2 Category Representation and Recognition Evolvement

We here list two more aspects of the recognition process, the aspect of structural variability independence and the aspect of viewpoint independence (Palmer, 1999). With these two aspects in mind, we characterize previous and current vision systems and it will allow us to better outline the systematics of our approach.

2.1 Structural Variability Independence

We have already touched the aspect of structural variability independence in the previous chapter. Here we take a refined look at it. Figure 4 shows different instances of the category 'chair', with the goal to point out the structural variability existent within a category. We intuitively classify the variability into three types:

a) *Part-shape variability*: the different parts of a chair - leg, seat and back-rest - can be of varying geometry. The legs' shape for example can be cylindrical, conic or cuboid, sometimes they are even slightly bent. The seating shape can be round or square like or of any other shape, so can the back-rest (compare chairs in figure 4a).

b) *Part-alignment variability*: the exact alignment of parts can differ: the legs can be askew, as well as the back-rest for more relaxed sitting (top chair in figure 4b). The legs can be exactly aligned with the corners of the seating area, or they can meet underneath it. Similar, the back-rest can align with the seating area exactly or it can align within the seating width (bottom chair in figure 4b).

c) *Part redundancy*: there are sometimes parts that are not necessary for categorization, as for example the arrest or the stability support for the legs (figure 4c). Omitting these parts does still lead to proper categorization.

Despite this variability, the visual system is able to categorize these instances: the process operates independent of structural variability. A chair representation in the visual system may therefore not depend on exact part shapes or exact alignments of parts. It may neither contain any structures that are not absolutely necessary for categorization. The corresponding category representation would therefore be something very loose and flexible. The degree of looseness would depend on the degree of variability found in a category. For example, the category chair certainly requires a larger degree of looseness than the category book or ball.

2.2 Viewpoint Independence

Another aspect of recognition is its viewpoint independence. We are able to recognize an object from different viewpoints despite the dif-

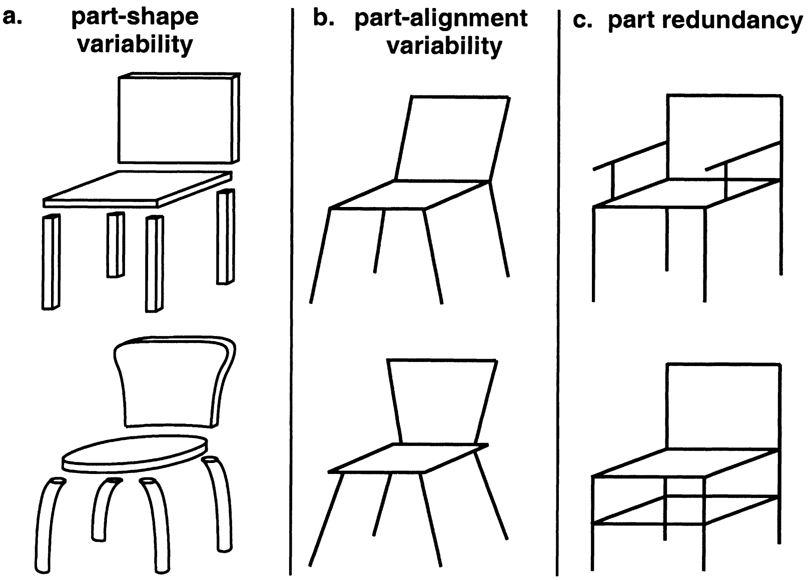

Figure 4: Intuitive classification of structural variability in the category chair. a. Part-shape variability. b. Part-alignment variability. c. Part redundancy. The category representation must be something loose and flexible.

ferent 2D appearance of the object's structure for any given viewpoint. The viewpoints of an object can be roughly divided into *canonical* and *non-canonical* (Palmer et al., 1981). Canonical viewpoints exhibit the object's typical parts and its relations, like the chairs seen in the left of figure 5. In contrast, non-canonical viewpoints exhibit only a fraction of the object's typical parts or show the object in unexpected poses, and are less familiar to the human observer, like the chairs seen in the right of figure 5.

