Semantic Question-Answering with Video and Eye-Tracking Data: AI Foundations for Human Visual Perception Driven Cognitive Film Studies

Jakob Suchan and **Mehul Bhatt**
Human-Centred Cognitive Assistance Lab. — hcc.uni-bremen.de
Cognitive Vision — www.cognitive-vision.org
University of Bremen, Germany
jsuchan, bhatt}@cognitive-vision.org

Abstract

We present a computational framework for the grounding and semantic interpretation of *dynamic visuo-spatial imagery* consisting of video and eye-tracking data. Driven by cognitive film studies and visual perception research, we demonstrate key technological capabilities aimed at investigating attention & recipient effects vis-a-vis the motion picture; this encompasses high-level analysis of subject's visual fixation patterns and correlating this with (deep) semantic analysis of the dynamic visual data (e.g., fixation on movie characters, influence of cinematographic devices such as *cuts*). The framework and its application as a general AI-based assistive technology platform —integrating vision & KR— for cognitive film studies is highlighted.

1 Introduction

Research in visual perception is predominantly an empirical or evidence-based research initiative aimed at the formation or confirmation of hypotheses, theories etc. In recent years, eye-tracking has emerged as an increasingly powerful means for analysing visual and visuo-locomotive human behaviour in general settings, as well as in specialised areas of everyday life and professional activity. Within eye-tracking based visual perception research, statistical data analytics and complex data visualisation have received significant interest in both academia and industry [Blascheck *et al.*, 2014]; this is typically done in synchrony with manual questionnaire based subject-experimenter interactions, think-aloud protocols etc. As for eye-tracking methodology itself, a key emphasis and primary concern from a technological perspective has been on computational and algorithmic foundations aimed at evaluating the distribution and dynamics of eye-movement patterns [Holmqvist *et al.*, 2011]. Our research extends these lines of work, but is a departure from dominant approaches in its focus on *high-level semantic interpretation, qualitative analysis, and multi-modality* at the interface of AI, HCI, and Visual-Spatial Computing:

▶ *Assistive technologies (applications).* from the applied perspective of human-centred cognitive assistive technologies for evidence-based studies in human perception, we present an AI based computational backbone –encompassing computer vision and KR methods– for next-generation software and services in (eye-tracking driven) visual perception research.

▶ *Integrating Vision and KR.* from the theoretical perspective of vision and KR research, we focus on developing general methods for the intergation of visual processing with (logic-based) declarative reasoning about space and motion in the context of constraint logic programming.

The key emphasis in this paper is on human-centred semantic interpretation and qualitative analysis of multi-modal perceptual data encompassing vision and eye-tracking. Whereas visual perception provides a compelling applied backdrop for the development and demonstration of vision and KR-centric general methods and tools for visuo-spatial computing, the broader orientation of the particular line of research (presented in this paper) is geared toward tighter integration of KR with state of the art in computer vision, contributing to the agenda of what has been attributed as *cognitive vision* at the interface of *language, logic, and artificial intelligence* [Cohn *et al.*, 2003; Vernon, 2008; Bhatt *et al.*, 2013b]. This, we posit, impacts several AI application areas (e.g., vision and robotics) beyond the focus of this paper.

Cognitive Film Studies (CFS) Cognitive studies of *the moving image* —film, digital media etc— has emerged as an area of research at the interface of disciplines as diverse as aesthetics, psychology, neuroscience, film theory, and cognitive science.[1] Within CFS, the role of *mental activity of observers* (e.g., subjects / spectators) has been regarded as one of the most central objects of inquiry [Nannicelli and Taberham, 2014; Aldama, 2015; Sobchack, 2004]. Principal research questions addressed pertain to the systematic study and generation of evidence that can characterise and establish correlates between principles for the synthesis of the moving image, and its cognitive (e.g., embodied visuo-auditory, emotional) recipient effects on observers [Suchan and Bhatt, 2016].

Our technological focus within CFS is on the high-level analysis of subject's visual fixation or saccadic eye-movement patterns whilst watching a film and correlating this

[1] Society for Cognitive Studies of the Moving Image (SCSMI). http://scsmi-online.org.

with semantic analysis of the visuo-auditory data (e.g., fixation on movie characters, influence of cinematographic devices such as *cuts* and sound effects on attention etc).

Integrated Vision and KR for Visual Perception This paper focusses on an integration of computer vision and KR for semantic question answering with video and eye-tracking data in the domain of film. We present a formal model and general methods & tools focussing on (F1–F3):

(F1). **Visual Processing** an integrated pipeline for visual processing of video and eye-tracking data from the viewpoint of high-level feature extraction encompassing spatio-temporal gaze data clustering, people tracking, and (for the film domain) identification of scene structure, camera movements, and character identity.

