, stable, and have intrinsic orientation and scale. As edge responses have a poorly defined scale, we use a redundant strategy of describing those regions at multiple scales.

Most of the existing local approaches mentioned in Section 4.4 typically encode the visual information with a relatively high accuracy even at smaller scales. In general, this prevents from generalizing across category members, which could match only in global appearance, or even only by structural similarity. The local features are therefore designed to provide spatial, scale, and appearance anchoring by not being too informative about the local image content, while still providing essential cues on orientation and scale.

5.4 Perceptual learning of local appearance

We characterize the invariant regions using a sparse codebook. Sparse coding as an codebook learning/discovery has been most notably promoted by Olshausen and Field [108], Bell and

Sejnowsky [7], and Hyvarinen et al. [58], who analyzed the codebooks that can represent natural images. In contrast to their work, where all of the patterns that appear in nature are regarded as equally probable, we use the Independent Component Analysis of pre-attended patterns, i.e. the regions of interest detected by the invariant region detector, and, due to the implicit scale and orientation of thee regions, derive much smaller and more informative codebooks.

5.5 Gestalt binding of low level features

To reduce the complexity, we adjoin local groupings of features with congruent edge responses at different octaves of the *scale space*. This grouping is determined according to the laws of the *Gestalt* theory: *proximity*, *similarity*, *closure*, and *simplicity*, which can be summarized in the idea of a good continuation. Gestalt rules have been widely studied, and several frameworks use them implicitly or explicitly [132]. Most interestingly, in the seminal work of Fidler and Leonardis [36], it is shown that learned spatial statistics of hierarchical compositionality of objects largely reflects the gestaltists principles of organization.

5.6 Hierarchical synchronous feature binding

Between-feature geometry is gradually integrated in a way that is inherently local, but slowly acquires a global character, and therefore reflects a bottom–up compositionality. This is achieved by a binding process of multiples of features into $n-$plets, which represent geometrical relations between n local regions, preserving their weak appearance characteristics.

The motivation for our hierarchical architecture is to efficiently bind appearance and structure in a way that gradually encompasses the object, starting from its locally coherent bindings and extending to the whole object. Due to the local to global binding process, the high level bindings are spatially grounded in their low level subordinate activations. This differs from hierarchical models where the spatial responses are abstracted by polling across spatial dimension to achieve invariance to position.

5.7 Coarse to fine matching

We implement a coarse-to fine matching strategy which has many predecessors. Notably, the most related architectures are those of Dickinson et al. [134; 135; 136], or Todorovic

and Ahuja [150]. However, in contrast to Dickinson's work, we do not explicitly model the global scene structure, but rather construct in each event of recognition in a constructive and opportunistic manner. In contrast to Todorovic, we do not learn the configurations of local parts. Furthermore, we use a more dense low level representation. Hierarchical models as the one by Fidler and Leonardis [34; 35; 36], or Riesenhuber and Poggio's H-MAX architecture [124], start at the pixel level and derive specific intermediate parts that are used to represent categories.

For reasons that will become evident through the next chapter, we decided to base the representation on a discrete characterization, and we research towards a framework that is essentially non-statistical. The optimistic results point to possible solutions that differ significantly from the current statistical approaches to cognitive modeling.

5.8 Categorization as successful construction of high level features

Low level percepts are combined in a way to trigger activities at higher levels of the hierarchy, and, although the activations at separate levels can be used to partly reconstruct the signal, they do not play the role of representing some part of the object in the sense of its symbolic value in the act of comparison. Activations, or features, are triggered only to construct a viable solution to a structural match between the stimuli and the prototype, and are in this view predictable only in a broad perspective.

As the system typically finds several thousands "solutions" that explain an object, an external observer could not find stable patterns of activations across a category, as they in general differ form one exemplar to the other. Higher order features are therefore emerging and transient, and the computing units serve merely as support in a process of active recombination of stimuli. In this way, the levels of processing are agnostic to the type of signals that are being binded, which opens up various possibilities to include e.g. contextual information, or to feed activity data into the hierarchy.

This could serve as a basis for modeling active models of perception, as for example the model by Noë [104], who proposes that seeing is an act of *actively reaching*, therefore more analogous to touch than to standard models of vision, or the earlier model of O'Regan and Noë [110]. Bar [6] proposed a predictive mechanism that continuously generates predictions

that approximate the relevant future based on associative processing. Analogies can be based on similarity on various levels, which includes perceptual, conceptual, or goal oriented dimensions. Analogies then trigger basic representations in memory (which can also vary in dimensions), and these representations have to be verified against perceptual data, a process which could be performed as proposed in our work.

Although iconic representations were often criticised, and are currently overshadowed by feature based representations, it was lately proposed that they could play a role in enactive models of cognition. In particular, the role of *imagination* in cognition was only sparsely researched, and was too often regarded only as part of *fantasizing*, or as taking part in problem solving. The idea that imagination could support categorization is easily confronted by questions on the mechanism that would trigger the correct mental image that is compared to the scene. Although there is no definite answer to this, the problem seems solvable if one envisions perception as a continuous formulating of hypotheses about the world.

5.9 Biological plausibility

We see our contribution primarily as a part of the mosaic in a sketch of a larger cognitive system, which simultaneously implements other cognitive functionalities, in particular *intentionality* and *attention*. Most of these functionalities have roots in study and observation of biological cognitive systems, and are tested with computational models that exhibit some level of *biological plausibility*. Note that the development of artificial cognitive system does not require its design to be inspired by nature. There are however strong reasons to believe that this difficult undertaking has more chances to succeed if it considers at least some of the lessons from nature, and if it implements biomimetic subsystems. One of the strong arguments for biomimetic cognition is that, by developing a non-biomimetic cognitive system, the end product could never be evaluated for its cognitive capabilities that are known to us as human beings. If there would be no one-to-one mapping between cognitive functionalities of man and machines, the existence of e.g. introspection or awareness would require a testing where these functionalities would be compared to that of natural cognitive systems. However, up to now, there has been no agreement on how would such a test be designed.[1] Of course, nature can be mimicked at different levels of functionality. Some connectionist models for example attempt to model the neural substrates of cognition, and concentrate on

[1]Notable attempts of designing a test which evaluates a system's "intelligence" run from René Descartes [21] to Alan Turing's test [154] and John Searle's Chinese room experiment [130].

properties and connectivities of computing units, modeling thus neurons an neuronal layers. Other approaches might target properties of biological systems that are exhibited at a higher level of abstraction, for example by replicating the cognitive modularity, by modeling the activities of functional units as a response to objects and events, or by implementing optimization or emergent criteria that are believed to be biologically relevant. While our approach adopts several ideas that arise from studies of biological cognitive systems, such as hierarchical processing and distribute computation, it also relies on heavy parallelism, which can be today found only in biological computational systems.

This chapter described our framework for hierarchical structure matching. Next, we describe the overall hierarchy, beginning with the bottom levels of detection of local regions, followed by the procedure of supervised feature binding at higher levels of the hierarchy. Details on learning the matching parameters based on a prototype, and other aspects of the architecture are introduced with representative examples.

Part II

Hierarchical framework for categorization

6

Detection and characterization of local features

In this chapter we describe the first stage of image processing that results in a set of local features that form a basis for hierarchical binding and hierarchical matching. This stage comprises the detection of invariant regions, the learning of the codebooks, and the gestalt grouping.

6.1 Problem definition

At the very early stage, the image is nothing but a collection of discrete values of pixels that carry intensity and color information, or a response of local light receptors on the eye's retina. In this chapter we describe how this low level signal can be abstracted to a more manageable representation, that is far less redundant than the original pixel values. A possible bypass to a characterization at a *pixel-level*, is a pre-selection of a limited number of local *foci of attention*, or local regions, that encompass enough characteristic information, and are stable enough to be consistently identified across view. In searching an ideal attention operator, we want meet the following requirements:

structural informativeness - the operator has to find regions that can be characterized by a characteristic scale and a unique orientation

repeatability across category - the operator has to consistently find the same regions across different exemplars of a category

viewpoint invariance - the operator has to exhibit some level of invariance to viewpoint changes

density and redundancy - the operator shall yield a relatively large number of responses that redundantly describe the object; this increases the robustness and increases the probability of a match between objects of same category that share only a subset of features.

To characterize the regions of interest yielded by the operator, we aim at an operator that would assign one of the codebook entries for each of the regions. The codebook should be small, therefore the operator shall discriminate only the very basic appearance characteristics. We aim at:

low informative value - the low information value is needed to prevent false negative matches at the resolution where no significant differences between objects are expected.

sparse response - the probability density function of a given local region class shall be constant, yielding thus an efficient coding scheme and minimizing the errors in assigning the class label.

inversion invariance - the operator shall be invariant to inversion of background and foreground intensity values.

We therefore aim at a basic feature level that would abstract the appearance information and concentrate on scale, orientation, and viewpoint invariant regions, that densely characterize the structure and the appearance of the object. Each of the features should be assigned to an appearance class, and this assignment shall reflect the very basic appearance properties in order to reject only the most diverse local structures.

6.2 Regions of interest

In the first stage of processing, regions of interest (ROI) are detected that form the basis for subsequent structural grouping. As it was already discussed in Chapter 4, many ROI detectors exist that give qualitatively and quantitatively different results. As an example, the difference of Gaussians (DoG) detector [87] selects rather sparse areas with good localization properties. The Kadir–Brady detector [64] focuses on centers of high entropy areas. Regions that lie on edges are often omitted, as 1) they do not contain discriminative information on appearance and 2) localization is unreliable or impossible. While such detectors do provide a solid basis for appearance based recognition, they are in a large part independent on the structural properties of objects. E.g., while homogeneous areas or symmetry will be consistently discovered, object boundaries or other typical shape elements will be omitted.

When the aim is to find salient information that describes *shape*, edge based features can be used [156].

We design the detector to obtain salient responses on regions with informative appearance and simultaneously on structurally significant regions, such as edges. To find local regions at different scales, we search for keypoints in a scale space of image $I(x,y)$. As it is described in [85; 87], the scale space of an image is defined as a function $L(x,y,\sigma)$, that is produced as a convolution of the input image with a Gaussian kernel $G(x,y,\sigma)$ of a variable scale:

$$L(x,y,\sigma) \quad = \quad G(x,y,\sigma) * I(x,y) \quad . \tag{6.1}$$

At each octave, starting from the original image size $(c \times r)$ increased by a factor of two, we calculate three scales per octave, and after each octave is processed, we down-sample the Gaussian images by a factor of two. In this way we obtain $(\log \frac{\min(c,\, r)}{\log 2} - 2)$ octaves, where the local extrema in scale space are identified [87]. As in [87], we reject the extrema that reside in neighborhoods of low contrast, however, in contrast to [87], we retain the extrema that satisfies the *edginess* criteria. The probability that an extrema lies on an edge is computed using a 2×2 second-order Hessian matrix

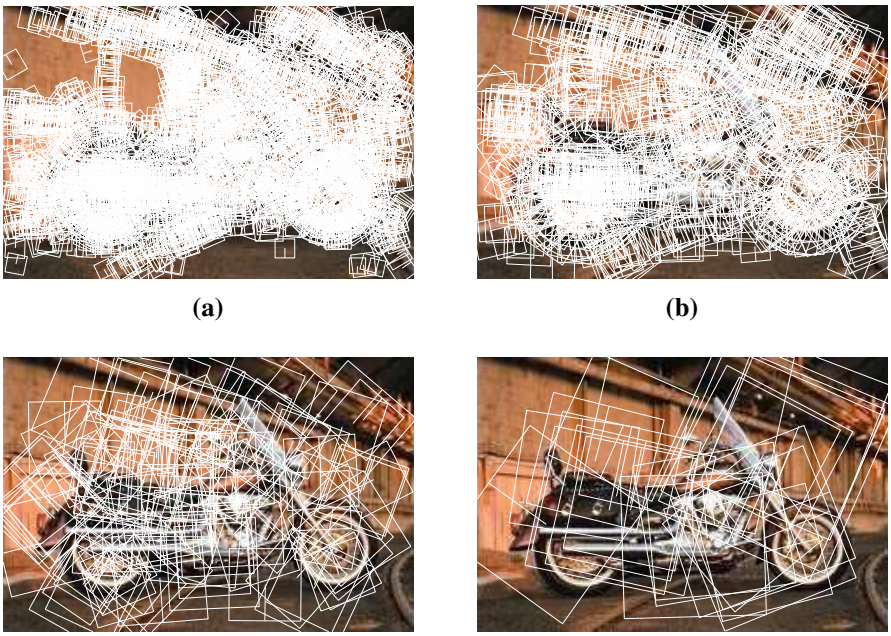
$$\mathbf{H} = \begin{bmatrix} I_{xx} & I_{xy} \\ I_{xy} & I_{yy} \end{bmatrix}, \tag{6.2}$$

and the criteria $\frac{Trace(\mathbf{H})^2}{Det(\mathbf{H})} \geq \frac{(r+1)^2}{r}$ proposed by Harris and Stephens [56], where r is the ratio between the eigenvalues of $\mathbf{H}$. In this way, we obtain a number of local edge-centered regions, which are characterized by the scale and the octave on which they were detected. As responses on edges tend to be dense, we reduce their number by a local inhibition step, which inhibits edge responses within a fixed radius of k pixels. The orientation of each local region is calculated as the direction where the histogram of gradients for the region area reaches a global maximum [87].

The resulting local regions K_i are characterized by their center ${}^{c}\mathsf{K}_i$, orientation ${}^{o}\mathsf{K}_i$, and radius ${}^{r}\mathsf{K}_i$, which is calculated from their generic scale and octave [87]. Regions discovered at the first four octaves can be seen in Figure 6.1.

6.3 Learning a sparse codebook

The function of regions of interest is to offer spatial support for features at higher levels of the hierarchy. At the same time, their invariant properties provide scale and rotation invariance.



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

Figure 6.1: Keypoints - Keypoints K_i discovered at first (a), second (b), third (c) and fourth octave (d). Rectangles encompass the area of the local region and indicate its generic orientation.

The appearance information that the regions encompass can be characterized by descriptors, calculated on normalized patches. In a non-hierarchical framework, this information could be fed to a classification framework: for recognition, invariant descriptors that selectively represent a highly discriminating feature, e.g., a typical pattern, can be memorized and retrieved during recognition. For categorization, a more elaborate feature selection strategy has to be adopted: for example, a restricted codebook of features can be constructed, by e.g. vector quantization, to represent typical local features representative of a category [80]. To reduce the dimensionality, one could describe the local regions in low-dimensional appearance subspaces [66].

In spirit of a coarse-to-fine hierarchical description, we encode the local appearance with a small codebook. As it will become evident later, this decision results in a massive amount of false positive matches at lower levels of the hierarchy, however it offers a good basis for the discovery of consistencies between the prototype and the stimuli at higher levels. It also provides a balance between specificity and robustness; if the low level features would be too specific, only a low number of matches would be found, and these matches would not provide enough support for forming of hypothetical structural couplings at higher levels.

The aim is therefore to use a small codebook which still provides enough information to

discriminate between a number of basic classes of patterns. Furthermore, we want to minimize the uncertainty of discrimination between the basic classes, i.e. to derive a classification criteria with an as sparse response as possible. We also want this layer to be adaptive - i.e., we want to be able to learn the basic codebook from one or more examples of a category. In this way, the discrimination ability will be fitted to a restricted set of possible instances of local regions that typically emerge for exemplars of a certain category.

In practice, a sparse codebook could be achieved by e.g. vector quantization. However, inspired by the properties of neurons in the primary visual cortex, several researchers speculated that the optimal coding could be achieved by an unsupervised learning of a factorial components coding of visual features, where the factorial code would satisfy several requirements: in particular it would reduce the redundancy and enforce statistical independence. Techniques such as Principal Component Analysis (PCA) [66; 102; 106] can be used to remove the dependencies of a second order, i.e. to decorrelate the signal. Field [37] however proposed a *minimal entropy coding* as the principle that best accounts for the neuronal organization. Such a coding, which results in maximally sparse responses to a set of filters, can be computed by a learning algorithm based on nonlinear information maximization, which is a form of solving the *Independent Component Analysis* (ICA).

Olshausen and Field [108], Bell and Sejnowsky [7], and Hyvarinen et al. [58], among others, have demonstrated how to compute optimal sparse factorial codings for natural images, and compared the results to evidence from neuroscience. When trained over random patches extracted form natural images, the ICA results in a codebook that resembles the receptive fields of cells in the primate primary visual cortex. Lewicki and Olshausen [84] compared the ICA codebooks to other classes of factorial codes, such as Gabor, PCA, Haar, Fourier, or Daubechies filters, and demonstrated a superior performance of ICA in terms of coding efficiency.

As we want to derive an optimal sparse coding of visual features encompassed within local regions, we calculate the independent components *only on patches defined by local regions*. The advantage of this localized codebook is twofold. Firstly, as patches are oriented according to the direction of the maximum of the histogram of gradients, and have an intrinsic scale, they already implicitly encode a large amount of variance that would otherwise have to be captured by the codebook. Secondly, the selectivity of the detector ensures that only a very limited subset of all possible patches enters the computation, therefore further diminishing the variance in appearance that has to be encoded. If we calculate the ICA codings only on

one category, we obtain a codebook of components that clearly encodes some of the typical aspects of the exemplars in the category.

To calculate the ICA codebook, we use the *FastICA* algorithm that is part of the toolbox `imageica` [58]. Let the size normalized image vectors $\mathbf{x}$ (resized image patches in vector form) that enter the computation be sorted in an image matrix $\mathbf{X}$ with zero mean. An image patch $\mathbf{x}_i$ can be represented by a linear superposition of sources $\mathbf{b}_i$,

$$\mathbf{x}_i = \sum_{j=1}^{N} a_{ij}\mathbf{b}_j \tag{6.3}$$

weighted by coefficients a_{ij}. The matrix form of Equation 6.3 is

$$\mathbf{x} = \mathbf{A}\mathbf{b} \ . \tag{6.4}$$

If $\mathbf{u}$ represents the source as recovered form the original signal the mapping from the signal $\mathbf{x}$ to $\mathbf{u}$ can be written as $\mathbf{u} = \mathbf{W}\mathbf{x}$, where $\mathbf{W} = \mathbf{A}^{-1}$, if $\mathbf{A}$ forms an invertible linear system. ICA calculates such $\mathbf{W}$ and $\mathbf{b}$ that the sources are statistically independent,

$$f_{\mathbf{b}}(\mathbf{b}) \approx \prod_j f_{b_j}(b_j) \ . \tag{6.5}$$

Another assumption is that all of the mixing sources shall be non-Gaussian; this leads to efficient optimization algorithms, where non-Gaussianity measures, such as *kurtosis*, are used to estimate the mixtures. As $\mathbf{W}$ is the filter matrix which is used for calculation of the vector product, it represents the hypothetical computational units in network models of mammalian brain.

In practice, the image data is preprocessed by centering, whitening, and dimensionality reduction. This removes the linear dependencies and therefore simplifies the computation of ICA; by reducing the dimensionality, fewer parameters have to be estimated, and by the implicit orthogonalization, we also orthogonalize the estimation problem. In order to whiten the image data, the covariance matrix of $\mathbf{X}$ has to equal the identity matrix:

$$E\{\mathbf{X}\mathbf{X}^T\} = \mathbf{I} \ . \tag{6.6}$$

The data is whitened using *Singular Value Decomposition*, therefore by estimating $E\{\mathbf{X}\mathbf{X}^T\} = \mathbf{U}\mathbf{L}\mathbf{U}^T$, where $\mathbf{U}$ is the orthogonal matrix of *eigenvectors*, and $\mathbf{L}$ is the diagonal matrix of eigenvalues. In order to reduce the dimensionality, only k eigenvectors with the highest eigenvalues can be used. Data is therefore whitened by

$$\hat{\mathbf{x}} = \mathbf{U}_k\mathbf{L}_k^{-1/2}\mathbf{U}_k^T\mathbf{x} \ , \tag{6.7}$$

transforming Equation 6.4 to

$$\hat{\mathbf{x}} = \mathbf{U}_k \mathbf{L}_k^{-1/2} \mathbf{U}_k^T \mathbf{A}\mathbf{b} \tag{6.8}$$

$$\hat{\mathbf{x}} = \widehat{\mathbf{A}}\mathbf{b} \ . \tag{6.9}$$

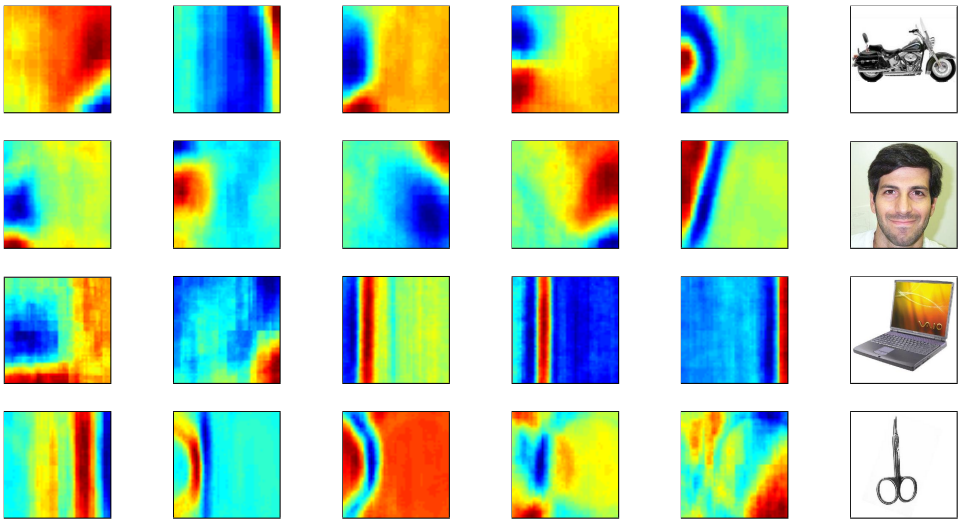


Figure 6.2: ICA codebooks - Codebooks derived by ICA for *Motorbikes*, *Faces*, *Laptop*, and *Scissors*.

The orthogonality of $\widehat{\mathbf{A}}$ results in less degrees of freedom in ICA estimation; $\widehat{\mathbf{W}}$ can be calculated by $\widehat{\mathbf{W}} = \widehat{\mathbf{A}}^T$. The FastICA algorithm finds the ICA components by iteratively finding the maximum of non-Gaussianity of $\widehat{\mathbf{w}}^T\hat{\mathbf{x}}$, where $\widehat{\mathbf{w}}$ is one of the independent components, using the measure of approximated negentropy in an approximative Newton iteration. Typically, we use approximately 5000 patches to calculate a codebook of 5 to 10 codes by reducing the dimension to $k = 40$. The iterative scheme typically converges in a few dozens steps.

Each of the regions K_i is characterized by the response to the set of filters: if the appearance content of the region responds with a maximum response for a filter $\mathbf{w}_m$, its class is ${}^w\mathsf{K}_i = m$.

To fine-tune the filters to be sensitive to a category of objects, we train for ICA filters on appearance patches of local regions of interest of the prototype image. Additional (correctly categorized) images can be used to refine the codebook in a continuous manner, as there is no higher level of the representation that would depend directly on the codebook. As it turns out, a low number of ICA components satisfactorily represents the variation across patches for most of basic categories of objects, such as for example motorbikes and faces. Nevertheless,

the ICA codebook manages to capture the distinctive features of a category, such as parts of wheels in motorbikes, or parts of scissors (Figure 6.2). For complex shapes, such as motorbikes, the components clearly depict frequent local structures, and some prominent features at a larger scale, such as parts of wheels. For simple shapes, such as scissors, we can observe some redundancy of ICA representation, as the distinctive structures could be represented well with an even lower number of filters.

6.4 *Gestalt* grouping

As it has been advocated by the *Gestalt* psychologists, the phenomenology of visual perceptual grouping can be attributed to a series of rudimentary principles which affect the perception at the early stage, thus reducing the inherent complexity of the image that falls on the retina to a series of *Gestalts*, i.e. percepts which have, in a typical *gestaltist* terminology, a higher informative value than is the sum of the informativeness of its constituent parts. The grouping adheres to laws of *proximity*, which binds percepts according to their spatial distance, *similarity*, which binds percepts that are similar, *closure*, which reinforces percepts that lead to continuity, and *simplicity*, which reinforces the simplest grouping hypothesis.

Algorithm 6.1: Gestalt binding

Input: Keypoints K, angle α

Output: Local regions H^0

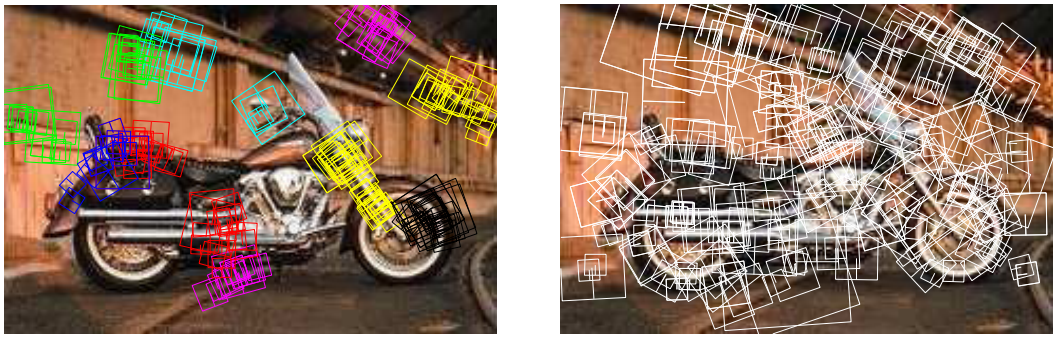
repeat
  Pick a pivot $\mathsf{K}_p : \mathsf{K}_p \notin \mathsf{C}$;
  repeat
    Find all $\mathsf{K}_j \in \mathit{neighborhood}(\mathsf{K}_p) : ({}^w\mathsf{K}_p = {}^w\mathsf{K}_j) \wedge (\mathsf{K}_j \notin \mathsf{C})$;
    Bind all $(\mathsf{K}_p, \mathsf{K}_j)$ that satisfy $abs({}^o\mathsf{K}_p - {}^o\mathsf{K}_j) \leq \alpha + \varepsilon$;
    Select a new pivot K_{n_0}, $m_0 = \arg\max_n ||\overline{{}^c\mathsf{K}_p {}^c\mathsf{K}_n}||_2$;
    $\mathsf{K}_p = \mathsf{K}_{n_0}$
  until *count of bindings in neighborhood satisfies criteria* ;
  Add all binded to $\mathsf{C}_{\{i\}}$;
until *all* K_i *visited* ;
From each $\mathsf{C}_{\{i\}}$ construct a region H_i^0;

In order to group the edge keypoints into significative fragments, we implemented a simple iterative algorithm which sequentially expands and shifts the receptive field until the number

of edge keypoints that meet a subset of the Gestalt criteria is above a certain threshold. We group the edge features according to:

- *Proximity*: Only edge peaks that fall within a receptive field R are grouped
- *Similarity*: Only edge peaks with the same sparse filter response class are grouped
- *Common fate*: Only edge peaks that have a neighbor within the receptive field that forms a pair of congruent angles of α radians when connected are grouped.

We perform the binding process for for a series of angles α. Typically, we set α to $\alpha \in \{0, 0.2, 0.4\}$ radians, with a tolerance of $\varepsilon = 0.1$ radians.



