# How Humans Describe Short Videos

Adrian Ion, Hurbert Hausegger, Walter G. Kropatsch, and Yll Haxhimusa*
Pattern Recognition and Image Processing Group
Institute of Computer Aided Automation
Vienna University of Technology
Favoritenstr. 9, A-1040, Vienna, Austria
{ion, hauseggh, krw, yll }@prip.tuwien.ac.at

## Abstract

*Recognition, manipulation and representation of visual objects can be simplified significantly by "abstraction". By definition abstraction extracts essential features and properties while it neglects unnecessary details. We have conducted two sets of experiments in order to relate abstraction levels used by humans when describing videos, to abstraction level categories used in computer vision. Experimental results show the high abstraction levels used and motivate cognitive vision research towards this direction.*

## 1 Introduction

Cognitive vision is certainly one of the youngest concepts that exist in nowadays computer vision related research. Some expectations regarding it's properties exist but a generally accepted clear delineation of what the term should mean and how we could verify it's existence is missing. Most of the opinions regarding it, have as a basis the one known entity to posses these capabilities, the human.

The *Research Roadmap of Cognitive Vision* [9], presents this emerging discipline as "a point on a spectrum of theories, models, and techniques with computer vision on one end and cognitive systems at the other".

Thinking of the previous and searching for a proper representation for a cognitive vision system, a conclusion drawn is that a good starting point for a representation would bring together the following:

- enable easy extraction of data for human comparison;
- bridge together high and low level abstraction data used for cognitive and computer vision processes.

*This paper was supported by the Austrian Science Fund under grant FSP-S9103-N04 and FWF-P18716-N13.

Bases for human-like qualitative spatial and temporal representation and reasoning allready exists. Interval calculus [1] is used in systems that require some form of temporal reasoning capabilities. In [1] 13 interval-interval relations are defined: 'before', 'after', 'meets', 'met-by', 'overlaps', 'overlapped-by', 'started-by', 'starts', 'contains', 'during', 'ended-by', 'ends' and 'equals'. In [8], motivated by the work in [1, 4, 5], an interval calculus-like formalism for the spatial domain, the so called region connection calculus (RCC) was presented. The set of $8$ region-to-region relations defined in [8] ($RCC-8$) are: 'is disconnected from', 'is externally connected with', 'partially overlaps', 'is a tangential proper part of', 'is non-tangential proper part of', 'has a tangential proper part', 'has non-tangential proper part', and 'equals'. A more expressive calculus can be produced with additional relations to describe regions that are either inside, partially inside, or outside other regions ($RCC-15$).

In computer vision, object representations have spanned from prototypical models (generic/class based) to exemplar-based (appearance/template based) with each of them best suited for different applications. Certainly one of the main challenges of cognitive vision will be to extract the abstract features required for reasoning while bridging the representational gap between the output of low level image processing modules (e.g. segmentation) and the "parts" of a generic model [6].

After all this, it is clear that one of the main things one has to address when thinking of cognitive vision, is a proper internal representation which should be obtained by extracting abstract image features and should be usable to reason and communicate in a human way. To address this issue, we have conducted a set of preliminary experiments regarding human description of videos and tried to relate abstraction levels used by them, to abstraction levels from computer vision.

In Section 2 we give a brief presentation of abstraction with details about computer vision. Section 3 presents the

experiments, the results and their interpretation. The paper ends with the outlook (Section 4) and conclusions (Section 5).

## 2 Abstraction levels used in computer vision

Abstraction simplifies recognition, manipulation and *representation* of visual objects, since it selects essential features and properties while it neglects unnecessary details. Two types of unnecessary details can be distinguished: redundancies and data of minor importance.

Details may not be necessary in different contexts and under different objectives which reflect different types of abstraction. In general, three different types of abstraction are distinguished [7]:

**isolating abstraction:** important aspects of one or more objects are extracted from their original context.

**generalizing abstraction:** typical properties of a collection of objects are emphasized and summarized.

**idealizing abstraction:** data are classified into a (finite) set of ideal models, with parameters approximating the data and with (symbolic) names/notions determining their semantic meaning.

These three types of abstraction have strong associations with well known tasks in computer vision: recognition and object detection tries to *isolate* the object from the background; perceptual grouping needs a high degree of *generalization*; and classification assigns data to *"ideal"* classes disregarding noise and measurement inaccuracies. Such generalization allows to treat all the elements of a general class in the same way. When applied successively, the three types of abstraction imply a hierarchical structure with different levels

- of concepts for representing knowledge about the world, e.g. the conceptual hierarchy in [2],
- of representation,
- of processing stages, e.g. hierarchies of invariance in cognition [3], and
- in the complexity of processing images.

