The Recognition of Sketches as a Test Case for Complex Computational Cognition – Position Paper –

Angela Schwering¹ and Ulf Krumnack² and Helmar Gust² and Kai-Uwe Kühnberger²

¹ University of Münster, Institute for Geoinformatics
angela.schwering@uni-muenster.de
² University of Osnabrück, Institute of Cognitive Science
{kumnack,gust,kkuehnbe}@uos.de

Abstract. In order to enable machines to operate intelligently in their environment, it is important that they do not only collect sensory input about their environment, but also recognize and understand objects. Analogical reasoning is considered fundamental for many complex cognitive processes. In this paper, we present an experiment which gives empirical support of our hypothesis that object recognition and concept formation rely fundamentally on analogical similarities. Similar object sketches with the same structure are recognized faster and more frequently than similar object sketches with different structure. Afterwards, we introduce our analogy-making framework Heuristic-Driven Theory Projection (HDTP) and explain how HDTP can be used for object recognition.

1 Introduction

In order to enable machines to operate intelligently in our world, it is important that they do not only collect sensory input and observe the environment, but also recognize and understand it. The correct classification of perceived objects allows a machine to use its background knowledge about the world to reason on it. This paper concentrates on sketches of objects and investigates techniques how machines can recognize new stimuli and classify them according to their ontological knowledge. We examine how concepts change over time and develop an analogy-based approach for learning and revising conceptual knowledge and for explaining the creation of new and abstract knowledge.

Realizing perception on a machine requires an appropriate language for describing spatial objects and environments. It must be possible to capture the geometry of all elements in a scene and the spatial relations between them. Furthermore, the representational formalism must be adaptable to change representations of the same scene according to the different perceptions in varying contexts. Recognition requires the ability of comparing new stimuli to existing stimuli in the memory. In particular for spatial objects, the structural composition of the defining elements is very important. Analogical mapping is used
to compare two stimuli—a new stimulus and a well-known stimulus—for structural similarities. In a recognition task, the well-known stimulus can be a typical instance of a concept or the specification of a concept from memory.

The model of computational cognition proposed in this paper uses knowledge gained through recognition tasks to learn new and revise old concepts. The two main mechanisms for learning constitute learning via transfer and learning by abstraction [10]. Once a new stimulus is successfully classified, either additional knowledge about the concept can be transferred to the newly classified stimulus, or features observed about the new stimulus can be transferred and integrated in the existing concept description. This additional knowledge leads to a richer and more precise concept description. Moreover, the comparison process aligns analogous elements in both stimuli, i.e. reveals the commonalities of both stimuli at an abstract level. These analogous commonalities describe the essential characteristics defining a concept.

This paper is structured as follows: in Section 2, psychological evidence is provided that structural changes of a visual stimulus do influence object categorization of humans stronger than non-structural changes. Section 3 proposes some ideas for a model of object recognition based on an analogy engine. Section 4 provides a vision how adaptations of representations for analogy-based stimulus recognition can be used for learning new concepts. Section 5 concludes the paper.

2 Object Categorization and Structural Alignment

2.1 The Experiment

A lot of common everyday objects are made up of several, distinct components. The same is true for the kitchen stove depicted by the line drawing in Figure 1. Some components typical for the outward appearance of such a stove have been highlighted in grey color. Obviously, these core elements are spatially related to each other entailing individual relationships between them. It is now possible to describe these relationships in a qualitative manner. Commonly used spatial
relations are topological, directional, or metric relations [1] and may involve other qualities such as symmetry and repetition of elements.

When applying this general idea to the stove in Figure 1, its highlighted components might be regarded as separate regions with certain underlying topological relations. The four hotplates on top could be regarded as four disjoint regions all of which are in turn situated inside Area 1. Underneath, Area 2 contains six disjoint temperature regulators. Similar relationships can be found as to the front handle and the spy window both of which are disjoint and situated within another area (Area 3) on the stove’s foreside. Furthermore, the lateral Area 4 directly meets Area 2, and so forth.

To investigate the role of structured representation in human object recognition, an experiment was set up, in which subjects had to recognize line drawings of different objects. 3 132 line drawings were selected for the experiment. Of these, 72 functioned as filler items, whereas the remaining 60 drawings acted as the so-called ”basic” experimental stimuli. The latter served as a basis for the development of four additional variations, namely two versions of non-structural modifications and two versions of structural modifications (cf. Figure 2). Generally speaking, each experimental condition was conceptualized as a pair of two experimental stimuli, henceforward referred to as item pairs.