In our daily lives we see many objects primarily from canonical viewpoints, because the objects happen to be in certain poses: chairs are generally seen on their legs or cars are generally on their wheels. Canonical viewpoints can certainly be recognized within a single glance (Potter, 1975; Thorpe et al., 1996; Schendan et al., 1998). In contrast, non-canonical viewpoints are rare and one can assume that the recognition of non-canonical viewpoints requires more processing time than a single glance. Recognizing a non-canonical viewpoint may consist of a short visual search using a few saccades, during which textural details are explored; or the perceived structure of the object is transformed in some way to find the appropriate category (Farah, 2000). Behavioral evidence from a person with visual agnosia suggests that non-canonical views are indeed something unusual (Humphreys and Riddoch, 1987a). The person is able to recognize objects in daily live without problems, yet struggles to comprehend non-canonical views of objects given in a picture. This type of visual disorder was termed perceptual categorization deficit, but pertains to the categorization of unusual (non-canonical) views only. One may conclude from this case, as Farah does, that such views do not represent any real-world visual tasks.

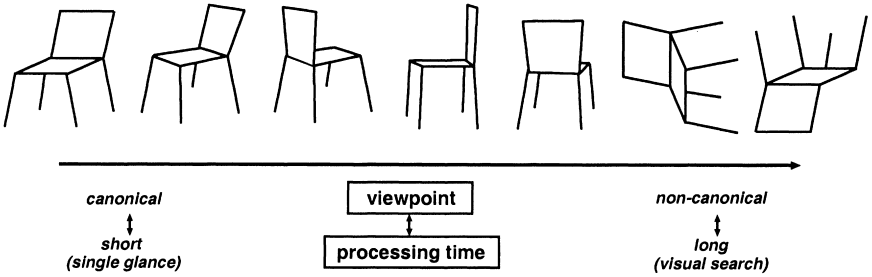

Figure 5: Different poses and hence viewpoints of a chair. Viewpoint and processing time are correlated. Left: canonical views that are quickly recognizable. Right: non-canonical views that take longer to comprehend, possibly including a saccadic visual search.

2.3 Representation and Evolvement

We now approach the heart of the matter: how are we supposed to represent categories? Ideally, the design of a visual system starts by defining the nature of *representation* of the object or category, for example the object is described by a set of 3D coordinates or a list of 2D features. This representation is sometimes also called the object model. In a second step, after defining the representation, a suitable reconstruction method is contrived that extracts crucial information from the image, which in turn enables the corresponding category. One may call this object *reconstruction* or *evolvement*. Such approaches were primarily developed from the 60's to the 80's, but are generally not extendable into real-world objects and gray-scale images. Recent approaches have taken a heuristic approach, in which the exact representation and evolvement is found by testing.

Most of these systems - whether fully designed or heuristically developed - start with some sort of contour extraction as the first step, followed by classifying contours and relating them to each other in some way to form higher features, followed by finding the appropriate category. We here review mainly Artificial Intelligence (computer vision) approaches and some psychological approaches . Neural network approaches are mentioned in the next chapter.

2.3.1 Identification Systems

Early object recognition systems aimed at identifying simple building blocks from different viewpoints. Because they intended to do that precisely, the object model was defined as a set of corner points specified in a 3D coordinate system. Robert devised such a system performing this task in roughly three steps (figure 6,(Robert, 1965)): Firstly, contours were extracted and 2D features formed. Secondly, these extracted 2D features were matched against a set of stored 2D features that would point towards a specific object. Finally, each of those object models, whose 2D features were successfully matched in the second step, were matched against the contours, determining so the object identity. With the identified object model it is possible to find the object's exact pose in 3D space.

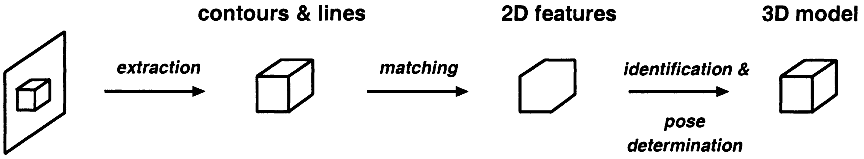

Figure 6: Roberts identification and pose determination system. The object was represented as a set of 3D coordinates representing the corners of a building block. Recognition evolved firstly by extracting contours and lines, followed by a matching process with stored 2D features, followed by eventual matching some of the possible models against the image.

Many later systems have applied this identification and pose determination task to more complex objects using variants, elaborations and refinements of Roberts' scheme (Brooks, 1981; Lowe, 1987; ULLMAN, 1990; Grimson, 1990). Some of them are constructed to serve as vision systems for part assembly in industry performed by roboters. Some of them are able to deal with highly cluttered scenes, in which the object identity is literally hidden in a mixture of lines. These systems do so well with this task, they may even surpass the performance of an untrained human 'eye'.