(a) **(b)**

Figure 6.3: Gestalt features at H^0 - Features K_i are grouped by Algorithm 6.1. (a) shows some of the feature groupings that combine into features H^0. (b) depicts all H^0 features.

Based on the output of Algorithm 6.1, we define region of interest as circumvent areas encompassing each of the resulting groupings. The regions inherit the class parameters of constituent edge keypoints; therefore ${}^{w}\mathsf{H}^0 = {}^{w}\mathsf{K}_i$. The orientation and the scale of the region is calculated from an eigen-representation of the spatial extent of subordinate features. The orientation ${}^{o}\mathsf{H}^0$ of the region equals the direction of the eigenvector with the smallest eigenvalue in a decomposition of the covariance matrix that encodes the shape covered by the subordinate features. The size of the feature is proportional to the ratio of the corresponding eigenvalues, ${}^{r}\mathsf{H}^0 = \lambda_1/\lambda_2$, and the center ${}^{c}\mathsf{H}^0$ is the median of x and y coordinates of grouped keypoints.

Regions of interest now cover cocentric edge regions, parts of linear regions, and the remaining non-edge regions that were not binded by detected by Algorithm 6.1. An example of all regions at H^0 in an image can be seen in Figure 6.3.

♢

At this point, a set of local features represents the information in the image, encompassing a coarse information on local appearance, local orientation and scale. Once this information has been derived, it can be used to construct more complex and more discriminative features. In the next Chapter, we therefore describe in detail the hierarchical binding of features.

7

Hierarchical binding

At this stage, the informative value of features at H^0 is negligible for any matching or recognition task. In the following sections, we present the hierarchical binding of features that introduces between–feature geometry, while promoting the characteristics of the regions at the lower levels.

7.1 Problem formulation

We want to construct a coarse to fine, local to global progression of features that would eventually encompass the whole object and could be used to efficiently indicate structurally matching parts of objects from the same category. Descriptions should first encode the local between–feature geometry, and would gradually encompass and describe the structural properties of wider localities. One possible solution is a hierarchical binding, where, at each level of the hierarchy, a certain number of features bind to construct a more complex feature. We designate the hierarchical binding process to start at level H^0 with a binding $\mathsf{H}^0 \rightarrow \mathsf{H}^1$, and proceeds for an arbitrary number of levels. At each level, features at the lower level locally bind with features at the same level that fall within a feature's *receptive field*. Features at a higher level are therefore supported by n binded features at the subordinate level, forming thus a feature n-plet. In theory, one could bind n-plets for an arbitrary n, however our initial experiments indicated that 2-plets or 3-plets are the choices to consider for most object categories. The binding of triplets was successful only for simple structured objects, as triplets exhibit a significant distinctiveness already at the second level of binding. We therefore concentrate on binding of *twoplets*, or pairs of features.

7.2 Geometric attributes

The basic geometric attributes that describe the relative positioning of two circular regions are depicted on Figure 7.1. From two regions at level H^k, we derive a new region at level H^{k+1}, which is defined by the following attributes:

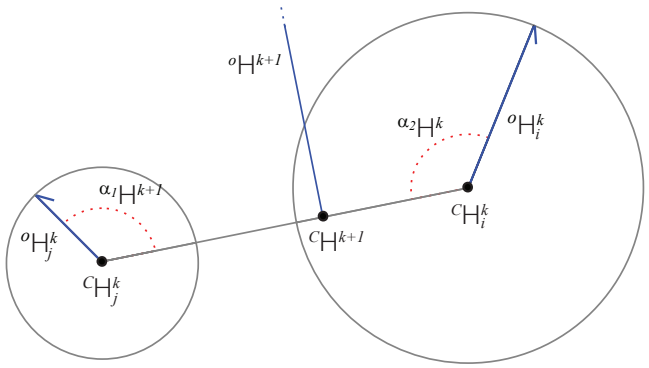


Figure 7.1: Between–feature geometic attributes in a *twoplet* - At binding of level H^k, features at H^k are binded as a novel twoplet at H^{k+1}. The geometric attributes describe the ordering, the relative size, the relative length, and the inner angles with their relative orientation.

1. *Ordering*: binded features H_i^k and H_j^k are ordered according to their size, so that

$$^{r}\mathsf{H}_i^k > {}^{r}\mathsf{H}_j^k$$

2. *Relative size*: $^{s}\mathsf{H}^{k+1}$ equals the logarithmic ratio

$$^{s}\mathsf{H}^{k+1} = log\left(1 + \frac{^{r}\mathsf{H}_i^k}{^{r}\mathsf{H}_j^k}\right)$$

3. *Relative length*: $^{l}\mathsf{H}^{k+1}$ equals the logarithmic ratio

$$^{l}\mathsf{H}^{k+1} = log\left(1 + \frac{||\overline{^{C}\mathsf{H}_i^k\,{}^{C}\mathsf{H}_j^k}||_2}{^{r}\mathsf{H}_i^k}\right)$$

4. *Angles to connector*: the inner angles $^{\alpha_1}\mathsf{H}^{k+1}$ and $^{\alpha_2}\mathsf{H}^{k+1}$ between vectors $^{o}\mathsf{H}_i^k$ and $^{o}\mathsf{H}_j^k$ and $\overline{^{C}\mathsf{H}_i^k\ {}^{C}\mathsf{H}_j^k}$.

5. *Relative inner angle orientation*: $^{or}\mathsf{H}^{k+1} \in \{0,1,2,3\}$, depending on the relative quadrant position of $^{\alpha_1}\mathsf{H}^{k+1}$ and $^{\alpha_2}\mathsf{H}^{k+1}$.

This set of properties assures a description of a feature twoplet that is invariant to changes in scale, orientation, and position. This implies that at every stage of binding, the local descriptors are indifferent to a particular global constraint on object position, orientation and size. They are also agnostic to their position within the object, as they are defined purely in terms of their immediate local surroundings. The radius ${}^{r}\mathsf{H}^{k+1}$, the center ${}^{c}\mathsf{H}^{k+1}$, and the orientation ${}^{o}\mathsf{H}^{k+1}$ of the receptive region at H^{k+1} are calculated by finding the circumferent circle that encompasses the two constitutive regions at level H^{k}. These parameters are entirely extrinsic, meaning that they are describing an instantiation of a feature, but are not used to encode its fundamental properties, and they are used solely for the calculation of the intrinsic parameters at the next level.

7.3 Geometric signatures

In conceptual representations [44], and in constructivist epistemology, percepts are often represented in a dichotomous psychological space (Figure 7.2). According to Kelly [67], such spaces abandon the cartesian notions of distance and continuum, but rather represent constructs as axis, or perpendicular planes, that divide the psychological space into dichotomies such as dim / bright, old / young etc. A scale system is used to organize the axis and to compensate for lack of distance measures. Combinations of constructs in psychological space then correspond to multiple axes of reference, and the planes representing their distinctions intersect to define regions of the space corresponding to composite constructs. In our example, relative size (2), relative length (3), angles to connector (4), and angle orientation (5) represent constructs, where dichotomies can be regarded as opposites between larger and smaller (2, 3), acute versus obtuse (4), and same versus opposite (5). We therefore choose to characterize the geometric properties by discrete descriptors ${}^{G}\mathsf{H}$, obtained by a quantization of the continuous values of the conceptual space of geometric properties.

The quantization can be either fixed or learned (see Section 8.3). In order to estimate the extent of the parameter space of the attributes, we estimate the distribution of the parameters by running a learning sequence of bindings through all the levels. The continuous parameter space is then sampled into a number of bins, where the number of bins can be used to increase or decrease the specificity of the level. As the distributions are essentially unimodal, we can estimate the sampling parameters from the probability distributions for each of the parameter separately. Parameters ${}^{\alpha_1}\mathsf{H}$ and ${}^{\alpha_2}\mathsf{H}$ exhibit a near-normal distribution $\mathcal{N}({}^{\alpha}\overline{\mathsf{H}}, \sigma_{{}^{\alpha}\mathsf{H}})$; we therefore center the bins an their mean value ${}^{\alpha}\overline{\mathsf{H}}$. Parameters ${}^{s}\mathsf{H}$ and ${}^{l}\mathsf{H}$ exhibit a near-exponential distribution, we therefore center the bins on the *mode* of the distribution, which can be easily estimated by an intermediate dense discrete sampling.

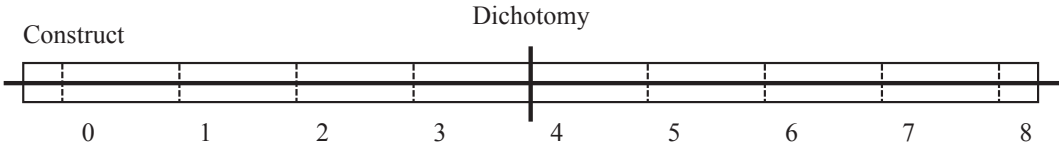


Figure 7.2: Geometry of psychological space - According to Kelly [67], the psychological space can be imagined as a geometry of dichotomies, where a *construct* on an *axis of distinction* represents a basis for categorizing in a para-cartesian hyperspace with relatively stable scalar axes (here numbered as `[0]` to `[7]`).

The geometric attributes of a twoplet H_i^k can be encoded using a single discrete signature

$$ {}^{G}\mathsf{H}_i^k \;=\; \Big[\, \mathsf{Q}_s({}^{s}\mathsf{H}_i^k),\, \mathsf{Q}_l({}^{l}\mathsf{H}_i^k),\, \mathsf{Q}_\alpha({}^{\alpha_1}\mathsf{H}_i^k),\, \mathsf{Q}_\alpha({}^{\alpha_2}\mathsf{H}_i^k),\, {}^{or}\mathsf{H}_i^k \,\Big] \;, \tag{7.1} $$

where $\mathsf{Q}(\,\bullet\,)$ denotes quantization. For instance, the sequence ${}^{G}\mathsf{H} =$`[23310]` describes the local geometry between features where the relative size falls into equivalence class `[2]`, relative length into equivalence class `[3]`, angles to connector to classes `[3]` and `[1]`, while `[0]` indicates the layout of the inner angles.

Let us denote the first step of binding as a transition from the level of regions of interest to a first set of twoplets as $\mathsf{H}^0 \rightarrow \mathsf{H}^1$, the second step as $\mathsf{H}^1 \rightarrow \mathsf{H}^2$, and in general $\mathsf{H}^k \rightarrow \mathsf{H}^{k+1}$. The starting level is therefore processed through levels $\mathsf{H}^1 \ldots \mathsf{H}^N$. At each of the level, features at the lower level group in pairs that reflect the geometric properties of the constituent features. With the progression through levels, the information captured by twoplets encompasses progressively stronger geometrical constraints, that cover progressively larger areas of the image. The inherited appearance and geometric information can be encoded as a compact feature signature

$$ {}^{D}\mathsf{H}_i^0 \;=\; {}^{w}\mathsf{H}_i^0 \tag{7.2} $$

$$ {}^{D}\mathsf{H}_m^{k+1} \;=\; {}^{G}\mathsf{H}_m^{k+1} \;||\; {}^{D}\mathsf{H}_i^k \;||\; {}^{D}\mathsf{H}_j^k;\; k > 0; \tag{7.3} $$

which gets longer with each level. As all the information can be accessed by backtracking the hierarchy, this representation is redundant, however it can be efficient in feature matching, and therefore plays an important part in the implementation. For example, feature signatures can be represented both as strings or as positive natural numbers, supporting therefore a fast comparison and efficient data structures and indexing [1]. Features at first five levels could be therefore characterized with the following signatures:

[1] The fast string search is implemented in PERL.

${}^{D}\mathsf{H}^{0}$	2
${}^{D}\mathsf{H}^{1}$	$11232-21$
${}^{D}\mathsf{H}^{2}$	$17341-11232-21-12111-33$
${}^{D}\mathsf{H}^{3}$	$22441-17341-11232-21-12111-33-11124-40111-12-32221-30$
${}^{D}\mathsf{H}^{4}$	09232
...	$22441-17341-11232-21-12111-33-11124-40111-12-32221-30$
...	$11122-30111-12221-21-32111-24-43121-20501-12-30223-31$
${}^{D}\mathsf{H}^{5}$	15210
...	$22441-17341-11232-21-12111-33-11124-40111-12-32221-30$
...	$11122-30111-12221-21-32111-24-43121-20501-12-30223-31$
...	$73777-23822-23218-33-23886-01-54552-23891-02-23853-00$
...	$22123-22123-54355-10-29222-22-12365-53433-32-34131-11$

In this example, the base level descriptor H^0 carries information on the appearance class `[2]`, while level H^1 inherits the appearance classes of two constituent regions at H^0 `[21]` and adds the geometric relations between the two `[11232]`.

At this point, a set of high level features represents the information in the image, encompassing a detailed information on local appearance, and a local-to global progression of structural properties. Once this information has been derived, it could be matched for similarity; for example high level features of a prototype could be matched against high level features derived from an exemplar, or from an unknown testing image. However, as it will become clear in the next chapter, we employ a different strategy: the matching process is designed in a way to gradually, in a hierarchical fashion, construct only those features that match at a certain level of complexity, therefore constructing only high level features that represent *any* of the possible similarities that potentially exist between two entities.

8

Hierarchical matching

In this chapter, we arrive to the core of our architecture: the process of synchronous matching, where prominent features at level H^{k+1} are actively constructed by a synchronous procedure that uses a prototype image as a model that guides the construction.

8.1 Problem formulation

Hierarchical feature binding typically requires a level-dependent representation of relevant category–specific features. For example, in the HMAX model [124], the Gaussian summation on S units is analogous to weights on the inputs of simple processing units (resembling thus the synapses of simple cells in the visual cortex [29]), and therefore determines the promotion of features to the next level. Alternatively, a set of learned features are represented at each level and participate in a hierarchical voting or classification scheme, where each of the features can be shared among categories [34; 35; 36; 150]. Such mechanisms also provide the plasticity that is needed to learn the representation by observing a number of training exemplars from a certain category. Object recognition or categorization is consequently defined as a classification task which discriminates between sets of activated features. The classifier can be trained on high level features[36; 124], or, where applicable, on features from multiple levels of the hierarchy [35; 151].

In our approach, we investigate a different strategy: instead of learning a multilevel representation of discriminant features, we actively construct high level features by a *synchronous binding*, *matching*, and *inhibition* between the image which is being interpreted and the exemplar image of a prototype. In that way, no high level information has to be modeled in advance, but is rather constructed ad-hoc for each categorization task. The representation consists only of a raw image which conditions the construction, and the hierarchy promotes

only features that can be binded and matched in both images. Consequently, the similarity can be evaluated by the number (the strength) of high level features that were successfully constructed. If a number of high level features can be constructed, there is a high probability that some similarity between the prototype image and the interpreted image is found - note that, as the similarity is not based on a predefined set of features, it can capture a wide array of different *kinds of* similarities that potentially exist between category members, and can accommodate different types of local features .

8.2 Synchronous matching and inhibition

The reason to resort to a synchronous matching and inhibition is twofold: firstly, each of the members of the same category will generate a large amount of features at each level of the hierarchy, but only a small fraction of those features will be common to more than one member. Let us simplify this analysis by supposing that the receptive fields extend over the whole image area. Since within the binding process the number of candidate features grows exponentially, the probability that two features in different images match becomes quickly extremely low. As an example, if at level H^1 only 50% of prototype and exemplar features match, the chance of binding two matching features is $p_m \approx 0.25$ in each image. If, within 25% of features at H^2, the chance of matching is again 50%, only $p_m \approx 0.125$ of features are candidates for matching at the next level. The higher we go up the hierarchy, the higher has to be the number of generated features, as with a lower percentage of hypothetical matches the system is more prone to the dissimilarity of the matched images in the first phase. Due to the exponential nature of hierarchical binding, the number of features that get promoted has therefore to be constrained, or the probability of selecting the right features becomes negligible.

It is therefore crucial to generate at each level only the features that will result in a potential match at a higher level.

Secondly, synchronous inhibition allows to implement an *attentive search* strategy, where the receptive field in the image that is being interpreted adapts to the constraints imposed by the prototype.

Let $\widehat{\mathsf{H}}^k$ be the features at level k derived from a representative, non-occluded image of a member of a category, and H^k the features from the image being interpreted. Let F^k be the set of matches $\mathsf{F}^k = \left\{ \left(\widehat{\mathsf{H}}_i^k, \mathsf{H}_j^k \right) : \ {}^D\widehat{\mathsf{H}}_i^k = {}^D\mathsf{H}_j^k \right\}$. Synchronous matching is performed as follows: for each matching pair of features F^k, a receptive field around $\widehat{\mathsf{H}}_i^k$ is initiated, covering

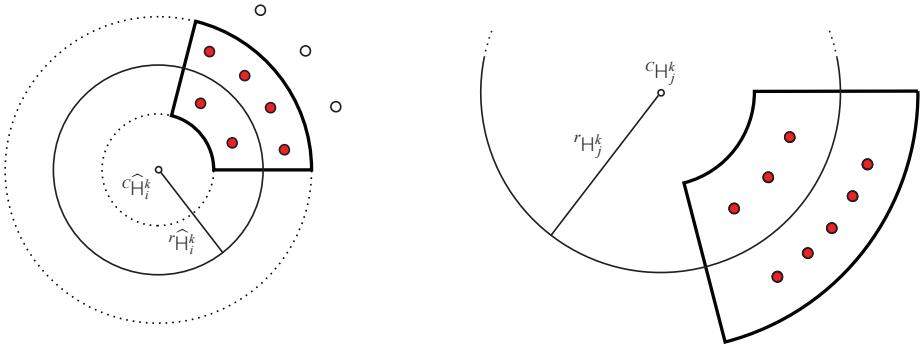


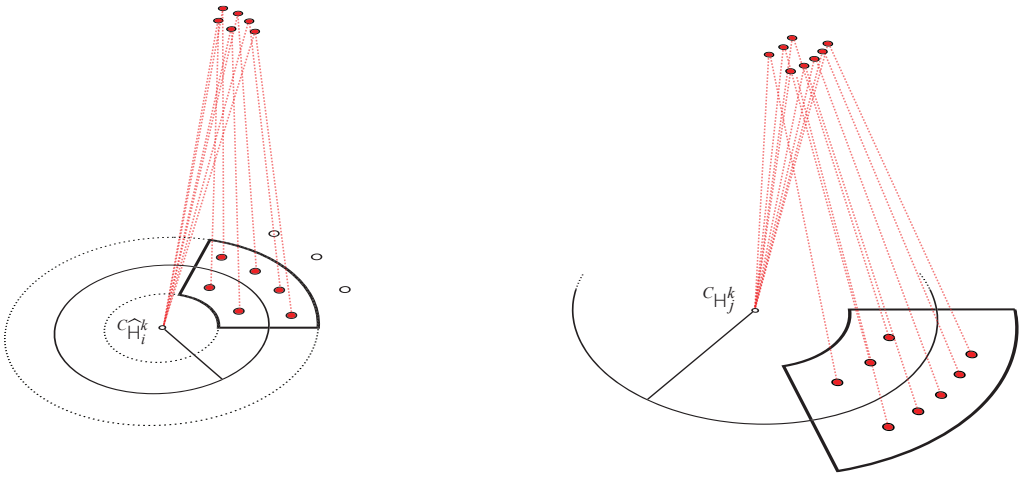
Figure 8.1: Receptive field - (a) *fanOut* features were found in the shaded area between the minimal and maximal radii around $^{C}\widehat{H}_i^k$. (b) The corresponding shape and dimension of the receptive field is searched for matching features around $^{C}H_i^k$.

a disk–shaped area spanning from r_{min} to r_{max}, where both the minimal and the maximal radii are proportional to the radius of the supporting region $\widehat{H}_i^k$. The receptive field then first *expands* and then *contracts*, until a predefined number *fanOut* of candidate neighboring features falls within the receptive field's area. This ensures a homogeneous binding across areas with diverse spatial density of features. As all of the neighboring features that were found between the inner and the outer radius fall within the shaded area, and this area can be uniquely characterized by its relation to the generic feature's orientation and scale, we need to search only a corresponding receptive area around each of the candidate features in the interpreted image.

The receptive field of H_j^k is then initiated to reflect the spatial properties of the receptive field in the prototype. For example, Figure 8.1b depicts the receptive field of H_j^k, which is rotated and scaled according to the relative scale and orientation properties of $\widehat{H}_i^k$ in Figure 8.1a. Within this receptive field, only features H_l^k, for which a match $(\widehat{H}_i^k, H_l^k) \in F^k$ exist, are considered as candidates. After all the candidates are initiated, matching features at H^{k+1} are constructed [1]. As a result, this process generates only the candidates that will lead to potential matches at higher levels, focusing the hierarchy tree only on meaningful bindings, and freeing up the processing capacity (Figure 8.2).

Synchronous binding of features at $H^k \rightarrow H^{k+1}$ is summarized in Algorithm 8.1. The input

[1] Alternatively, both levels can be fully binded using fixed receptive fields, followed by a step of matching. This however prevents a parallel execution on an arbitrary number of processing units.



(a) **(b)**

Figure 8.2: Binding in receptive fields - (a), (b) Features in receptive fields are binded to features at H^{k+1}.

to the algorithm are features $\widehat{\mathsf{H}}^k$ and H^k, and an index table F^k that indicates the matching pairs of features at level k. The parameters *inner* and *outer* define the starting values for the inner and outer boundary of the receptive field, which is initialized according to ${}^r\mathsf{H}_i^k$. The parameter $fanOut$ defines the desired minimum number of candidate neighboring features that have to fall within the receptive field area. The parameters mostly depend on the type of imagery (resolution, level of detail), and are not specific with respect to object categories.

The output of Algorithm 8.1 are features at $\widehat{\mathsf{H}}^{k+1}$, H^{k+1}. Simultaneously with the binding, an indexing array with pointers to matching pairs F^{k+1} is generated and fed to the next level. As matching is already performed in the biding, $\widehat{\mathsf{H}}^{k+1}$ and H^{k+1} already comprise only features that are members of at least one matching pair at $k+1$. It can be easily shown that, since F^k comprise all matches at level k, all of the possible matches that can be found under the locality constraints of the receptive field are constructed by binding. E.g., it is not possible that two matching descriptors appear at level $k+1$ and that their matching pair was not discovered within the local neighborhood, as for each feature H^{k+1}, the constituent features at H^k already had to match and were therefore involved in binding. It is however possible there are matching features outside the receptive field area, however these are typically potential false matches, and are therefore omitted.

Figures 8.3 to 8.11 illustrate the synchronous hierarchical matching for a study example,

Algorithm 8.1: Hierarchical binding

Input: Level k regions $\widehat{\mathsf{H}}^k$ and H^k, matching pairs F^k, *inner*, *outer*, $fanOut$

Output: $\widehat{\mathsf{H}}^{k+1}$, H^{k+1}, F^{k+1}

foreach $\mathsf{F}^k_u = \left\{ \left(\widehat{\mathsf{H}}^k_i, \mathsf{H}^k_j \right) \right\}$ **do**

- $rMin = {}^r\widehat{\mathsf{H}}^k_i * inner$;
- $rMax = {}^r\widehat{\mathsf{H}}^k_i * outer$;
- $\widehat{\mathsf{A}} = \left\{ \widehat{\mathsf{H}}^k_l : \ {}^c\widehat{\mathsf{H}}^k_l \in \widehat{aField}(rMin, rMax) \right\}$;
- **while** $card\left(\widehat{\mathsf{A}}\right) < fanOut \ \wedge \ rMin > 0$ **do**
  - **if** *outer boundary within image* **then**
    - $rMax = rMax + eps$;
  - **else**
    - $rMin = rMin - eps$;
  - $update\left(\widehat{\mathsf{A}}\right)$;
- $\widehat{\mathsf{C}}^{k+1} = bind\left(\widehat{\mathsf{A}}, \widehat{\mathsf{H}}^k_i\right)$;
- $\widehat{aField} = convexHull\left(\widehat{\mathsf{A}}\right)$;
- transform $\widehat{aField}$ to $aField$;
- $\mathsf{A} = \left\{ \mathsf{H}^k_m : \ \left({}^C\mathsf{H}^k_m \in aField\right) \ \wedge \ \left(\exists\, \widehat{\mathsf{H}}^k_n \in \widehat{\mathsf{A}} : \ \left(\widehat{\mathsf{H}}^k_n, \mathsf{H}^k_m \right) \in \mathsf{F}^k \right) \right\}$;
- $\mathsf{C}^{k+1} = bind\left(\mathsf{A}, \mathsf{H}^k_j\right)$;
- $\mathsf{F}^{k+1} = \mathsf{F}^{k+1} \cup match\left(\widehat{\mathsf{C}}, \mathsf{C}\right)$;
- $\widehat{\mathsf{H}}^{k+1} = \widehat{\mathsf{H}}^{k+1} \cup \left\{ \widehat{\mathsf{C}}^{k+1}_m : \ \widehat{\mathsf{C}}^{k+1}_m \in \mathsf{F}^{k+1} \right\}$;
- $\mathsf{H}^{k+1} = \mathsf{H}^{k+1} \cup \left\{ \mathsf{C}^{k+1}_m : \ \mathsf{C}^{k+1}_m \in \mathsf{F}^{k+1} \right\}$;

where a segmented side view of a motorbike is given as a prototype, and the task is to estimate whether there is a structurally similar object, potentially a motorbike, on an interpreted image. The $\widehat{\mathsf{H}}^k \rightarrow \widehat{\mathsf{H}}^{k+1}$ displays bindings in the prototype image and in positioned on the top, while $\mathsf{H}^k \rightarrow \mathsf{H}^{k+1}$ bindings are depicted at the bottom part of the figures. Red lines depict connections that represent active features, while yellow connections stand for features that did not successfully bind (i.e. could not generate a match) in the last stage of binding, and were therefore inhibited. H^0 and $\widehat{\mathsf{H}}^0$ with the binded features at H^1 and $\widehat{\mathsf{H}}^1$, respectively.