(F2). **Space - Motion - Histories** a framework for the semantic interpretation of dynamic visuo-spatial imagery encompassing video and eye-tracking data; here, we especially highlight one aspect of the framework concerned with ontologically and computationally elevating perceptual and analytical entities like *moving objects, areas of attention* and *interest, visuo-perceptual saliency, heatmaps* as primitive spatio-temporal objects that can be qualitatively and declaratively reasoned about within constraint logic programming.

(F3). **Semantic Question-Answering** running examples of the underlying constraint logic programming implementation with sample queries in the context of a film & eye-tracking dataset.[2] The examples focus on question-answering pertaining to the *geometry of a scene* [Suchan and Bhatt, 2016] (from a cinematographic viewpoint) in synergy with visual attention predicates related to eye-tracking.

The overall framework (Fig. 2) includes several modules and a pipeline needed for the semantic analysis of visual perception: eye movement and corresponding video datasets are obtained from experiments in visual perception and processed for qualitative spatio-temporal analysis and semantic interpretation. The key modules in the pipeline include the general declarative representations and the inference and query capability based on constraint logic programming. In the backdrop of (F1–F3), we demonstrate the manner in which the integrated visual computing and KR foundations may be applied for the development of human-centred assistive technology supporting high-level interpretation and qualitative analysis. As one instance, we illustrate how results may be used on-demand with question answering, or via a (semantic) database that can be used for applications such as natural language summarisation of experiments.

2 Visual Processing: Perception — Scene Structure

Visuo-spatial semantics for cognitive film studies (from the viewpoint of this paper) include *scene objects* (people, objects in the scene), *cinematographic aids* (camera movement,

[2]Our dataset consists of a total of 31 (eye-tracked) subjects, involving 16 scenes (per subject) from 12 films, with each scene ranging between 0 : 38 minute to max. of 9 : 44 minutes in duration). Eye-movement data is collected using using the Tobii X2-60 Eye Tracker at a rate of 60 Hz.

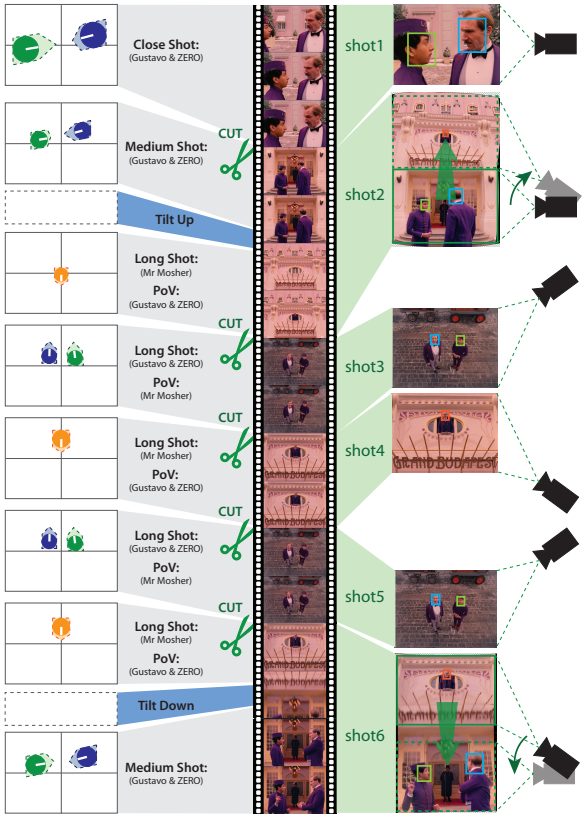

Figure 1: Cinematographic Scene Structure
The Grand Budapest Hotel (2014, Director: Wes Anderson)

shot types, cuts and scene structure), and *perceptual artefacts* (eye-tracking / gaze points, areas of attention). In the following, we summarise the visual processing module(s) of Fig. 2 with respect to the cinematographic scene structure of Fig. 1 and Alg. 1.

Perceptual Artefacts Visual attention may be estimated based on the dynamics and distribution of eye movement data [Holmqvist *et al.*, 2011]. Gaze data can be grouped for an individual, or may be aggregated from multiple subjects, to *Areas of Attention* (AOA), via the calculation of eye movement primitives, e.g. *scan-path* of single spectator including detection of gaze types such as saccadic movement, fixations, smooth pursuit etc; *heat maps* based on aggregate gaze; *clustering* of gaze points. We estimate regions of high attention for a group of people using density based clustering on the gaze points of all participants at a single time point. We also estimate subject attention by calculating a heat map from the gaze points, in a static way, using all gaze points at one time point, and additionally dynamically, using motion compensated gaze points for consecutive time points: (1) estimate the motion in the video data at the position of the gaze point based on Lucas-Kanade *optical flow* [Lucas and Kanade, 1981]; (2) afterwards the heat map is generated by weighted addition of the gaussian for the motion compensated gaze points for n consecutive time points.

