Reconstructing mental object representations: A machine vision approach to human visual recognition

EROL OSMAN\textsuperscript{1}, ADRIAN R. PEARCE\textsuperscript{2}, MARTIN JÜTTNER\textsuperscript{1} and INGO RENTSCHLER\textsuperscript{1}

\textsuperscript{1}Institute of Medical Psychology, University of Munich, Germany
\textsuperscript{2}School of Computing, Curtin University, Perth, Australia

Received 23 August 1999; revised 15 March 2000; accepted 29 March 2000

Abstract—This paper introduces a new approach to assess visual representations underlying the recognition of objects. Human performance is modeled by CLARET, a machine learning and matching system, based on inductive logic programming and graph matching principles. The model is applied to data of a learning experiment addressing the role of prior experience in the ontogenesis of mental object representations. Prior experience was varied in terms of sensory modality, i.e. visual versus haptic versus visuo-haptic. The analysis revealed distinct differences between the representational formats used by subjects with haptic versus those with no prior object experience. These differences suggest that prior haptic exploration stimulates the evolution of object representations which are characterized by an increased differentiation between attribute values and a pronounced structural encoding.

Keywords: Object recognition; representation; recognition-by-parts; graph matching.

1. INTRODUCTION

The question concerning the quality of internal representations underlying human object recognition is still unresolved. Prominent computational theories have postulated, on the one hand, that objects are mentally represented by three-dimensional (3D), object centered, part-based descriptions (Marr and Nishihara, 1978; Biederman, 1987). On the other hand, more recent studies have argued for representations of 3D objects in terms of multiple, viewer centered, two-dimensional (2D) views, among which the visual system interpolates if necessary (Tarr and Pinker, 1989; Poggio and Edelman, 1990; Ullman and Basri, 1991; Bülthoff and Edelman, 1992).

In a comparative study using signal-detection analysis Liu \textit{et al.} (1994) have found that the internal representations humans use for object recognition may be characterized by falling between the extremes of a 2D and 3D format. However,
their analysis did aim to delimit human performance rather than making details of these representations explicit. Here we present a novel method to reconstruct properties of visual representations underlying object recognition, by modeling human performance by a recognition system adopted from computer vision. In contrast to previous approaches (see e.g. Poggio and Edelman, 1990; Bülthoff and Edelman, 1992) the objective here is not just to fit recognition performance in terms of the global error rate but to reproduce the specific pattern of confusions, i.e. classification matrices, between objects in the recognition task at hand.

The machine vision system we are using is CLARET (consolidated learning algorithm using relational evidence theory — Pearce and Caelli, 1999). CLARET is based on inductive logic programming, attribute generalization and graph matching principles employing a recognition-by-parts paradigm. It provides a structural object representation in terms of components and relational rules. Within the field of machine vision CLARET has been successfully applied to classical object-recognition problems such as the interpretation of handwritten characters (Pearce and Caelli, 1999). In the current context, CLARET will serve us as a computational model to analyze the structure of mental object representations concerning the relative importance of different attributes, the degree of differentiation between attribute values and the relational depth of rules.

Using this technique we have investigated in a supervised learning paradigm, how various forms of prior knowledge and the presence of depth information influence learning speed, recognition performance and the ability of spatial generalization. The objects to be learned were composed of spheres, which allowed for an easy and unambiguous segmentation into spatially bounded components. Based on the behavioural data we compare the structure of the underlying mental object representations, as reconstructed by CLARET, and discuss their differences in relation to the various learning conditions.

2. OBJECT LEARNING EXPERIMENT

The learning set consisted of three objects, constructed and displayed on a SGI O2 computer using the Open Inventor software package. Each object was composed of four spheres, with three of them forming an isosceles triangle and the fourth being placed perpendicularly above the center of one of the base spheres (Fig. 1). Object 2 and object 3 were mirror symmetric to each other. 2D views were generated as perspective projections of the objects onto the screen plane of the computer display. For the training views, the viewing sphere was sampled in 60 deg steps; views redundant due to object symmetry were eliminated. This resulted in 22 views in total (6 views for object 1, 8 views for object 2 and object 3). In a similar way, 2D test views were generated by sampling the viewing sphere in 30 deg steps, leading to 83 views (21 for object 1, 31 for both object 2 and 3). Because of the sampling interval, 19 of the test views were old views (views used during training, which were not considered during evaluation of the results), whereas the rest of them were