At $\mathsf{H}^0 \rightarrow \mathsf{H}^1$ (Figure 8.3), we start with ≈ 400 H^0 features. We set a $fanOut$ of 30, which means that every feature will generate H^1 candidates by binding with 30 nearest neighbors. The figure depicts the ≈ 4000 successful bindings. Some of the matching pairs are depicted

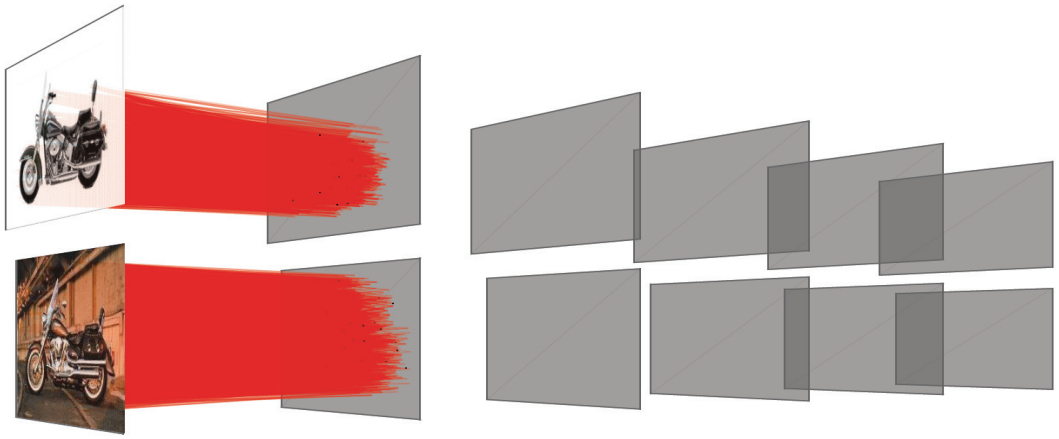


Figure 8.3: Synchronous hierarchical matching at $\mathsf{H}^0 \rightarrow \mathsf{H}^1$ - Top row represents successful bindings $\widehat{\mathsf{H}}^0 \rightarrow \widehat{\mathsf{H}}^1$ of features in the prototype image, while the bottom row represents bindings $\mathsf{H}^0 \rightarrow \mathsf{H}^1$ in the interpreted image. Red lines depict connections that represent active features.

in Figure 8.4. At this point, the number of spurious matches is high. However there is no information about position, scale, and orientation of objects we search for (considering they are present), so we have to retain all generated features. Due to a low information value, the feature pool is extremely noisy, and does not exhibit any bias that could be used to hypothesise about viewing parameters. The match in Figure 8.4a illustrates the invariance of matching to scale and orientation. Figure 8.4b depicts a match that is locally correct, however it is not valid with respect to the configuration of parts, and will most probably not generate successful matches at higher levels.

After the $\mathsf{H}^1 \rightarrow \mathsf{H}^2$ binding step, features at level H^2 in Figure 8.5, already encompass significant portions of the prototype object. Features at H^1 that were not successfuly binded within their local neighborhood are now depicted in yellow; these features are inhibited from growing. The information contained within a feature signature is now still rudimentary – note that at this point, the descriptor can be encoded with approximately 50 bits of information.

FigureFigure 8.6 depicts some of the matches generated at H^2. For better visibility, we depict matches as configurations of subordinate features at H^0; regions of same color belong to the same H^1 twoplet. The number of H^2 features is ≈ 2500, same is the number of matching pairs F^2. Figure 8.6a depicts an incorrect match to the background, which will be most probably elimnated in further matchings. Figure 8.6b match is locally correct, however it does not comply to the global object structure, and will be also probably eliminated at a higher level of matching.

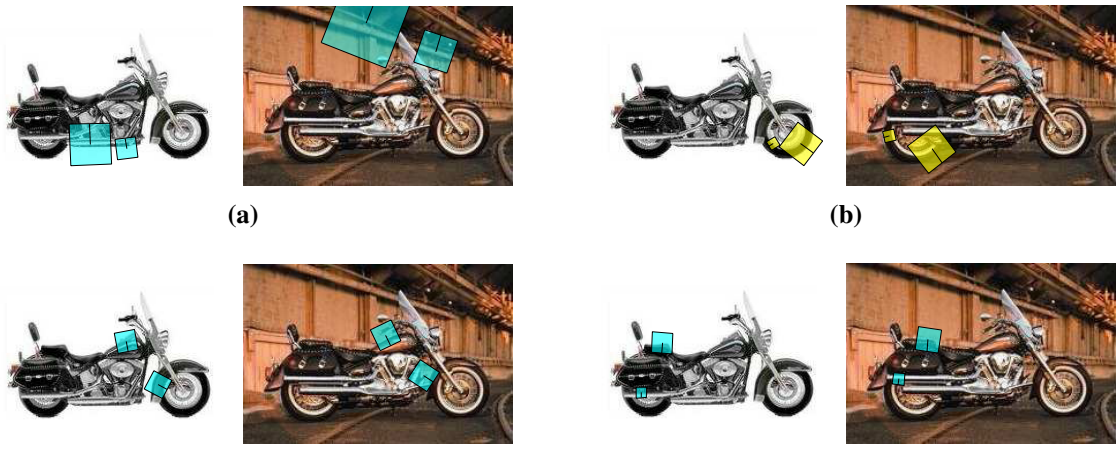


Figure 8.4: Matches at H^1 - (a) illustrates the invariance of matching to scale and orientation. (b) depicts a match that is locally correct, however it is not valid with respect to the configuration of parts in the object, and will therefore not be able to generate successful true matches at higher levels

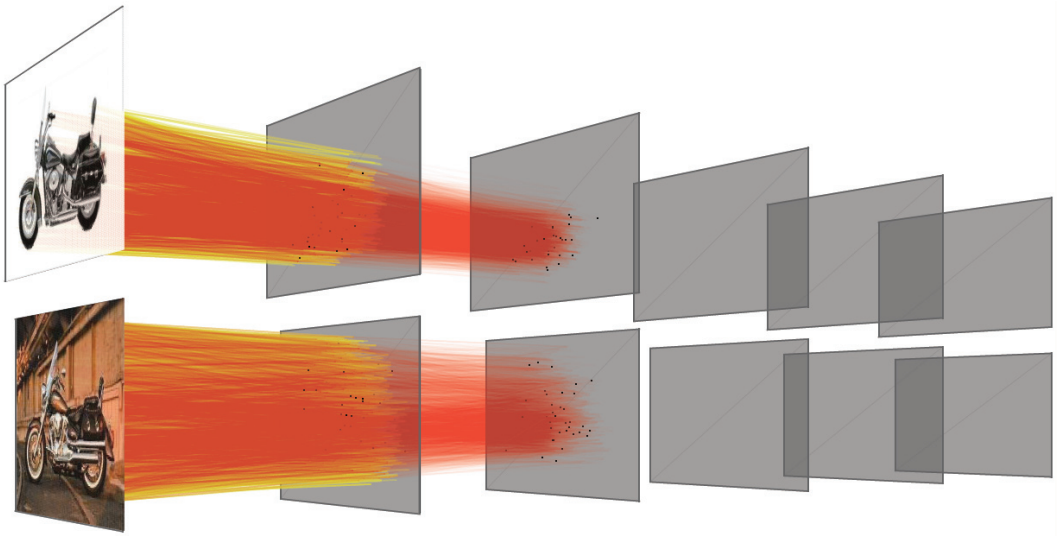


Figure 8.5: Synchronous hierarchical matching at $H^1 \rightarrow H^2$ - Top row represents successful bindings $\widehat{H}^1 \rightarrow \widehat{H}^2$ of features in the prototype image, while the bottom row represents bindings $H^1 \rightarrow H^2$ in the interpreted image. The connectivity trees of active features are colored in red, while connectivity trees of inhibited features features that could not generate a match in the last stage of binding are depicted in yellow.

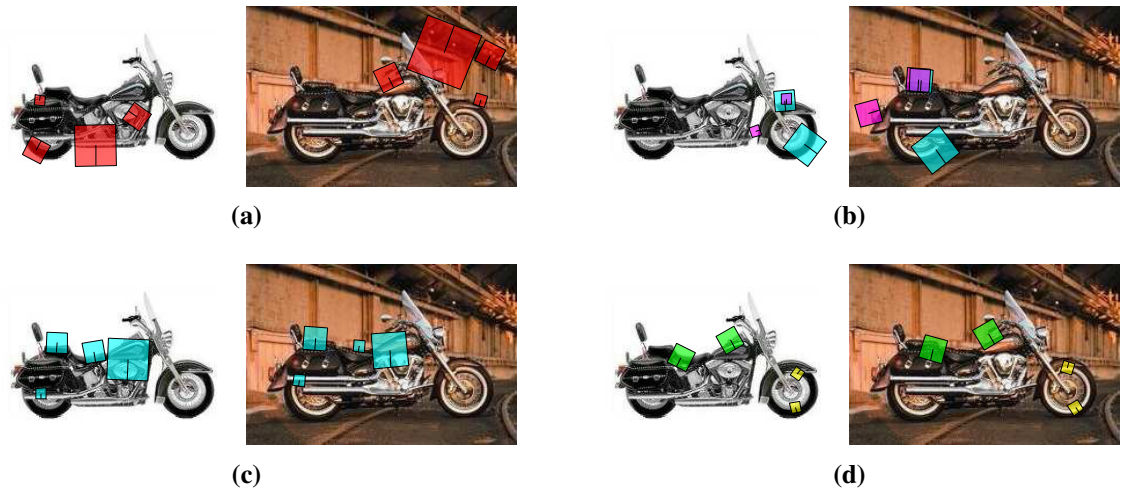


Figure 8.6: Matches at H^2 - (a) an incorrect match to the background will be most probably eliminated in further matchings. (b) match is locally correct, however it does not comply to the global object structure, and will be also probably eliminated.

At H^3, a large part of the features already expands to an area larger than half of the object. The level at which this happens depends largely on the spatial resolution of the images. Note that in images with a higher resolution, the scale invariant local features would also cover larger areas, and, consequently, the initial receptive fields would be larger. The size of the image should therefore not directly influence the number of levels required to resolve an image. However, in practice, it is expected that more detail in larger images results in a more fine representation, which would require more levels of binding. As it can be seen in Figure 8.7, a large part of the background features at lower levels is at this point inhibited (connections in yellow), which means that the features are already of a high enough complexity that makes the probability of a false positive match in the background low. Note also how the centers of the regions converge toward the center of the object, which results in densely populated neighborhoods. However, as the probability of a matches to a single feature is now low, only a few features that fall within a receptive field will generate matches across images.

In Figure 8.8, one can observe how the quantized geometric signatures allow some degree of freedom, which results in elasticity of the matched structure. We can also observe how the matches already cover a large part of the object.

Figure 8.9 illustrates the binding $H^3 \rightarrow H^4$, and Figure 8.10 illustrates some of the matches at H^4. We explicitly show examples of deformed structural matches that can be generated,

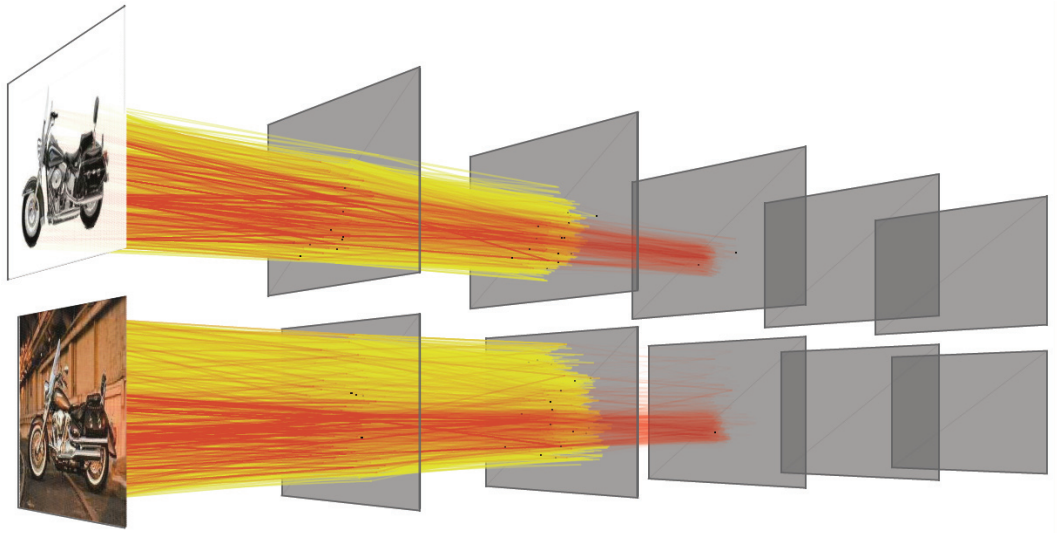


Figure 8.7: Synchronous hierarchical matching at $H^2 \rightarrow H^3$ - Top row represents successful bindings $\widehat{H}^2 \rightarrow \widehat{H}^3$ of features in the prototype image, while the bottom row represents bindings $H^2 \rightarrow H^3$ in the interpreted image. The connectivity trees of active features are colored in red, while connectivity trees of inhibited features features that could not generate a match in the last stage of binding are depicted in yellow.

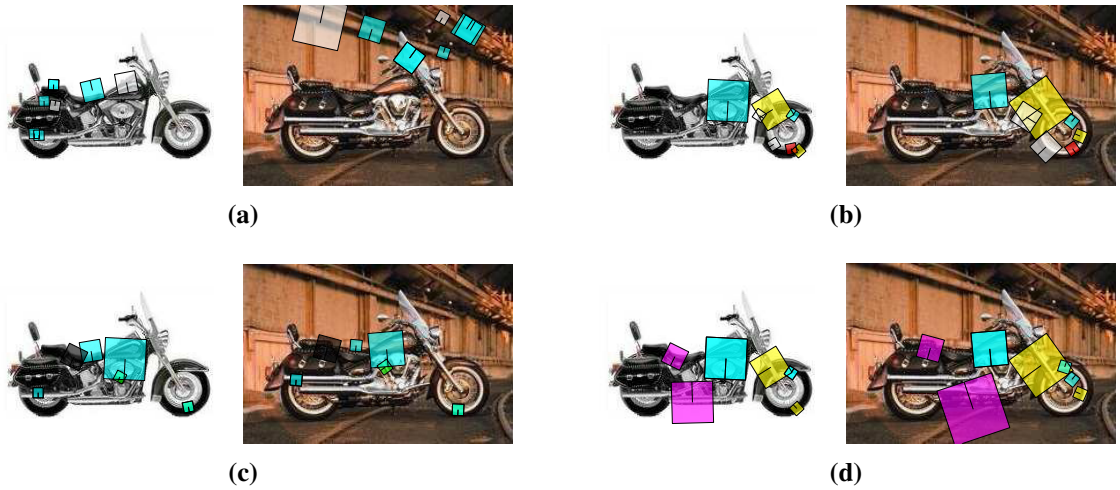


Figure 8.8: Matches at H^3 - The quantized geometric signatures allow some degree of freedom, which results in elasticity of the matched structure. Most of the matches already cover a large part of the object.

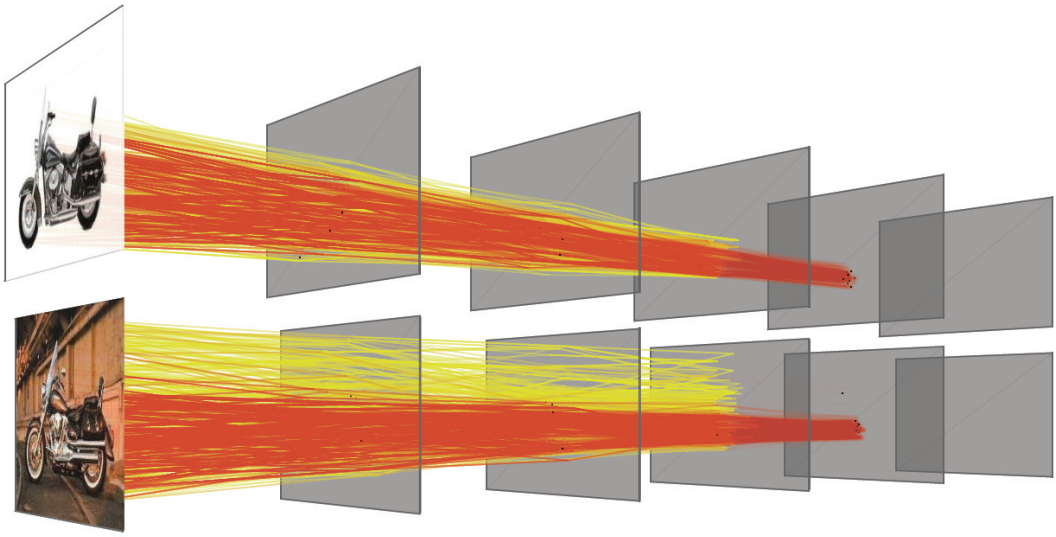


Figure 8.9: Synchronous hierarchical matching at $\mathsf{H}^3 \rightarrow \mathsf{H}^4$ - Top row represents successful bindings $\widehat{\mathsf{H}}^3 \rightarrow \widehat{\mathsf{H}}^4$ of features in the prototype image, while the bottom row represents bindings $\mathsf{H}^3 \rightarrow \mathsf{H}^4$ in the interpreted image. The connectivity trees of active features are colored in red, while connectivity trees of inhibited features features that could not generate a match in the last stage of binding are depicted in yellow.

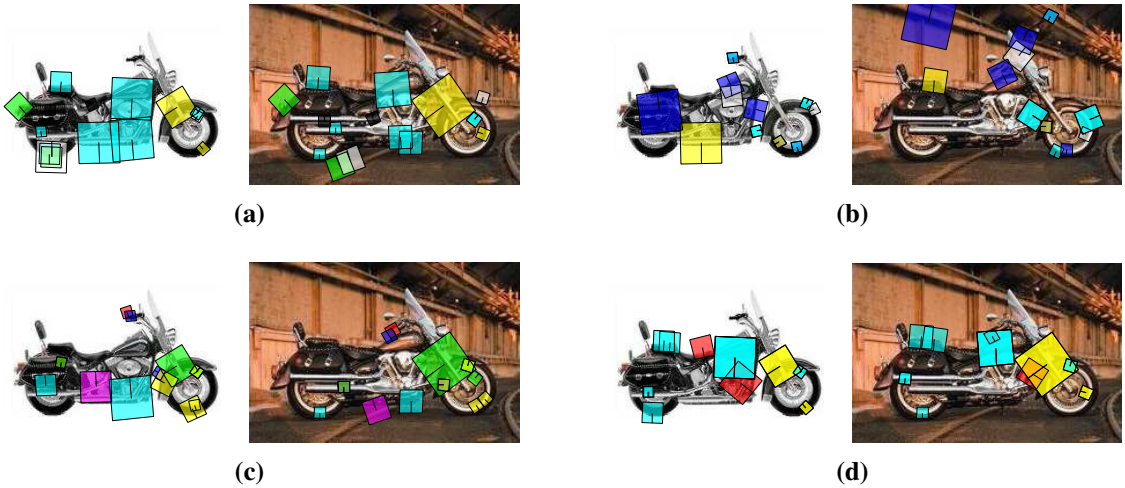


Figure 8.10: Matches at H^4 - As we do not constrain the absolute size and orientation of feature at levels that are lower than H^{k-1}, some deformed structural matches can be generated, as e.g. in (a) and (c).

e.g. in Figure 8.8a and in Figure 8.8c. Finally, Figure 8.11 and Figure 8.12 illustrate the last stage of processing. The positive matches are now consolidated and each of them (≈ 2500) stands for one of the structural matches between all features in the subtree down to H^0. The centers of regions are clustered near the center of the object, and the receptive fields roughly encompass the area of the object in the image. In Figure 8.11, the final receptive fields are explicitly drawn to illustrate the "strength" of the matching. In Figure 8.12, final matches are shown, and, for comparison, examples of matches for the category of *Cars* are included in Figure 8.13.

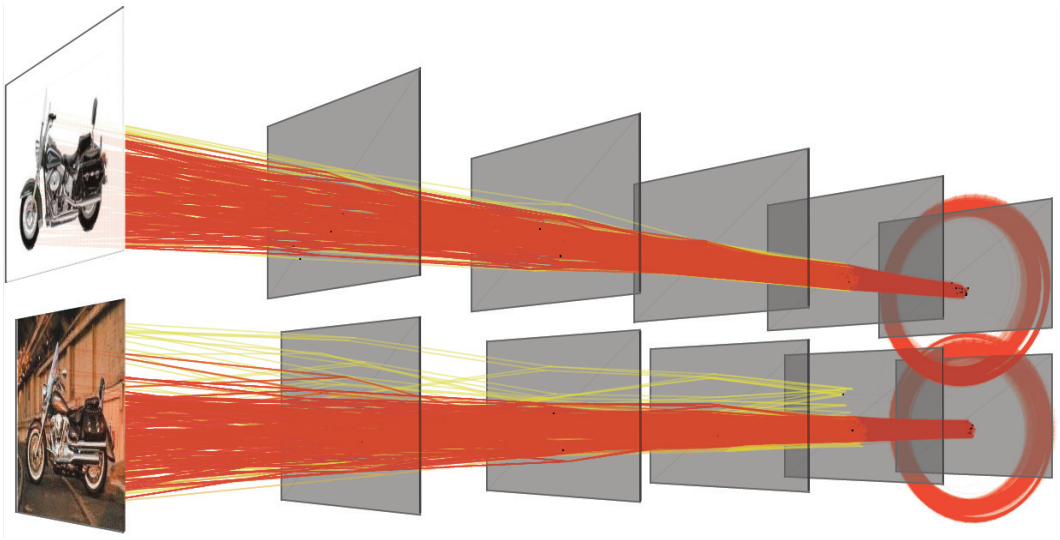


Figure 8.11: Synchronous hierarchical matching at $\mathsf{H}^4 \rightarrow \mathsf{H}^5$ - Top row represents successful bindings $\widehat{\mathsf{H}}^4 \rightarrow \widehat{\mathsf{H}}^5$ of features in the prototype image, while the bottom row represents bindings $\mathsf{H}^4 \rightarrow \mathsf{H}^5$ in the interpreted image. Red lines depict connections that represent active features, while yellow connections stand for features that did not successfully binded (i.e. could not generate a match) in the last stage of binding, and were therefore inhibited. The red circles at level five denote the extension of the $\rightarrow \mathsf{H}^5$ and H^5 twoplets, respectively. The number of succesfully matched features at H^5 is approximately 2500 in both images.

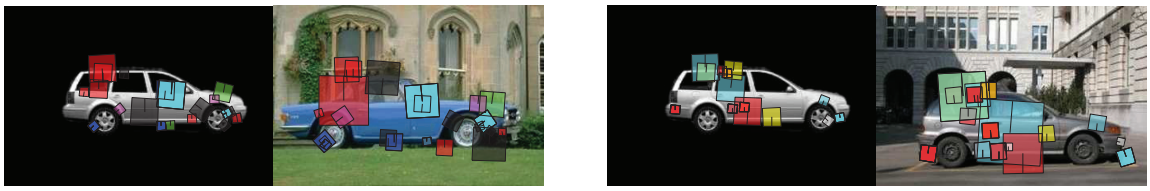
8.3 Category learning and recognition

The hierarchical matching framework uses two mechanisms to tune the matching process by learning the parameters of the hierarchy. In Section 6.3, we already described how we learn a sparse codebook for a specific category. Here we describe the learning of the parameters to construct a geometric conceptual space tuned to a category.

As it was stated in Section 7.3, the geometric attributes get quantized to a number of discrete

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

Figure 8.12: Matches at H^4 - Matches at this level cover the whole object area. Centers of features are clustered around the center of the object, and the receptive fields encompass the whole area of the object.



(a) **(b)**

Figure 8.13: Matches at H^5 - *Car* - Matches at H^5 for *Car* category.

quantization intervals. Imagine two categories of objects which significantly differ in shape. Within hierarchical matching of objects from the same category, the geometric attributes of the binded features will have a different probability distribution for each of the categories. For most of the categories, we can assume that the distributions are essentially unimodal, so we can estimate the category related sampling parameters from the probability distributions for each of the geometric attributes, and for each of the levels of the hierarchy. Parameters ${}^{\alpha_1}H$ and ${}^{\alpha_2}H$ exhibit a near-normal distribution $N({}^{\alpha}\overline{H}, \sigma_{{}^{\alpha}H})$. We therefore center the bins on the mean value ${}^{\alpha}\overline{H}$. Parameters ${}^{s}H$ and ${}^{l}H$ exhibit a near-exponential distribution. We therefore center the bins on the *mode* of the sample distribution, which can be easily estimated by a dense discrete sampling.

These parameters are estimated by a simple learning step: we match the exemplar image to itself, and sample the emerging distributions from the sets of binded features at each level.

From the sampled distributions we estimate the mean value and the mode, together with the minimum and the maximum value. The quantization intervals are therefore set as follows:

	center	quantization interval
Relative size	$\text{mode}(p({}^{s}\mathsf{H}))$	$\min(p({}^{s}\mathsf{H})) \leq x \leq \text{mode}(p({}^{s}\mathsf{H})) - \min(p({}^{s}\mathsf{H}))$
Relative length	$\text{mode}(p({}^{s}\mathsf{H}))$	$\min(p({}^{l}\mathsf{H})) \leq x \leq \text{mode}(p({}^{l}\mathsf{H})) - \min(p({}^{l}\mathsf{H}))$
Angles to connector	$\text{mean}(p({}^{\alpha}\mathsf{H}))$	$\min(p({}^{\alpha}\mathsf{H})) \leq x \leq \max(p({}^{\alpha}\mathsf{H}))$

All twoplets whose parameters exceed the sampling interval are labeled as invalid, and can not participate in matching.

The framework is now ready to structurally match the exemplar to a prototype by Algorithm 8.2.

Algorithm 8.2: Hierarchical matching

Input: Image I and prototype image $\widehat{I}$
Output: number of matching features at level N, or at highest nonempty level

```
Calculate K and K̂;
Calculate local features H^0 and Ĥ^0 by Algorithm 6.1;
for k = 0...N−1; k++ do
    index matches F^k;
    bind Ĥ^k, H^k to Ĥ^{k+1}, H^{k+1}, F^{k+1} by Algorithm 8.1;
```

In the previous section we already showed a successful matching of a prototype image of the *Motorbike* category matched to a scene where a motorbike is present (Figure 8.11). The number of generated matchings was ≈ 2500. In Figure 8.14, we match the same exemplar image of a motorbike with an image where the motorbike is not present. At the first two levels of binding, the number of candidate twoplets is still significant, as low level structures are matched with sub–parts in both objects or in the background. As the complexity of features increases, the inhibition retains only features that match at a higher scale of complexity. In this case, bindings at higher levels can not be generated, and the process terminates at H^4 (with a total of 74 $\widehat{\mathsf{H}}^4$ features and 102 H^4 features), as no matches can be generated at H^5.

In Figure 8.15, we trained the system to recognize side views of cars. When matching the same image as in Figure 8.11 to a prototype of the car category, bindings at higher levels can not be generated, and the process terminates at H^4, as no matches can be generated at H^5.

Note that the features that were successfully generated at H^4 aggregate mostly parts that are similar, e.g., wheels, or the horizontal structure of the roof.

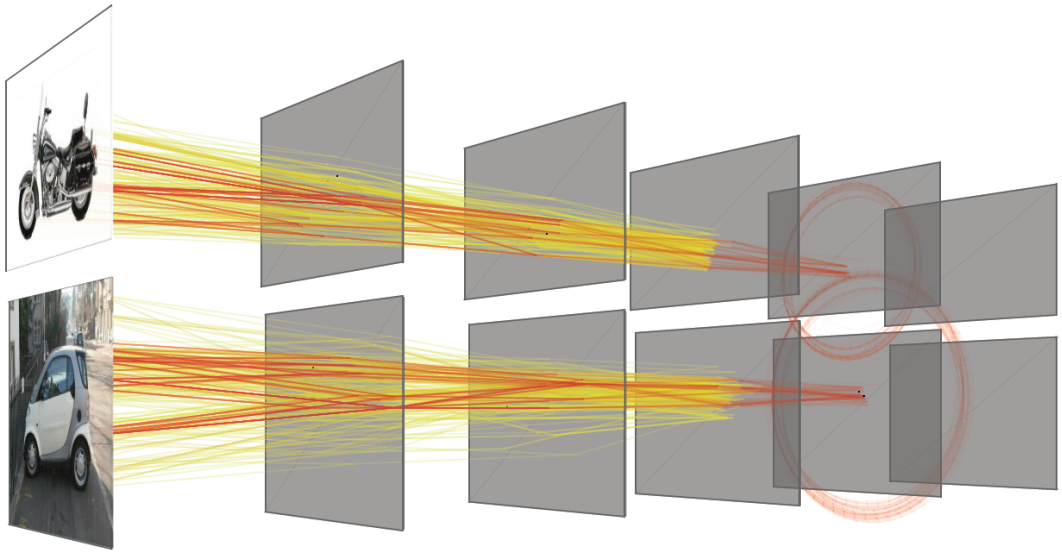


Figure 8.14: Failed synchronous matching at $H^4 \rightarrow H^5$ - When matching the exemplar image to an image where the motorbike is not present, bindings at higher levels can not be generated, and the process terminates at H^4 (with a total of 74 $\widehat{H}^4$ features and 102 H^4 features), as no matches can be generated at H^5.