Basically, a single experimental trial was composed of a source image stimulus and a subsequent target image stimulus. First, the source stimulus was shown and all subjects were expected to name the object that they thought to have identified in the black and white line drawing by an oral answer. Then, subjects had to press the keyboard’s down-arrow key to call up the target image. In preparation for the imminent stimulus, a fixation cross with a duration of 250 ms was shown in the middle of the monitor prior to the occurrence of the target image. Finally, the target image stimulus appeared for maximally 650 ms. This time, the subjects’ task consisted in deciding as quickly as possible by pressing the ”yes” or ”no” button whether the object they were just seeing was an instance of the same concept as the object they had named in the step before.

Due to the five experimental conditions, we created equally many stimulus lists that counterbalanced item pairs and conditions. Each subject saw 36 filler item pairs, 12 MAT item pairs, 12 NS1 item pairs, 12 NS2 item pairs, 12 S1 item pairs, and 12 S2 item pairs yielding 96 experimental trials in total. Figure 2 specifies the modified versions of the original stimulus.4

The interested reader is referred to [20] for a complete presentation of the experiments.

4 75 native German subjects, 50 females and 25 males, volunteered for the experiment and confirmed normal or corrected normal vision. The vast majority of participants consisted of undergraduate students who were enrolled in Psychology or Cognitive Science at the University of Osnabrück. The mean age was 23.2 years, ranging from age 18 to age 58. The experiment was conceptualized and generated with the aid of the software suite E-Prime 2.0 by Psychology Software Tools Inc.
MAT: The match condition was conceptualized as an item pair with identical source and target images. Solely the 60 basic experimental stimuli served as basis to set up this condition. Furthermore, this condition served as a baseline with respect to the reaction time measurements and required a clear "yes" response from the subjects.

NS1: This condition entailed the movement of significant picture elements. These manipulations were not taken for a structural change since it was made sure that the topological relationships between the manipulated and unaffected picture elements remained untouched. It was anticipated that the subjects would show a high tendency to give a "yes" response.

NS2: This condition entailed the resize of picture elements without moving them to another position. Simple resize was not taken for a structural change as long as the topological relations between the manipulated and other picture elements remained constant. It was anticipated that the subjects would show a high tendency to give a "yes" response.

S1: As for the first structural condition, it exclusively implicated the removal and/or addition of selected picture elements. Adding to or removing significant elements from the overall scene was regarded as a clear structural change. It was decided to accept both a "yes" and a "no" response as "potentially correct".

S2: The second structural condition likewise implied the movement of significant picture elements as with condition NS1. However, this time a structural change was deliberately caused by moving selected elements into another area. Alternatively, this condition involved the resize of desired picture elements as with condition NS2. Both "yes" and "no" were accepted as potentially correct answers.

Fig. 2. The types of stimuli used in the experiment: Match condition, non-structural condition I (NS1), non-structural condition II (NS2), structural condition I (S1), and structural condition II (S2).

2.2 Results

For the goals of this paper, it suffices to find evidence for the assumption that humans would need more time to recognize structurally manipulated objects compared to non-structurally manipulated objects. As a consequence, it was decided to combine both non-structural (NS1 & NS2) as well as the two structural conditions (S1 & S2), essentially because of their strong relatedness. In doing so, the overall number could be reduced to a quantity of three experimental conditions: match condition "MAT"; condition “NSCOM” (the combination of the former two non-structural conditions); condition “SCOM” (the combination of

---

5 A detailed presentation of the results with separate treatment of all conditions can be found in [20].
<table>
<thead>
<tr>
<th>Condition</th>
<th>RT in ms (Std. Dev)</th>
<th>ACC in %</th>
<th>Yes / No Ratio in %</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAT</td>
<td>618 (147)</td>
<td>95.6</td>
<td>—</td>
</tr>
<tr>
<td>NSCOM</td>
<td>708 (182)</td>
<td>—</td>
<td>82.1 / 17.9</td>
</tr>
<tr>
<td>SCOM</td>
<td>752 (200)</td>
<td>—</td>
<td>61.3 / 38.7</td>
</tr>
</tbody>
</table>

Table 1. Descriptive statistics results - analyses by subjects (“Yes” and ”No” responses).

the former two structural conditions). The relevant reaction times per subject were summed up and averaged afterwards. The same holds for the “yes”/“no” response ratios yielding the numbers shown in Table 1.