In all cases, abstraction drops certain data items which are considered less relevant. Hence the *importance* of the data needs to be computed to decide which items to drop during abstraction. The importance or the relevance of an entity of a (discrete) description must be evaluated with respect to the purpose or the goal of processing. The system may also change its focus according to changing goals after knowing certain facts about the actual environment, other aspects that were not relevant at the first glance may gain importance. Representational schemes must be flexible enough to accommodate such attentional shifts in the objectives.

In current computer vision research, multiple abstraction levels can be identified. They span from the low, image-based (pixels) to the high, object, model, and topology based. Table 1 shows the main abstraction categories and some of their properties.

## 3 Experiments

To motivate the research in the direction of qualitative spatial representation and reasoning, we have conducted a small set of experiments focusing on human description of videos.

### 3.1 The first set - 7 subjects, 2 videos

As a first step, 7 students (mother tongue German, descriptions made in German) were shown 2 videos ("two cups"[1] and "yellow ball"[2], see Figure 1) containing 2 identical black cups, a ball, a table (support for the cups and ball), and a hand that acts only on the cups by changing their position (on the table by pushing/shifting and in the air by picking up and holding). The 2 videos are approximately 15 and 31 seconds long.

A description of the experiments is as follows: each of the students were given a piece of paper (size A5) and told that 2 videos will be shown to them, which they should describe. After watching each video, a limited amount of time was given to describe it. No other clues were given. Of course, one can say, that seeing the hand hiding the ball using one of the cups is enough for a human (knowing the game) to focus on the ball. Which is most probably true, and can be seen on the produced descriptions. But this just enforces the hypothesis that humans focus on a given task and do not give attention to details not related do it.

One of the first things that should be mentioned is that there were two constraints for the descriptions: one intended, which was the time allowed to write the description (2 of the descriptions are not finished), and the other one noticed, the space available on the paper for the description (more then 70% of the descriptions use up all the allocated half of the A5 paper). We can easily associate the 2 constraints with allowed processing time and available memory, and notice that humans do very well on adapting to them.

[1] http://www.prip.tuwien.ac.at/Research/FSPCogVis/Videos/Two-Cups.mpg

[2] http://www.prip.tuwien.ac.at/Research/FSPCogVis/Videos/Yellow-Ball.mpg

**Table 1. Abstraction levels in computer vision**

| | addressing and axis | entities | neighborhood |
|---|---|---|---|
| image-based | 2D (row, column) | pixel | 4,8-neighborhood |
| appearance(view) -based | nmD eigenvectors | appearance | distance in eigenspace |
| part-based | part-whole relation | properties, parts | semantics |
| object/model-based | name, location | sub-objects/models | |
| scene-based | (x,y,z,t,...) | objects | spatio-temporal semantics |
| topology-based | relational paths | topology domain | explicitly encoded |

**Figure 1. The first frame of the 2 videos used in the first experiment.**

All of the participants (except one) have produced narrative descriptions, with very short sentences of the form *object action direction/position*, focusing on the movement of the cups and on the position of the ball at the end of the videos. The remaining participant has used a bullet-ed list with subsections and very schematic description.

All descriptions follow a two section pattern[3]:

1. **initial configuration**: contains the 3 objects initially visible, 2 cups {Tasse, Schale}, 1 ball {Ball}, and for 8 descriptions also the spatial arrangement using words like "left" {links}, "right" {rechts}, and "center" {Mitte},

2. **actions**: short sentences of the form *object action direction/position* using "left cup", "right cup", "ball", and "hand" {Hand} to identify objects, a whole variety of verbs for actions (e.g. "move" {bewegen}, see Table 2 for more examples) and and expressions to identify positions (e.g. "between two *objects*" {zwischen}, see Table 3 for more examples).

"left" and "right" are the relational/positional words with the highest appearance (about 25 times each), followed by center/middle (less then 10 times).

[3]In this paper, words are given in the form "English" {German translation 1, German translation 2, etc.}.

