A Topology-Based Concept for Contraction in Spatiotemporal Space [1)]

Adrian Ion, Yll Haxhimusa, and Walter G. Kropatsch
Pattern Recognition and Image Processing Group
Institute for Computer-Aided Automation
Vienna University of Technology
{ion,yll,krw}@prip.tuwien.ac.at

Abstract:
A concept relating story-board description of video sequences with spatio-temporal hierarchies build by local contraction processes of spatio-temporal relations is presented. Object trajectories are curves in which their ends and junctions are identified. Junction points happen when two (or more) trajectories touch or cross each other, which we interpret as the "interaction" of two objects. Trajectory connections are interpreted as the high level descriptions.

1 Introduction

Even though there is no generally accepted definition of cognitive vision yet, presumptions about the cognitive capabilities of a system can be made by comparing it's results with that of an entity, already 'known' and accepted to have these capabilities, the human. Also, 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'. A conclusion drawn from the previous, is that a good starting point for a representation would bring together the following: **i)** enable easy extraction of data for human comparison; **ii)** bridge together high and low level abstraction data used for cognitive and computer vision processes.

After 'watching' a video of some complex action, one of the things, that we would expect a cognitive vision system to do, is to be able to correctly answer queries regarding the relative position of occluded objects. Let us take the video[2)] given by a simple scenario of two black cups and a yellow ball and describe the scene in simple English words (see the description in Table 1). The description contains: **objects**: hand, cup, ball; **actions**: grasp, release, move, etc., and **relations**: to-the-left, to-the-right, etc. While observing a dynamic scene, an important kind of information is that of the change of an object's location, i.e. the change of topological information. In most of the cases, this

[1)]This Work was supported by the Austrian Science Foundation under grants P14445-MAT, P14662-INF and FSP-S9103-N04.

[2)]http://www.prip.tuwien.ac.at/Research/FSPCogVis/Videos/Sequence_2_DivX.avi

kind of change is caused by an active object (e.g. agent: hand, gravity, etc) acting on any number of passive objects (e.g. cup, ball, etc.).

From all the work done in the domain of qualitative spatial and temporal information we enumerate the following works: 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, 5, 6], an interval calculus-like formalism for the spatial domain, the so called region connection calculus (RCC) was presented. The set of 8 region-region base 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 ($RCC-15$). Different graph based representations have been used to describe the changes/events in a dynamic scene. In [4] graphs are used to describe actions (vertices represent actions). Graphs are also used in [2], but here vertices represent objects. Balder [2] argues that arbitrary changes can be best described by state approach: the state of the world before and after the change characterizes the change completely. The Unified Modeling Language, in its state diagram, also defines a graph based representation for tracking temporal changes.

In the following section we describe a spatio-temporal story board representation, and in Section 2, a concept of spatial and temporal contraction is presented. A short discussion follows in Section 3, and we end by presenting the conclusions.

1.1 Spatiotemporal Story Board of a Film

The scene history is a description of the actions and spatial changes in the scene. It should depict the spatiotemporal changes in the scene, in a way that could be used to create a human-like description. For this we propose a graph based representation where vertices represent spatial arrangement states and edges represent actions (see Figure 1a). Each vertex contains a topological description of the spatial arrangement of the objects in the scene, that results through a transition from a previous state, by applying the actions that link it to the current. What we refer to as objects are actually detected relevant visual entities, which in the ideal case would be objects or, groups of objects in a "special" physical relation e.g occluding, containing, etc. Vertices are added when the topological description of the spatial arrangement changes. There are no vertices that contain (identify) the same topological description (scene state). If the scene enters a state, which has a topological description identical to one of the descriptions already identified by a vertex in the scene history graph (it has been in the same state in the past), then an edge/edges from the vertex identifying the previous state, to the existing vertex should be added.