8.3.1 Verification by backprojection

Features that are promoted to H^5 typically encompass a large enough area to cover most of the object. By backtracking the corresponding low level features at $H^1 - H^4$, we could in theory further verify the similarity between the prototype and the match. As the only "pure appearance" information that enters the hierarchy is the class specific response to the sparse set of ICA basis at level H^0, which, due to the small codebook, imposes only a weak constraint on local appearance, we could test all the matching pairs of patches in the tree of one high level match with a stronger similarity measure, e.g., by normalized correlation.

To illustrate the visual similarity of parts that are discovered by matching, we match a motorbike prototype to a motorbike image (Figure 8.16a), and a car prototype to a car image (Figure 8.16b). Figure 8.17 and Figure 8.18 depict the corresponding patches (all derived by backtracking one of the several high level matches) at levels H^2 and H^3. It can be seen from the figures that the parts are correctly matched, although there is a significant variation in their appearance.

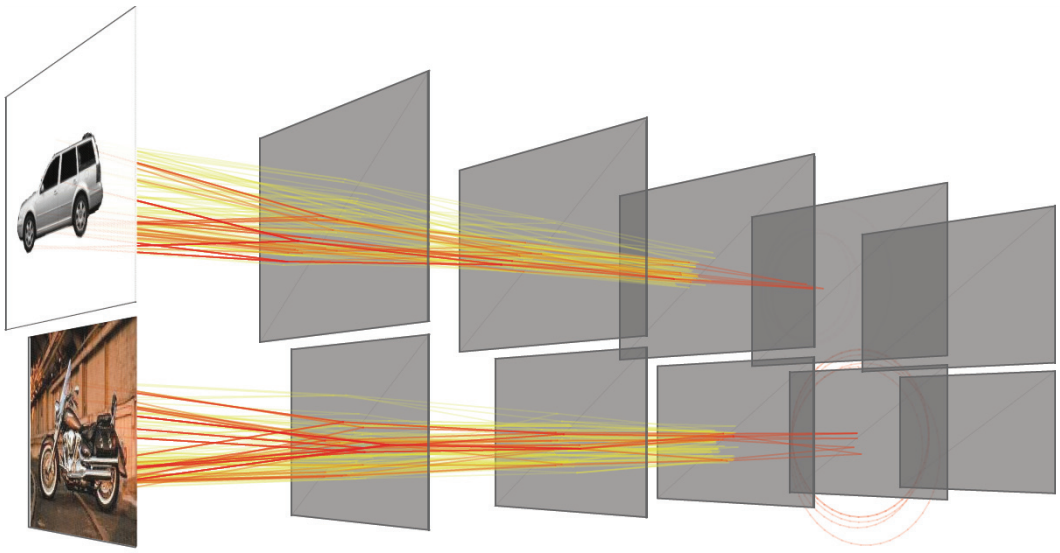


Figure 8.15: Failed synchronous matching at $\mathsf{H}^4 \rightarrow \mathsf{H}^5$ - When matching the same image to an exemplar of the car category, bindings at higher levels can not be generated, and the process terminates at H^4, as no matches can be generated at H^5.

The experiments showed that the verification by normalized correlation improves the ratio between false and true positives only in case of categories with exemplars that are visually similar, and in moderately occluded images. As there are typically many matching pairs in the tree with radically different appearance, we had to use a robust measure: one of the most successful strategies was to accept only matches for which a certain ratio of partial matches exceeds a relatively high similarity threshold. In future work, we plan to experiment with more robust similarity measures, for example based on distance between histograms of gradients. Nevertheless, the matches obtained by backtracking open up interesting possibilities, as they are essentially dense structural matches that do not match only the global structure, but indicate a much finer resolution of matching parts between objects that are significantly different.

8.4 Complexity analysis and implementation details

In this section we evaluate the computational complexity of the hierarchical matching framework and give some details on implementation and execution times.

Let n_k be the number of features at level $\widehat{\mathsf{H}}^k$, and m_k the number of matching pairs of features, i.e., the number of pairs in F^k. In this analysis, we assume that both images have approximately the same number of features n_k. To enable a fast search of features that

(a)

(b)

Figure 8.16: Motorbike and car prototype – scene pair - (a) Prototype and occluded exemplar images for the *Motorbikes* category . (b) Prototype and occluded exemplar images for the *Cars* category.

fall within a receptive field, features are represented in a *k*d-tree structure, which, for n_k features, requires $O(n_k \log^2 n_k)$ of processing time. The core of the hierarchical matching framework is the hierarchical binding, matching and inhibition described by Algorithm 8.1. Algorithm 8.3 presents a shortened version of the algorithm, abstracted to the time-critical tasks. The following parameters influence the execution time of the algorithm: The number of matching pairs m_k defines the number of performed synchronous bindings. The target number of neighboring features that bind with a single feature $fanOut$ defines the expected number of local bindings. Finally, the parameters *inner* and $outer$ set the initial extension of the receptive field $\widehat{aField}$.

Let f_k denote the desired number of local bindings, $f_k = fanOut$. The algorithm queries the *k*d-tree for features in the receptive field area $\widehat{aField}$; the receptive field then extends or contracts until it encompasses at least f_k features. The number of iterations depends on the distribution of features around the pivot feature, and on the level of contraction / expansion. In practice, it is possible to set the parameters *inner*, $outer$, and eps in order to achieve a negligible number of iterations, typically under five. The complexity of area query in the *k*d-tree is $O(n_k^{1/2} + p)$, where p is the number of retrieved features, which can be approximated by f_k.

Once the f_k neighborhood features are found, the features are binded to the pivot feature in the prototype image. As there are f_k candidate features, binding requires at most $O(f_k)$

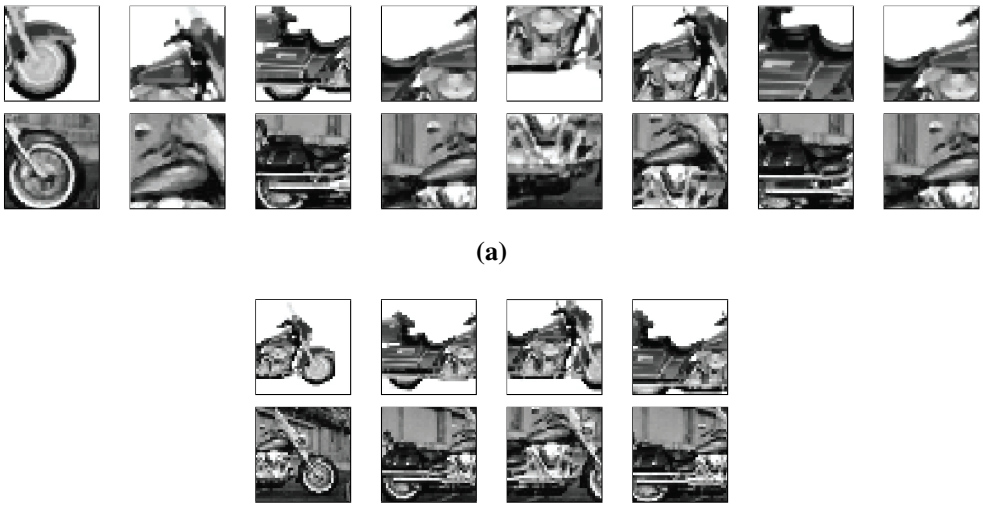


Figure 8.17: Patch representations of matched receptive fields at H^2 and H^3 - motorbike - (a) Local patches defined by receptive fields of matched features at H^2 between an exemplar and an occluded motorbike image. (b) Local patches defined by receptive fields of matched features at H^3 between an exemplar and an occluded motorbike image. For better clarity, local patches are depicted in absolute orientation. The appearance content of the patches can used for similarity based verification by back–projection.

operations.

At this point, the receptive field of the interpreted image is initialized and queried for features, which requires at most $O(n_k^{1/2} + f_k)$ of processing time. Binding of features in the query images requires at most $O(f_k)$ operations.

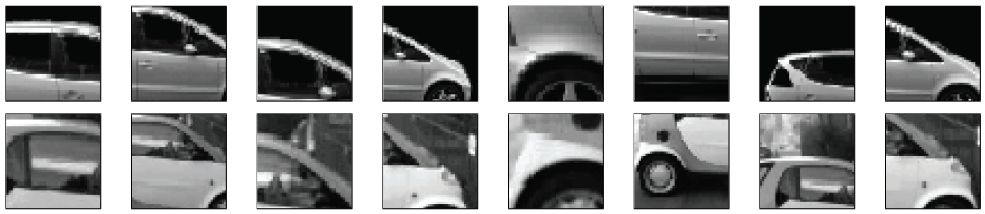
Once the features are binded, we match the feature descriptors (see Section 7.3) in both images. In binding, features are directly inserted to a hash table. As the number of matched features is usually relatively low, the matching using the hashed keys consumes approximately $O(f_k)$ of processing time.

In total, we have to perform m_k iterations to process all the local neighborhoods. As each of the iterations in independent, it can be processed on a separate computing unit. If we have P computing units, we have to perform m_k/P operations. The asymptotic complexity of the algorithm is therefore

$$O(n_k \log^2 n_k) + \frac{m_k}{P} O(2n_k^{1/2} + 5f_k) \text{ , or}$$

$$O(n_k \log^2 n_k + m_k n_k^{1/2} + m_k f_k) \text{ .}$$

The computational complexity of the part of the algorithm that can be computed on a dis-



(a)

(b)

Figure 8.18: Patch representations of matched receptive fields at H^2 and H^3 - car - (a) Local patches defined by receptive fields of matched features at H^2 between an exemplar and an occluded car image. (b) Local patches defined by receptive fields of matched features at H^3 between an exemplar and an occluded car image. For better clarity, local patches are depicted in absolute orientation.

tributed processsing unit is

$$O(n_k^{1/2} + f_k) \ ,$$

and is dependent on the target number of local bindings $f_k = fanOut$, and on the number of features n_k. It is evident that the computational complexity is linear with f_k. As n_k is usually a magnitude larger than f_k, the query of the receptive field, although of a fractional power complexity, still presents a performance bottleneck of the algorithm.

The framework is implemented in MATLAB, C and PERL, and runs on an 8-core PC with 8GB of memory. We use the MULTICORE toolbox to distribute the computational load to multiple processors. Levels with ≈ 1000 features and a fanout of ≈ 100 typically require 200 - 800 ms for a local binding step. Using five processing cores, a typical level requires around 100 seconds, and the whole hierarchy around 500 seconds. While this performance is far from any requirements for practical use, the computational load can be distributed on an arbitrary number of processors.[1]

$\diamondsuit$

[1]For example, BlueGene/L offers 32,768 processors and has been recently used in a simulation of a 55 million neurons rat-scale cortical model [4].

Algorithm 8.3: Hierarchical binding (task breakdown)

Input: $\widehat{\mathsf{H}}^k$, H^k, F^k, *inner*, *outer*, $fanOut$
Output: $\widehat{\mathsf{H}}^{k+1}$, H^{k+1}, F^{k+1}

```
1 foreach F_u^k do
2     while card(Â) < fanOut do
3         expand and query âField;
4     bind Ĉ^{k+1};
5     query aField;
6     bind C^{k+1};
7     match(Ĉ, C);
```

This finishes the description of our framework. In the next part, we describe the experimental evaluation, and give a critical overview and an outlook of future work.

Part III

Experimental evaluation

9

Experimental evaluation

We evaluate our framework on two domains: the first set of experiments assesses the performance of the hierarchical categorization on segmented images of objects that come from multiple categories, and the task is to find the closest match to a given image. The second set of experiments assesses the performance of hierarchical categorization of objects in occluded images, where the task is to decide whether the test image contains an object of the same category as is the prototype image or not. In evaluation, we use two popular databases. For our fist set experimental setup, we use the *ETH80* image database with 8 categories, 10 objects per category and 41 images per exemplar with clear background and from different viewpoints. Our second set of experiments is performed on a subset of *Caltech-101* image database, which contains several occluded images for each of the 101 object categories.

9.1 Object categorization

In this first set of experiments, we evaluate the potential of our method to discriminate between different categories of objects. The experiments are performed on images with a clear background, therefore assessing the performance of categorization of segmented objects. As our framework estimates the similarity between the prototype image and the test image, we use one of the images of segmented objects as the prototype image. Images of other objects are then tested for similarity.

9.1.1 ETH80 database

To provide a transparent evaluation, we use one of the popular image databases that includes multiple categories of objects. The *ETH80* image database contains 8 different image categories. For each of the categories, 10 different exemplars are represented form 41 different

Figure 9.1: ETH-80 object database - Eight exemplars for eight categories of the ETH Zurich image database ETH-80. Objects are depicted in a canonical orientation $azimuth = 68°$, $pan = 180°$.

views around the half view-sphere. The objects are depicted in front of an uncluttered blue background, which makes the database suitable for chroma-keying, and segmentation masks are available, which makes it possible do process only the information related to the object. Figure 9.1 depicts all the exemplars form all eight categories in a canonical orientation ($azimuth = 68°$, $pan = 180°$).

The database is freely available at [1].

9.1.2 Methodology

We test the categorization for a canonical orientation, which is a side-view of the object with an $azimuth = 68°$ and $pan = 180°$. We select one prototype from the ten exemplars as a model for each of the eight categories. For each of the exemplars in a category, the training stage is performed using the prototype and the eight remaining exemplars:

- The ICA filters ($n = 5$) are trained using the prototype eight remaining in a category. We train the codebook using an extended set of exemplars mainly because of the rel-

atively small number of regions of interest generated by some of the objects in the database.

- The prototype is used to grow features at each of the levels sequentially; the statistics of the generated features is used to tune the geometric parameters.

The prototype is therefore matched to all other exemplars in a leave-one-out fashion, meaning that the we omit the exemplar image being interpreted from the training stage. For each matching round we record the scores at five levels of hierarchy.

The nature of the database led to a modification of the original algorithm: for some of the exemplars which were practicably indistinguishable by shape, the number of matches at intermediate levels was too high to be promoted to the next level, as this would lead to a combinatorial explosion. We therefore had to do a random cut of the features (the upper limit at L^3 and L^4 was set to 5000 features). We noticed that this anomaly happens only when non-occluded and very similar images are matched, as, in cases where the features bind also with background features, or with features that do not generate potential matches, an equilibrium between binding, matching and inhibition emerges naturally.

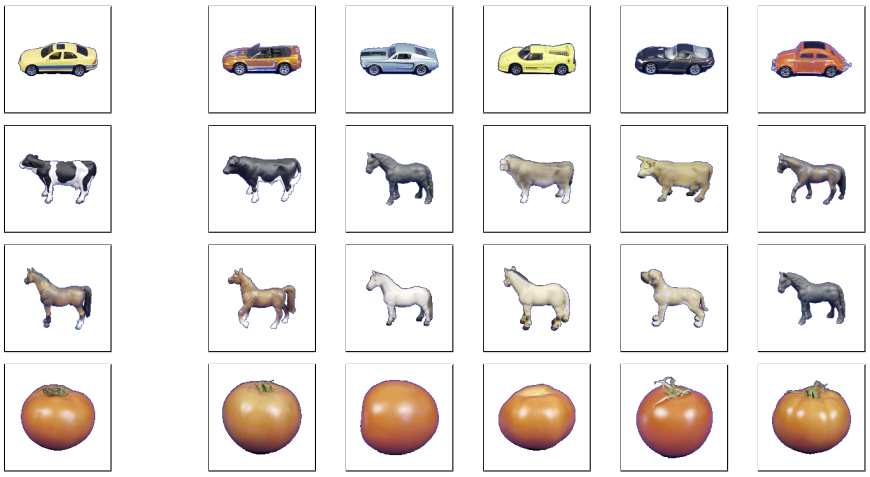


Figure 9.2: Exemplars sorted by the number of generated features - Prototype images with the first five exemplars sorted by the number of generated features at H^5.

9.1.3 Evaluation

The dataset was used in a similar context in [135]. Due to restrictions of our approach (we use only a prototype in a canonical orientation as a model), the results can not be compared directly. Although we do not test for all the orientations, we nevertheless demonstrate that

a reliable categorization can indeed be achieved. Furthermore, the experimental data shows how the discrimination between categories is progressively built through the hierarchy.

The resulting scores are first evaluated on the capability of the method to recognize the most similar exemplar as a member of the same category. If the exemplar with the maximum number of generated features at H^5 is a member of the same category as the prototype, we label the outcome as positive (similarly to [135]). This criteria yields a 100% recognition score, where the method was able to differentiate event between categories that exhibit a significant smilarity in shape, for example apples and tomatoes. Figure 9.2 shows the prototype images with the first five exemplars, sorted by the number of generated features at H^5.

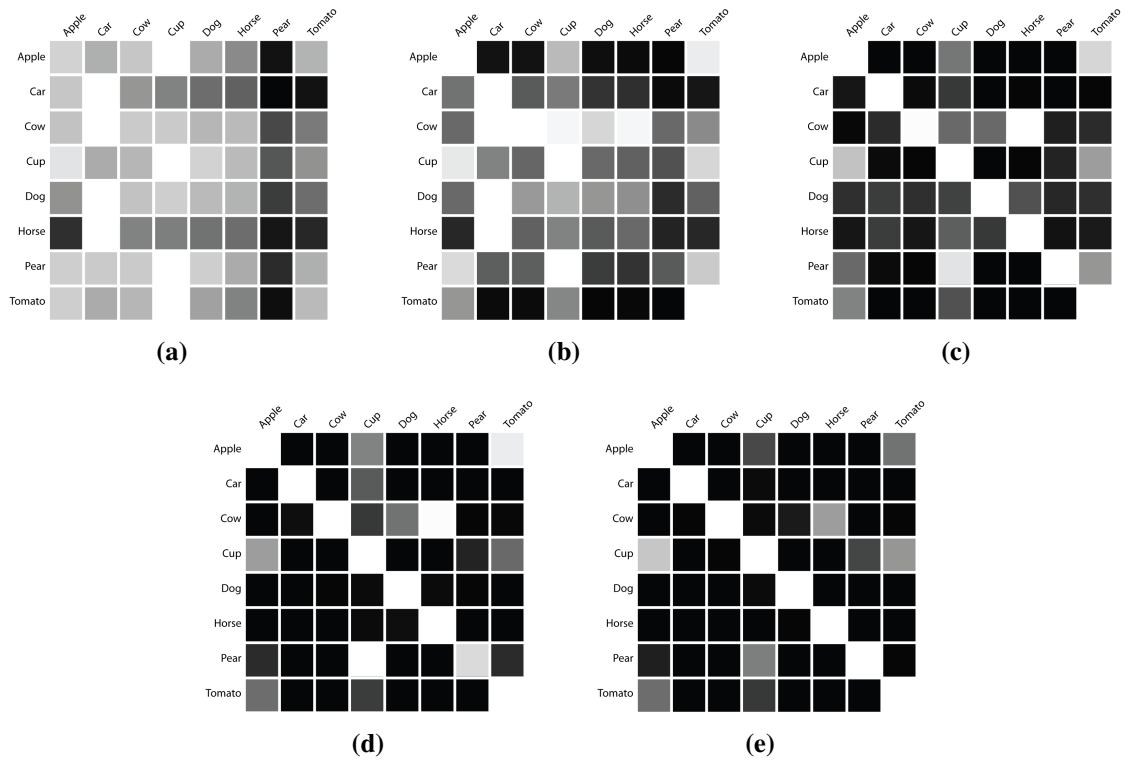


Figure 9.3: Confusion matrices for H^1 - H^5 - The confusion matrices represent the mean number of matched features for each of the five levels (matrices (a) to (e)) of the hierarchy , and for all eight categories. Each row represents the target category (i.e, the category of the prototype), while each of the columns represents mean scores for the exemplar objects, grouped by category.

Next, we evaluated the overall discrimination that arises as the result of binding, matching and selection through the hierarchy levels. For doing that, we calculated the confusion matrices based on *mean number of matched features* for each of the five levels of the hierarchy. As it can be seen in Figure 9.3, the discrimination between categories is poor at the

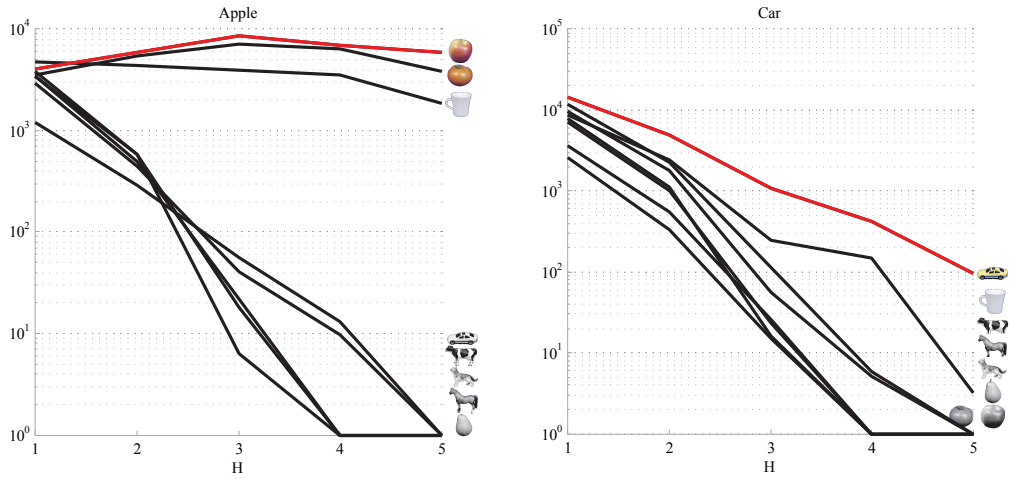


Figure 9.4: Average scores for H^1 - H^5, *Apple* and *Car* categories - The plots depict the average number of matched features for exemplars of each of the eight categories when matched against (a) *Apple* and (b) *Car* prototypes. The results for the query category are denoted by red lines. The icons at the right follow the rank of the categories at level H^5. The icons that stand for categories that scored 0 at any level before H^5, or at H^5, are depicted in grayscale.

lowest levels, while at the higher levels differentiations emerge, first between categories that are completely different in shape, and later between categories that are similar in shape.

As one can observe, there is a pronounced similarity between simply structured objects (*Apple*, *Cup*, *Tomato*), and four-legged animals (*Horse*, *Dog*, *Cow*). Note also that the categories contain exemplars that represent a large departure form the category appearance, e.g. the lying cow, which introduce some bias in the mean measure. Nevertheless, it is clear that at level H^5, a clear distinction between categories is made.

Figures 9.4 to 9.7 further illustrate the average number of matched features at the five levels of the hierarchy for each of the categories separately. The plots depict the average number of matched features for exemplars of each of the eight categories when matched against *Apple* (Figure 9.4a), *Car* (Figure 9.4b), *Cow* (Figure 9.5a), *Cup* (Figure 9.5b), *Dog* (Figure 9.6a), *Horse* (Figure 9.6b), *Pear* (Figure 9.7a), and *Tomato* (Figure 9.7b) prototypes. The results for the category of the prototype are denoted by red lines. The icons at the right follow the rank of the categories at level H^5. The icons that stand for categories that scored 0 at any level of the hierarchy are depicted in grayscale.

The somehow varying quantity of generated features across categories has several causes. In some part, it is influenced by the number of local features that are discovered at H^0.

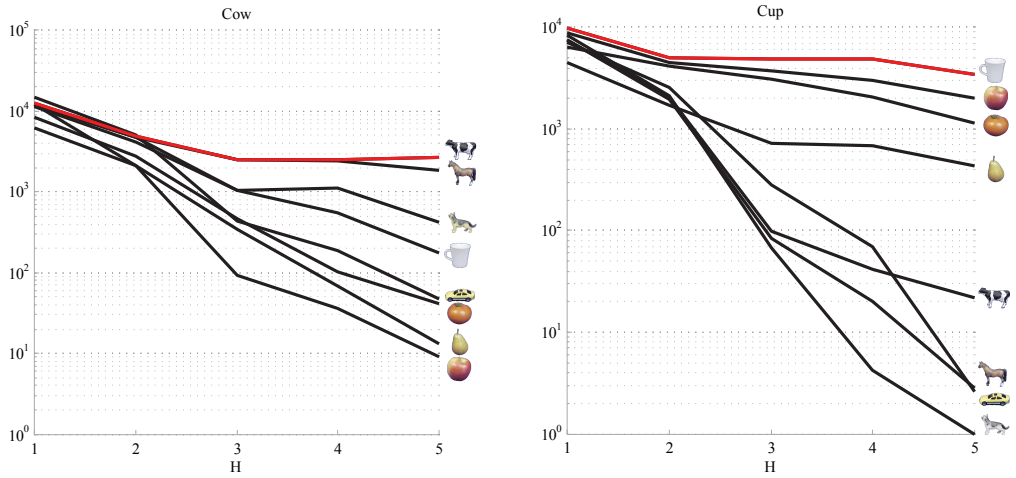


Figure 9.5: Average scores for H^1 - H^5, *Cow* and *Cup* categories - The plots depict the average number of matched features for exemplars of each of the eight categories when matched against (a) *Cow* and (b) *Cup* prototypes. The results for the category of the prototype are denoted by red lines. The icons at the right follow the rank of the categories at level H^5. The icons that stand for categories that scored 0 at any level of the hierarchy are depicted in grayscale.

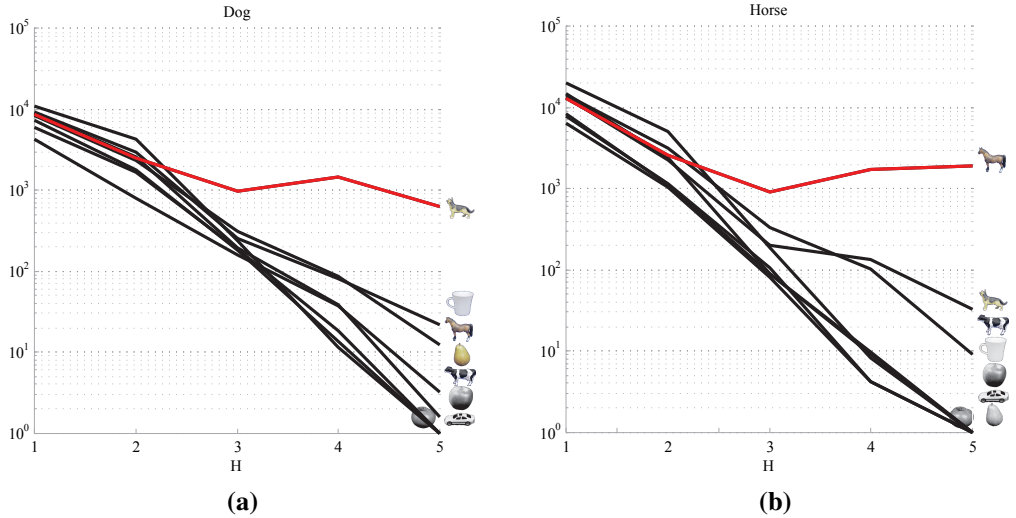


Figure 9.6: Average scores for H^1 - H^5, *Dog* and *Horse* categories - The plots depict the average number of matched features for exemplars of each of the eight categories when matched against (a) *Dog* and (b) *Horse* prototypes. The results for the category of the prototype are denoted by red lines. The icons at the right follow the rank of the categories at level H^5. The icons that stand for categories that scored 0 at any level of the hierarchy are depicted in grayscale.