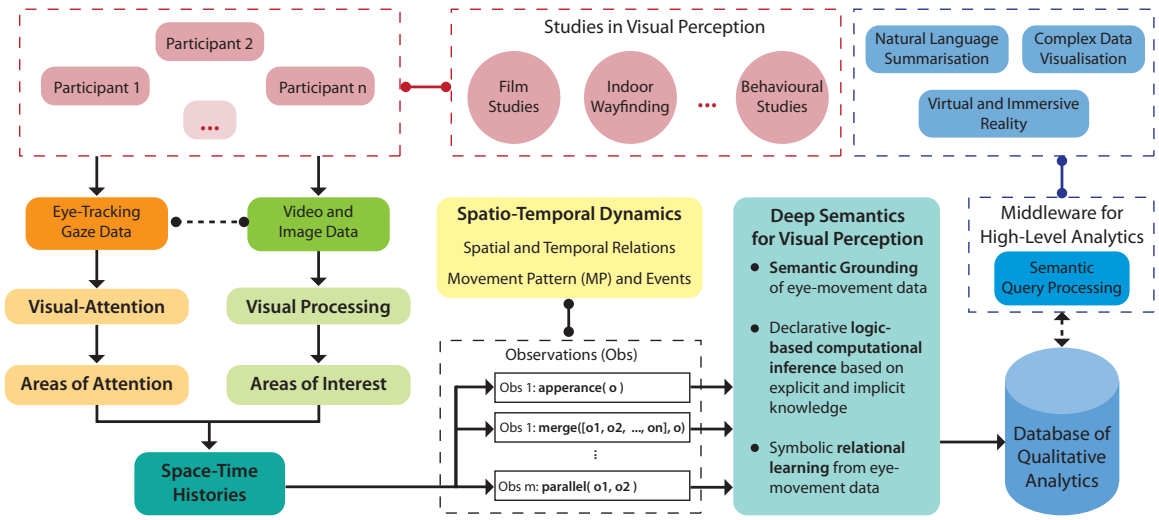

Figure 2: Semantic interpretation — Question Answering — Summarisation for Studies in Visual Perception

Scene Structure Computer vision (CV) research has resulted in a variety of methods for detecting humans, body structure, interactions [Hoai and Zisserman, 2014; Bojanowski *et al.*, 2013; Laptev and Pérez, 2007], as well methods for estimating facing directions [Marin-Jimenez *et al.*, 2014], or recognising the identity of characters in movies [Tapaswi *et al.*, 2012]. The low-level visual processing algorithms that we utilise for high-level semantic analysis are founded in state-of-the-art outcomes for detection and tracking of *people, objects, and motion* [Farnebäck, 2003; Dalal and Triggs, 2005; Felzenszwalb *et al.*, 2010; Rodriguez-Molina and Marin-Jimenez, 2011; Jia *et al.*, 2014].

Analysing the structure of the scene involves **identifying cuts**, i.e., segmenting [Apostolidis and Mezaris, 2014] the scene into its basic elements. This results in single shots, which are used for further cinematographic analysis of the scene. Subsequently, **estimation of camera movement** (i.e., up, down, left, right, forward, backward) is based on Fernaback's dense optical flow [Farnebäck, 2003]; estimating the *horizontal* and *vertical* camera movement is done by calculating the average movement of all sample points in the x and the y direction. For estimating *forward* and *backward* movement, we normalise the direction of movement for each sample point with respect to the centre of the frame and calculate the average movement for the normalised samples. We use *histograms of oriented gradients (HOG)* [Dalal and Triggs, 2005] for **face detection** and *deformable part models* (DPM) [Felzenszwalb *et al.*, 2010; Rodriguez-Molina and Marin-Jimenez, 2011] to **detect people and upper bodies**. For **tracking**, we use *particle filters* for each potential track in the scene. We use optical flow [Lucas and Kanade, 1981] and color histograms to track the movement of the detected entities. Thus, we obtain space-time histories for all detected entities in the scene (Fig. 4, and Alg. 1). Finally, for **character identification**, we use *Convolutional Neural Networks (CNN)* based deep learning as implemented and made available in the Caffe framework [Jia *et al.*, 2014]; we train the network on pictures of the faces of the characters in the movie, to associate the character names to the extracted people tracks, obtained by the detection and tracking algorithms.