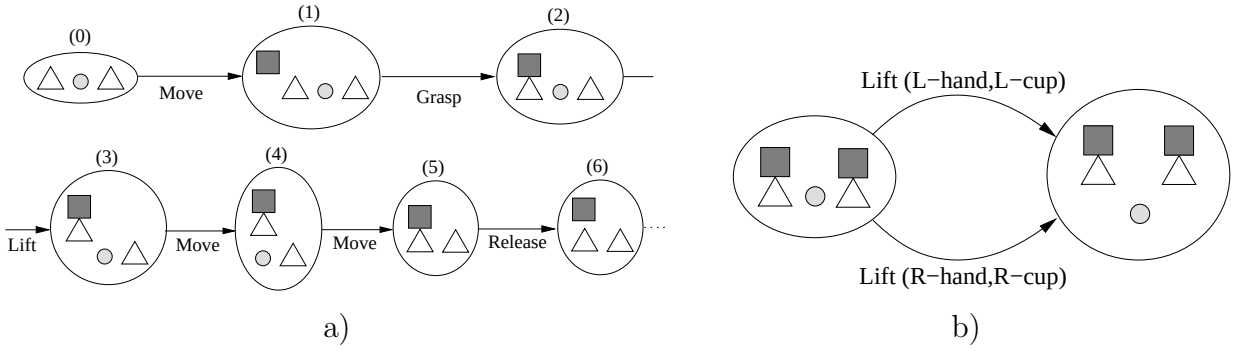

Figure 1: a) History graph. b) Parallel actions. □ Hand, ◯ Ball, △ Cup.

Edges are associated with actions and identify the type/class of the action. Also, each edge links to the objects (from the source and destination state vertex) involved in this particular action. If an object taking part in the action cannot be identified as one of the known objects, a new instance should be created and the edge linked to it. Later on, through reasoning, the new created instance, can be identified as a previously known object or a new one (or some presumption can be made, using certain criteria). In case of simultaneous actions, more than one edge is used to connect 2 vertices. Each edge should describe the actions that happened in parallel. (Figure 1b) shows how to describe 2 hands lifting 2 cups at the same time). Although it is small, the possibility of being able to go from one state directly to another, by using 2 different actions, exists. We enumerate 2 possibilities for avoiding confusion in such a case. One of them is to add an additional vertex, identifying a pseudo state and breaking one of the actions in 2 sub-actions and avoiding having the mentioned situation, and the other would be to group parallel edges together, such that each group contains the actions that have to be taken (have been taken) to change the system from one state to the other.

2 Contraction in Spatiotemporal Space

The idea here is to contract in 3D (2D space + time) along 'the trajectory' of the movements. Every frame could be represented by a region adjacency graph (RAG) or combinatorial map. In order to stretch this into time, the RAGs of all frames of the video have to be taken, in chronological order and matched to each other, i.e. the RAG at time t is matched with the one in $t + 1$ and so on.

In this sense we could define a 'trajectory' of each region. This trajectory becomes a curve in 3D and with the techniques analogous with that of contraction of a 2D curve pyramid in [7], we can contract regions adjacent along this curve to produce the more abstract representation of the scene, e.g. where the movement started, where it ended etc. (Figure 2).

If the analyzed scene has a structured background, then, depending on it's granularity, this is enough to detect movement using only topological information. This will, of course, increase the number

