STRUCTURAL ALIGNMENT OF VISUAL STIMULI INFLUENCES HUMAN OBJECT CATEGORIZATION

Rolf Stollinski
rstollin@uos.de

Angela Schwering
aschweri@uos.de

Kai-Uwe Kühnberger
kkuehnbe@uos.de

Ulf Krumnack
krumnack@uos.de

Institute of Cognitive Science,
University of Osnabrück,
Albrechtstraße 28, 49076 Osnabrück (Germany)

ABSTRACT

This paper describes an experiment which investigated subjects’ reactions during an object categorization task with structurally and non-structurally modified objects and attempts to examine more closely the role of structural comparison in this connection. In our experiment, subjects reacted faster and agreed more often in those trials where the second object was a non-structurally modified variation of the first object than in trials where the second object was structurally modified. These findings suggest that structural commonalities and differences influence object categorization supporting the hypothesis that analogy can be used as an explanation for certain cognitive processes involved in human object recognition.

INTRODUCTION

Every day we encounter new and familiar objects in our environment. Our visual system perceives the incoming information and computes descriptions of the visual structure (Marr, 2000). One of the fundamental tasks we are confronted with day by day is therefore to go beyond what is perceptually available, organize the visual information with respect to our prior experiences, and make sense of the objects we see (Day & Gentner, 2007). Via comparison to already familiar objects we are able to recognize them and understand the world around us.

Even though this perceptual mechanism is accomplished so quickly and effortlessly, object categorization and recognition are highly complex and sophisticated processes. In the last years, research has investigated more closely the role of analogy in such cognitive processes. Analogical thinking and reasoning have often been identified as the core in many cognitive abilities such as everyday and abstract scientific reasoning (Markman & Gentner, 2001), learning and categorization (Gentner & Kurtz, 2005; Goldstone, 1994), reminding and organizing memory (Kokinov & Petrov, 2001; Kroger et al., 2004), creativity (Indurkhya, 2006), and so forth. Like other researchers (Lovett et al., 2006; Yaner & Goel, 2007), we argue that analogical comparison provides a possible explanation of object categorization and recognition. Furthermore, analogical comparison is structurally sensitive: The structural representations of two objects are not only compared for similarities between the components of objects, but also between the relations of components (Gentner, 1983; Markman & Gentner, 2001). Along with the findings of Krawczyk, Holyoak, and Hummel (2004), we furthermore suggest that relational structure is a dominant determinant, both in analogical mapping and inference generation.
In this paper, we provide empirical evidence that the cognitive mechanisms of object categorization and recognition rely fundamentally on analogies as well. Subjects were presented each time two successive objects in the course of a single experimental trial. In comparison to the first object, the second one was either identical, non-structurally varied, structurally varied, or a completely different object. We supposed that humans would be faster at classifying objects which have been varied without changing its structural representation than recognizing objects with a structural variation. Additionally, structural variations were assumed to decrease the similarity between two objects even more than non-structural variations, and thus we expected a higher level of agreement in those experimental trials with non-structurally varied objects.

The remainder of this paper is structured as follows: In the next section, we outline the notion of structure and structural change as to the visual stimuli used in the experiment. The following sections describe the pre-experimental phase, the actual experiment, and discuss the results with respect to our hypotheses. Finally, we summarize the main findings and outline directions for future work.

THE STRUCTURE OF VISUAL STIMULI

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 unquestionably possible to describe these relationships in a qualitative manner. Commonly used spatial relations are topological, directional, or metric relations (Cohn & Renz, 2007) and may involve other qualities such as size, order, and quantity.

Egenhofer and Franzosa (1991) presented a formal framework to describe topological spatial relations between regions. Those relations were defined in terms of the intersections of the boundaries, exteriors and interiors of two regions. With reference to Figure 2, it can be seen that for two spatial regions there are eight distinct topological relations with equally many linguistic terms adopted for them. The relation “contains” equates to the exact inverse of “inside” just as “covers” is the inverse of “coveredBy”. It is worth mentioning in this context that the relations “equal” and “overlap” were not of importance for the current study.