3 Space, Motion, Histories

Commonsense spatial, temporal, and spatio-temporal relations and patterns (e.g.,"left", "overlap", "during", "between", "separation", "collision") serve as powerful abstractions for the spatio-linguistic grounding of visual perception and embodied action & interaction [Bhatt *et al.*, 2013a; Suchan *et al.*, 2014]; such spatio-linguistic primitives constitute the basic ontological building blocks of **visuo-spatial computing** in diverse areas, especially those involving the *processing and interpretation* of potentially large volumes of highly *dynamic spatio-temporal data* and commonsense reasoning about space, action, and change [Bhatt, 2012]:

Notation: Spatial and temporal objects may be abstracted with *primitives* such as *regions*, *points*, *oriented points*, *line segments*. We use a first-order language with sorts for: objects: $\mathcal{O} = \{o_1, o_2, ..., o_i\}$; space-time primitives (regions, points etc): $\mathcal{E} = \{\varepsilon_1, \varepsilon_2, ..., \varepsilon_i\}$; time points: $\mathcal{T} = \{t_1, t_2, ..., t_i\}$; 1D intervals: $\Delta = \{\delta_1, \delta_2, ..., \delta_i\}$; fluents: $\Phi = \{\phi_1, \phi_2, ..., \phi_i\}$; actions and events: $\Theta = \{\theta_1, \theta_2, ..., \theta_i\}$. The spatial configuration of objects in the scene is represented using n-ary *spatial relations* $\mathcal{R} = \{r_1, r_2, ..., r_n\}$ of a particular logic of space / time. $\Phi = \{\phi_1, \phi_2, ..., \phi_n\}$ is a set of propositional and functional fluents, e.g. $\phi(\varepsilon_1, \varepsilon_2)$ denotes the spatial relationship between ε_1 and ε_2. We use functions that map from the *object* to the corresponding *spatial primitive* – extend: $\mathcal{O} \times \mathcal{T} \mapsto \varepsilon_\phi$ where $\mathcal{O}$ is the *object* and ε_ϕ is the *spatial primitive* denoting a spatial property of the *object* at time t. Predicates holds-at(ϕ, r, t) and holds-in(ϕ, r, δ) are used to denote that the fluent ϕ has the value r at time t, resp. in time interval δ. Accordingly, we use occurs-at(θ, t), and occurs-in(θ, δ) to denote that an *event* or *action* θ occurred at a *time point* t or in an *interval* δ.

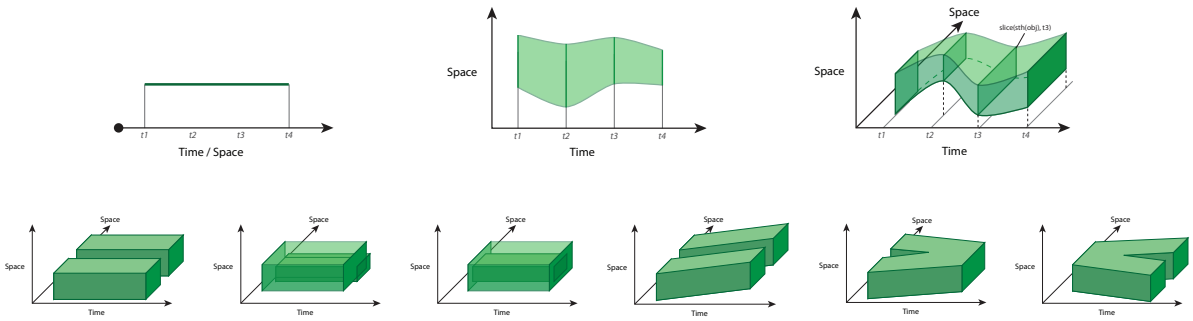

Figure 3: Commonsense Spatial Reasoning with Spatio-Temporal Entities. Illustrated are: Space-Time Histories, and Spatio-Temporal Pattern and Events, i.e. discrete, overlapping, inside, parallel movement, merge, and split

Space and Time Spatial and temporal relations are used to represent the perceived dynamics in a scene. The spatio-temporal domain is modelled using the mereotopological relations of the RCC8 fragment of the RCC calculus [Randell *et al.*, 1992], which consists of the eight base relations $\mathcal{R}_{\text{top}} \equiv \{\text{dc}, \text{ec}, \text{po}, \text{eq}, \text{tpp}, \text{ntpp}, \text{tpp}^{-1}, \text{ntpp}^{-1}\}$, the positional relations using the rectangle algebra which uses the relations of Allen's interval algebra [Allen, 1983] $\mathcal{R}_{\text{interval}} \equiv \{\text{before}, \text{after}, \text{during}, \text{contains}, \text{starts}, \text{started\_by}, \text{finishes}, \text{finished\_by}, \text{overlaps}, \text{overlapped\_by}, \text{meets}, \text{met\_by}, \text{equal}\}$, for representing position for each dimension (horizontal and vertical) separately. We use ordering relations $\{<, =, >\}$ to compare properties of spatial objects, i.e. size and distance. Further, Allen's intervals algebra is used for representing temporal relations between events and actions, where we consider time points to be intervals where the start point is equal to the end point.