All of these systems work on the identity level (figure 3, chapter 1). They do not categorize and therefore do not deal with structural variability and have in some sense 'only' dealt with the viewpoint independence aspect. They have been applied in a well defined environment with a limited number of objects. The real world however contains an almost infinite number of different objects, which can be categorized into different levels. The structural variability that one then faces therefore demands different object representations and possibly a different recognition evolvement.

The construction of such pose-determining systems may have also influenced some psychological research on object recognition, which attempts to uncover that humans recognize objects from different viewpoints by performing a similar transformational process as these computer vision systems do (e.g. see (Tarr and Bulthoff, 1998; Edelman, 1999) for a review).

2.3.2 Part-based Descriptions

Part-based approaches attempt to describe objects by a set of forms or 'parts', arranged in a certain configuration: it is also called a structural description approach (figure 7).

Guzman suggested a description by 2D features (Guzman, 1971). In his examples, an object is described by individual shapes: For example, a human body is described by a shape for the hand, a shape for the leg, a shape for the foot and so on. These shapes were specified only in two dimensions. Figure 7 shows a leg made of a shape for the leg itself and a shape for a shoe. Guzman did not specifically discuss the aspect of structural variability independence, but considered that objects can have deformations like bumps or distortions and that despite such deformations the visual system is still able to recognize the object correctly. In order to be able to cope with such deformations, he proposed that representations must be sort of 'sloppy'. This aspect of 'deformation independence' is actually not so different from the aspect of structural variability independence.

Binford came up with a system that measures the depth of a scene by means of a laser-scanning device (Binford, 1971). His objects were primarily expressed as a single 3D volume termed 'generalized cones', which were individual to the object. For example the body of a snake is described as one long cone (Agin and BINFORD, 1976). Reconstruction would occur by firstly extracting contours, followed by determining the axis of the cones using a series of closely spaced cone intersections. The example in figure 7 shows a snake, which is represented by a single, winding cone. Binford did not specifically address the structural variability aspect.

Binford's system likely influenced Marr's approach to represent animal bodies by cylinders (Marr and Nishihara, 1978). The human body for example would be represented as shown in figure 7. Similar to Binford, Marr planned to reconstruct the cylinders by finding their axes: firstly, surfaces of objects are reconstructed using multiple cues like edges, luminance, stereopsis, texture gradients and motion, yielding the 2.5D 'primal sketch' (Marr, 1982); secondly, the axis would be reconstructed and put together to form the objects. Marr did not specifically address the aspect of structural variability either, but cylinders as part representations would indeed swallow a substantial amount of structural variability. The idea to reconstruct surfaces as a first step in recognition was emphasized by Gibson (e.g. (Gibson, 1950)).

Pentland described natural objects like trees with superquadrics like diamonds and pyramidal shapes (Pentland, 1986) (not shown in figure 7).

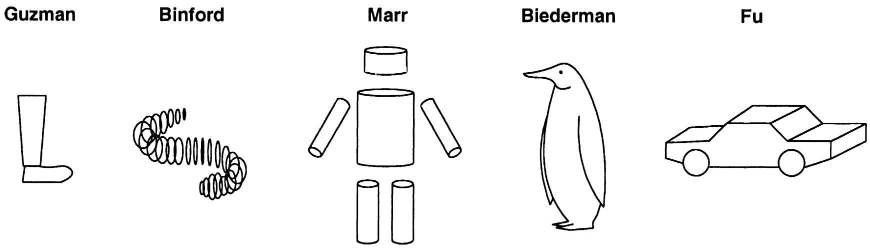

Figure 7: Object representations by parts. Guzman: individual 2D shapes. Binford: 'generalized cones'. Marr: cylinders. Biederman: geons. Fu: surfaces. Loosely redrawn from corresponding references given in text.

Fueled by the idea of a representation by 3D volumes, Biederman proposed an even larger set of 'parts' for representation, like cylinders, cuboids and wedges, 36 in total, which he called 'geons' (Biederman, 1987). The example in figure 7 shows a penguin made of 9 different such geons. To account for the structural variability, Biederman suggested that a category representation may contain interchangeable geons for certain parts. This may however run into a combinatorial explosion for certain categories, especially the ones with a high structural variability. The evolvement of the geons and objects would start with finding firstly vertex features.