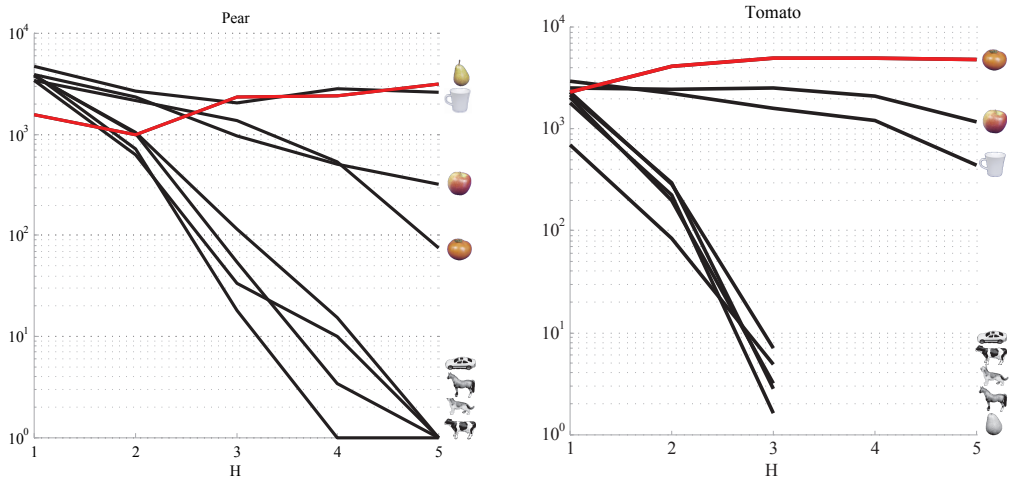


Figure 9.7: Average scores for H^1 - H^5, *Pear* and *Tomato* categories - The plots depict the average number of matched features for exemplars of each of the eight categories when matched against (a) *Pear* and (b) *Tomato* prototypes. The results for the category of the prototype are denoted by red lines. The icons at the right follow the rank of the categories at level H^5. The icons that stand for categories that scored 0 at any level of the hierarchy are depicted in grayscale.

Another important factor is the structural complexity of the category exemplars, as more structural complexity usually means also more variation, and consequently a lower number of matched features. This however leads to a wider receptive field (which depends also on the density of features); it is therefore hard to analytically predict the outcome of growing. As already mentioned, the matching of very similar objects had to be artificially bounded, which is also visible in Figures 9.4a, 9.5b, and 9.7b, where the number of matches at intermediate levels H^3 and H^4 were cut to 5000 matches whenever the total number of matches exceeded this number.

9.2 Object categorization by matching in occluded scenes

In this section we report experimental results on object categorization in arbitrary images, i.e. images where objects appear against a cluttered background, occluded and in different scales. As the primary function of the architecture is structural matching rather than finding category related features, we evaluate the categorization performance on a number of categories that fulfill the following criteria:

- the category has one or more generic views which can be considered in testing

- the members of the category have a consistent structure as a whole, and that structure can be efficiently represented by one example only.

Experiments were performed by using a segmented image as a model, a positive testing set of images depicting objects of the same category against background, and a negative testing set of images that do not depict objects of that category. Most of the categories are taken from the *Caltech 101* database [2; 30]. For testing, we first manually segment one of the images that will represent a the model for that category. The next step is to learn the ICA filters and derive the low level features. Categorization is then performed by selecting a random subset of images from the same category, and an equivalent number of images from all the other categories. We grow a five level hierarchy, and for each image we record the number of features that are generated at the highest level. As the number of features at the highest level reflects the similarity between the model and the object found in the test image, we use it as a similarity measure based on which we decide whether the interpreted image contains the object in question or not. We evaluate the success of the similarity measure by the *Receiver operator characteristic* (ROC).

9.2.1 Caltech-101 database

The *Caltech-101* database [2; 30] is, together with the smaller *Caltech*, and the extended *Caltech-256* database, one of the most used databases of object categories. Pictures of objects are sorted to 101 categories, where the number of images per category is between 40 to 800. Most categories are represented by approximately 50 images. The resolution of images varies, but is roughly between 200×200 and 350×350 pixels. Object outlines are also available, which makes possible to manually segment the objects form the background; we used this information to produce the prototype images.

Objects in *Caltech-101* are often represented in different modalities (photographs, line art, cartoons, clip arts, drawings), however they are often depicted in the center of the image, and under moderate variation in scale and viewpoint. The occlusion is often not present in the image, or is not problematic. While viewpoint and scale can not be considered extremely hard in this database, there is often a drastic variation in appearance of exemplars of the same category. Nevertheless, Fei-Fei et.al [30] reported good categorization results using only a few training examples.

In testing, we were constrained by the time complexity of the algorithm when running on

a serial machine; this limitation allowed for extensive testing only on a small subset of categories. Another constraint is posed by the characteristics of images in *Caltech 101*: as already mentioned, while the variation in scale and viewpoint are typically moderate, other properties of the images vary greatly, and span from high resolution clip arts to low resolution photographs with significant compression artifacts. This causes significant problems in the detection of local features, where our algorithm is designed to perform well under reasonably favorable conditions, such as good contrast, resolution, and low level noise. We therefore use only a small subset of categories: *Motorbikes*, *Airplanes*, *Revolver*, *Dolphin*, *Rooster* and *Faces*. As we are not primarily interested in image retrieval, this restricted set of categories still offers a valuable evaluation that is relevant for the general task of object categorization.

The database can be downloaded at [2].

9.2.2 Methodology

We test the categorization by using one prototype from each of the categories as a model, and in turn match randomly selected images from disjunct categories to estimate whether there is a similar object in the image. For each of the prototypes (for each category), a training is performed to adapt the hierarchy parameters to the model's properties, similarly as in subsection 9.1.2:

- The ICA filters ($n = 5$) are trained using a training subset of manually segmented exemplars. We start by extracting keypoints from the prototype image, and we add keypoints from other training images from the same category until we reach a total of 5000 local patches. ICA filters are then calculated from the centered and whitened image matrix. The training subset of images does not contain the testing images.

- The prototype is used to grow features at each of the levels sequentially; the statistics of the generated features is used to tune the level's parameters.

For testing, we randomly select a subset of 40 (or less, if not available) positive images of the same category as the given prototype, and the same number of random negative exemplars from randomly selected disjunct categories. The prototype is then matched to all testing images. For each matching round we record the score at the highest level of hierarchy.

9.2.3 Evaluation

For each of the tested categories we calculated the *Receiver operator characteristic* (ROC), and rendered a ROC curve. ROC was estimated using a sliding threshold on the number of H^5 features that are successfully matched between the prototype and the image.

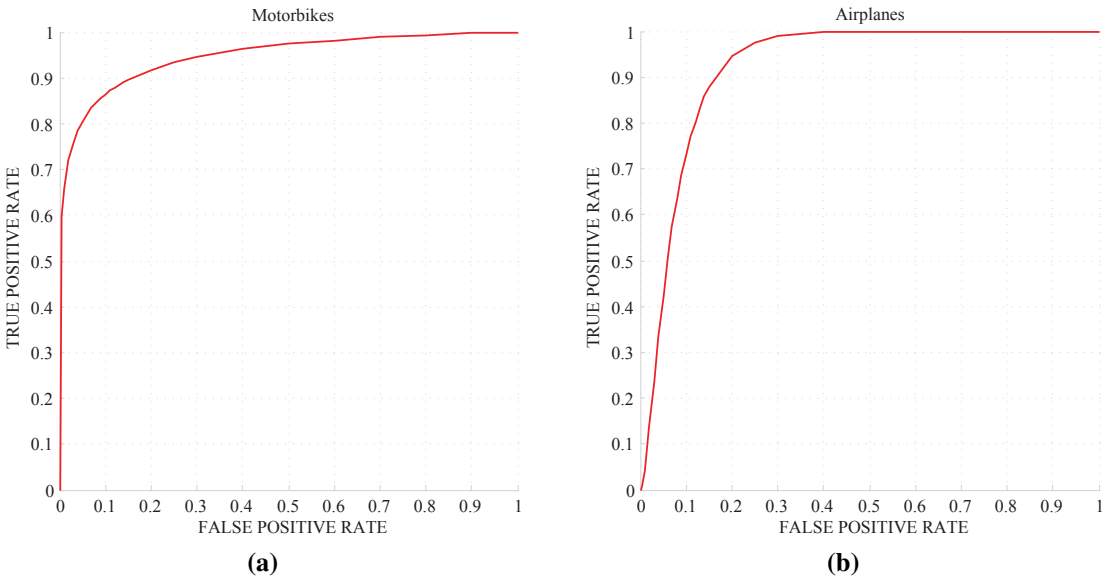


Figure 9.8: ROC curves - *Motorbikes* and *Airplanes* categories - (a) ROC curve for the *Motorbikes* category. (b) ROC curve for the *Airplanes* category.

The Caltech-101 *Motorbikes* category depicts side views of motorbikes. As many of the images depict motorbikes against a clear background, we decided to omit these from the testing set, as they yield a relatively high response in comparison to cluttered images; their inclusion would therefore heavily bias the evaluation and would overestimate the effectiveness of the method. Nevertheless, the ROC curves in Figure 9.8a reveal that the method is successful in adequately grading the images with motorbikes, and that the probability of false positives is relatively low. Whenever a high level exhibited a significant activity, the object was correctly localized, i.e. the receptive fields at the highest level correctly outlined the object. Furthermore, with each of the positive outcomes, we also obtained correspondent parts at different levels of the hierarchy. As correspondences can be established also at the lowest level, it would be possible to develop efficient segmentation algorithms at the pixel level, however this is out of scope of this work.

Good results on this category are mostly due to the rich structure that characterizes motor-

(a)

(b)

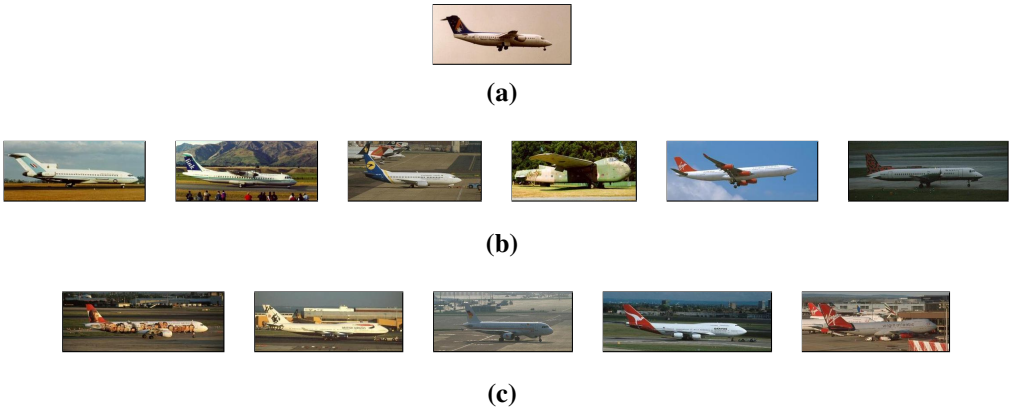
(c)

Figure 9.9: Categorization performance on Caltech-101 *Motorbike* category - (a) depicts the prototype *Motorbike* image. (b) depicts some of the *Motorbike* images with the highest number of matched features, while images in (c) generated a relatively low number of matched features (while those few features correctly identified the object).

bikes. The information in the model image generates enough high level features that clearly discriminate the structural characteristics against of the motorcycle to those that can be found in background. The failure modes can be deduced from the positive images with a low confidence (Figure 9.9c). Firstly, the method responds poorly to noisy images and to images with poor contrast between the object and the background. This is mostly due to a lack of local regions that coincide with the model image. Secondly, although the flexibility allowed by the quantization of geometric features, the method can tolerate only a small level of affine transformations, and responds poorly to objects with out of camera axis rotation with respect to the prototype image.

The Caltech-101 *Airplanes* category depicts side views of airplanes. As the category has less structural features than the *Motorbikes* category, we expected a far worse performance. The method was however efficient in discovering most of the airplanes, as can be also seen from the ROC curve in Figure 9.8b. Problems were encountered only with images of airplanes set against a background with many lines parallel to the body, with images with heavily painted airplanes, and with airplanes with dimensions (length) significantly different than the prototype airplane. In the former case, spurious matches were formed between the background and the body of the airplane which often drifted the final solution to a more consistent (less occluded, or larger) section of the background that had the same elongated structure as the prototype airplane. Figure 9.10b depicts some of the images that sported the highest num-

bers of matched features, while images in Figure 9.10c yielded a relatively low number of matched features.



(a)

(b)

(c)

Figure 9.10: Categorization performance on Caltech-101 *Airplanes* category - (a) depicts the prototype *Airplanes* image. (b) depicts some of the *Airplanes* images with the highest number of matched features, while images in (c) generated a relatively low number of matched features.

The Caltech-101 *Revolver* category depicts side views of revolvers. Although our method successfully discovers most of the revolvers, some images with considerably less local features produce a low score. These images get often over-voted by negative examples that have considerably more local features. This shows the major drawback of our method, i.e. strong hallucination. Figure 9.11a depicts the ROC curve, while Figure 9.12 depicts the most successful and the less successful matches.

Categorization scores on Caltech-101 *Dolphin* category appear worse than results on other categories. However, after examination, it was clear that the bad performance is due to a small number of images that exhibit a large amount of noise and poor figure/background contrast. These images did not form any responses at the highest level of the hierarchy, and are therefore sorted at the very tail of responses. This can be seen at the sharp downturn at the right end of ROC curve, meaning that it would be impossible to achieve a maximum recognition rate until we set the threshold to zero. Figure 9.11b shows the ROC curve, while Figure 9.13 shows the *Dolphin* prototype (Figure 9.13a), dolphins retrieved with a high level of matches (Figure 9.13b), and some of the unsuccessful matches (Figure 9.13c).

Caltech-101 *Rooster* category depicts roosters in side view. This category exhibits a rich response at the local feature level, and side views of roosters are characteristic enough to be easily discernible from other categories (Figure 9.14a, Figure 9.15).

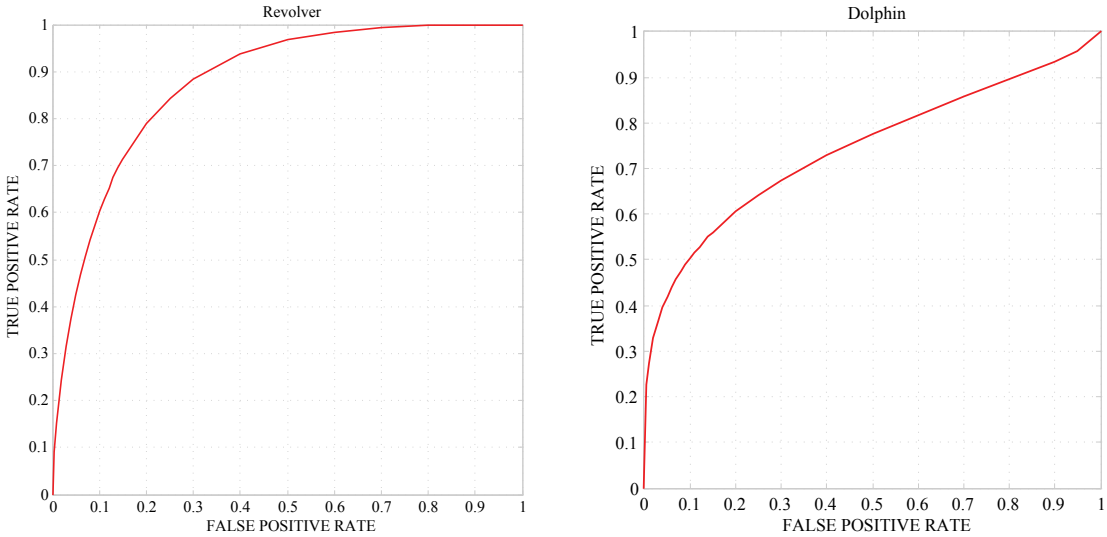


Figure 9.11: ROC curves - *Revolver* and *Dolphin* categories - (a) ROC curve for the *Revolver* category. (b) ROC curve for the *Dolphin* category.

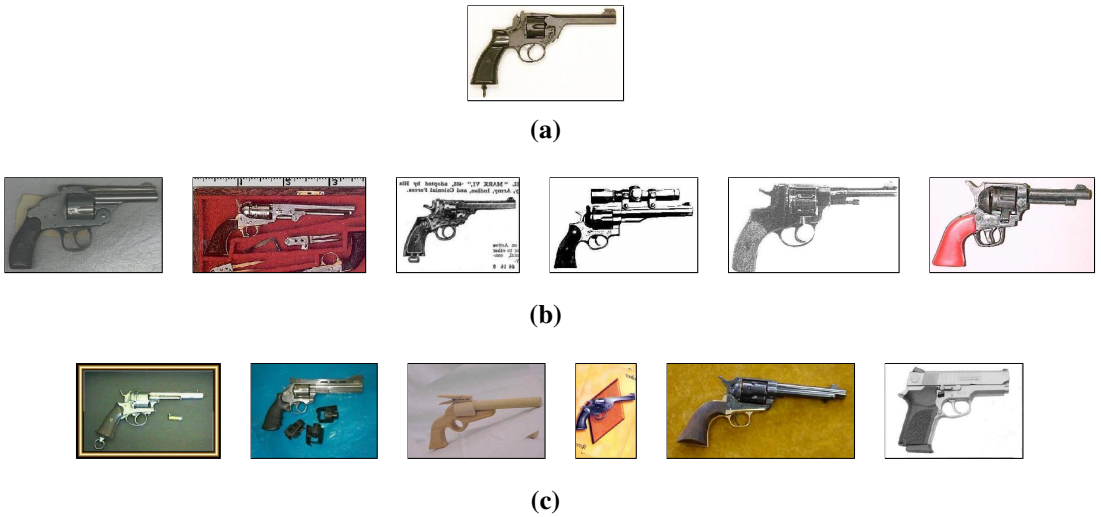


Figure 9.12: Categorization performance on Caltech-101 *Revolver* category - (a) depicts the prototype *Revolver* image. (b) depicts some of the *Revolver* images with the highest number of matched features, while images in (c) generated a relatively low number of matched features.

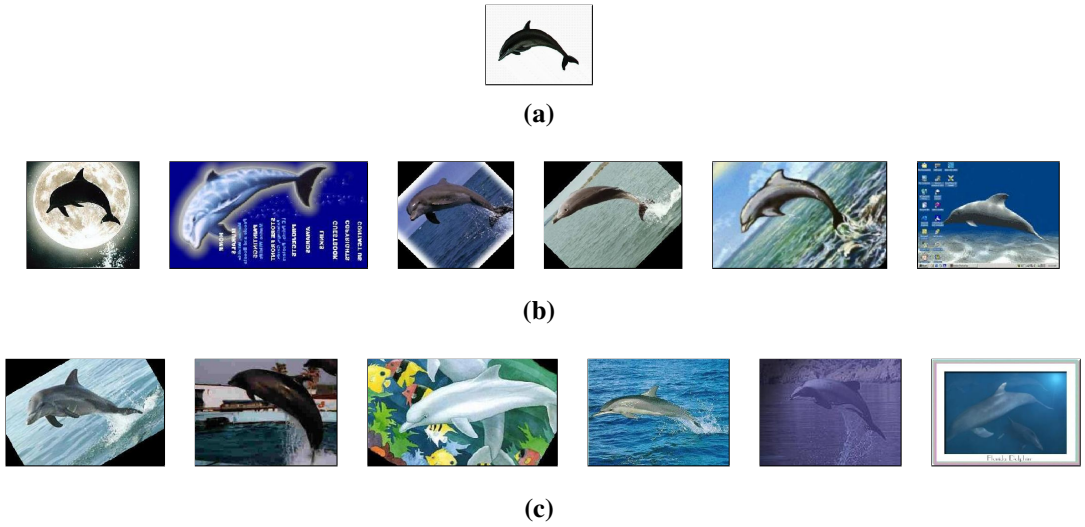


Figure 9.13: Categorization performance on Caltech-101 *Dolphin* category - (a) depicts the prototype *Dolphin* image. (b) depicts some of the *Dolphin* images with the highest number of matched features, while images in (c) generated a relatively low number of matched features.

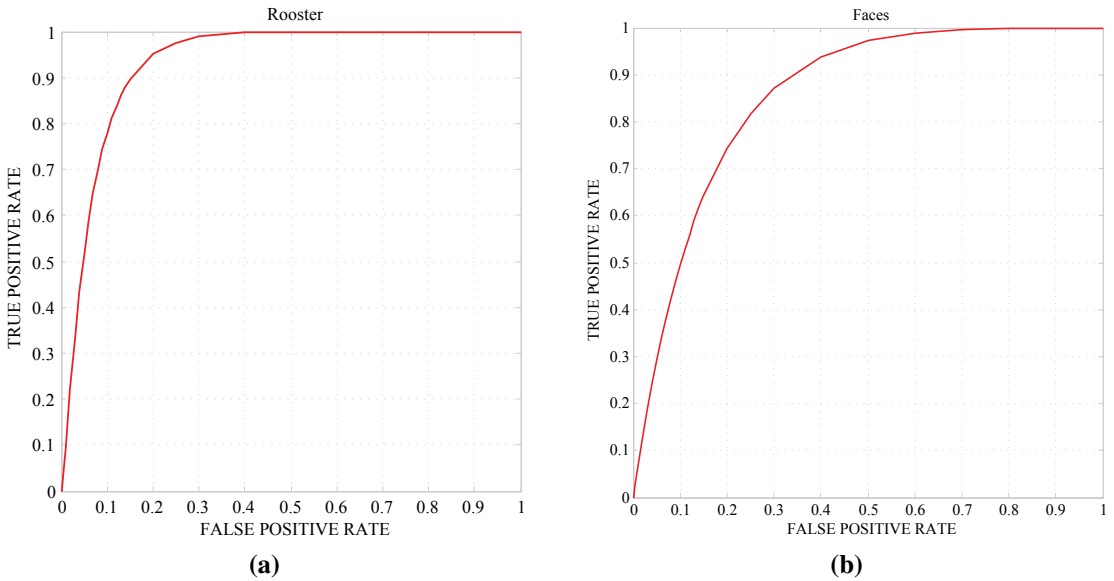


Figure 9.14: ROC curves - *Revolver* and *Dolphin* categories - (a) ROC curve for the *Revolver* category. (b) ROC curve for the *Dolphin* category.

(a)

(b)

Figure 9.15: Categorization performance on Caltech-101 *Rooster* category - (a) depicts the prototype *Rooster* image. (b) depicts some of the *Rooster* images with the highest number of matched features.

The category *Faces* depicts frontal snapshot of faces in occluded scenes with a slight variation in scale. Initially, we observed a large disproportion between the number of L^0 features that lie on borders of the head's silhouette and the few L^0 features that respond to nose, eyes and mouth areas. Due to this anomaly, structures in the background, where an oval shape could be hallucinated often generated a far larger number of successful matches that the true positive features. This was partially due to a non-consistency in the scale of features at the border of the face, where some edges were interrupted by background features, which resulted in a scale that was significantly different of that in the model. To correct for this anomaly, we set the scale of all L^0 features to an approximately same size, which dramatically improved the results. Note however that by doing this, we lost invariance to scale changes. The ROC curve for the modified algorithm can be seen in Figure 9.14b.

Figure 9.16 and Figure 9.17 depict some of the matches at level H^5 for the *Motorbikes* and *Faces* category, respectively. Although, in terms of category detection, all of the matches can be considered as successful, and all of them are drawn from a high number of successful matches, the matchings can be used to identify some potential failure modes.

9.3 Discussion

One of the problems we encountered is the increasing variability of spatial configuration of features at the bottom levels when they participate in high level bindings. At each level of binding, the geometric features are quantized into a number of intervals, which define the bias that is allowed for a certain geometric property. Consequently, the features that are binded with a succession of borderline values (therefore exhibiting a large deviation

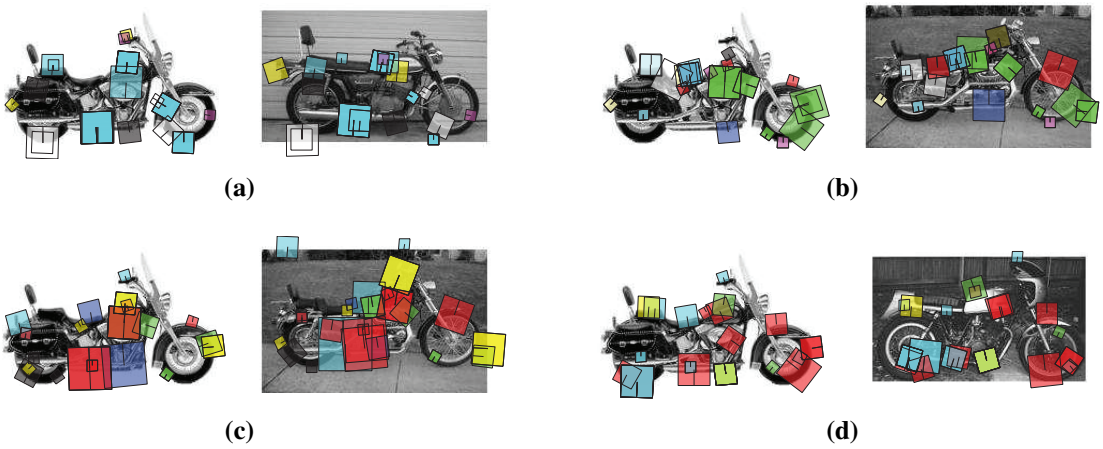


Figure 9.16: Matches - *Motorbikes* - Some matches between prototype and positive test images at level H^5 in *Motorbikes* category. (a) most of the low level matches on the object are correctly retrieved, with the exception of the position of the handlebar. (b) the seat is estimated as being lower than in the prototype image. (c) the background features made it through the geometric constraints and were promoted to the highest level. (d) dissimilar objects are correctly matched.