Space-Time Histories These are regions in space-time [Muller, 1998] (depicted in Fig. 3). The space-time history sth of an object o is given by the function sth: $\mathcal{O} \mapsto \mathcal{E} \times \mathcal{T}$, which maps the object to its appearance in space and time. $\text{sth}(o, \delta) = (\varepsilon_1, \varepsilon_2, \varepsilon_3, ..., \varepsilon_n)$, where ε_1 to ε_n denotes the spatial primitive representing the object o at the time points t_1 to t_n. Space-time histories serve as basic primitives to represent and reason about the spatio-temporal dynamics in a perceived scene, by defining movement patterns (dynamic spatio-temporal relations), and actions and events, based on the perceived object movement. We define movement relations based on changes in object positions.

$$\begin{aligned}&\text{holds-in}(moving(o), true, \delta) \supset \text{during}(t_i, \delta) \wedge \text{during}(t_j, \delta) \wedge \\ &\quad \text{before}(t_i, t_j) \wedge (\text{position}(o, t_i) \neq \text{position}(o, t_j)).\end{aligned} \tag{1}$$

$$\begin{aligned}&\text{holds-in}(stationary(o), true, \delta) \supset \text{during}(t_i, \delta) \wedge \text{during}(t_j, \delta) \wedge \\ &\quad \text{before}(t_i, t_j) \wedge (\text{position}(o, t_i) = \text{position}(o, t_j)).\end{aligned} \tag{2}$$

Accordingly, *growth* and *shrinkage* of an object is defined based on the changes in size of an object, in one or more dimensions.

$$\begin{aligned}&\text{holds-in}(growing(o), true, \delta) \supset \text{during}(t_i, \delta) \wedge \text{during}(t_j, \delta) \wedge \\ &\quad \text{before}(t_i, t_j) \wedge (\text{size}(o, t_i) < \text{size}(o, t_j)).\end{aligned} \tag{3}$$

$$\begin{aligned}&\text{holds-in}(shrinking(o), true, \delta) \supset \text{during}(t_i, \delta) \wedge \text{during}(t_j, \delta) \wedge \\ &\quad \text{before}(t_i, t_j) \wedge (\text{size}(o, t_i) > \text{size}(o, t_j)).\end{aligned} \tag{4}$$

Algorithm 1: $SceneSemantics(O, PA, \Delta_S)$

Data: Visuo-Spatial input data: $scene\ objects\ (O)$, and $perceptual\ artefacts\ (PA)$ for each time point in $\mathcal{T}$; temporal intervals of detected $shots\ (\Delta_S)$.

Result: Set of Space-Time Histories ($\mathcal{STH}$) which constitute the dynamics of spatial objects in the scene.

```
1  STH_{PA,O} ← ∅
2  for pa ∈ PA do
3      sth_pa ← ∅
4      for t ∈ T do
5          sth_pa ← sth_pa ∪ pa_t
6      STH_PA ← STH_PA ∪ sth_pa
7  for δ ∈ Δ_S do
8      for obj ∈ O do
9          sth_obj ← ∅
10         for t ∈ δ do
11             sth_obj ← sth_obj ∪ extend(obj, t)
12         STH_O ← STH_O ∪ sth_obj
13 STH ← STH_O ∪ STH_PA
14 return STH
```

Movement Pattern (MP) describe spatio-temporal dynamic, by combining arbitrary spatial and temporal relation. The space of possible movement patterns is huge and there are many patterns that are useful to describe visuo-spatial phenomena. E.g. the following pattern describes that one object moves inside another object.

$$\begin{aligned}&\text{holds-in}(\text{inside}(o_i, o_j), true, \delta) \supset \\ &\quad \text{holds-in}(\text{moving}(o_i), true, \delta) \wedge \text{holds-in}(\text{moving}(o_j), true, \delta) \wedge \\ &\quad \text{holds-in}(\phi_{\text{top}}(o_i, o_j), \{tpp, ntpp, eq\}, \delta).\end{aligned} \tag{5}$$

Relative Movement of objects, such as *approaching* and *receding*, is defined based on changes in distance between objects. E.g. *approaching* is defined as follows:

$$\begin{aligned}&\text{holds-in}(\text{approaching}(o_i, o_j), true, \delta) \supset \text{during}(t_i, \delta) \wedge \text{during}(t_j, \delta) \wedge \\ &\quad \text{before}(t_i, t_j) \wedge (distance(o_i, o_j, t_i) > distance(o_i, o_j, t_j)).\end{aligned} \tag{6}$$

Complex movement patterns are defined by combining different spatio-temporal aspect, e.g. a pattern describing that two objects are moving parallel to each other could then be defined as follows:

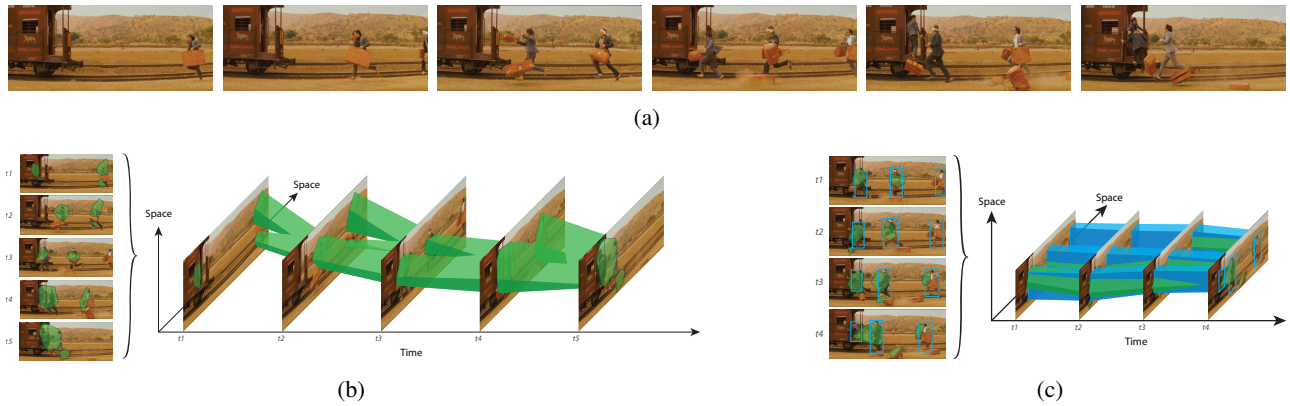

Figure 4: Space-Time Histories: (a) a scene from Darjeeling Limited (2007, Director: Wes Anderson), (b) clustering and association of gaze points, and (c) combined attention clusters and people tracking related by RCC-8 mereotopological primitives

$$\begin{aligned}&\text{holds-in}(\text{parallel}(o_i,o_j),true,\delta) \supset \text{during}(t_i,\delta)\wedge\text{during}(t_j,\delta)\wedge\\&\quad\text{before}(t_i,t_j)\wedge(distance(o_i,o_j,t_i)=distance(o_i,o_j,t_j))\wedge\\&\quad\text{holds-in}(\phi_{\text{top}}(o_i,o_j),dc,\delta).\end{aligned}\tag{7}$$

Actions and Events describe processes that change the spatio-temporal configuration of objects in the scene, at a time point t or in a time interval δ; these are defined by the involved spatio-temporal dynamics in terms of changes in the status of st-histories caused by the action or event, i.e. the description consists of spatio-temporal relations and movementpatterns of the involved st-histories, before, during and after the action or event.

► **Appearance and Disappearance** describes the cases where the existence status of an object changes, i.e. the time point, where the st-history starts to exists, resp. ends to exist.

$$\begin{aligned}&\text{occurs-in}(\text{appearance}(o),\delta)\supset\\&\quad\text{starts}(t_i,\delta)\wedge\text{finishes}(t_j,\delta)\wedge\text{meets}(t_i,t_j)\wedge\\&\quad\text{holds-at}(\text{exists}(o),false,t_i)\wedge\text{holds-at}(\text{exists}(o),true,t_j).\end{aligned}\tag{8}$$

$$\begin{aligned}&\text{occurs-in}(\text{disappearance}(o),\delta)\supset\\&\quad\text{starts}(t_i,\delta)\wedge\text{finishes}(t_j,\delta)\wedge\text{meets}(t_i,t_j)\wedge\\&\quad\text{holds-at}(\text{exists}(o),true,t_i)\wedge\text{holds-at}(\text{exists}(o),false,t_j).\end{aligned}\tag{9}$$

► **Movement Events** describe changes in the spatial state of the space-time histories, due to movement of individuals in the scene, e.g. *crossing* describes the events that two objects, i.e. st-histories of detected persons cross each other. This happens, for example, when the movement of two persons crosses each other.

$$\begin{aligned}&\text{occurs-in}(\text{crossing}(o_i,o_j),\delta)\supset\\&\quad(\text{holds-at}(\phi_{\text{orient}}(o_i,o_j),left,t_i)\wedge\text{holds-at}(\phi_{\text{orient}}(o_i,o_j),right,t_j))\vee\\&\quad(\text{holds-at}(\phi_{\text{orient}}(o_i,o_j),right,t_i)\wedge\text{holds-at}(\phi_{\text{orient}}(o_i,o_j),left,t_j))\wedge\\&\quad\text{starts}(t_i,\delta)\wedge\text{finishes}(t_j,\delta)\wedge\text{meets}(t_i,t_j).\end{aligned}\tag{10}$$

Complex interactions, e.g. a person passing in front, or behind another person, or a person passing between two persons, can be described by combining multiple actions and events. We define a range of actions and events, for describing the dynamics of human interactions, visual attention, and cinematography (Fig. 5).