These part-based approaches have never really been successfully applied to a large body of gray-scale images. One reason is, that it is computationally very expensive to extract the volumetric information of each single object part. Another reason is that the contour information is often fragmentary in gray-scale images and that this incomplete contour information does not give enough hints about the shape of 3D parts, although Marr tried hard to obtain a complete contour image (Marr, 1982). Instead of this 3D reconstruction, it is cheaper and easier to interpret merely 2D contours, as Guzman proposed it. Fu has done that using a car as an example (figure 7): the parallelograms that a car projects onto a 2D plane, can be interpreted as a surface (Lee and Fu, 1983). Still, an extension to other objects could not be worked out.

Furthermore, in most of these part-based approaches, the representations are somewhat chosen according to human interpretation of objects, meaning a part of the recognition system corresponds to a part in a real object, in particular in Guzman's, Marr's and Biederman's approach. But these types of parts may be rather a component of the semantic representation of objects (figure 2, right side). As we pointed out already, the perceptual representations we look for, do not need to be that elaborate (figure 2, left side). Nor do they need to rely on parts.

2.3.3 Template Matching

In a template matching approach, objects are stored as a 2D template and directly matched against the (2D) visual image. These approaches are primarily developed for detection of objects in gray-scale images, e.g. finding a face in a social scene or detecting a car in a street scene. Early attempts tried to carry out such detection tasks employing only a 2D luminance distribution, which was highly characteristic to the category. To find the object's location, the template is slid across the entire image. To match the template to the size of the object in the image, the template is scaled. Because this sliding and scaling is a computationally intensive search procedure, the developers of such systems spend considerable effort in finding clever search strategies.

Recent attempts are getting more sophisticated in their representations (Amit, 2002; Burl et al., 2001). Instead of using only the luminance distribution per se, the distribution is nowadays tendentially characterized by determining its local gradients, the differential of neighboring values. This gradient profile enables a more flexible matching. Such a vision system would thus run first a gradient detection algorithm and the resulting scene gradient-profile (or landscape) is then searched by the object templates. In addition, an object is often represented as a set of sub-templates representing significant 'parts' of objects. For instance, a face is represented by templates for the eyes and a template for the mouth. In some sense, these approaches move toward more flexible representations in order to be able to cope with the structural variability existent in categories. These systems can also perform very well, when the image resolution is low. In comparison, in such low resolution cases, a human would probably recognize the object rather with help of contextual information, that means that neighboring objects facilitate the detection of the searched object. Such contextual guidance can take place with frames.

2.3.4 Scene Recognition

The first scene recognition systems dealt with the analysis of building blocks like cuboids and wedges depicted in line drawings, so-called polyhedral scenes. Guzman developed a program that was able to segment a polyhedral scene into its building blocks (Guzman, 1969).

His study trailed a host of other studies refining and discussing this type of scene analysis (Clowes, 1971; Huffman, 1971; Waltz, 1975). The goal of such studies was to determine a general set of algorithms and rules that would effectively analyze a scene. However the explored algorithms and representations are difficult to apply to scenes and objects in the real world because their structure is much more variable.

Modern scene recognition attempts aim at the analysis of street scenes depicted in gray-scale images. A number of groups tries to form representations for objects made of simple features like lines and curves, and of a large set of rules connecting them (e.g. (Draper et al., 1996)). Evolvement would occur by a set of control feedback loops, searching for the correct match. These groups have faced the structural variability aspect and addressed it as follows: when they are confronted with a large variability, they 'sub-categorize' a basic-level category, moving thus toward an increasing number of 'templates'.

Many of these systems intend to recognize objects from gray-scale images that have a relatively low resolution. In these images, objects can appear very blurred and it is very difficult and probably even impossible to perform proper recognition without taking context into account, as the developers realized. The human observer has of course no problem categorizing such images, thanks to the power of frames that can provide rich contextual information. We have more on the subject of scene recognition in chapter 11.

2.4 Recapitulation

We summarize the different approaches with regard to their type of representations - whether they are specified in 2D or 3D - and their method of reconstruction (figure 8).

Some artificial intelligence approaches focused on object representations specified in a 3D dimensional coordinate system and they attempted to reconstruct the constituent 3D parts directly from the image, like Binford's and Marr's approach, as well as Brook's identification system (figure 8a). Roberts' and Lowe's system also represent objects in 3D, but evolvement was more direct by going via 2D features (figure 8b). Scene recognition approaches search for representations using merely simple 2D features and extensive feedback loops for matching (figure 8c). The most direct recognition systems are the template matching systems, which can be roughly labeled as 2D-2D systems (figure 8d). We also assign neural networks (NN) to that category, because many of them aim at a feature matching in some sense (chapter 3, section 3.1). The single arrow should indicate that evolvement is either direct (in case of templates) or continuous (for neural networks). Figure 8e refers to spatial transformations which we will also discuss in chapter 3.