**Table 2. Action description.**

| English | German |
|---|---|
| pick up | aufholen, aufheben |
| raise | aufheben, heben, hochheben |
| move | bewegen |
| shift | schieben |
| slide | schieben |

**Table 3. Positional description.**

| English | German |
|---|---|
| over | auf, über |
| behind | hinten |
| place of | auf den Platz |
| between two | zwischen |
| left | links |
| right | rechts |
| center | mitte |

One of the descriptions refers to one of the cups as "the cup with the ball" for all the time the cup is hiding the ball. On the other hand, the bullet-ed list descriptions (2, made by the same person) refer to all objects as "object" (colours

are used at the initial configuration description, but only there). They contain no more than 3 actions to describe all the changes in the video and clearly state the final outcome. There is a description containing the wrong result, and some contain interesting hypothesis like the diameter of the ball in centimetres and the gender of the person that the hand belongs to.

After looking at the descriptions, the main observations are that:

- all the participants focused on the "implicit" problem statement (where is the ball?) and most of them basically ignored the hand;
- objects that cannot be identified easily by aspect are referred to using positions relative to the scene limits or relative to other objects;
- if the result of a position change is an interaction with another object, then this is used to describe the action, if not, then the final position is used and described relatively to the scene limits or relative to other objects using qualitative measures (left, right, front, middle, etc.);
- the descriptions focus on interaction between the objects, that could be considered relevant for the task.

In Figure 2 you can see the abstraction levels of the expressions (minimum number of consecutive words that made sense) used by the humans and their respective ratio. Appearance- and part-based most probably do not appear because of the content of the videos. An example of the description done by a subject is shown in Figure 3.

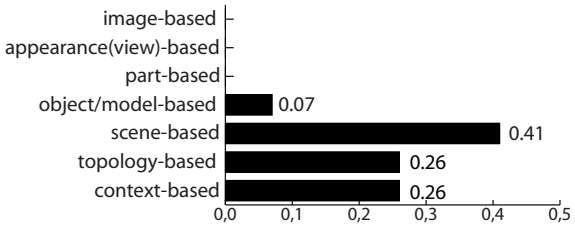


**Figure 2. Abstraction levels found in the human video descriptions - first experiment.**

### 3.2 The second set - 20 subjects, 2 videos

Intrigued by the results from the first set of experiments we have decided to continue and conducted a second set. This time a little bigger group was used, consisting of 20 subjects. Again two similar videos have been shown, but this time they also differed in the appearing objects (one had additional objects) not just in the actions. They both show a table with a calender on it ("Video1"[4] and "Video3"[5], see Figure 4) and a set of typical office objects (markers, boxes, PostIt's, cup) which are manipulated by a hand (markers and PostIt's put inside boxes, and cup, hidden behind calender, moved around on the table).

A number of 9 questions, from very specific (e.g. which logo appears on the calendar? - 3 options where given) to very general (e.g. what happens in the 2 videos?) have been divided in two groups. One group of questions was given before the videos were shown and the other one after. Also, part of the subjects got at the beginning the group of questions the others got at the end and vice-versa. The aim of this division was to show the role of focus on solving these simple tasks (especially on the question with the logo), but unfortunately it seems that the two videos were not complicated enough to require the subjects to strictly filter out all the details irrelevant for the a priori known questions. The logo was most probably correctly remembered because it was on the calender, which was once used to hide a marker, and which was on the table i.e. was in the important part of the scene (nobody mentioned the cabinet or the floor). Again, there were space and time constraints and the subjects have made the descriptions in German (different subjects were used than in the first experiments).

From all the 9 questions, we focus here on the results from only one of them (number 3,"what happens in the two videos?" which continues the line of human-video-description experiments presented in Section 3.1. For an example, see Figure 7). All the observations from the first set of experiments, regarding the way the actions and changes were described are verified again in the descriptions produced this time.

Motivated by the previous, we decided to try to find a relation between abstraction levels in (cognitive) computer vision and abstraction levels in human descriptions, and the usage of each category. In order to find such a relation, we have taken each description and tried to assign each expression to one of the abstraction categories.

Unfortunately, but expected, additional categories were required because some description did not fit properly into any of the abstraction level categories used in computer vision. Especially regarding redundancy, humans tend to omit a lot of things which they consider obvious. E.g. *The marker is hidden behind the calendar, and then the Post-It is hidden inside the box. After that, the marker and Post-It are taken out*. As can be seen in the previous description, where the objects are taken out from, seems obvious and

[4] http://www.prip.tuwien.ac.at/Research/FSPCogVis/Videos/video-emt-1.avi

[5] http://www.prip.tuwien.ac.at/Research/FSPCogVis/Videos/video-emt-3.avi