On that basis, a 1 (source image) × 3 (target image type: MAT, NSCOM, SCOM) factorial analysis of variance (ANOVA) including repeated measures was conducted on the response latencies by subjects and by items. Only data points that were maximally two standard deviations away from their corresponding mean were taken into account to reduce the quantity of outliers in the first place. A confidence interval of 95% was consistently used.

As a result, the main effect for target image type was highly significant in the analysis by subjects (F1) and by items (F2) with $F_1(1.61, 112.56) = 87.51$, $p < .001$ (Huynh-Feldt corrected); $F_2(2, 110) = 69.15$, $p < .001$. Concerning the main effect for list, it was only significant in the analysis by items, $F_1(4, 70) = 1.52$, $p > .72$; $F_2(4, 55) = 7.50$, $p < .001$. By contrast, the two-way interaction between list and target image type was not significant at all with $F_1(8, 138) = 1.21$, $p > .30$; $F_2(8, 108) = 2.00$, $p > .05$.

In support of these initial results, several pairwise comparisons (MAT vs. NSCOM; MAT vs. SCOM; NSCOM vs. SCOM) were carried out. In all these pairwise comparisons, the main effect for target image types was highly significant in the analysis by subjects and by items.

2.3 Discussion

The experiment provides two results that are relevant for the discussion in this paper. First, the relation of “yes”/“no” responses shows that the degree of recognition is significantly higher if the structure of the visual stimulus is not changed (NSCOM), compared to the cases where it is changed (SCOM). This indicates that subjects are more willing to accept an object as belonging to a category, if its relational structure stays intact. Second, the reaction time is shorter in these cases, indicating that the task is cognitively less complex if a structural match of stimuli can be found.

Both results back the claim, that object recognition seems to be based, at least partly, on matching structural representations of the provided stimuli. A cognitive plausible model of object recognition should therefore incorporate such representations and matching mechanisms. In the rest of the paper, we sketch a model for recognizing visual stimuli that is driven by analogical mapping and that furthermore allows to introduce a learning mechanism based on recognition.
3 Analogy-Based Recognition of Visual Stimuli

The model we propose is based on Heuristic-Driven Theory Projection (HDTP), a formal framework to compute analogies. This section gives a brief introduction to HDTP, that will focus on those aspects relevant to the intended application. A more comprehensive description of HDTP and its features can be found in [19].

3.1 Syntactic Basis of HDTP

Classically, an analogy is established between two domains of knowledge, called source and target domain. By discovering corresponding structures in both domains, an analogical relation can be constructed. Such a relation can be used to identify commonalities and differences between the domains. Furthermore, gaps discovered in one domain can be filled by transferring knowledge from the other domain, based on the analogical relation. Such analogical inferences, though possibly incorrect from a logical point of view, can be a basis to explain certain aspects of cognitive phenomena like creativity and learning.

HDTP provides a formal framework to compute analogical relations and inferences, for domains represented in first-order logic. Both, source and target domain, are given by axiomatizations, i.e. finite sets of first-order formulae. The basic idea is to associate pairs of formulae from the domains in a systematic way. HDTP uses anti-unification to identify common patterns in formulae. In anti-unification, two formulae are compared and the most specific generalization subsuming both formulae is identified. As a result, besides the generalized formula a pair of substitutions is computed, that expresses the analogical relation between the two formulae.

This process of generalization by anti-unification can be iteratively applied to formulae of the two axiomatizations. However, it might be the case that for some axiom no good corresponding axiom exists on the other side. Nevertheless, there might still exist a good formula in the theory spanned by the axiomatization, i.e. among the formulae that can be derived from the axioms. In this case, HDTP will try to prove such a formula. This process can be considered as a kind of re-representation [11], since the originally given axiomatization is adapted to match the needs of the analogy considered. As a consequence HDTP does not compute an analogy between two specific axiomatizations, but between the theories spanned by these axiomatizations.

HDTP distinguishes between domain knowledge (facts and laws holding for the source or the target domain) and background knowledge, which is true across domains. The background knowledge is of special importance in the context of re-representation, as it may be used to derive further formulae in the two domains, which then can be used again for generalization.

Uncovered parts of the source and the target domain, i.e. formulae that are not part of the analogical relation and therefore cannot be derived from the generalized formulae, are candidates for analogical transfer. The established analogical relation is used to translate these formulae. If the result does not lead
to a contradiction in the other domain, it can be considered as an analogical inference, i.e. new knowledge that might be added to the axiomatization of that domain.