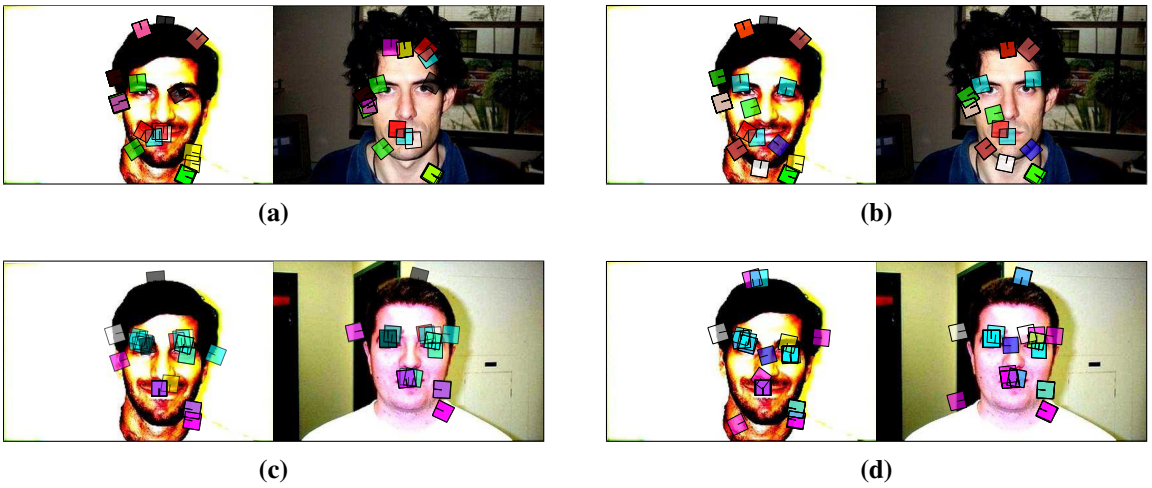


Figure 9.17: Matches - *Faces* - Some matches between prototype and positive test images at level H^5 in *Faces* category. (a), (b) most of the low level matches on the object are correctly retrieved; the size of the face is underestimated, which results in matching the top of the head in the prototype with the hairline in the test image.

from the "center" value of a bin in the geometric conceptual representation), accumulate more deviations through the hierarchy, which result in matching of moderately deformed structures. An example of such matching can be seen e.g. in Figure 9.16c,where a H^1 feature is matched to a background structure that lies far off the object, and partially in Figure 9.17a, Figure 9.17b and Figure 9.17d, where the scale of parts of the face are overestimated.

Another problem of our method is *hallucination*, which is most evident in cases where an image with a complex background is tested for the presence of an object with a simple structure. This problem can be also seen as a consequence of *Ugly Duckling Theorem* ([163], Section 2.4). For example, faces, or body shapes can be easily hallucinated in natural patterns, which often happens in human vision. This phenomena was extensively studied in psychology, and is usually attributed to the dynamics and the interplay of bottom-up binding and top-down expectation in the visual hierarchy [53]. In our case, top-down expectations are represented by the prototype and the tuning of the hierarchy to capture the variation of features in the prototype. As nature of the hierarchy is such that only one prototype is given, as the goal is to find *any* match, there is usually not enough information in the model to support further checking of consistency or other mechanisms that would reject spurious detections. Consequently, a large number of hypothetical matching features have to be promoted from the bottom levels to the top levels. For shapes that are globally simple, or where the features defining a simple outline overweight the features that define the more detailed structures (as, for example, in faces), perceptually simple structures in background are more effective in producing matches at the bottom levels, and, supported by a large number of hypothesis, either steer the match at the highest level away form the true object, or hallucinate an object where there is none.

The distances computed between two objects, i.e., their number of matched top level features, are absolute distances, based on the number of their subordinate features that enter the matching process. Therefore, the distance between two complex, but similar, objects will exceed the distance between two similar objects with fewer features. Also, the distance between a simple object and a noisy background could exceed the distance between two simple objects. The distance between two objects could be normalized by a measure of each object's complexity, such as for example the average number of features at H^1, or the overall number of matches [112].

Another possible remedy would be to learn the structural significance patterns over a category. We however found that, with a large number of exemplars in a category, and with

the imminent large number of possible feature couplings, this is not as straightforward as it seems. Furthermore, the introduction of a statistical framework is not in spirit of our investigation.

When the system of categorical perception is embedded in a cognitive framework, several other sources of information can be used to establish a stable interpretation of the scene. For example, the framework presented in [113] uses an attentive mechanism based on contextual constraints that focus the processing on contextually relevant areas and eliminate the spurious detections. Several other mechanisms of contextual processing were proposed that greatly downplay the importance of invariance (to noise, scale etc.) in recognition; however most of them require a rich source of contextual information, which is typically available only in active, embodied systems. In other words, there is a huge difference between requirements and preconditions that arise in *artificial cognitive vision* and in *image retrieval*. In our opinion, the fact that image retrieval does not operate on objects, but rather on pictorial representations of objects, was not stressed in the vision community to a sufficient degree.

Several researchers stressed the relevance of tests that are performed on so called databases of "natural" images [118; 121]. As we already mentioned, the images in Caltech-101 database have only a slight intra-category variance in viewing parameters, and most of the objects are depicted in a standard setting, i.e., they are usually photographed [121]. Another important issue is however that the image background often complies with the category, which basically favors many of the recognition methods that score state-of-the-art results. As Pinto et al. showed, a very simple V1-like model with no special machinery to tolerate image variation performed remarkably well on Caltech-101, surpassing the performance of most state-of-the-art methods. However, when the same model was tested on a set of images with systematical changes in viewing parameters (position, scale, in-plane rotation and depth-rotation), the performance of the same model degraded quickly, both in images with a scene background, as in images with white noise background [118].

10

Conclusion

In this thesis we presented a new framework for visual object categorization by hierarchical matching. The main contribution is a novel approach for visual categorization of objects by synchronous hierarchical matching to a prototype, where high level matches between an object and a prototype are gradually discovered through several steps of binding, selection and inhibition. We demonstrated how categorization can be achieved without an excessive collection of evidence, allowing thus to discover objects of a certain category in images of unknown objects and in occluded scenes by matching to a single prototype that represents the category.

The experimental results proved that our method is successful in matching objects that belong to the same category (defined by a human agent), and that the number of high level matches is a good indicator of appearance and structural similarity between two objects. The major novelty with respect to classical methods that also operate on local features is that we can retrieve, besides the global matches, also a large number of consistent subordinate matches at each level of the hierarchy; we can therefore estimate structural similarities at different levels of detail. It is also important that we achieve this without an explicit part-based representation—the matched parts are constructed ad-hoc. Nevertheless, if integrated in a cognitive system, the matches offer a solid support for "meta-categorization", i.e. for an introspective analysis of why two objects match, although no semantic or sub–semantic parts are identified.

Another novelty is the hierarchical ad-hoc binding, matching and inhibition heuristics, that can be seen as a greedy approach for finding similarities in structure. Its major advantage is that it operates on limited neighborhoods to find local consistencies, that it directs attention by modeling conditioning the receptive field, and that it then gradually expands to encompass

the whole model. As the local bindings are independent, the hierarchy can be implemented in a parallel computing environment using an arbitrary number of basic computational units.

As the categories are not learned by an extensive collecting of evidence, but rather by tuning the hierarchy with respect to the prototype, and as the pictorial representation of the prototype is nearer to a raw physical phenomena than to an abstracted learned representation, one can draw interesting parallels to the paradigms of enactive cognition. We do not learn what is a category, but rather how to do the matching (by tuning the conceptual space). The prototype based representation seems a natural approach to categorization: for example, we are clearly able to match two objects from a category that we encounter for the first time (e.g. matching two different cars in the same scene by estimating their similarities), and we can estimate their similar parts, which is not the case for many of the statistical and representational frameworks.

A novel contribution is also the way that we calculate the low level codebook. Sparse codebooks have been until now computed only on random patches from "natural scenes", and the derived filters were compared to receptors in biological vision systems. These filters therefore modeled a pre-attentional stage of processing. Our codebooks are calculated on the regions of interest, and therefore model the appearance information on oriented and scaled regions. The local ICA approach also differs from similar methods [66], as it results in a maximally sparse response vector for each of the local regions. Although we learn category–specific codebooks, one could also learn a general codebook, while still keeping the number of codebook entries relatively low.

We also discussed the drawbacks and the failure modes. Hallucination is a serious problem when dealing with image retrieval; it is a much lesser problem in a cognitive framework, with additional information at hand. Nevertheless, we feel that a more careful design of the retrieval of low level features would dramatically improve the performance. Note that the framework is agnostic to the type of ROI detector, so local regions of different types, such as contours or blobs, could be integrated in the early processing stage.

Another drawback is the promoting and amplification of the local errors through the hierarchy. We plan to address this problem by implementing a redundant discretization of the conceptual space, and by conditioning the binding by a cost function that would penalize borderline values.

An obvious limitation of our framework is that it recognizes only a canonical view of the

object. This touches the core debate on representation in vision, namely how to represent different views of objects. We claim that a system that is capable of recognizing a small set of canonical views, and that exhibits some tolerance to changes in orientation, can be efficient in most situations. In an embodied cognitive system, ambiguities can be resolved by shifting the viewpoint to select a more informative view. We therefore plan to specifically analyze and improve the projective and the affine invariance of our method.

To answer the question on how multiple categories can be represented, we first plan to embed the framework in a contextual framework, which will prime for a certain categories of objects. Further, we plan to experiment with synchronous matching to multiple prototypes. Furthermore, we plan to investigate a hierarchical system of prototypes, where scenes are first matched to general categories (basic shapes), and then to more specific prototypes.

We also plan to extend our method so that it could operate with rendered 3-D prototypes; we speculate that if bindings are initially generated over the whole viewing sphere, they would, by the process of match and inhibition, automatically converge and "vote" for a certain viewpoint.

Lastly, we plan to test the matching framework on other modalities.

References

[1] http://www.vision.ethz.ch/projects/categorization/eth80db.html. 112

[2] http://www.vision.caltech.edu/Image_Datasets/Caltech101/Caltech101.html. 118, 119, 154

[3] Amit, Y. and Geman, D., A computational model for visual selection. *Neural Computation*, 11(7):1691–1751, 1999. 50

[4] Ananthanarayanan, R. and Modha, D.S., Anatomy of a cortical simulator. In *Supercomputing 07: Proceedings of the ACM/IEEE SC2007 Conference on High Performance Networking and Computing*, 2007. 106

[5] Ballard, D., Animate vision. *Artificial Intelligence Journal*, 48:57–86, 1991. 25

[6] Bar, M., The proactive brain: using analogies and associations to generate predictions. *Trends in Cognitive Sciences*, 11(7):280–289, 2007. 18, 41, 67

[7] Bell, A.J. and Sejnowski, T.J., The 'independent components' of natural scenes are edge filters. *Vision Research*, 37(23):3327–3338, 1997. 66, 77, 152

[8] Belongie, S., Malik, J., and Puzicha, J., Shape matching and object recognition using shape contexts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(4):509–522, 2002. 58

[9] Berg, A., Berg, T., and Malik, J., Shape matching and object recognition using low distortion correspondences. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 26–33 vol. 1, 2005. 58, 59, 65

[10] Biederman, I., Recognition-by-components: A theory of human image understanding. *Psychoogical Review*, 94:115–147, 1987. 26, 45, 50

[11] Bouchard, G. and Triggs, B., Hierarchical part-based visual object categorization. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 710–715vol.1, 2005. 50, 51

[12] Brooks, R.A., Intelligence without reason. In J. Myopoulos and R. Reiter, editors, *Proceedings of the 12th International Joint Conference on Artificial Intelligence (IJCAI-91)*, pages 569–595, Morgan Kaufmann publishers Inc.: San Mateo, CA, USA, Sydney, Australia, 1991. 25, 31

[13] Chen, D.Y., Tian, X.P., Shen, Y.T., and Ouhyoung, M., On visual similarity based 3d model retrieval. In P. Brunet and D. Fellner, editors, *EUROGRAPHICS 2003*, 2003. 54

[14] Chum, O. and Zisserman, A., An exemplar model for learning object classes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2007. 59

[15] Crandall, D., Felzenszwalb, P., and Huttenlocher, D., Spatial priors for part-based recognition using statistical models. In *CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1*, pages 10–17, IEEE Computer Society, Washington, DC, USA, 2005. 48

[16] Cross, A. and Hancock, E., Graph matching with a dual-step EM algorithm. *IEEE Transactions on Pattern Analysis and MAchine Intelligence*, 20(11):1236–1253, 1998. 47

[17] Csurka, G., et al, Visual categorization with bags of keypoints. In *ECCV International Workshop on Statistical Learning in Computer Vision*, 2004. 48

[18] Damasio, A.R., Synchronous activation in multiple cortical regions: a mechanism for recall. *Seminars in the Neurosciences*, 2:287–296, 1990. 40

[19] Deng, H., et al, Principal curvature-based region detector for object recognition. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1–8, 2007. 47

[20] Descartes, R., *Treatise of Man*. Prometheus Books, 2003. First published 1629. 40

[21] ———, *A Discourse on the Method*. Oxford University Press, 2006. First published in 1637. 68

[22] Dickinson, S., et al, Object categorization and the need for many-to-many matching. In *DAGM05*, page 501, 2005. 55

[23] Dickinson, S.J., Pentland, A.P., and Rosenfeld, A., 3-d shape recovery using distributed aspect matching. *IEEE Trans. Pattern Anal. Mach. Intell.*, 14(2):174–198, 1992. 46

[24] Dreyfus, H.L., *What Computers Still Can't Do: A Critique of Artificial Reason*. MIT Press, 1992. 31

[25] Duncan, J., Humphreys, G., and Ward, R., Competitive brain activity in visual attention. *Curr Opin Neurobiol*, 7(2):255–261, 1997. 38

[26] Edelman, S., Representation, similarity, and the chorus of prototypes. 1994. 19

[27] ———, Representation is representation of similarity. *Behavioral and Brain Sciences*, 21:449–498, 1998. 20, 21

[28] ———, *Representation and Recognition in Vision*. MIT Press, 1999. 21, 26, 49

[29] Farah, M.J., *The Cognitive Neuroscience of Vision*. Blackwell Publishing, 2000. 23, 28, 29, 40, 41, 89

[30] Fei-Fei, L., Fergus, R., and Perona, P., Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *Computer Vision and Pattern Recognition Workshop, 2004 Conference on*, pages 178–178, 2004. 48, 118

[31] Fergus, R., Perona, P., and Zisserman, A., Object class recognition by unsupervised scale-invariant learning. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 2, pages II–264–II–271vol.2, 2003. 49

[32] ———, A sparse object category model for efficient learning and exhaustive recognition. *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, 1:380–387 vol. 1, 2005. 49

[33] Ferrari, V., Fevrier, L., Jurie, F., and Schmid, C., Groups of adjacent contour segments for object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(1):36–51, 2008. 49

[34] Fidler, S., Berginc, G., and Leonardis, A., Hierarchical statistical learning of generic parts of object structure. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 1, pages 182–189, 2006. 50, 53, 67, 89

[35] Fidler, S., Boben, M., and Leonardis, A., Similarity-based cross-layered hierarchical representation for object categorization. In *CVPR*, 2008. 53, 67, 89

[36] Fidler, S. and Leonardis, A., Towards scalable representations of object categories: Learning a hierarchy of parts. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1–8, 2007. 50, 53, 66, 67, 89

[37] Field, D.J., Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America A*, 27:2370–2393, 1987. 77

[38] Fodor, J.A. and Pylyshyn, Z.W., Connectionism and cognitive architecture: A critical appraisal. *Cognition*, 28:3–71, 1988. 29, 34

[39] Foerster, H.V., *Observing Systems*. Intersystems Publications, second edition, 1982. 31

[40] Freeman, W.J. and Skarda, C.A., Brain organization and memory cells, systems & circuits. chapter Representations: Who needs them?, pages 375–380, Oxford University Press: NewYork, 1990. 31

[41] Froese, T. and Ziemke, T., Enactive artificial intelligence. euCognition white paper, 2007. 31

[42] Frome, A., Singer, Y., Sha, F., and Malik, J., Learning globally-consistent local distance functions for shape-based image retrieval and classification. In *IEEE 11th International Conference on Computer Vision, ICCV 2007*, pages 1–8, 2007. 59

[43] Fukushima, K., Miyake, S., and Ito, T., *Artificial neural networks: theoretical concepts*, chapter Neocognitron: a neural network model for a mechanism of visual pattern recognition, pages 136–144. IEEE Computer Society Press, Los Alamitos, CA, USA, 1988. 50

[44] Gärdenofors, P., *Conceptual spaces: the geometry of thought*. MIT Press, 2004. 17, 20, 30, 37, 85

[45] van Gelder, T.J., Dynamic approaches to cognition. In F. Wilson and F. F. Keil, editors, *The MIT Encyclopedia of Cognitive Sciences*, pages 244–246, MIT Press, 1999. 32

[46] Gibson, J.J., *The Ecological Approach to Visual Perception*. New Jersey and London: Lawrence Erlbaum Associates, Publishers., 1979. 25

[47] Granlund, G.H., An associative perception-action structure using a localized space variant information representation. In *AFPAC '00: Proceedings of the Second International Workshop on Algebraic Frames for the Perception-Action Cycle*, pages 48–68, Springer-Verlag, London, UK, 2000. 55

[48] Granlund, G.H. and Moe, A., Unrestricted recognition of 3d objects for robotics using multilevel triplet invariants. *AI Mag.*, 25(2):51–67, 2004. 53

[49] Grauman, K. and Darrell, T., The pyramid match kernel: Discriminative classification with sets of image features. In *ICCV '05: Proceedings of the Tenth IEEE International Conference on Computer Vision*, pages 1458–1465, IEEE Computer Society, Washington, DC, USA, 2005. 48

[50] Gray, C., König, P., Engel, A., and Singer, W., Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties. *Nature*, 338(6213):334–337, 1989. 38

[51] Grimson, W. and Lozano-Perez, T., Model-based recognition and localization from tactile data. In *Robotics and Automation. Proceedings. 1984 IEEE International Conference on*, volume 1, pages 248–255, 1984. 46

[52] Grimson, W. and Lozano Perez, T., Localizing overlapping parts by searching the interpretation tree. *PAMI*, 9(4):469–482, 1987. 47

[53] Grossberg, S., How hallucinations may arise from brain mechanisms of learning, attention, and volition. *Journal of the International Neuropsychological Society*, 6:579–588, 1999. 127

[54] Harnad, S., The symbol grounding problem. *Physica D*, 42:335–346, 1990. 34

[55] ———, Handbook of categorization in cognitive science. In H. Cohen and C. Lefebvre, editors, *Summer Institute in Cognitive Sciences on Categorisation*, chapter To Cognize is to Categorize: Cognition is Categorization, pages 19–43, Elsevier Science Ltd, Oxford, 2005. UQaM Summer Institute in Cognitive Sciences on Categorization. 30 June - 11 July 2003. 23

[56] Harris, C. and Stephens, M., A combined corner and edge detector. In *Proceedings of The Fourth Alvey Vision Conference, Manchester*, pages 147–151, 1988. 75, 151

[57] Heidegger, M., *Being and Time*. Harper Perennial, 2008. 31

[58] Hyvarinen, A. and Hoyer, P., Emergence of complex cell properties by decomposition of natural images into independent feature subspaces. In *Artificial Neural Networks, 1999. ICANN 99. Ninth International Conference on (Conf. Publ. No. 470)*, volume 1, pages 257–262vol.1, 1999. 66, 77, 78, 152, 153

[59] J. Shotton, A. Blake, R.C., Multi-scale categorical object recognition using contour fragments. *IEEE Trans. on PAMI*, 30(7):1270–1281, 2008. 49

[60] Jastrow, J., The mind's eye. *Popular Science Monthly*, 54:299–312, 1899. 33

[61] Jin, Y. and Geman, S., Context and hierarchy in a probabilistic image model. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 2145–2152, 2006. 50, 52

[62] Jogan, M. and Leonardis, A., Robust localization using an omnidirectional appearance-based subspace model of environment. *Robotics and Autonomous Systems*, 45(1):51–72, 2003. 35, 47

[63] Jolicoeur, P., Gluck, M., and Kosslyn, S., Pictures and names: making the connection. *Cognitive Psychology*, 16(2):243–275, 1984. 16

[64] Kadir, T. and Brady, M., Scale saliency: a novel approach to salient feature and scale selection. In *Visual Information Engineering, 2003. VIE 2003. International Conference on*, pages 25–28, 2003. 74

[65] ———, Saliency, scale and image description. *Int. J. Comput. Vision*, 45(2):83–105, 2001. 47

[66] Ke, Y. and Sukthankar, R., Pca-sift: a more distinctive representation for local image descriptors. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pages II–506–II–513Vol.2, 2004. 76, 77, 130, 160

[67] Kelly, G.A., *The Psychology of Personal Constructs*. Routledge, 1991. 85, 86

[68] Keselman, Y., Shokoufandeh, A., Demirci, M., and Dickinson, S., Many-to-many graph matching via metric embedding. In *CVPR03*, pages I: 850–857, 2003. 55, 56

[69] Keselman, Y. and Dickinson, S., Generic model abstraction from examples. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(7):1141–1156, 2005. 55

[70] Kim, G., Faloutsos, C., and Hebert, M., Unsupervised modeling of object categories using link analysis techniques. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2008. 58

[71] Kim, W.Y. and Kak, A.C., 3-d object recognition using bipartite matching embedded in discrete relaxation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 13(3):224–251, 1991. 54

[72] Kimia, B.B., Tannenbaum, A.R., and Zucker, S.W., Shapes, shocks, and deformations, I: The components of shape and the reaction-diffusion space. *IJCV*, 15(3):189–224, 1995. 54

[73] Kinsbourne, M., Integrated field theory of consciousness. In *Consciousness in Contemporary Science*, page 239256, 1988. 40

[74] Kosslyn, S., Ball, T., and Reiser, B., Visual images preserve metric spatial information: Evidence form studies of image scanning. *Journal of Experimental Psychology: Human Perception and Performance*, 4:47–60, 1978. 42

[75] Kosslyn, S., *Image and Mind*. Harvard University Press, Cambridge, MA, 1980. 42

[76] Lacey, A., Individuals. In *A Dictionary of Philosophy*, Routledge, third edition edition, 1996. 15

[77] ———, Universals and particulars. In *A Dictionary of Philosophy*, Routledge, third edition edition, 1996. 14

[78] Laeng, B., Zarrinpar, A., and Kosslyn, S.M., Do separate processes identify objects as exemplars versus members of basic-level categories? Evidence from hemispheric specialization. *Brain and Cognition*, 53(1):15–27, 2003. 15

[79] Lee, T.S. and Mumford, D., Hierarchical bayesian inference in the visual cortex. *J Opt Soc Am A Opt Image Sci Vis*, 20(7):1434–1448, 2003. 53

[80] Leibe, B., Leonardis, A., and Schiele, B., Robust object detection with interleaved categorization and segmentation. *International Journal of Computer Vision*, 77(1-3):259–289, 2008. 49, 76

[81] Leonardis, A. and Bischof, H., Robust recognition using eigenimages. *Computer Vision and Image Understanding - Special Issue on Robust Statistical Techniques in Image Understanding*, 78(1):99–118, 2000. 47

[82] Leordeanu, M. and Hebert, M., A spectral technique for correspondence problems using pairwise constraints. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 2, pages 1482–1489Vol.2, 2005. 58

[83] Leordeanu, M., Hebert, M., and Sukthankar, R., Beyond local appearance: Category recognition from pairwise interactions of simple features. In *Proceedings of CVPR*, 2007. v, 58, 59

[84] Lewicki, M.S. and Olshausen, B.A., A probabilistic framework for the adaptation and comparison of image codes. *Journal of the Optical Society of America A*, 16(7):1587–1601, 1999. 77

[85] Lindeberg, T. and Eklundh, J., Scale detection and region extraction from a scale-space primal sketch. In *Computer Vision, 1990. Proceedings, Third International Conference on*, pages 416–426, 1990. 4, 47, 56, 75, 151

[86] Lowe, D.G., *Perceptual Organization and Visual Recognition*. Kluwer Academic Publishers, Norwell, MA, USA, 1985. 46

[87] ———, Object recognition from local scale-invariant features. In *Proc. of the International Conference on Computer Vision, Corfu*, IEEE Computer Society, 1999. 4, 47, 74, 75, 151, 152

[88] Loy, G. and Zelinsky, A., Fast radial symmetry for detecting points of interest. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(8):959–973, 2003. 47

[89] Ma, W.C., Wu, F.C., and Ouhyoung, M., Skeleton extraction of 3d objects with radial basis functions. In *Shape Modeling International, 2003*, pages 207–215, 2003. 54

[90] Makadia, A., Visontai, M., and Daniilidis, K., Harmonic silhouette matching for 3d models. In *3DTV Conference, 2007*, pages 1–4, 2007. 54

[91] Malafouris, L., Before and beyond representation: Towards an enactive conception of the Palaeolithic image. In C. Renfrew and I. Morley, editors, *Material beginnings: a global prehistory of figurative representation*, 2007. 42, 61

[92] Malisiewicz, T. and Efros, A.A., Recognition by association via learning per-exemplar distances. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, 2008. 2, 18, 60, 65

[93] Marr, D., *Vision*. W. H. Freeman, San Francisco, CA, 1982. 25, 26, 27, 45, 50

[94] Matas, J., Chum, O., Urban, M., and Pajdla, T., Robust wide baseline stereo from maximally stable extremal regions. *Image and Vision Computing*, 22(10):761–767, 2004. 47

[95] Maturana, H.R. and Varela, F.J., *Autopoesis and Cognition: The Realization of the Living*. D. Reidel Publishing Company: Boston, 1980. 31

[96] M.Carcassoni and Hancock, E., Spectral correspondence for point pattern matching. *Pattern Recognition*, 36:193–204, 2003. 55

[97] McGee, K., Enactive cognitive science. part 1: Background and research themes. *Constructivist Foundations*, 1(1):19–34, 2005. 31

[98] ———, Enactive cognitive science. part 2: Methods, insights, and potential. *Constructivist Foundations*, 1(2):73–82, 2006. 31

[99] Merleau-Ponty, M., *Phenomenology of Perception*. Routledge, 2002. 14, 31

[100] Mikolajczyk, K. and Schmid, C., Scale and affine invariant interest point detectors. *International Journal of Computer Vision*, 60(1):63–86, 2004. 4, 47

[101] Mundy, J., Object recognition in the geometric era: A retrospective. In *CLOR06*, pages 3–28, 2006. 46

[102] Murase, H. and Nayar, S., Visual learning and recognition of 3-d objects from appearance. *IJCV*, 14(1):5–24, 1995. 47, 77

[103] Newell, A. and Simon, H.A., Computer science as empirical inquiry: symbols and search. *Commun. ACM*, 19(3):113–126, 1976. 29

[104] Noë, A., *Action in Perception*. MIT Press, Cambridge, MA, 2004. 25, 67

[105] Ohbuchi, R., Minamitani, T., and Takei, T., Shape-similarity search of 3d models by using enhanced shape functions. In *Theory and Practice of Computer Graphics, 2003. Proceedings*, pages 97–104, 2003. 54

[106] Oja, E., *Subspace methods of pattern recognition*. Research Studies Press, 1983. 77

[107] Olshausen, B.A. and Field, D.J., How close are we to understanding v1? *Neural Computation*, 17:1665–1699, 2005. 30

[108] Olshausen, B.A. anf Field, D., Emergence of simple–cell receptive field properties by learning a sparse code for natural images. *Nature*, (381):607–609, 1997. 65, 77, 152

[109] Opelt, A., Pinz, A., and Zisserman, A., Incremental learning of object detectors using a visual shape alphabet. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 1, pages 3–10, 2006. 49, 65