4 Semantic Question-Answering: Moving Image and its Reception

From the viewpoint of semantic question-answering for the analysis of the visual reception of the moving image, consider the instances in (Q1–Q3) reflecting the kinds of Q/A capabilities necessary from the viewpoint of cognitive film studies:

Q1. how is the spectator attention shifting, when the camera is moving / after a cut / during a long shot?

Q2. which movement / characters / objects is the spectators attention following in a spatio-temporal sense?

Q3. are there individual or aggregate regularities with respect to the shift in spectator attention at a certain time?

As one use-case, consider again the scene depicted in Fig. 4; using our framework, it is possible to define (manually, or using other UI means) high-level rules and execute queries in the logic programming language PROLOG to reason about spectator attention; details follow:

► **Attention Predicates and Queries (sample).** The set of *rules* characterising different kinds of attention and fixation behaviours via-a-vis video analysis is in principle extensive, and open-ended. Some examples include:

- attn_on(Obj, Int) – attention Att is overlapping or covering object Obj during time interval Int
- attn_following(Att, Obj, Int) – attention Att is following the movement of object Obj during time interval Int
- attn_shift(Att, T) – attention Att shifts at time point T
- attn_focusing(Att, Int) – attention Att becomes more focused during the time interval Int

We illustrate some select sample encodings given the backdrop of Q/A needs such as in (Q1–Q3). The following attention predicate is true if the space-time history of an object is topologically connected, i.e. inside or overlapping, with the space-time history of attention.

```
attn_on(Obj, Int) :-  sth(Obj, ST_Obj),
     sth(aggregate_aoa(spectator_set(gp_list)), ST_AOA),
     holds_in(inside(ST_Obj, ST_AOA), Int);
     holds_in(overlapping(ST_AOA, ST_Obj), Int).
```

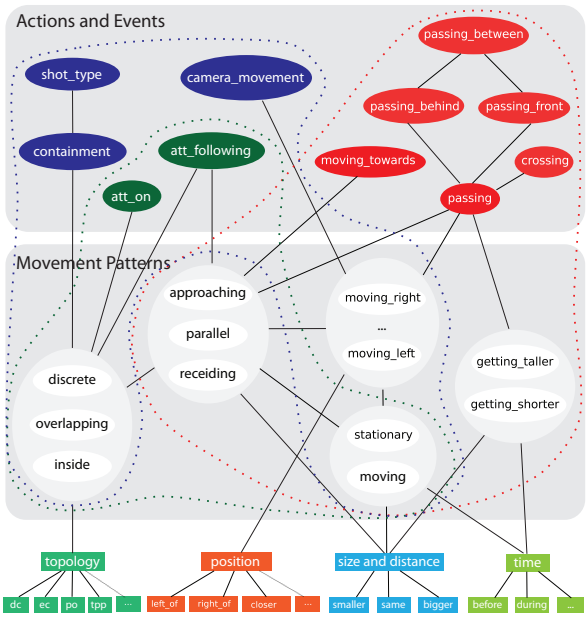

Figure 5: Interaction Taxonomy

Given the above rule, a query where the spatio-temporal history of the character *Jack* is compared with the aggregated Area of Attention of all participants would be the following:

```
?- Int = interval(_, _), attn_on(jack, Int).
```

The query results in all time intervals during which spectator attention is on the character *Jack*:

```
Int = interval(5, 30);
...
```

One could also analyse the dynamics of spectator attention based on movement patterns and events. For instance, consider the st-histories of Fig. 4b: here, a rule determining how the attention follows the objects in the scene is:

```
attn_following(Att, Obj, Int) :-  sth(Obj, ST_Obj),
     sth(aggregate_aoa(spectator_set(gp_list)), Att),
     occures_in(following(Att, ST_Obj), Int).
```

This can be used to query objects the attention is following:

```
?- Int = interval(_, _), attn_following(_, Obj, Int).
```

This results in the objects the attention is following, i.e., the main characters of the scene:

```
Obj = jack,
Int = interval(5, 30);
Obj = francis,
Int = interval(13, 30);
```

```
Obj = peter,
Int = interval(18, 30);
...
```

Further, one could formulate a query to determine **what happened** when the areas of attention following *Jack* and *Francis* merged?

```
?- Int = interval(_, _), TP = timepoint(_),
|    sth(jack, st_jack), sth(francis, st_francis),
|    attn_following(ST_AOA_1, st_jack, _),
|    attn_following(ST_AOA_2, st_francis, _),
|    occures_at(merge([ST_AOA_1, ST_AOA_2], _), TP),
|    occures_in(Obs, Int), time(TP, Int, during).
```

The result of the query is that *Francis* is approaching *Jack* when the respective areas of attention merge:

DARJEELING LIMITED (2007) | VISUAL ATTENTION.
Director. Wes Anderson

This scene involves Francis, Jack, and Peter. The analysis focusses on the influence of CHARACTER MOVEMENT and CAMERA TRACKING on visual fixation.