Figure 1: The line drawing of a typical kitchen stove.

When applying this general idea to the stove in Figure 1, its grey highlighted components might be regarded as separate regions with certain underlying topological relations. That is, the four hotplates on top could, for example, 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 foaside. Furthermore, the lateral Area 4 directly meets Area 2, and so forth.
Structural Alignment of Visual Stimuli influences Human Object Categorization

In sum, these numerous topological relationships in their entirety amount to a certain perceptual structure, in this case to the common structure of a typical stove. A change in the common structure of a given object can easily be achieved by altering and thus violating the existing topological relationships. One has to be careful though because not all visible manipulations are necessarily equivalent to a structural change. For example, we did not regard directional or metric manipulations as a clear structural change in the first place provided that the existing topological relations remained untouched at the same time. Additionally, we were aware that certain image variations (cf. the hotplates in Figure 3, Variation 4) can easily lead to severe changes in the functional properties of the underlying object (Lin & Murphy, 1997). Nonetheless, a wide range of our experimental stimuli remained functionally intact despite the non-structural and structural topological changes that were applied to them. Furthermore, in cases where changes of functionality did occur, they were not restricted to either non-structural or structural changes, i.e. they occurred in both classes. It was therefore decided to suspend a closer examination of object functionality within the scope of the current study and to incorporate this issue in our future work.

Recapitulating, all these considerations and approaches served as a starting point for the conception of the current study, its underlying experimental design, and the corresponding visual stimuli.

**PRE-EXPERIMENTAL PHASE: ACQUISITION OF THE STIMULUS MATERIAL**

The primary goal of the pre-experimental phase was to obtain a suitable collection of well-recognizable visual stimuli in the form of black and white line drawings (Figure 1). The drawings were either obtained from freely accessible internet databases or from other offline sources and were edited afterwards as desired. All object stimuli were purposefully selected and, for example, chosen from the domains of animals, technical devices, buildings, or home appliances.

Altogether 16 volunteers were invited to name and rate in written form 150 of such drawings each of which portrayed a common everyday object. While the naming task was intended to give some feedback about an object’s recognizability, the subjective Likert-scale ratings (Likert, 1932) ought to give some useful clues as to an object’s general quality. Only the 132 best rated and correctly identified objects were taken over to the experimental phase for further editing.

**EXPERIMENT: METHODS**

**Subjects**

75 native German subjects, 50 females and 25 males, volunteered for the experiment and confirmed normal or corrected normal vision. 70 subjects were right-handed, the other five were left-handed. 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. If required, volunteers received one participation hour for their attendance, however no remuneration or course credit was granted.
Materials

The experiment was conceptualized and generated with the aid of the software suite E-Prime 2.0 by Psychology Software Tools Inc. It was run on a conventional PC with Microsoft Windows XP Professional SP3 as operating system. A conventional 21 inch CRT monitor (Hewlett Packard P1120) was connected to the PC via the analogue interface in order to display the stimuli. The monitor’s refresh rate was set to 75 Hz. A resolution of 1280 by 1024 pixels was consistently used. Responses were recorded via a keyboard. Both the left-arrow and the right-arrow key interchangeably served as the “yes” or the “no” response button.

To construct the final set of stimuli, we utilized the 132 black and white line drawings from the pre-experimental phase. Of these line drawings, 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 (Figure 3). In this way, a set of altogether 300 experimental stimuli was acquired: 240 variations entailing specific topological manipulations plus the underlying 60 basic experimental stimuli. The subsequent paragraphs are intended to shed light on the existing experimental conditions and to dwell on the therein applied manipulations.

Generally speaking, each experimental condition was conceptualized as a pair of two experimental stimuli, henceforward referred to as item pairs. Each pair in turn comprised a certain source image and a specific target image (Figure 4). Besides a match condition, there were four more experimental conditions which could be subdivided into two related classes with two conditions each, henceforward referred to as non-structural conditions and structural conditions, respectively.

**Match condition:** The match condition (MAT) 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 (Figure 4, top row). Furthermore, this condition served as a baseline with respect to the reaction time measurements and required a clear “yes” response from the subjects.