[110] O'Regan, J.K. and Noë, A., A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences*, 24(5):883–917, 2001. 25, 32, 33, 67

[111] Palmer, S., Rosch, E., and Chase, P., Canonical perspective and the perception of objects. In *Attention and Performance IX*, pages 135–151, 1981. 19

[112] Pelillo, M., Siddiqi, K., and Zucker, S., Attributed tree matching and maximum weight cliques. In *Image Analysis and Processing, 1999. Proceedings. International Conference on*, pages 1154–1159, 1999. 55, 127

[113] Perko, R., Wojek, C., Schiele, B., and Leonardis, A., Probabilistic combination of visual context based attention and object detection. In *International Workshop on Attention in Cognitive Systems*, Lecture Notes in Computer Science, pages 166–179, 2008. 128

[114] Piaget, J., *The psychology of intelligence*. Routledge, second edition, 2002. First English edition published 1950 by Routledge & Kegan Paul. 31

[115] Piater, J. and Scalzo, F., Unsupervised learning of visual feature hierarchies. In *Machine Learning and Data Mining in Pattern Recognition*, volume 3587 of *Lecture Notes in Computer Science*, pages 243–252, 2005. 53

[116] ———, Unsupervised learning of dense hierarchical appearance represe. In *18th International Conference on Pattern Recognition ICPR 2006*, volume 2, pages 395–398, 2006. 53

[117] Piater, J., Scalzo, F., and Detry, R., Vision as inference in a hierarchical markov network. In *Twelfth International Conference on Cognitive and Neural Systems*, 2008. 53

[118] Pinto, N., Cox, D.D., and Dicarlo, J.J., Why is real-world visual object recognition hard? *PLoS Computational Biology*, 4(1):e27+, 2008. 128

[119] Plaut, D. and Farah, M.J., Visual object representation: Interpreting neurophysiological data within a computational framework. *Cognitive Neuroscience*, 4:320–343, 1990. 28

[120] Ponce, J. and Chelberg, D., Finding the limbs and cusps of generalized cylinders. In *Robotics and Automation. Proceedings. 1987 IEEE International Conference on*, volume 4, pages 62–67, 1987. 45

[121] Ponce, J., et al, Dataset issues in object recognition. In *Towards Category-Level Object Recognition*, pages 29–48, Springer, 2006. 128

[122] Pylyshyn, Z., *Seeing and Visualizing: It's not what you think*. MIT Press, Pyl, 2003. 36, 41

[123] Quiroga, R.Q., et al, Invariant visual representation by single neurons in the human brain. *Nature*, 435(7045):1102–1107, 2005. 27

[124] Riesenhuber, M. and Poggio, T., Models of object recognition. *Nature Neuroscience Supplement*, 3:1199–1203, 2000. 50, 51, 67, 89

[125] Roberts, L.G., Machine perception of three-dimensional solids. *Optical and Electro-optical Information Processing*, pages 159–197, 1956. 45

[126] Rosch, E., Natural categories. *Cognitive Psychology*, 4:328–350, 1973. 19

[127] Rosch, E., et al, Basic objects in natural categories. *Cognitive Psychology*, 8:382439, 1976. 16, 19, 32, 65, 148

[128] Russell, B., *The Analysis of Mind*. Allen and Unwin, 1921. 42

[129] Schmid, C. and Mohr, R., Local greyvalue invariants for image retrieval. *PAMI*, 19(5):530–535, 1997. 47

[130] Searle, J., Minds, brains, and programs. *Behavioral and Brain Sciences*, 3:417–424, 1980. 68

[131] Sebastian, T., Klein, P., and Kimia, B., Recognition of shapes by editing their shock graphs. 26(5):550–571, 2004. 55

[132] Selinger, A. and Nelson, R.C., A perceptual grouping hierarchy for appearance-based 3d object recognition. *Computer Vision and Image Understanding*, 76(1):83–92, 1999. 49, 51, 65, 66

[133] Shapiro, L.S. and Brady, J.M., Feature-based correspondence: an eigenvector approach. *Image and Vision Computing*, 10(5):283–288, 1992. 55

[134] Shokoufandeh, A., et al, Indexing hierarchical structures using graph spectra. *IEEE Transactions on Pattern Analysis and Machine Intelligence, special issue on Syntactic and Structural pattern Recognition*, 27(7):1125–1140, 2005. 55, 56, 57, 66

[135] ———, The representation and matching of categorical shape. *Computer Vision and Image Understanding*, 103(2):139–154, 2006. 56, 57, 59, 66, 113, 114

[136] Siddiqi, K., Shokoufandeh, A., Dickinson, S., and Zucker, S., Shock graphs and shape matching. *International Journal of Computer Vision*, 30:1–24, 1999. 54, 55, 66

[137] Singer, W. and Gray, C.M., Visual feature integration and the temporal correlation hypothesis. *Annu Rev Neurosci*, 18:555–586, 1995. 40

[138] Sivic, J., et al, Discovering object categories in image collections. In *Proceedings of the International Conference on Computer Vision*, 2005. 48

[139] Sloman, S.A., Love, B.C., and Ahn, W.K., Feature centrality and conceptual coherence. *Cognitive Science: A Multidisciplinary Journal*, 22(2):189–228, 1998. 17

[140] Štimec, A., Jogan, M., and Leonardis, A., Unsupervised learning of a hierarchy of topological maps using omnidirectional images. *International Journal of Pattern Recognition and Artificial Intelligence (IJPRAI)*, 22(4):639–665, 2008. 35

[141] Storkey, A.J. and Williams, C.K., Image modeling with position-encoding dynamic trees. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(7):859–871, 2003. 52

[142] Sudderth, E., Torralba, A., Freeman, W., and Willsky, A., Learning hierarchical models of scenes, objects, and parts. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 2, pages 1331–1338Vol.2, 2005. 48, 50

[143] Tanaka, K., Neuronal mechanisms of object recognition. *Science*, (262):685–688, 1993. 38

[144] ———, Mechanisms of visual object recognition: Monkey and human studies. *Current Opinion in Neurobiology*, (7):523–529, 1997. 38

[145] Taraborelli, D., Feature binding and object perception. does object awareness require feature conjunction? In *10th Annual Meeting of the European Society for Philosophy and Psychology - ESPP 2002*, Lyon, 2002. 38, 39

[146] Tarr, M.J. and Black, M.J., A computational and evolutionary perspective on the role of representation in vision. *Computer Vision, Graphics, and Image Processing. Image Understanding*, 60(1):65–73, 1994. 25, 36

[147] Terzopoulos, D. and Metaxas, D., Dynamic 3d models with local and global deformations: deformable superquadrics. In *Computer Vision, 1990. Proceedings, Third International Conference on*, pages 606–615, 1990. 45

[148] Thomas, N.J., Mental imagery. In E.N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Fall 2008. 42, 43

[149] Thomasson, A., Categories. In E.N. Zalta, editor, *The Stanford Encyclopedia of Philosophy (Fall 2004 Edition)*, 2004. 13, 14, 15

[150] Todorovic, S. and Ahuja, N., Extracting subimages of an unknown category from a set of images. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 1, pages 927–934, 2006. 52, 67, 89

[151] ———, Learning subcategory relevances to category recognition. In *Proc. IEEE Comp. Soc. Conf. Computer Vision and Pattern Recognition (CVPR 2008), Anchorage, AL*, 2008. 52, 89

[152] Treisman, A., Feature binding, attention and object perception. *Philos Trans R Soc Lond B Biol Sci.*, 353:1295–1306, 1998. 38

[153] Treisman, A. and Gelade, G., A feature-integration theory of attention. *Cognitive Psychology*, 12:97–136, 1980. 38

[154] Turing, A., Computing machinery and intelligence. *Mind*, 59:433–460, 1950. 68

[155] Turk, M. and Pentland, A., Face recognition using eigenfaces. In *Proc. Computer Vision and Pattern Recognition, CVPR-91*, pages 586–591, 1991. 47

[156] Tuytelaars, T. and Van Gool, L., Matching widely separated views based on affine invariant regions. *Interantional Journal on Computer Vision*, 59(1):61–85, 2004. 47, 65, 75

[157] Ullman, S., *High-level vision*. The MIT Press, 1996. 49

[158] Ullman, S. and Epshtein, B., Visual classification by a hierarchy of extended fragments. In J. Ponce, M. Hebert, C. Schmid, and A. Zisserman, editors, *Toward Category-Level Object Recognition*, volume 4170 of *Lecture Notes in Computer Science*, pages 321–344, Springer, 2006. 50

[159] Underwood, S. and Coates, C.L., J., Visual learning from multiple views. *Computers, IEEE Transactions on*, C-24(6):651–661, 1975. 46

[160] Varela, F.J., Thompson, E.T., and Rosch, E., *The Embodied Mind: Cognitive Science and Human Experience*. MIT Press, 1992. 31

[161] Vernon, D., Metta, G., and Sandini, G., A survey of artificial cognitive systems: Implications for the autonomous development of mental capabilities in computational agents. *IEEE Transactions on Evolutionary Computation*, 11(2):151–180, 2007. 29, 34

[162] Viola, P. and Jones, M., Rapid object detection using a boosted cascade of simple features. In *CVPR 01*, 2001. 48

[163] Watanabe, S., *Pattern Recognition, Human and Mechanical*. John Wiley & Sons, 1985. 19, 127

[164] Weber, M., Welling, M., and Perona, P., Towards automatic discovery of object categories. In *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, volume 2, pages 101–108vol.2, 2000. 49

[165] Wittgenstein, L., *Tractatus Logico-Philosophicus*. Routledge & Kegan Paul Ltd, London, 1961. First published in 1913. 18, 42

[166] ———, *Philosophical Investigations*. Macmillan, New York, 1973. Translated by G. E. M. Anscombe. 1, 18, 33, 148

[167] ———, *Philosophical Grammar*. University of California Press, Berkeley and Los Angeles, 1978. 1, 18, 33, 148

Appendix A

Razširjeni povzetek

A.1 Uvod

V disertaciji predlagamo novo metodo za vizualno prepoznavo in kategorizacijo objektov, ki je osnovana na hierarhičnem odkrivanju intenzitetnih in strukturnih ujemanj med interpretirano sliko in prototipno sliko objekta. Z iskanjem ujemanj lahko kategoriziramo objekte z izčrpnim iskanjem podobnosti med objekti; kategorije objektov torej niso opredeljene z naborom značilk, ki bi omogočale optimalno kategorizacijo, temveč kot mrežni sistem podobnosti med objekti.

Osnovni princip delovanja temelji na sinhronem in hierarhičnem iskanju ujemanj, pri čemer se visokonivojska ujemanja konstruirajo pragmatično, v več nivojih združevanja, izbire in inhibicije. Model uspešno združuje kategorizacijske metode na lokalnih značilkah ter hierarhičen pristop združevanja, in odpravlja siceršnjo odvisnost lokalnih metod od lokalnih intenzitetnih vzorcev, saj omogoča postopno gradnjo značilk, ki postajajo diskriminativne šele na stopnji, ko že opisujejo večji del objekta.

Motivacija Kategorizacija je eden od središčnih problemov kognitivne znanosti. Raziskave na področju umetnih kognitivnih sistemov kategorizacijo pogosto obravnavajo kot tisti temeljni mehanizem, ki omogoča implementacijo osnovnih kognitivnih funkcionalnosti, saj podatke abstrahira in jih predstavi v ustreznem sistemu kategorij. To velja tudi za podatke, katere umetni sistemi pridobivajo iz senzorjev. Zaznavanje s sposobnostjo kategorizacije senzorične informacije imenujemo kategorično zaznavanje. Vizualna kategorizacija je podproblem kategoričnega zaznavanja, ki obravnava problem kategorične zaznave vidne informacije.

Osnovna funkcija vizualne kategorizacije je torej kategorizacija posameznega vidnega dražljaja. Kategorizacija pri tem navadno omogoča vzpostavitev višjenivojske opredelitve dražljaja, denimo kot poimenovanje objekta z imenom kategorije objekta, ali pa kot bolj splošno umestitev dražljaja v nek sistem kategorij glede na generične značilnosti, ki ne opredeljujejo le konkretnih lastnosti zaznanega dražljaja, temveč so lastne večjemu številu entitet.

Vizualna kategorizacija objektov omogoča uvrstitev opazovanega objekta v nek sistem kategorij objektov. Pri tem je posebej značilna razlika med procesoma prepoznave objekta in kategorizacije. Pri prepoznavi objekta so pomembni predvsem tisti atributi, po katerih lahko predmet hitro prepoznamo kot že videnega, obenem pa ga lahko zanesljivo razlikujemo od ostalih, podobnih predmetov. Lahko bi torej trdili, da je kategorizacija problem, podrejen prepoznavi: če lahko prepoznamo vse objekte, in če poznamo kategorije, katerim določeni objekti pripadajo, potlej lahko vse objekte tudi kategoriziramo. Vendar pa je v resnici problem kategorizacije mnogo težji, saj moramo v novem objektu, katerega opazujemo prvič, prepoznati tiste lastnosti, na podlagi katerih ga lahko uvrstimo v določeno kategorijo. Kategorije torej ne opredelimo z naštevanjem posameznih članov, temveč z opredelitvijo ali z odkrivanjem lastnosti, ki omogočajo ugotavljanje podobnosti za objekte znotraj kategorije, in na podlagi katerih jih lahko razlikujemo od predmetov v tujih kategorijah. Pri tem seveda ni jasno, katere značilnosti so ustrezne za opredelitev določene kategorije. Prvotni računski modeli kategorizacije so te značilnosti opredelili vnaprej, z uporabo abstraktnih modelov, denimo geometrijskih oblik. Sodobne raziskave pa v veliki meri obravnavajo modele, ki temeljijo na statističnem modeliranju. Značilnosti kategorij se torej naučijo z opazovanjem velikega števila pozitivnih in negativnih primerov.

V tem delu proučujemo možnost kategorizacije na podlagi kanonične slikovne predstavitve objekta kot prototipa, pri čemer se pripadnost kategoriji odloča glede na podobnosti, ki se konstituirajo za vsak posamezen postopek interpretacije. Kategorije so tako določene le kot mreža podobnosti [166; 167], pripadnost kategoriji pa kot podobnost prototipu glede na število ugotovljenih ujemanj [127]. Iskanje ujemanj je zasnovano kot hierarhičen, distribuiran in sinhron postopek iskanja ujemanj med prototipom in interpretirano sliko v omejenih lokalnih območjih. Lokalna območja se v postopoma širijo, dokler ne opišejo ujemajočih se struktur na nivoju objekta. Pri tem nas predvsem zanima možnost iskanja ujemanj brez uporabe naučenega znanja o višjenivojskih strukturah, ki so značilne za kategorijo. Vizualno učenje opredelimo kot učenje razpršenega slovarja nizkonivojskih lokalnih filtrov, ter kot učenje geometrijskih konceptov, ki vplivajo na konstrukcijo višjenivojskih opisov.

Originalni prispevki k znanosti Originalne prispevke k znanosti lahko strnemo v naslednjih točkah:

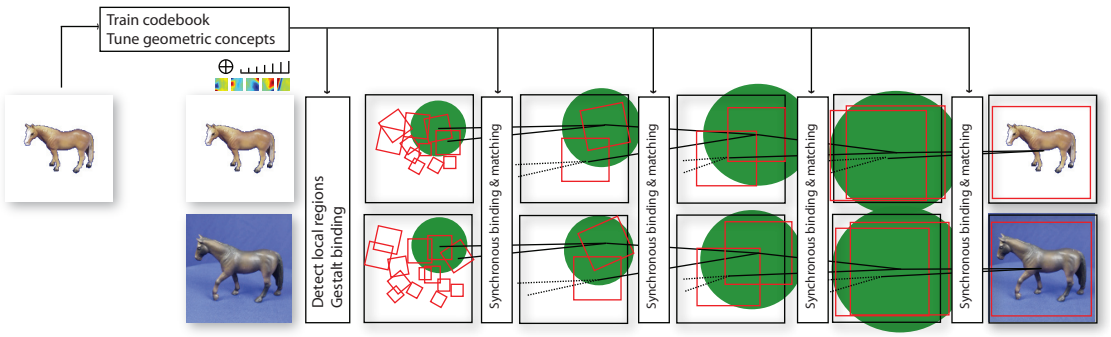
- Predlagamo nov pristop k hierarhičnemu iskanju ujemanj, ki je zasnovan kot distribuiran in sinhron proces. Pristop temelji na združevanju potencialnih ujemanj v lokalnih območjih in inhibicije opisnikov za katere ne najdemo ujemanja. Lokalnost in distribuiranost pristopa omogoča, da sicer kompleksen problem razdelimo na podprobleme, ki zaradi sinhronosti procesiranja prototipnega modela in interpretirane slike konvergirajo k ustrezni rešitvi. Hierarhičnost procesiranja omogoča postopno izbiro obetavnih ujemanj, ki tvorijo osnovo za višjenivojska ujemanja.
- Predlagamo novo metodo za učenje razpršenih slovarjev za predstavitev lokalnih regij. Metoda temelji na analizi neodvisnih komponent množice zaplat, ki opisujejo regije. Tako naučeni slovarji zagotavljajo maksimalno razpršen odzivni vzorec, kar omogoča optimalno klasifikacijo lokalnih regij.
- Predlagamo postopek učenja parametrov hierarhije glede na geometrijske lastnosti prototipa.
- Višjenivojski opisniki so v hierarhiji konstruirani pragmatično in niso del naučene predstavitve kategorije. Dokazujemo torej, da lahko objekte učinkovito kategoriziramo tudi brez eksplicitnega modeliranja višjenivojskih struktur in delov objekta. Predlagamo tudi hipotezo, da je iskanje podobnosti primernejši mehanizem za kategorizacijo kot učenje na primerih.
- Lastnosti predlagane metode pritrjujejo vzpostavitvenim (*enactive*) teorijam zaznavanja in nudijo možno interpretacijo problema imaginacije v kognitivnih sistemih.
- Metodo evaluiramo na dveh slikovnih zbirkah. Pokažemo, da metoda učinkovito kategorizira objekte, ter da poleg tega omogoča tudi odkrivanje večjega števila ujemanj na različnih nivojih podrobnosti.

A.2 Povzetek

Predlagana metoda uporablja hierarhično iskanje ujemanj med prototipom in interpretirano sliko na podlagi strukturnih in fotometričnih opisnikov, katerih kompleksnost narašča z nivoji hierarhije. Združevanje se prične na nivoju lokalnih regij, ki opisujejo informativna in stabilna področja v sliki. Opisnike v lokalnih področjih nato v hierarhičnem postopku

združujemo v pare, ki preko nivojev hierarhije pridobivajo in prenašajo informacijo o strukturnih značilnostih lokalnih konstelacij. Lokalne soseščine se pri tem sistematično povečujejo, dokler ne pokrivajo celotne površine objekta. Združevanje značilk je uravnavano s sprotnim iskanjem ujemanj in z inhibicijo tistih značilk, ki ne vodijo k potencialnim ujemanjem na višjem nivoju. V nasprotju s hierarhičnimi modeli prepoznave, ki eksplicitno modelirajo hierarhično predstavitev kategorije objektov, temelji naš pristop na sprotni konstrukciji hierarhičnih značilk.

Diagram na sliki A.1 prikazuje poenostavljeno shemo poteka kategorizacije. Zavoljo preglednosti je na sliki je prikazan le del hierarhičnega drevesa povezovanj dveh začetnih lokalnih značilk, od najnižjega do najvišjega nivoja. Učenje kategorije ne predpodstavlja obširnega zbiranja znanja o kategoriji objektov, temveč je implementirano kot sprotno učenje lastnosti prototipa, in kot učenje in prilagajanje parametrov združevanja, kar je na shemi prikazano kot prva stopnja kategorizacije. Značilke se združujejo v pare v okviru lokalnih področij, ki z vsakim nivojem rastejo glede na velikost sestavljenih zančilnic. Če je bilo iskanje ujemanj uspešno, lokalna področja na končnem nivoju obsegajo celotno površino prototipa, in ustrezno ujemajočo površino interpretirane slike.



Slika A.1: Shema modela za hierarhično iskanje ujemanj - Učenju na podlagi prototipa sledi več nivojev združevanja, iskanja ujemanj in inhibicije. Procesi potekajo sinhrono, v okviru lokalnih soseščin prototipa in na interpretirani sliki. Z rdečo barvo so uokvirjena lokalna področja, z zeleno pa področja združevanja. Zavoljo preglednosti je na sliki prikazan le del hierarhičnega drevesa povezovanj dveh začetnih lokalnih regij.

Detekcija lokalnih regij Z detekcijo stabilnih lokalnih regij zagotovimo osnovo za kasnejše združevanje slikovne informacije. Klasične metode prepoznavanja objektov, ki temeljijo na podlagi lokalnih regij, uporabljajo informativne opisnike, ki so načrtovani z namenom minimizacije števila napačnih ujemanj. Ker se lahko objekti iz določene kategorije

na lokalnem nivoju zelo razlikujejo, morajo te metode pridobiti in modelirati veliko informacije; določanje lokalnih lastnosti, ki so bistvene za kategorijo, hkrati pa dovolj diskriminatorne, pa je težko, kar še posebej velja pri modeliranju večjega števila kategorij. Pričujoči pristop lokalno fotometrično informacijo uporabi le kot osnovo, na kateri postopno gradi vedno bolj informativne značilke, ki dejansko opisno moč pridobijo šele na višjih nivojih opisa posameznega objekta. Kljub temu pa morajo nizkonivojske značilke ustrezati naslednjim kriterijem:

strukturna informativnost
značilke naj imajo lastno karakteristično velikost in enolično določeno smer;

ponovljivost
lokacije značilk so ponovljive za objekte posamezne kategorije;

odpornost na spremembo gledišča
značilke naj bodo stabilne ob manjši spremembi gledišča;

gostost in redundanca
značilke naj bodo dovolj goste in naj redundantno opisujejo značilne lokalne strukture.

Lokalna območja poiščemo z detektorjem lokalnih regij, ki je zasnovan na podlagi analize diferenčnih slik v prostoru meril [85; 87]. Sliko $I(x,y)$ tako preslikamo v prostor meril $L(x,y,\sigma)$, do katerega pridemo s konvolucijo slike z Gaussovim jedrom $G(x,y,\sigma)$ spremenljive velikosti:

$$L(x,y,\sigma) = G(x,y,\sigma) * I(x,y) \tag{A.1}$$

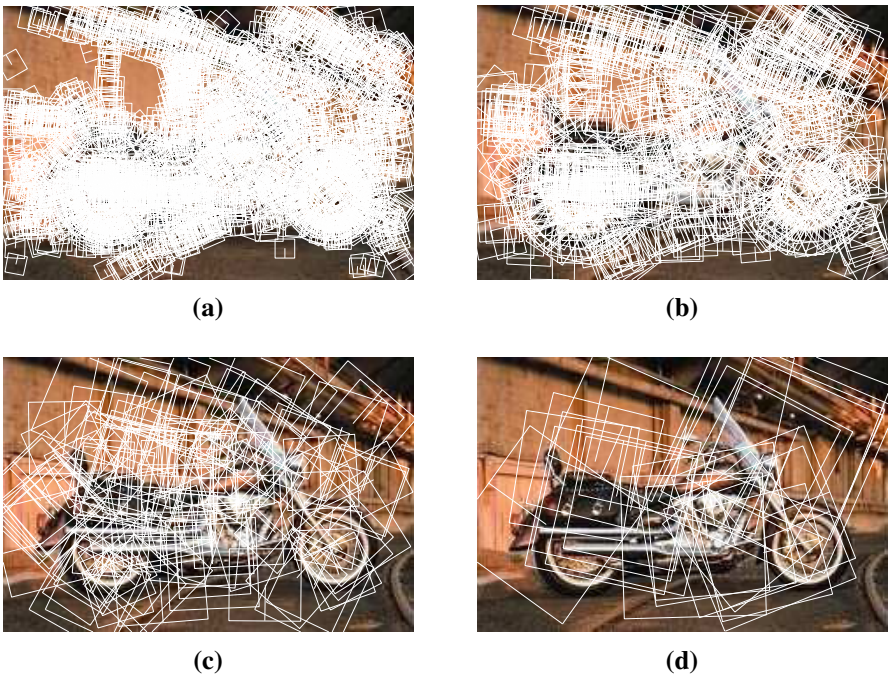
$$G(x,y,\sigma) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2}. \tag{A.2}$$

Za vsako oktavo, začenši z velikostjo slike $(c \times r)$, pridobimo tri diferenčne slike na različnih merilih. Vsako naslednjo oktavo pridobimo z vzorčenjem na polovično velikost. V hierarhiji slik velikosti $(\log \frac{\min(c,\, r)}{\log 2} - 2)$ oktav nato poiščemo ekstreme, ki nakazujejo središčne koordinate potencialnih kandidatk za lokalne regije [87]. V nasprotju z postopkom, opisanim v [87], kandidatom za lokalne regije pridružimo tiste, ki ustrezajo kriteriju za prisotnost robu. Ustreznost takih regij ocenimo na podlagi Hessove matrike

$$\mathbf{H} = \begin{bmatrix} I_{xx} & I_{xy} \\ I_{xy} & I_{yy} \end{bmatrix}, \tag{A.3}$$

in Harrisovega kriterija $\frac{Trace(\mathbf{H})^2}{Det(\mathbf{H})} \geq \frac{(r+1)^2}{r}$ [56], kjer je r razmerje med lastnimi vrednostmi matrike $\mathbf{H}$.

Velikost posamezne regije je določena z merilom, na katerem je bila regija detektirana. Usmerjenost regije je določena kot smer, kjer histogram lokalnih gradientov zavzame maksimalno vrednost [87].



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

Slika A.2: Lokalne regije - Lokalne regije K_i, detektirane na prvih štirih oktavah. Pravokotniki ponazarjajo velikost regij in njihovo usmerjenost.

Lokalne regije K_i so torej opisane z lokacijo ${}^{c}\mathsf{K}_i$, usmerjenostjo ${}^{o}\mathsf{K}_i$, ter velikostjo ${}^{r}\mathsf{K}_i$. Slika A.2 prikazuje regije, detektirane na prvih štirih oktavah.

Učenje slovarja Lokalne regije so sedaj opisane le s strukturnimi značilnostmi regije. Opis videza regije podamo na podlagi slovarja filtrov, katerih maksimalno razpršeni odzivi določajo razred lokalne regije. Število razredov, in s tem velikost slovarja, je namenoma majhno, saj se želimo izogniti natančnemu opisu, ki bi preprečilo konstrukcijo visokonivojskih ujemanj.

Nabora filtrov se naučimo z analizo neodvisnih komponent (*Independent Component Analysis, ICA*) slikovne matrike učnih regij, ki jih pridobimo s prototipne slike, ali pa iz večih slik objektov določene kategorije. V literaturi se slovarji ICA pogosto obravnavajo kot model nevronskih receptorjev v primarnem vidnem korteksu [7; 58; 108], vendar pa učenje slovarja

vedno temelji na naključno izbranih regijah naključnih slik. V nasprotju s pristopi v literaturi v tem delu računamo slovar le na informiranih področjih, ki so bili že izbrani v postopku detekcije. Slovar se zato izogne modeliranju redundantne informacije, in se osredotoči na opise značilnih struktur z generično smerjo in velikostjo, ki je določena že z algoritmom detekcije. Če poleg tega učenje omejimo na predmete ene kategorije, lahko že manjši slovar opiše značilne dele objektov (Slika A.3).

