Noname manuscript No.
(will be inserted by the editor)

Systematicity and Compositionality in Computer Vision

Germán Martín García · Simone Frintrop · Armin B. Cremers

Received: date / Accepted: date

Abstract The systematicity of vision is a topic that has been discussed thoroughly in the cognitive science literature; however, few accounts of it exist in relation to computer vision (CV) algorithms. Here, we argue that the implications of the systematicity of vision, in terms of what behaviour is expected from CV algorithms, is important for the development of such algorithms. In particular, the fact that systematicity is a strong argument for compositionality should be relevant when designing computer vision algorithms and the representations they work with. In this paper, we discuss compositionality and systematicity in CV applications and present a CV system that is based on compositional representations.

Keywords Systematicity · Compositionality · Computer Vision

1 Systematicity and Compositionality

In their seminal paper [8], Fodor and Pylyshyn address the question of systematicity of cognition. Systematicity is the property by which related thoughts or sentences are understood. Anyone able to understand the sentence "John loves the girl" should be able to understand the related sentence "The girl loves John". This can be explained because both sentences are syntactically related. It is because there is a structure on the sentences that language, and thought, exhibit systematic behaviour. The compositionality principle states that the meaning, or the content, of a sentence is derived from the semantic contribution of its constituents and the relations between them [11]. It is because *John*, *the girl*, and *loves* make the same semantic contribution to the sentence "John loves the girl", and to "The girl loves John", that we are able to systematically understand both of them. In the case of language, systematicity is achieved by a compositional structure of constituents. In general, systematicity is a strong argument for compositionality [11]: we are able to understand an immense number of sentences which we have never seen before.

This can be extended to vision: we are able to make sense of scenes we have never seen before because they are composed of items we know. The systematicity of vision is defended by several authors. Already in [8], Fodor and Pylyshyn foresee that systematicity is probably a general property of cognition that is not limited to verbal capabilities. In the cognitive science literature, there are several arguments that support that vision is systematic [2,12]: *"if a subject is capable of visually representing a red ball then he must be capable of representing: i) the very same red ball from a large number of different viewpoints (and retinal inputs); ii) a number of similar red balls [...]; and iii) red objects and ball-shaped objects in general."* [2].

In this paper, we are concerned with the sort of systematic behaviour that should be expected when a scene is observed from different points of view: a systematic CV algorithm should be able to determine the visual elements that compose the images and find the correspondences between them over time. Some authors claim that systematicity in vision can be achieved without having compositionality [7,6]. However, the models they provide have not shown to be applicable in real world CV problems. We argue that from a computer

Germán Martín García · Simone Frintrop · Armin B. Cremers
Institute of Computer Science III, Universität Bonn
53117 Bonn, Germany
Tel.: +49 228 73 4530
Fax: +49 228 73 4382
E-mail: {martin,frintrop,abc}@iai.uni-bonn.de

scientist point of view, recurring to compositionality is beneficial when designing CV algorithms.

2 Compositionality in Computer Vision Algorithms

The systematicity problem is rarely addressed in computational models of vision. In [6], the authors acknowledge that structural descriptions are the preferred theory about human vision that allows for viewpoint abstraction and novel shape recognition. In the structural approaches to vision, the visual information is explained in terms of atomic elements and the spatial relations that hold between them [5]. One example is the Recognition-by-Components theory of Biedermann [3]. In this theory, object primitives are represented by simple geometric 3D components called *geons*. However, extracting such primitive elements from images is by no means a trivial task in CV. Approaches that attempt to extract such primitives to explain the visual phenomena are hard to realise in practice, and according to Andreopoulos & Tsotsos there is no method that works reliably with natural images [1]. Here, we suggest to generate such primitive elements by grouping mechanisms realised by segmentation methods which are well investigated in CV. In the following section, we propose a computer vision system that bases on such perceptually coherent segments to represent scenes in a compositional way.