### 3.2 A Formal Language to Represent Spatial Objects

We now apply the ideas of HDTP to the processing and recognition of visual stimuli. In that setting, source and target are both from the same domain, i.e. the domain of visual stimuli. For the representation of visual stimuli, we distinguish between flat and structured representations. A flat representation covers all features of a stimulus. It may be realized as a list of primitive visual elements. A structured representation captures regularities of a stimulus, like symmetry, iterations, Gestalt groupings etc. It furthermore comprises geometrical and topological relations. A structured representation can be build from a flat representation according to a certain set of rules.

The application of HDTP as a framework for object recognition requires the development of a suitable language to represent spatial objects, the ability to adapt these representations such that analogous structures between the source and the target object become visible, and finally a mechanism for analogy-based learning of concepts. As a consequence the language has to meet two major requirements: it must describe all elements in a spatial scene with respect to the aspects relevant in human perception, but it must describe as well the spatial relationships which are important to compare and recognize objects. To reflect human perception, the language must comprise significant perceptual features, but also vocabulary to specify visual structures. When the human visual sensory system observes a spatial object, it transforms the unstructured information into a structured representation of coherent shapes and patterns. Human perception tends to follow a set of Gestalt principles: stimuli are experienced as a possibly good Gestalt, i.e. as regular, simplistic, ordered, and symmetrical as possible. Therefore the language focuses on basic Gestalt principles of perception, i.e. it allows for groupings according to the principle of similarity, the principle of proximity, closure, and good continuation.

The second requirement refers to spatial features: the geometry of elements in a scene and their spatial relations have to be represented in a way that allows...
for cognitively plausible reasoning. Common calculi for qualitative spatial reasoning such as RCC 8 for topological relations [14] and TPCC calculus [12] or neighborhood-based approaches [6, 15] for directional relations are integrated in the formal language. Lately, various approaches have been developed to describe visual stimuli and detect analogous structures, such as the Languages of Perception [2], GeoRep and the sketch mapping tool nuSketch [4, 5], Galatea and the Proteus analogy model [3], and an analogy model for comparison of topographic maps [13]. However, none of them can cope with the reasoning on both, the representation of perceptual aspects and the spatial relations of relatively complex figures.

In [17], we developed first steps towards a language for representing simple figures in geometric proportional analogies. Figure 3 shows an example of a formal language representing a stove. On the left is an unstructured representation of the stove listing its primitive elements (in this case lines and round elements). On the right is a structured representation of a stove: The four connected lines are represented as closed shape (in this case a polygon). The four hotplates are grouped together according to the Gestalt principle of similarity and proximity. The topological relation inside and the directional relation above are captured as well. The groups of hotplates are inside the polygon \( p_2 \) and polygon \( p_2 \) is above polygon \( p_1 \). In the following section, we explain how HDTP automatically adapts the unstructured representation to form a structured one.

3.3 Adaptation of the Representation for Analogy-Based Stimulus Recognition

The cognition of spatial objects involves the construction of a consistent and meaningful overall picture of the environment. Gestalt Psychology argues that human perception is holistic: instead of collecting every single element of a spatial object and afterwards composing all parts to one integrated picture, we experience things as an integral, meaningful whole. The whole contains an internal structure described by relationships between the individual elements.

To apply HDTP, a visual stimulus is described via a set of axioms specifying the features of all elements at a basic level. A set of perception rules and rules
Closed Shape (adapted from Gestalt principle)
\[
\text{lineConnection}(A, B) \ :- \ \text{line}(A, (X, Y), (X, Y))\), \ \text{line}(B, (X, Y), (X, Y))\).
\]

Topological Relation proper part (adapted from RCC8)
\[
\text{regionConnection}(X, Y) \ :- \ \text{region}(X), \ \text{region}(Y), \ \text{not}(\text{disjoint}(X, Y))\).
\]

\[
\text{part}(X, Y) \ :- \ \text{regionConnection}(Z, X), \ \text{not}(\text{regionConnection}(Z, Y))\).
\]

\[
\text{properPart}(X, Y) \ :- \ \text{part}(X, Y), \ \text{not}(\text{part}(Y, X))\).
\]

\[
\text{overlap}(X, Y) \ :- \ \text{part}(Z, X), \ \text{part}(Z, Y).
\]

Fig. 5. Adaptation rules are stored in the background knowledge and define how unstructured descriptions can be re-represented to structured ones.

for spatial reasoning form the background knowledge of the system. The set of all formulae that can be inferred from the axioms comprises all possible re-representations of the same visual stimulus, but at different structural levels. The initially flat representation can be transformed into a structured one by means of logical inference.