In case of the identification systems, the representation and evolve-

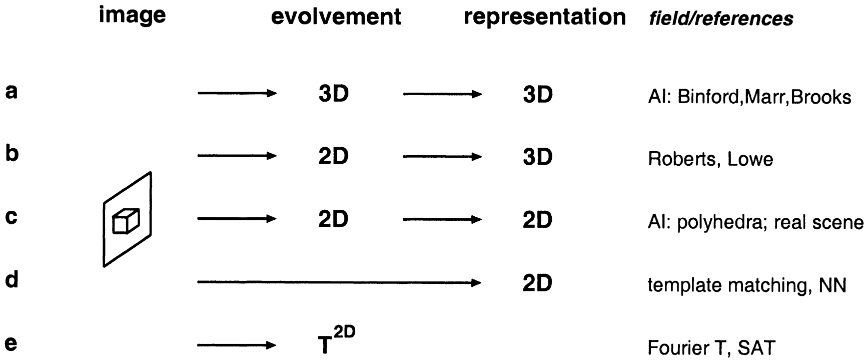
image
evolvement
representation
field/references
a → 3D → 3D AI: Binford,Marr,Brooks
b → 2D → 3D Roberts, Lowe
c → 2D → 2D AI: polyhedra; real scene
d → 2D template matching, NN
e → T^{2D} Fourier T, SAT

Figure 8: Summary of recognition systems, roughly ordered by evolvement strategies and representation type. Top (a): pure 3D approaches, the model as well as reconstruction occurred via 3D volumes. Bottom (f): representation and evolvement involving spatial transformations.

ment was defined beforehand. This worked well because the range of objects was finite and their environment was often well defined. The part-based approach also defined representation and evolvement ahead, but this has not led to general applicable systems. Their type of representations seemed to correspond to a human interpretation of objects and may therefore serve better as a cognitive representation (right side of figure 2). Because successful representations are difficult to define, approaches like template matching and scene recognition employ an exploratory approach.

2.5 Refining the Primary Engineering Goal

Given the large amount of variability, it is difficult to envision a category representation made of a fixed set of rigid features. Our proposal is to view a category representation as a loose structure: the shape of features as well as their relations amongst each other is to be formulated as a *loose skeleton*. The idea of loose representations has already been suggested by others. 1) Ullman has used fragmented template representations to detect objects depicted in photos (Ullman and Sali, 2000). 2) Guzman has also proposed that a representation needs to be loose (Guzman, 1971). He developed this intuition - as mentioned before - by reflecting on how to recognize an object, that has deformations like bumps or distorted parts. He termed the required representation as 'sloppy'. 3) Results from memory research on geographical maps suggests that human (visual) object representations are indeed fragments: Maps seem to be remembered as a collage of different spatial descriptors (Bryant and Tversky, 1999). Geographical maps are instances of the identity level (figure 3): Hence, if even an instance of an identity is represented as a loose collage, then one can assume that basic-level category representations are loose as well, if not even much looser. Loose representations can also provide a certain degree of viewpoint invariance. Because the structural relations are not exactly specified, this looseness that would enable to recognize objects from slightly different viewpoints. We imagine that this looseness is restricted to canonical views only. Non-canonical views likely trigger an alternate recognition evolvement, for instance starting with textural cues.

At this point we are not able to further specify the nature of representations, nor the nature of recognition evolvement. We will do this in our simulation chapters (chapters 5, 7 and 8). Because it is difficult to define a more specific representation and evolvement beforehand, our approach is therefore exploratory like the template and scene recognition systems, but with the primary focus on the basic-level categorization process. The specific goal is to achieve categorization of canonical views. Non-canonical views are not of interest because they are rare (section 2.2). Thus, the effort has to go into finding the neuromorphic networks that are able to deal with the structural variability. Furthermore, this system should firstly be explored using objects which are depicted at a reasonable resolution. Once this 'motor' of vision, the categorization process, has been established, then one would refine it and make it work on low-resolution gray-scale images or extend it to recognition of objects in scenes.