As with the match condition, all of the following non-structural and structural conditions contained a basic experimental stimulus as the source image. However, this time the target image was invariably made up of a visually manipulated version of the source image.
Non-structural condition I: The first non-structural condition (NS1) always entailed the movement of significant picture elements (Figure 4). These manipulations were not taken for a structural change since it was made sure at all times 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.

Non-structural condition II: The second non-structural condition (NS2) invariably entailed the resize of chosen picture elements without moving them to another position or into another area (Figure 4). Simple resize was not taken for a structural change either as long as the topological relationships between the manipulated and other picture elements remained constant. Due to the rather minor effects of these manipulations on the whole image, it was likewise anticipated that the subjects would show a high tendency to give a “yes” response.

Structural condition I: As for the first structural condition (S1), it exclusively implicated the removal and/or addition of selected picture elements (Figure 4). Adding to or removing significant elements from the overall scene was regarded as a clear structural change due to the evident violation of the underlying topological relationships. Because of the disruptive severity of these image manipulations, it could not be determined with reasonable certainty how the subjects would react to these kinds of stimuli. Therefore, it was decided to accept both a “yes” and a “no” response as “potentially correct”. Ultimately, the results of the experiment ought to unveil any response tendencies.

Structural condition II: The second structural condition (S2) likewise implied the movement of significant picture elements as with condition NS1 (Figure 4). However, this time a structural change was deliberately caused by
moving selected elements into another area as, for example, the hotplates that were shifted from the top to the stove’s left lateral section (Figure 3, Variation 4).

Alternatively, this condition involved the resize of desired picture elements as with condition NS2. On this occasion though, size was either increased or reduced in order to cause again a deliberate violation of existing topological relationships. The latter manipulations were applied in just a few cases, predominantly if a change in orientation or the relocation of picture elements was not applicable to obtain a structural change. Similar to condition S1, the severity and salience of the mentioned image manipulations did not allow for a definite prediction as for the subjects’ responses. In consequence of that fact, both “yes” and “no” were accepted as potentially correct answers.

**Filler Items:** At last, the 72 filler items formed 36 unique filler item pairs which required a clear “no” response from the subjects since the source and the target image were consistently dissimilar (Figure 5).

**Design**

Due to the five experimental conditions, we created equally many stimulus lists that counterbalanced item pairs and conditions. It was ensured that each list included a different one of the five possible combinations for each object. All five lists were accompanied in addition by the same set of 36 unique filler item pairs. As each subject saw one list, this produced a 1 (list) × 1 (source image) × 5 (target image: MAT, NS1, NS2, S1, S2) setup, with “target image” as the sole within-subjects variable and “list” as the only between-subjects variable. Consequently, 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.

**Procedure**

At first, subjects were requested to carefully read the onscreen instructions. Once they confirmed that they had understood and internalized the instructions, they were given the possibility to perform several practice trials. Afterwards, the actual experiment began.

Basically, a single experimental trial was composed of a source image stimulus and a subsequent target image stimulus. For each stimulus there existed one particular task that the subjects had to accomplish in a consecutive fashion. These tasks were aimed at strengthening the conceptual representations of the two related stimuli within a trial.

With respect to the source image stimulus, all subjects were expected to name the object that they thought to have identified in the black and white line drawing by giving a single oral answer (in this way, it could be checked whether or not the source image was correctly identified). They were allowed to take as much time as they needed to be able to give a carefully considered answer. After the completion of the oral naming task, subjects had to press the keyboard’s down-arrow key to call up the target image once they were ready to do so. 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 and was maximally visible for 650 ms on the monitor. 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. To give a more comprehensible example, on the assumption that the subjects had just identified the source image object in Figure 4 as a “stove”, they were supposed to decide thereafter whether the target image again showed a stove in their opinion.

The time frame to give a response was limited to 3000 ms per trial (onset time frame = onset target stimulus) such that response latencies between 1 ms and 3000 ms were theoretically possible. Pressing either response button demarcated the immediate end of the current trial. The next trial followed automatically. The experiment ended after the completion of all 96 trials which took the subjects 35 minutes on the average.