In the recognition task, a new stimulus (target) is compared to a known stimulus (source). The source stimulus is described via a structured representation recalled from memory. The structural commonality between the flat representation of the target and the structured representation of the source is initially not obvious. To successfully classify a new stimulus, a mapping between the target stimulus and the source stimulus must be established, i.e. an analogous structure has to be established on the target stimulus. During the analogy-based mapping process the target must be re-represented such that common structures become visible. The re-representation process building a structure on the target side can be driven by heuristics motivated by human perception, like Gestalt principles.

Figure 5 shows adaptation rules as they can be found in the background knowledge: The first rule is applied to detect closed shapes such as a polygon and the second one is applied to compute topological relations such as inside. The re-representation process is driven by heuristics based on properties of human perception and by building a structure on the target side analogously to the structured stimulus on the source side. Here experimental data shall give the necessary insight for creating appropriate heuristics reflecting human strategies in spatial object recognition. The heuristics have a great influence on the efficiency of the whole computational approach.

4 Analogy-based Learning, Concept Formation, and Creativity

Similarity judgment is one of the most central constructs in cognitive processes. Organization of conceptual knowledge in memory, recognition of new stimuli, and learning hinge crucially on similarity comparisons [8]. In particular, the role of structural similarity in relational categories has been considered as important [7]. We argue that structural similarity as detected in analogies is particularly important for learning spatial concepts. Our approach for computational cognition
shall learn to classify spatial objects, i.e. the system shall be able to revise and refine its ontological knowledge during a training phase. Although researchers agree that analogy-making is central for human learning, there does not exist a comprehensive theory for analogical learning. Our own first ideas for a learning model based on HDTP were outlined in [18].

HDTP supports learning at two levels: analogical transfer and abstraction. Learning via analogical transfer means gaining new knowledge by applying additional knowledge about the source to the target. The system transfers knowledge about the concept (e.g. knowledge about the functionality) and applies this to the new stimulus. This enables the system to draw new inferences on the target side. Transfer also happens from the target to the source: the system observes characteristics about the new stimulus which leads to a revised concept definition. Learning via abstraction refers to the generalization process that is essential to derive abstract concept definitions. Existing approaches apply classical inductive learning which requires large set of data samples to create general laws. However, humans can generalize already over a small set of samples. Applying analogical comparison and describing structural commonalities at a general level is one possible way to make the essence defining a concept apparent. Reflecting this analogical generalization process is one of the strengths of HDTP [16]: during the analogical mapping, anti-unification automatically constructs a generalization for every aligned pair of formulae. This way, HDTP creates an explicit generalized theory over two theories – the source and the target theory. We argue that this generalized theory captures exactly the essential commonalities of the instances of a concept at an abstract level and therefore is an ideal mechanism for extracting the defining elements of a concept.

The following example illustrates how HDPT functions in concept formation and concept learning (cf. Figure 6). HDTP has a structural description of a stove in its knowledge repository. Presenting a new stove in a recognition task, HDTP detects the analogous structure and constructs a generalization containing the commonalities (i.e. common aspects about the geometry and spatial relation such as the temperature regulators being situated in the front polygon). The generalization represents the concept "stove" at an abstract level. If again a new stove is presented in a second recognition task (e.g. the third one in the above figure), it could be classified as a stove, however the new generalization is not so specific on the position of the temperature regulators. First steps towards this incremental analogy-based learning have been sketched in [9].

5 Conclusions

Analogies play a major role for cognition. We have shown empirically, that structural commonalities are important in object comparison and object recognition: In a recognition task, subjects have recognized sketches of non-structurally varied objects faster and easier than sketches of objects which were structurally varied. Then we have introduced an approach of using HDTP, a symbolic analogy-making framework, to compute analogies between sketches of objects. HDTP is a
Fig. 6. A structural comparison of these stoves reveals that all stoves sketches have the form of a 3D cube with four hotplates on top and a spy window at the front. Three sketches show stoves with temperature regulators at the front.

promising framework, because it supports adaptation and learning at an abstract level. Many times analogical structures are not visible per se, but result from a comparison and mapping task. HDTP combines the representation of basic elements in a sketch with background knowledge on human perception. Therefore, HDTP can reveal commonalities in different contexts and different perceptions. It re-represents an unstructured flat representation of a sketch and determines a structured representation of the target stimulus which possibly matches the structured representation of the source stimulus. Furthermore, HDTP compares structures of source and target stimuli and computes a generalization of the shared structures. This supports concept learning.

References