3 A Compositional Approach for Visual Scene Matching

Here, we present a compositional vision system that is able to represent a scene in terms of perceptually coherent components and the relations between them with help of a graph representation. A graph matching algorithm enables to match components between different viewpoints of a scene and, thus, enables a scene representation that is temporally consistent. In contrast to geons, our segments are easily extracted with standard segmentation algorithms; we use the well known Mean Shift segmentation algorithm [4]. Mean Shift produces a segmentation based on the proximity of pixels in spatial and colour spaces. We construct a graph where the nodes represent segments, and the edges the neighbourhood of segments. We use labelled edges, where the labels correspond to the relations between segments. These are of two types, *part of* and *attached to*, and can be obtained automatically from the image by simple procedures. To compute whether two segments that share a common border (*attached to* relation) it is enough to perform two morphological operations: first to dilate, and then intersect both segments. The remaining pixels will constitute the shared contour and will indicate that this relation is present. To find whether segment A is *part of* segment B is enough to check whether the outer contour of segment B is the same as the outer contour of the union of A and B.

Once the graphs are built, we can apply a graph matching algorithm to establish correspondences between nodes, and thus, between segments. Suppose we have two graphs $G_1 = (V_1, E_1, X_1)$ and $G_2 = (V_2, E_2, X_2)$, defined by a set of nodes V, edges E, and attributes measured on the nodes X. We want to find a labelling function f that assigns nodes from G_1 to nodes in G_2: $f : G_1 \rightarrow G_2$. We base our approach for matching on [13]. The authors propose a relaxation algorithm for graph matching that locally updates the label of each node based on an energy functional $\mathcal{F}$ defined on the labelling function f. By defining $\mathcal{F}(f)$ as the maximum a posteriori probability of the labelling given the measurements $\mathcal{F}(f) = P(f|X_1, X_2)$, and by applying Bayes' rule, we get:

$$P(f|X_1, X_2) = \frac{p(X_1, X_2|f)P(f)}{p(X_1, X_2)}. \quad (1)$$

Hereby, $p(X_1, X_2|f)$ is the appearance term that denotes the probability that the nodes of a given match f have certain attributes X_1 and X_2: we used colour average and dimensions of the minimum fitting rectangle as attributes. $P(f)$ is the structural term and is high if a matching preserves the structure of the graph; for this term to have a high value, if node A is mapped to A', then the neighbours of A should be mapped to the neighbours of A'. The algorithm works by iteratively assigning to each node u in G_1, the node v in G_2 that maximises Equation 1:

$$f(u) = \underset{v \in V_2}{\operatorname{argmax}}\ p(x_u, x_v|u, v)P(f). \quad (2)$$

We extended the original algorithm so that it is able to deal with directed graphs as well as with labelled edges. The labels represent the two different relations: *part of* and *attached to*. The directions of the edges denote the order in which the segments appear in the relation predicates, e.g., in the *part of* relation, the edge points towards the node that contains the other, and in the *attached to* the edge points towards the node that is either under or on the right side of the other. The details of the algorithm are out the scope of this paper and can be found in [9].