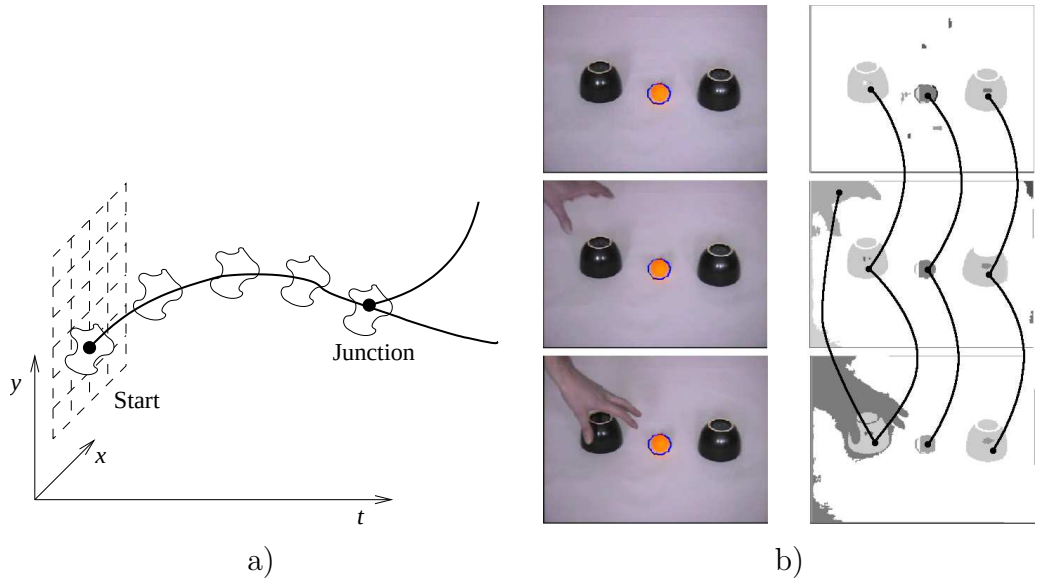

Figure 2: a), b) Trajectory of movements.

of consecutive frames that differ with respect to topological relations. To reduce the abundance of topological states, to a set containing the most relevant ones, a set of adaptive pyramids is used. There are no constraints regarding the time intervals between 2 consecutive states.

Principally, spatial contraction should focus on joining regions that share common properties, e.g. from physical ones, like being part of the same object, to contextual ones, like being important (of interest, foreground) or unimportant (not of interest, background) at that time. In [10] a method for building hierarchical image space partitions using Borůvka's minimal spanning tree algorithm is presented. The hierarchy is presented as a combinatorial pyramid with multiple resolutions, where each level is a 2D combinatorial map (a result of the algorithm is shown in Figure 3b). We used the idea of minimal spanning tree to find region borders quickly and effortlessly in a bottom-up way, based on local differences in a color space. The edge weighting criteria may differ, and can be extended, from just color distance to adding texture information or non physical information (e.g. expectations regarding a specific area, or importance of local details).

Temporal contraction should focus on identifying the key events/interactions in the scene and join the "unimportant" consecutive ones, creating a sparse description that still fully describes the changes under attention. Starting from the idea presented in [7] we define the following concepts for the time contraction: *the trajectories of (moving) objects (visual entities resulted from segmentation and tracked through the whole time span) represent curves connecting start, end and junction points*. Junction points happen when two (or more) trajectories touch or cross each other, which we interpret as the 'interaction' of two objects. Contraction can be done along the trajectory, preserving information about end points and junction points, and joining cells where no special interactions take place.

(a) 0 (16 384) (b) 23 (512) (c) 32 (65) (d) 39 (10) (e) 42 (4)

Figure 3: Some levels of the spatial partitioning of "Obj18__355": level (number of components).

2.1 Spacial contraction followed by temporal contraction

For each frame, whose topological description is different from the one of the previous frame, a space-contraction pyramid is build, that preserves only the spatial information required by the higher functionality levels (i.e reasoning) and by the time-contraction. A space-contraction pyramid is a pyramid where elements, from the same scene state, neighbored from a spatial point of view are contracted (e.g. Figure 3), and a time-contraction pyramid is a pyramid where elements, neighbored from a temporal point of view (consecutive scene states) are contracted.

To obtain the base level of the time-contraction pyramid from the set of space-contraction pyramids a matching step has to be performed (Figure 4a). Each 2 consecutive pyramids (from a chronological perspective) have to be matched, and the vertices that represent the same object/visual entity can be linked by an edge (if it is possible i.e. if the same object/visual entity exists in both structures - existed in both frames). If a certain object/visual entity, that exists in one of the pyramids, does not exists in the other (occlusion, moved out of the field of view, etc.), no connecting edge can be created, thus obtaining a trajectory endpoint. If similar entities disappear and reappear at different time intervals, it will be the job of the reasoning part to decide whether it was the same instance of the same class or not.