RESULTS

Table 1 displays the mean reaction times (RT) in milliseconds (standard deviations are stated in parentheses) as well as the response accuracies (ACC) for the match condition and the filler item pairs. Concerning the other four conditions, the particular “yes”/“no” response ratios are given.

Table 1: Descriptive statistics results - analyses by subjects. Note that the filler item pairs did not constitute a genuine experimental condition. As a consequence, they were excluded from further statistical analyses.

<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>NS1</td>
<td>723 (195)</td>
<td>---</td>
<td>77.9 / 22.1</td>
</tr>
<tr>
<td>NS2</td>
<td>693 (180)</td>
<td>---</td>
<td>86.3 / 13.7</td>
</tr>
<tr>
<td>S1</td>
<td>758 (209)</td>
<td>---</td>
<td>60.8 / 39.2</td>
</tr>
<tr>
<td>S2</td>
<td>745 (204)</td>
<td>---</td>
<td>61.8 / 38.2</td>
</tr>
<tr>
<td>Filler</td>
<td>592 (131)</td>
<td>97.6</td>
<td>---</td>
</tr>
</tbody>
</table>

Recollecting the phrased experimental hypotheses, one of the main goals of this study was to find evidence for the assumption that humans would need more time to recognize structurally manipulated objects as 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. The relevant reaction times per subject were summed up and averaged afterwards. The same holds for the “yes”/“no” response ratios yielding the numbers as shown in Table 2.

Table 2: Descriptive statistics results - analyses by subjects (“Yes” and “No” responses).

<table>
<thead>
<tr>
<th>Condition</th>
<th>RT (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>

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 the former two structural conditions).

That being said, 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.

Table 3: Descriptive statistics results - analyses by subjects / items (“Yes” responses only).

<table>
<thead>
<tr>
<th>Condition</th>
<th>RT by subjects (Std. Dev.)</th>
<th>RT by items (Std. Dev.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAT</td>
<td>619 (146)</td>
<td>615 (70)</td>
</tr>
<tr>
<td>NSCOM</td>
<td>701 (190)</td>
<td>665 (62)</td>
</tr>
<tr>
<td>SCOM</td>
<td>733 (206)</td>
<td>694 (85)</td>
</tr>
</tbody>
</table>

One might conjecture that the difference in RTs might be explainable by the difference
in the RTs between positive and negative responses. Nevertheless, it turns out (as shown in Table 3) that the relevant RTs per subject and per item restricted to “yes” responses show essentially the same characteristics as the overall data.

As a result, the main effect for target image type was highly significant in the analysis by subjects (F1) and by items (F2) with F1(1.61, 112.56) = 87.51, p < .001 (Huynh-Feldt corrected); F2(2, 110) = 69.15, p < .001. Concerning the main effect for list, it was only significant in the analysis by items, F1(4, 70) = .52, p > .72; F2(4, 55) = 7.50, p < .001. By contrast, the two-way interaction between list and target image type was not significant at all with F1(8, 138) = 1.21, p > .30; F2(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 as follows.

**MAT vs. NSCOM:** The main effect for target image type was highly significant in the analysis by subjects and by items with F1(1, 70) = 79.70, p < .001; F2(1, 55) = 76.95, p < .001. The main effect for list was only significant in the analysis by items, F1(4, 70) = .71, p > .59; F2(4, 55) = 5.94, p < .001. The two-way interaction between list and target image type was not significant at all with F1(8, 138) = 1.09, p > .37; F2(8, 108) = 2.43, p > .06.

**MAT vs. SCOM:** As to the main effect for target image type, it was highly significant in the analysis by subjects and by items with F1(1, 70) = 110.44, p < .001; F2(1, 55) = 118.57, p < .001. The main effect for list was not significant in the analysis by subjects, but in the analysis by items with F1(4, 70) = .59, p > .67; F2(4, 55) = 7.55, p < .001. The two-way interaction between list and target image type was not significant in both analyses with F1(4, 70) = 2.31, p > .07; F2(4, 55) = 2.46, p > .06.