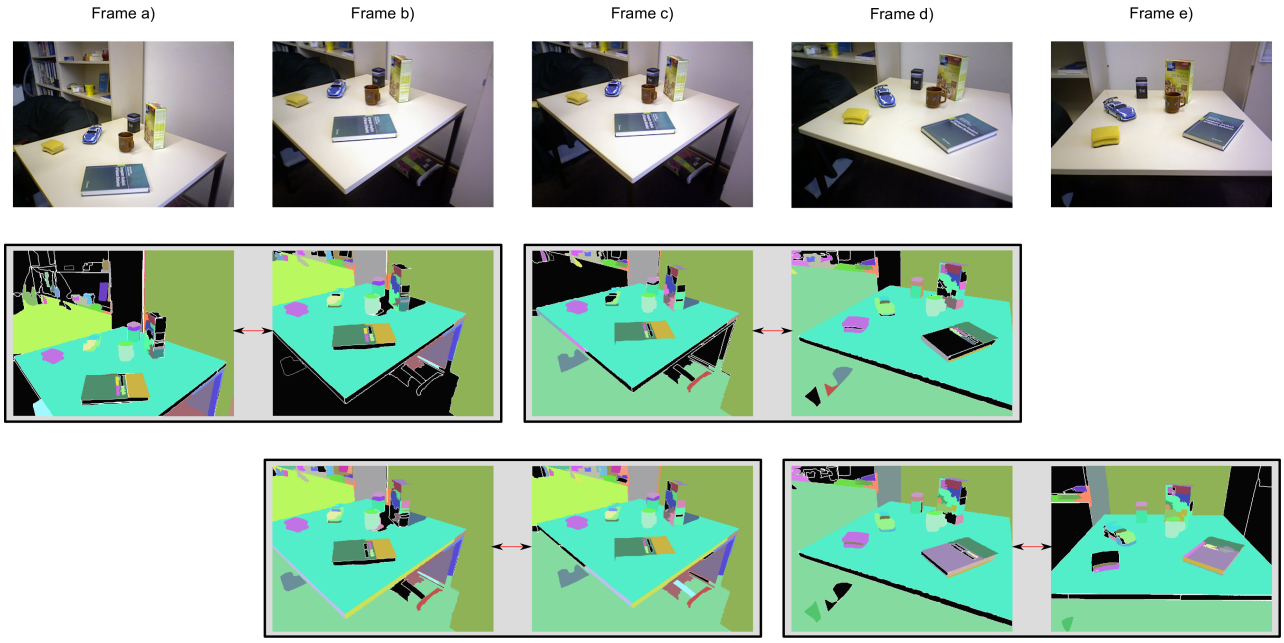

Fig. 1 First row: original non-consecutive images. Rows 2 & 3: results of the matching between the corresponding pair of frames. Matches are displayed with the same colours. Segments for which no match was found are shown in black.

We evaluated the algorithm on a real-world video sequence recorded at our office by matching pairs of consecutive and non-consecutive frames. In the first case, 84% of the segments were correctly matched, and in the second case, 57%. Some non-consecutive frames are shown in Figure 1: the matched segments are displayed with the same colour, and those that were missed are displayed in black. It can be seen that some missing matches originate from having non-repeatable segmentations over frames, i.e., the boundaries of the segments are not always consistent when the viewpoint changes (see, for example, the segmentation of the sponge in frames d) and e) in Figure 1). This is a known problem of image segmentation algorithms [10] that has two effects: a segment in frame 1 is segmented as two in frame 2, or the other way round. As a consequence, the graphs that are built on top of these segmentations are structurally different.

In future work, we will extend the matching algorithm so that merging of segments is performed. In the presented system, we show in an exemplar way how the concept of compositionality can be integrated into CV algorithms and, by making use of well-approved segmentation and graph-matching methods, a simple visual representation can be achieved that is coherent over time.

References

1. Andreopoulos, A., Tsotsos, J.K.: 50 years of object recognition: Directions forward. Computer Vision and Image Understanding (2013)
2. Aparicio, V.M.V.: The visual language of thought: Fodor vs. pylyshyn. Teorema: Revista Internacional de Filosofía **31**(1), 59–74 (2012)
3. Biederman, I.: Recognition-by-components: a theory of human image understanding. Psychological review **94**(2), 115 (1987)
4. Comaniciu, D., Meer, P.: Mean shift: a robust approach toward feature space analysis. Pattern Analysis and Machine Intelligence, IEEE Transactions on **24**(5), 603–619 (2002)
5. Edelman, S.: Computational theories of object recognition. In: Trends in Cognitive Science, pp. 296–304 (1997)
6. Edelman, S., Intrator, N.: (coarse coding of shape fragments) + (retinotopy) approximately = representation of structure. Spatial Vision **13**(2-3), 255–64 (2000)
7. Edelman, S., Intrator, N.: Towards structural systematicity in distributed, statically bound visual representations. Cognitive Science **27**(1), 73 – 109 (2003)
8. Fodor, J.A., Pylyshyn, Z.W.: Connectionism and cognitive architecture: A critical analysis. Cognition **28**(1), 3–71 (1988)
9. Garcia, G.M.: Towards a Graph-based Method for Image Matching and Point Cloud Alignment. Tech. rep., University of Bonn, Institute of Computer Science III (2014)
10. Hedau, V., Arora, H., Ahuja, N.: Matching images under unstable segmentations. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (2008)
11. Szabó, Z.G.: Compositionality. In: E.N. Zalta (ed.) The Stanford Encyclopedia of Philosophy, fall 2013 edn. (2013)
12. Tacca, M.C.: Seeing objects: the structure of visual representation. mentis (2010)
13. Wilson, R.C., Hancock, E.R.: Structural matching by discrete relaxation. IEEE Transactions on Pattern Analysis and Machine Intelligence **19**, 634–648 (1997)