The base level of the time-contraction pyramid contains a vertex for each of the frames in the source video, that differ in topological relations from the previous frame. Each vertex will contain the space-contraction pyramid for the region adjacency graph of the respective scene state. These vertices are linked together in a chronological manner i.e. each vertex is linked to the one of the previous and next frames. Also, as a result of the pyramid matching process mentioned before, the vertices from the consecutive space-contraction pyramids are linked together, showing the trajectories of the regions from the first through the last frame. E.g. take the topological descriptions for each frame and represent them in a 3D space, where one of the dimensions is time, and the other 2 are used to represent the planar RAGs. If for every 2 consecutive graphs, the vertices representing the same object/visual entity are linked together by an edge, then following these inter-state connection edges will produce the regions trajectory in 3D space.

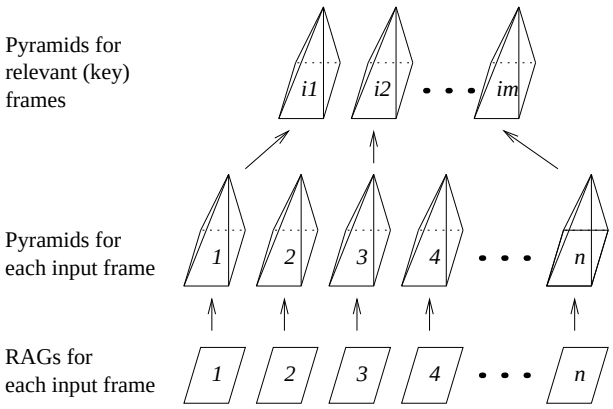

Figure 4: Space time contraction.

Each level of the time-contraction pyramid is a chronologically ordered list of space-contraction pyramids, each element describing the topological relations of a certain scene state. The space-contraction step reduces the spatial information in areas that are not of our interest. The purpose of the time-contraction pyramid is to skip the unnecessary frames caused by the presence of the structured background (which is needed for movement detection using only topological information).

2.2 Temporal contraction followed by spatial contraction

The base level of the time-contraction pyramid contains a vertex for each of the frames in the source video, that differ in topological relations from the previous frame (Figure 5b). Each of these vertices contains the RAG for the respective frames. Through a preliminary process of matching, each vertex in a RAG should be connected with the vertex(vertices), from the two neighboring graphs, that represent the same object/visual entity (if it is possible i.e. if the same object/visual entity exists in the neighboring RAGs frame). In other words, the base level of the pyramid is the discretized evolution of the region adjacency graph of the presented scene with the exception that identical consecutive states are merged into a single state.

If we would represent the base level structure in a N dimensional space (3D for 2D state descriptions + time) we would see that we have obtained curves representing the trajectories of the different regions analyzed. A line segment parallel to the time axis, will denote a static region through the respective time interval. Each level of the pyramid is made out of a sequence of region adjacency graphs. Each vertex in a region adjacency graph should be connected with the vertex(vertices), from the two neighboring graphs, that represent the same object/visual entity. With each new level added to the time contraction pyramid, the number of topological states decreases. After reducing the number of topological states, a contraction of topological information for each state can be considered (at this level the detail regarding the background should not be important any more).

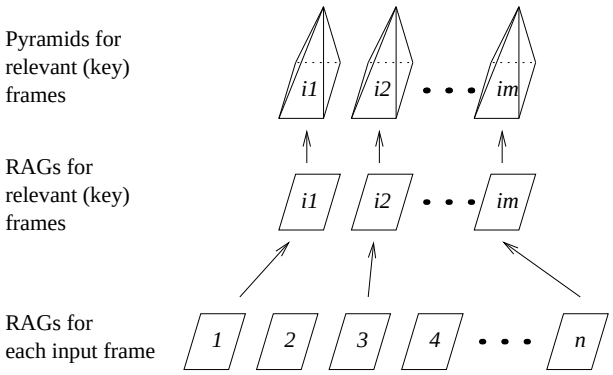

Figure 5: Time space contraction.