Slikovno matriko $\mathbf{X}$ sestavimo iz normaliziranih intenzitetnih vektorjev $\mathbf{x}$, ki opisujejo posamezne regije v generični smeri. Posamezen vektor $\mathbf{x}_i$ lahko predstavimo z linearno superpozicijo več izvorov $\mathbf{b}_i$,

$$\mathbf{x}_i \;=\; \sum_{j=1}^{N} a_{ij}\mathbf{b}_j \tag{A.4}$$

uteženo s koeficienti a_{ij}. Matrično obliko enačbe A.4 zapišemo kot

$$\mathbf{x} \;=\; \mathbf{A}\mathbf{b}\ . \tag{A.5}$$

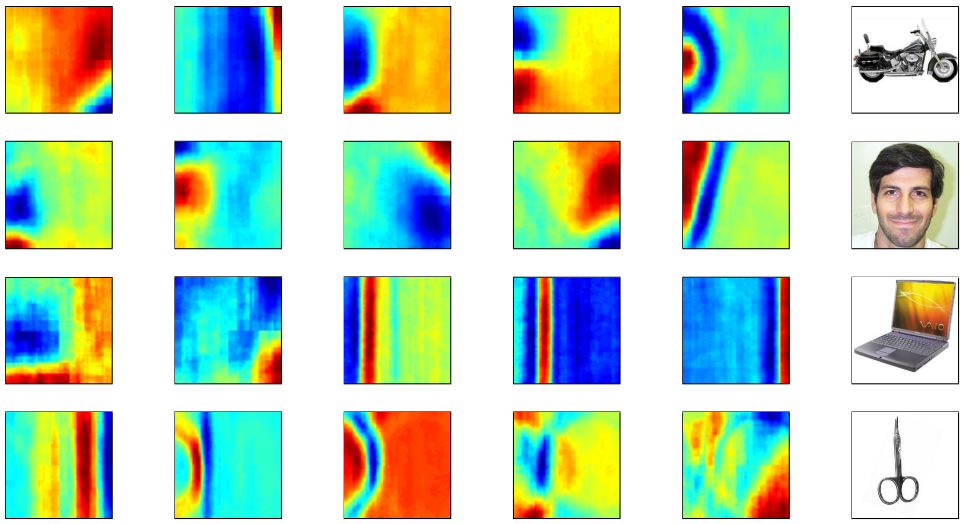
Če $\mathbf{u}$ označuje izvor, ki ga rekonstruiramo z originalnega signala, potem preslikavo $\mathbf{x}$ v $\mathbf{u}$ zapišemo kot $\mathbf{u} = \mathbf{W}\mathbf{x}$, kjer je $\mathbf{W} = \mathbf{A}^{-1}$, če $\mathbf{A}$ predstavlja invertibilen linearen sistem. Z analizo neodvisnih komponent izračunamo $\mathbf{W}$ in $\mathbf{b}$ tako, da pogojujemo statisično neodvisnost izvorov

$$f_{\mathbf{b}}(\mathbf{b}) \approx \prod_j f_{b_j}(b_j)\ . \tag{A.6}$$

Če predpodstavimo, da se statistična verjetnostna porazdelitev izvorov bistveno razlikuje od Gaussove porazdelitve, potem lahko za izračun izvorov uporabimo optimizacijske algoritme, ki ocenjujejo deviacijo porazdelitve od idealne Gaussove porazdelitve. Izračunana matrika $\mathbf{W}$ predstavlja slovar filtrov.

Izračun ICA lahko poenostavimo, če v podatkih odpravimo linearne odvisnosti ter zmanjšamo dimenzionalnost problema. Za izračun uporabimo algoritem FastICA [58]. Lokalne regije K_i dodelimo razredu ${}^{w}\mathsf{K}_i = m$, ki je določen z indeksom filtra $\mathbf{w}_m$ z najvišjo absolutno vrednostjo odziva.

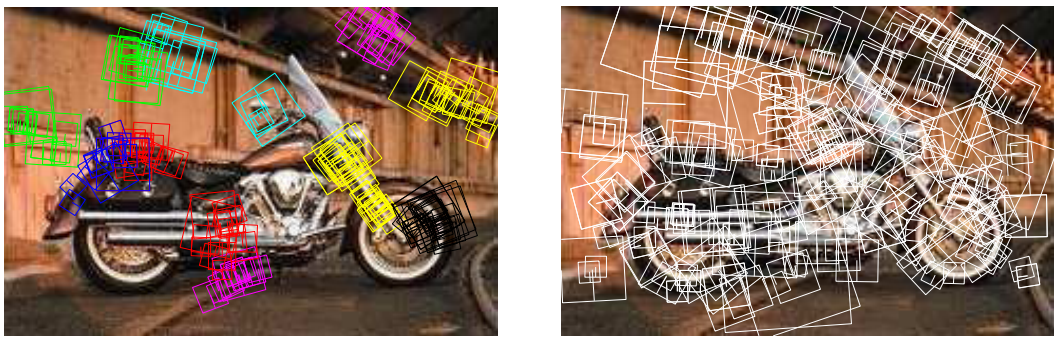
Gručenje *Gestalt* Lokalne značilke, ki so osredotočene na robnih točkah, se tipično pojavljajo v redundantnih, gostih gručah. Da bi zmanjšali redundantnost nizkonivojske predstavitve, smo implementirali postopek gručenja robnih regij, ki na podlagi principov *Gestalt* združi značilke v enotne regije, ki predstavljajo linearne strukture, ali pa področja enakomerne ukrivljenosti. Pri zasnovi algoritma smo uporabili tri principe *Gestalt*:



Slika A.3: Slovarji ICA - Slovarji ICA, naučeni na kategorijah *Motorbikes*, *Faces*, *Laptop*, in *Scissors* [2].

- *Bližina*: Gručijo se le značilke v omejenem območju R.
- *Podobnost*: Gručijo se le značilke, ki pripadajo istemu razredu.
- *Skupna usoda*: Gručijo se le značilke, katerih smer tvorijo kongruenten kot α radianov.

Algoritem A.1 izvedemo za vrednosti $\alpha \in \{0, 0.2, 0.4\}$ radianov in $\varepsilon = 0.1$. Slika A.4 prikazuje rezultat gručenja in pridobljene značilke. Na osnovi gručenja so nove značilke H^0 izpeljane iz lastnosti gruč K tako, da podedujejo razred ICA odziva ${}^w\mathsf{H}^0 = {}^w\mathsf{K}_i$, smer in velikost pa se izračunata iz geometrijskih lastnosti gruče.



(a) **(b)**

Slika A.4: *Gestalt* značilke H^0 - Značilke K_i gručene po algoritmu 6.1. (a) Nekaj gruč značilk K. (b) Vse značilke H^0.

Algoritem A.1: Gručenje *Gestalt*

Input: K, α
Output: H^0

repeat
  Izberi $\mathsf{K}_p : \mathsf{K}_p \notin \mathsf{C}$;
  repeat
    Poišči $\mathsf{K}_j \in \mathit{sosednost}(\mathsf{K}_p) : ({}^{w}\mathsf{K}_p = {}^{w}\mathsf{K}_j) \wedge (\mathsf{K}_j \notin \mathsf{C})$;
    Gruči $(\mathsf{K}_p, \mathsf{K}_j)$ ki ustrezajo $\mathit{abs}({}^{o}\mathsf{K}_p - {}^{o}\mathsf{K}_j) \leq \alpha + \varepsilon$;
    Nov pivot naj bo K_{n_0}, $m_0 = \arg\max_n ||\overline{{}^{c}\mathsf{K}_p {}^{c}\mathsf{K}_n}||_2$;
  until *število značilk v gruči dovolj veliko* ;
  Pridruži vse gručene v $\mathsf{C}_{\{i\}}$;
until *vsi* K_i *obiskani* ;
Iz vsake $\mathsf{C}_{\{i\}}$ izpelji značilko H_i^0;

Hierarhično združevanje Združevanje značilk poteka preko več hierarhičnih nivojev. Opisniki na H^0 predstavljajo le zelo skopo informacijo o regiji, njeni smeri in velikosti. Z združevanjem značilk pridobivajo opisniki dodatno informacijo o relativni geometrijski konfiguraciji posameznih značilk. Ker so geometrijske relacije podane relativno in na lokalni ravni, je način opisa neodvisen od globalnega koordinatnega sistema, kar omogoča popolno paralelizacijo procesov združevanja. Proces združevanja se prične s korakom $\mathsf{H}^0 \rightarrow \mathsf{H}^1$, in se v principu lahko izvede za poljubno število nivojev. Na vsakem nivoju se n značilk v lokalnem območju združi v n-terico. Ker združevanje večjega števila značilk hitro privede do prevelike informativnosti značilk, hierarhija združuje značilke le v pare.

Pare značilk opišemo z geometričnimi atributi, ki opisujejo relativno geometrično konfiguracijo para (Slika A.5). Uporabimo pet atributov:

1. *Urejenost*: značilke H_i^k in H_j^k uredimo, da velja

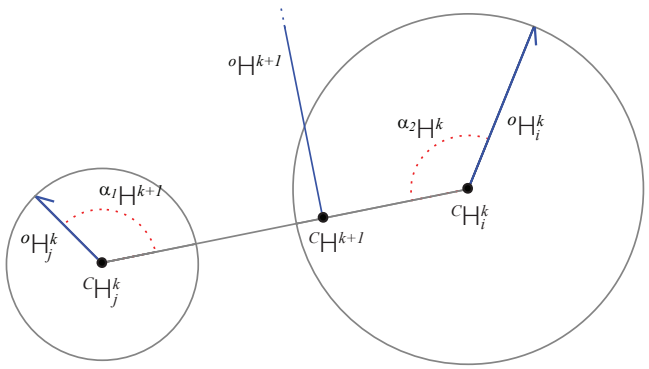
$$ {}^{r}\mathsf{H}_i^k > {}^{r}\mathsf{H}_j^k $$

2. *Velikost*:

$$ {}^{s}\mathsf{H}^{k+1} = log\left(1 + \frac{{}^{r}\mathsf{H}_i^k}{{}^{r}\mathsf{H}_j^k} \right) $$

3. *Razdalja*:

$$ {}^{l}\mathsf{H}^{k+1} = log\left(1 + \frac{||\overline{{}^{c}\mathsf{H}_i^k {}^{c}\mathsf{H}_j^k}||_2}{{}^{r}\mathsf{H}_i^k} \right) $$



Slika A.5: Geometrijski atributi para značilk - Geometrijski atributi para značilk pri združevanju H^k v H^{k+1}.

4. *Notranji kot*: $^{\alpha_1}\mathsf{H}^{k+1}$ in $^{\alpha_2}\mathsf{H}^{k+1}$ med $^{o}\mathsf{H}_i^k$ in $^{o}\mathsf{H}_j^k$ ter $\overline{^{C}\mathsf{H}_i^k\ {}^{C}\mathsf{H}_j^k}$.

5. *Kvadranta notranjih kotov*: $^{or}\mathsf{H}^{k+1} \in \{0,1,2,3\}$ glede na medsebojno lego $^{\alpha_1}\mathsf{H}^{k+1}$ in $^{\alpha_2}\mathsf{H}^{k+1}$.

Geometrijske atribute nato predstavimo v diskretiziranem konceptualnem prostoru. Diskretizacijo izpeljemo na podlagi naučenih unimodalnih distribucij, katere se naučimo za vsak nivo posebej. Geometrični atributi para H_i^k se tako zapišejo z diskretnim opisnikom

$$^{G}\mathsf{H}_i^k \quad = \quad \left[\, \mathsf{Q}_s(\,^{s}\mathsf{H}_i^k),\ \mathsf{Q}_l(\,^{l}\mathsf{H}_i^k),\ \mathsf{Q}_\alpha(\,^{\alpha_1}\mathsf{H}_i^k),\ \mathsf{Q}_\alpha(\,^{\alpha_2}\mathsf{H}_i^k),\ \,^{or}\mathsf{H}_i^k \,\right] \ , \tag{A.7}$$

kjer $\mathsf{Q}(\,\bullet\,)$ predstavlja diskretizacijo. Opisnik $^{G}\mathsf{H} =$`[23310]` naprimer predstavlja razmerje, kjer razmerje velikosti značilk pripada razredu `[2]`, razdalja razredu `[3]`, koti razredoma `[3]` in `[1]`, `[0]` pa opisuje lego notranjih kotov.

Naj $\mathsf{H}^k \rightarrow \mathsf{H}^{k+1}$ označuje nivoje združevanja. Značilke na vsakem od nivojev lahko opišemo z diskretnim zaporedjem, ki popolnoma opisuje podrejeni značilki ter geometrijska razmerja med njima:

$$^{D}\mathsf{H}_i^0 \quad = \quad ^{w}\mathsf{H}_i^0 \tag{A.8}$$

$$^{D}\mathsf{H}_m^{k+1} \quad = \quad ^{G}\mathsf{H}_m^{k+1} \,||\, ^{D}\mathsf{H}_i^k \,||\, ^{D}\mathsf{H}_j^k;\ k>0; \tag{A.9}$$

Sinhrono združevanje, iskanje ujemanj in inhibicija Naj bodo $\widehat{\mathsf{H}}^k$ značilke prototipa, H^k pa značilke interpretirane slike. Naj bo F^k indeks ujemanj

$$\mathsf{F}^k = \left\{ \left(\widehat{\mathsf{H}}_i^k, \mathsf{H}_j^k \right) :\ \ ^{D}\widehat{\mathsf{H}}_i^k = \,^{D}\mathsf{H}_j^k \right\} \ .$$

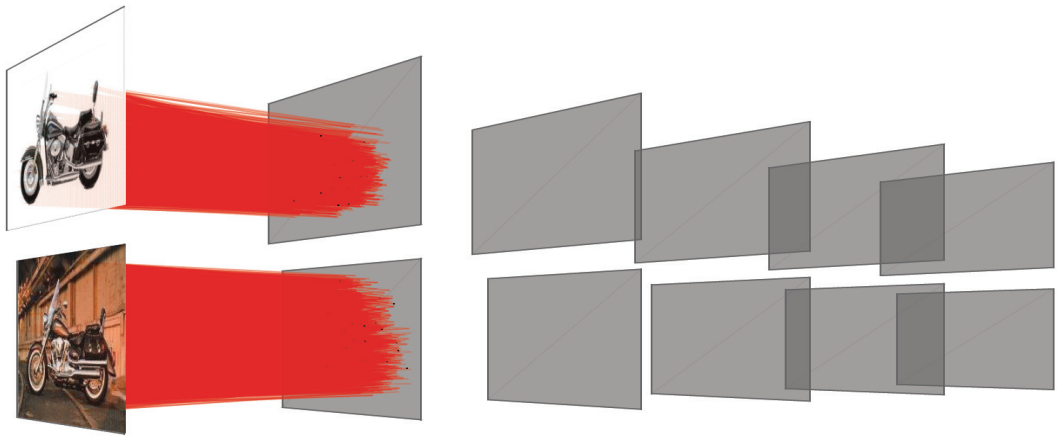
Za vsak par iz množice ujemanj F^k aktiviramo lokalno soseščino $\widehat{\mathsf{H}}_i^k$, v kateri poiščemo $fanOut$ kandidatov. Obliko soseščine nato preslikamo na nivo k interpretirane slike, kjer prav tako poiščemo in združimo kandidate za ujemanja. Lokalna ujemanja na H^{k+1} ohranimo, ostale značilke pa zavržemo.

Algoritem A.2 opisuje postopek lokalnega združevanja, iskanja ujemanj ter inhibicije. Izhod iz algoritma so značilke $\widehat{\mathsf{H}}^{k+1}$, H^{k+1}, ter indeks ujemanj F^{k+1}. Slika A.6 ponazarja združene značilke po prvem koraku $\mathsf{H}^0 \rightarrow \mathsf{H}^1$. Slika A.7 pa ponazarja končni rezultat, če predpodstavljamo pet nivojev združevanja.

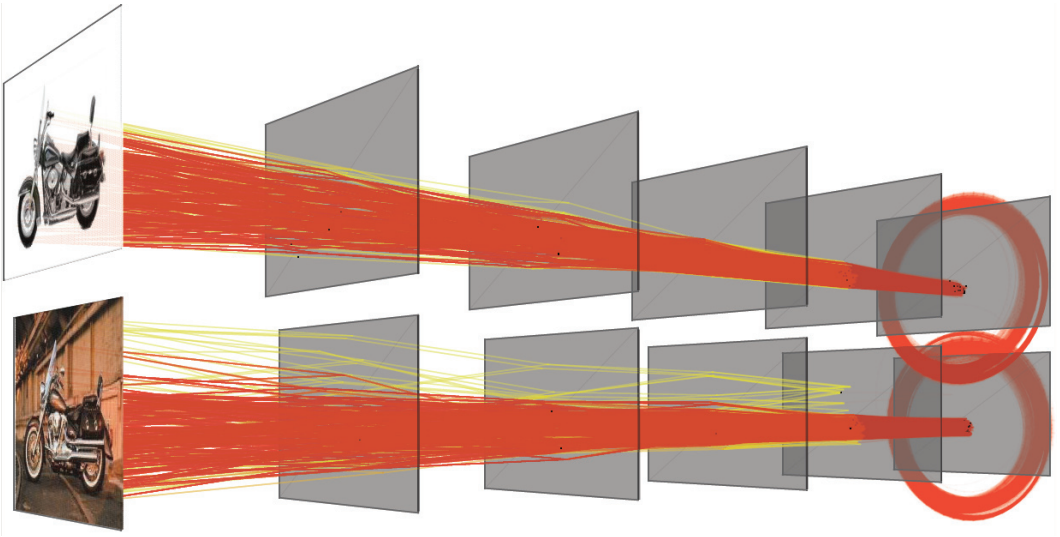
Algoritem A.2: Hierarhično združevanje, iskanje ujemanj in inhibicija

Input: $\widehat{\mathsf{H}}^k$, H^k, F^k, $inner$, $outer$, $fanOut$
Output: $\widehat{\mathsf{H}}^{k+1}$, H^{k+1}, F^{k+1}

foreach $\mathsf{F}_u^k = \left\{ \left(\widehat{\mathsf{H}}_i^k, \mathsf{H}_j^k \right) \right\}$ **do**
  $rMin = {}^r\widehat{\mathsf{H}}_i^k * inner$;
  $rMax = {}^r\widehat{\mathsf{H}}_i^k * outer$;
  $\widehat{\mathsf{A}} = \left\{ \widehat{\mathsf{H}}_l^k : \ {}^c\widehat{\mathsf{H}}_l^k \in \widehat{aField}(rMin, rMax) \right\}$;
  while $card\left(\widehat{\mathsf{A}}\right) < fanOut \ \wedge \ rMin > 0$ **do**
    if *soseščina znotraj slike* **then**
      $rMax = rMax + eps$;
    else
      $rMin = rMin - eps$;
    $update\left(\widehat{\mathsf{A}}\right)$;
  $\widehat{\mathsf{C}}^{k+1} = bind\left(\widehat{\mathsf{A}}, \widehat{\mathsf{H}}_i^k\right)$;
  $\widehat{aField} = convexHull\left(\widehat{\mathsf{A}}\right)$;
  preslikaj $\widehat{aField}$ v $aField$;
  $\mathsf{A} = \left\{ \mathsf{H}_m^k : \ \left({}^C\mathsf{H}_m^k \in aField\right) \ \wedge \ \left(\exists\, \widehat{\mathsf{H}}_n^k \in \widehat{\mathsf{A}} : \ \left(\widehat{\mathsf{H}}_n^k, \mathsf{H}_m^k \right) \in \mathsf{F}^k \right) \right\}$;
  $\mathsf{C}^{k+1} = bind\left(\mathsf{A}, \mathsf{H}_j^k\right)$;
  $\mathsf{F}^{k+1} = \mathsf{F}^{k+1} \cup match\left(\widehat{\mathsf{C}}, \mathsf{C}\right)$;
  $\widehat{\mathsf{H}}^{k+1} = \widehat{\mathsf{H}}^{k+1} \cup \left\{ \widehat{\mathsf{C}}_m^{k+1} : \ \widehat{\mathsf{C}}_m^{k+1} \in \mathsf{F}^{k+1} \right\}$;
  $\mathsf{H}^{k+1} = \mathsf{H}^{k+1} \cup \left\{ \mathsf{C}_m^{k+1} : \ \mathsf{C}_m^{k+1} \in \mathsf{F}^{k+1} \right\}$;



Slika A.6: Sinhrono združevanje $H^0 \rightarrow H^1$ - V zgornji vrstici so predstavljeni uspešno združeni pari $\widehat{H}^0 \rightarrow \widehat{H}^1$ značilk prototipa. Spodnja vrstica prikazuje uspešno združene pare $H^0 \rightarrow H^1$ v interpretirani sliki. Povezave v rdeči barvi predstavljajo aktivne pare.



Slika A.7: Sinhrono združevanje $H^4 \rightarrow H^5$ - V zgornji vrstici so predstavljeni uspešno združeni pari $\widehat{H}^4 \rightarrow \widehat{H}^5$ značilk prototipa. Spodnja vrstica prikazuje uspešno združene pare $H^4 \rightarrow H^5$ v interpretirani sliki. Povezave v rdeči barvi predstavljajo aktivne pare. Povezave v rumeni barvi predstavljajo inhibirane pare. Na končnem nivoju so prikazane tudi lokalne okolice značilk. Na nivoju H^5 je uspešno konstruiranih približno 2500 značilk.

Eksperimentalni rezultati Model smo preizkusili na problemu kategorizacije segmentiranih objektov v zbirki *ETH80*, ter kategorizacije in lokalizacije objektov na nesegmentiranih slikah objektov različnih kategorij *Caltech-101*. Rezultati dokazujejo uporabnost metode za kategorizacijo objektov v kanoničnem pogledu. Poleg same kategorizacije lahko metodo uporabimo za detekcijo potencialnih ujemanj med objekti na več nivojih podrobnosti.

A.3 Zaključek

Eksperimentalni rezultati so pokazali, da hierarhično iskanje ujemanj s prototipom uspešno poišče podobnosti med prototipom in pripadnikom kategorije (v eksperimentih je ta kategorija definirana s strani eksperimentatorja), ter da je število konstruiranih značilk na končnem nivoju lahko osnova za oceno raznovrstnih podobnosti med predmetoma, ter s tem tudi osnova za oceno pripadnosti isti kategoriji. Bistvega pomena je pri tem tudi možnost iskanja ujemanj na nižjih nivojih, vse do začetnega nivoja, saj to omogoča asociacijo raznovrstnih delov predmeta, ki sicer nimajo neposredne semantične vrednosti. Klasične metode za kategorizacijo objektov lahko globalno kategorizacijo pojasnijo le kot konfiguracijo prednaučenih delov, ali pa kot večje število ujemanj med seboj nepovezanih značilk.

Pomemben prispevek dela je tudi hierarhično pragmatično povezovanje značilk kot oblika hevrističnega reševanja problema iskanja podobnosti v strukturi. Povezovanje deluje nad lokalnih sošeščinah, kjer se med procesi v prototipu in procesi v interpretirani sliki sinhronizirajo preko vzpostavljanja oblike lokalnih soseščin, kar predstavlja nizkonivojski model pozornostnih mehanizmov. Neodvisnost lokalnih procesov omogoča masivno paralelizacijo algoritma na poljubnem številu procesnih enot.

Ker kategorije niso opredeljene z zbiranjem in učenjem vidnih vtisov, temveč z učenjem na podlagi enega samega prototipa, in ker je predstavitev v obliki slikovne reprezentacije objekta v kanoničnem pogledu bliže neposrednim kvalitetam objekta, kot je to v primeru abstrahiranih naučenih predstavitev, lahko vzpostavimo zanimive vzporednice s teorijami dinamičnega in vzpostavitvenega zaznavanja. Kategorija ni opredeljena kod odgovor na vprašane *Kaj*, temveč kot navodilo *Kako* lahko primerjamo objekt s prototipom. Prototipna predstavitev je tudi bolj naravna, saj lahko prepoznamo dva videna objekta kot podobna, tudi če njune kategorije ne poznamo.

Novost je tudi izračun slovarja kot nabora filtrov v obliki neodvisnih komponent. Razpršeni

slovarji so bili doslej učeni le na naključnih naborih zaplat, pri čemer so pridobljeni slovarji modelirali pred–pozornostni nivo procesiranja. Z učenjem slovarja filtrov na področjih, izbranih s pozornostnim mehanizmom, ki določa regije interesa ter jim določi lastno velikost in usmerjenost. Lokalno računanje neodvisnih komponent ICA se bistveno razlikuje od podobnih pristopov [66], saj omogoča maksimalno razpršene odzive na slovar, kar manjša nedoločenost klasifikacije odzivov.

Poglavitne pomanjkljivosti metode so izrazito haluciniranje ter propagiranje napak skozi nivoje hierarhije. Halucinacija se pojavlja predvsem pri iskanju ujemanj enostavnih predmetov na kompleksnem ozadju, kjer sistem lahko vzpostavi veliko število napačnih hipotez. Problem nameravamo odpraviti z dognanim načrtovanjem zgodnjih nivojev procesiranja, predvsem pa z vključitvijo različnih detektorjev stabilnih regij, ter kombiniranjem le-teh. Implementacije kognitivnih sistemov uporabljajo večmodalno senzorično informacijo in omogočajo vzpostavitev kontekstnega okvira, kar bistveno olajša odpravo napačnih pozitivnih detekcij. Propagiranje napake skozi nivoje hierarhije je predvsem posledica diskretizacije geometričnega konceptualnega prostora. Problem nameravamo rešiti z redundantno diskretizacijo ter z vključitvijo cene združevanja, ki bo kaznovala mejne vrednosti geometričnih lastnosti.

Problem predstavitve s prototipom v kanoničnem pogledu je predvsem nezmožnost prepoznave v pogledih, ki se bistveno razlikujejo od kanoničnega pogleda. Način predstavitve različnih pogledov je znan problem kognitivnih predstavitev. Vendar pa trdimo, da je v kontekstu utelešenih kognitivnih sistemov manjše število kanoničnih pogledov primerna predstavitev, saj se nedoločenost kategorizacije lahko zmanjša z aktivnim raziskovanjem. Kljub temu pa je potrebno prototip učinkovito prepoznati pri manjših odstopanjih od kanoničnega gledišča. Občutljivost na spremembo gledišča nameravamo temeljito analizirati in izboljšati. Problem simultane prepoznave večih kategorij nameravamo rešiti z umestitvijo v kontekstno informirano okolje, ki omogoča pogojevanje pričakovanih kategorij objektov za dano situacijo. Obetavna je tudi možnost simultanega sinhronega iskanja ujemanj z več prototipi hkrati. Raziskati nameravamo tudi možnost hierarhične predstavitve prototipov, kjer bi slike najprej primerjali z enostavnimi oblikami, nato pa z vedno bolj kompleksnimi prototipi.

Metodo nameravamo dodatno razširiti tako, da bo lahko uporabljala 3-D modele kot prototipe. Značilk, ki bodo primarno inicializirane na celotni sferi, bodo skozi proces hierarhičnega povezovanja, iskanja ujemanj in inhibicije usmerile procesiranje na obetavne smeri pogleda.