**NSCOM vs. SCOM:** Once more, the main effect for target image type was highly significant both in the analysis by subjects and by items with F1(1, 70) = 34.82, p < .001; F2(1, 55) = 15.90, p < .001. The main effect for list was only significant in the analysis by items, F1(4, 70) = .41, p > .80; F2(4, 55) = 3.40, p < .05. The two-way interaction between list and target image type did not yield any significance with F1(4, 70) = 1.52, p > .21; F2(4, 55) = 1.16, p > .34.

**DISCUSSION AND CONCLUSION**

Recapitulating, the conducted individual analyses on the response latencies by subjects and by items revealed results that met our expectations. Not only the match condition was each time significantly different from the other two conditions (NSCOM & SCOM), but also condition NSCOM differed significantly from condition SCOM in terms of the average response latencies.

Subjects needed significantly more time to react to both non-structurally and structurally manipulated object stimuli as opposed to the unaltered black and white line drawings in the match condition. In accordance with our experimental hypotheses, this result supports our prediction that it is generally harder to make a binary decision about non-structurally and structurally modified objects as reflected by the increased response latencies shown in Table 2. Additionally, the experiments showed that complete mismatch tasks (filler condition) can be solved very rapidly by the subjects.

A natural way to explain the reaction time difference between the MAT and the modified presentation of the stimuli is to assume that subjects try to match a structured representation of both stimuli. In the modified case, they need to adapt these representations. In the mismatch case, the structures are too different, such that match attempts do not take place.

On closer inspection of our results, it is evident that subjects furthermore needed significantly less time to respond to non-structurally manipulated object stimuli (NSCOM) as compared to structurally manipulated ones (SCOM). This particular outcome
Structural Alignment of Visual Stimuli influences Human Object Categorization

corresponds to our hypothesis that common structure plays an important role when it comes to recognizing everyday objects. While the topological relationships between the picture elements (Figure 1) remained unchanged as to all NSCOM object stimuli, the opposite was uniformly the case in terms of all stimuli belonging to the condition SCOM. It can thus be derived that the violation of existing topological relations must have been the crucial factor which finally gave rise to the average reaction time differences between NSCOM and SCOM.

Due to the fact that structural alignment processes require an association of substructures given the input and that structurally modified alignments need an adaptation method for the successful establishment of the association, the adaptation of the representations is more expensive in the non-structural case. Several frameworks for analogies and analogical reasoning (e.g. Schwering et al., 2009; Forbus et al., 2005) provide an explanation of this situation: the adapted (represented) alignment of relations requires more computational power and a more sophisticated algorithm for the establishment of an analogy than the alignment of attributes.

Appropriate results are found when examining closer the distributions of the “yes” and “no” answers across the experimental conditions (Table 2). As expected, subjects averagely showed an almost perfect accuracy performance in the match condition. More importantly, it can be clearly seen that there was an increasing tendency to give a “no” response with reference to the conditions NSCOM and SCOM. The non-structural changes as applied to all NSCOM object stimuli were sufficient to cause the subjects to give a “no” response in 17.9 % of all cases. This tendency is even more pronounced in respect of SCOM. The latter condition did not only yield the highest mean response latencies, but also the highest proportion of “no” answers with 38.7 %.

Concerning an explanation of these findings it is rather plausible to assume that the less adaptation that is required for establishing an analogy the easier it is to compute such an analogy. This results in a higher chance to dissociate two SCOM stimuli than two NSCOM stimuli, because more adaptation (representation) is required in the first case. In the extreme case of many adaptation steps, the resulting analogies tend to become less acceptable.

All in all, it can be concluded that both non-structural and structural alterations obviously lead to a decreased recognizability of everyday objects, and, by implication, to increasingly higher reaction times as well. It is worth pointing out again that structural changes seem to have an even stronger impact on the recognizability of objects than non-structural changes.

In future work, we intend to examine object functionality aspects in structural and non-structural changes, an issue that was omitted during creation of the stimuli in this study. Furthermore, we will add further possible factors for (non-)structural changes by considering not only topological relations in representations, but also non-topological relations (e.g. relative positions of objects).

REFERENCES