There are 2 ways that can be considered for doing this:

- contract each state independently (create a pyramid for each of the topological states at the top level of the time-contraction pyramid)
- contract all the graphs together (allow contraction kernels to span along more than one state graph)

3 Discussion

A simple, human language like description of a scene with two cups and a yellow ball is shown in Tab. 1. Even though the frame numbers are given, they are only for orientation purposes and can be easily eliminated from the description by putting the adverbial for example 'next', 'after that', etc. The previous description would be represented in the following way (Fig. 1a) in the resulting top level of both approaches. The initial configuration contains 3 objects: 2 cups and 1 ball. So we initialize the objects structure with the following: *cup(1)*, *ball* and *cup(2)*. (The numerical ids in parenthesis are present to distinguish the two cups, identification could be done in many other ways. Also in the same interest, vertices are numbered to identify different positions in time.) Vertex(0) in Fig. 1a depicts the initial configuration. The next vertices and edges are as follows: action *move*: creates object *hand* and adds *vertex(1)*; action *grasp*: links to objects *hand* and *cup(1)* and adds *vertex(2)*; action *lift*: links to objects *hand* and *cup(1)* and adds *vertex(3)*; and so on.

4 Conclusion

This paper presents a concept relating story-board description of video sequences with spatio-temporal hierarchies build by local contraction processes of spatio-temporal relations. Since object trajectories are connected curves we identify their ends and junctions and their connections as the high level descriptions.Junction points happen when two (or more) trajectories touch or cross each other, which we interpret as the 'interaction' of two objects. We propose to derive them similar

Frame	Description	Frame	Description
16–21:	hand from left	91:	grasps the same cup again
22:	grasps left cup	87–90:	releases it and moves up and down
27–30:	moves it over ball	85–86:	moves it to the right (but left of the right cup)
31:	releases cup	84:	grasps it
32:	grasp same cup (again)	76–77:	moves to the right cup
33–36:	shifts it to the left	75:	releases it
37:	releases cup	71–74:	shifts it to the right (but still to the left of the right cup
38–40:	moves to right cup	70:	grasps it
41:	grasps right cup	67–69:	moves to the left (most) cup
42–58:	shifts right cup in front of left cup (hiding left cup Fr 46–54) to the left of the original cup	66:	releases it
58:	releases cup	63–65:	shifts it to the right
59–61:	moves to the other cup	62:	grasps it
		...	...

Table 1: Scene description.

to curve pyramid in 2D [7]. For the implementation we plan to use the concept of combinatorial pyramids in 3D [3]. In [10] we used the concept of combinatorial maps in 2D for hierarchical spatial partitioning.

References

[1] J. Allen. An Interval-based Representation of Temporal Knowledge. In *Proc. 7th Inter. Joint Conf. on AI*, p:221–226, 1981.

[2] N. I. Balder. *Temporal Scene Analysis: Conceptual Descriptions of Object Movements*. PhD thesis, University of Toronto, Canada, 1975.

[3] L. Brun and W. G. Kropatsch. The Construction of Pyramids with Combinatorial Maps. Technical Report PRIP-TR-63, Pattern Recognition and Image Processing Group, TU Wien, Austria, 2000.

[4] A. Chella, M. Frixione, and S. Gaglio. Understanding Dynamic Scenes. *Artif. Intell.*, 123:89–132, 2000.

[5] B. Clarke. A Calculus of Individuals Based on Connection. *Notre Dame J. of For. Log.*, 23(3):204–218, 1981.

[6] B. Clarke. Individuals and Points. *Notre Dame J. of For. Log.*, 26(1):61–75,1985.

[7] W. G. Kropatsch. Property Preserving Hiearchical Graph Transformation. In C. Arricelli, L. Cordella, and G. Sanniti di Baja, editors, *Advances in Visual Form Analysis*, p:340–349, Singapore, 1998. World Scientific.

[8] D. Randell, Z. Cui, and A. Cohn. A Spatial Logic Based on Regions and Connection. In *Proc. 3rd Int. Conf. on Knowledge Representation and Reasoning*, p:165–176. Morgan Kaufmann, 1992.

[9] D. Vernon, editor. *A Reasearch Roadmap of Cognitive Vision (DRAFT Version 3.2)*. ECVision: The European Research Network for Cognitive Computer Vision Systems, 2004.

[10] A. Ion, Y. Haxhimusa, W.G. Kropatsch, and L. Brun. Hierarchical Image Partitioning using Combinatorial Maps. In proceedings of the 10th *Computer Vision Winter Workshop* 2005. ÖCG.