The scene involves one SHOT with a DOLLY TRACK of the Train from LEFT to RIGHT. DURING the SHOT, Jack enters the scene from the RIGHT APPROACHING TOWARD the Train; THEN Francis enters the scene from the RIGHT APPROACHING TOWARD the Train; THEN Peter enters the scene from the RIGHT APPROACHING TOWARD the Train.

Spectator eye-tracking data suggests fixation on the moving characters, and immediate MOVEMENT of attention to an appearing character.

Sample Analysis of Visual Fixation with Moving Objects (Fig. 4) **L1**

```
Obs = approaching(st_francis, st_jack),
Int = interval(25, 30),
TP = 28;
...
```

Hence, semantic Q/A becomes possible with spatio-temporal entities of visual attention as well as domain-specific perceptual elements; both categories exist as native entities within the (Prolog based) constraint logic programming framework.

Analytical Summarisation The declarative representations and the inference and query capability provided by the framework (Fig. 2) can be used as a basis for (language-based) analytical summarisation. Listing L1 is a select part of a summary corresponding to the scene in (Fig. 4); the summary has been generated using a (spatio-temporal feature based) natural language generator.[3] Note that the semantics for spatial, temporal, and behavioural information is grounded to relations in the underlying theory of space and motion. This manner of natural language based analytical summarisation of experiments –to the best of our knowledge– presents a novel user interaction paradigm and functional benchmark in visual perception research.

5 Summary

We presented a **visuo-spatial computing** framework consisting of integrated formal KR and low-level visual processing foundations, including the algorithms & data-structures, and resulting general methods & tools that serve as the computational backbone for next-generation software and services aimed at semantic interpretation and qualitative analytics (for visual perception studies). As examples, we focused on the capability to perform semantic Q/A about the dynamics of space-time histories and their mutual interactions within (constraint) logic programming.

This work is driven by a tighter integration of KR and computer vision; **cognitive vision** as an area of research has gained prominence, with recent initiatives addressing the topic from the perspectives of language, logic, and AI. There has also been recent interest from the computer vision community to synergise with cognitively motivated methods for perceptual grounding and inference with visual imagery. We posit that KR+Vision can serve a crucial role for the development of hybrid AI & cognitive interaction technologies where processing and human-centred semantic interpretation of dynamic visuo-spatial imagery are central.

[3] NLG [Reiter and Dale, 2000] is beyond the scope of this paper; we have used the specialised (PROLOG based) NL generator provided by [Suchan *et al.*, 2015].

References

[Aldama, 2015] Frederick Luis Aldama. The Science of Storytelling: Perspectives from Cognitive Science, Neuroscience, and the Humanities. *Projections*, 9(1):80–95, 2015.

[Allen, 1983] James F. Allen. Maintaining knowledge about temporal intervals. *Commun. ACM*, 26(11):832–843, 1983.

[Apostolidis and Mezaris, 2014] Evlampios E. Apostolidis and Vasileios Mezaris. Fast shot segmentation combining global and local visual descriptors. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014, Florence, Italy, May 4-9, 2014*, pages 6583–6587. IEEE, 2014.

[Bhatt *et al.*, 2011] Mehul Bhatt, Jae Hee Lee, and Carl Schultz. CLP(QS): A Declarative Spatial Reasoning Framework. In *Proceedings of the 10th International conference on Spatial information theory*, COSIT'11, pages 210–230, Berlin, Heidelberg, 2011. Springer-Verlag.

[Bhatt *et al.*, 2013a] Mehul Bhatt, Carl Schultz, and Christian Freksa. The 'Space' in Spatial Assistance Systems: Conception, Formalisation and Computation. In, *Representing space in cognition: Interrelations of behavior, language, and formal models. Series: Explorations in Language and Space*, Explorations in Language and Space. 978-0-19-967991-1, Oxford University Press, 2013.

[Bhatt *et al.*, 2013b] Mehul Bhatt, Jakob Suchan, and Carl Schultz. Cognitive Interpretation of Everyday Activities – Toward Perceptual Narrative Based Visuo-Spatial Scene Interpretation. In, *Computational Models of Narrative (CMN) 2013., a satellite workshop of CogSci 2013: The 35th meeting of the Cognitive Science Society.*, Dagstuhl, Germany, 2013. (OASIcs).

[Bhatt, 2012] Mehul Bhatt. Reasoning about Space, Actions and Change: A Paradigm for Applications of Spatial Reasoning. In *Qualitative Spatial Representation and Reasoning: Trends and Future Directions*. IGI Global, USA, 2012.

[Blascheck *et al.*, 2014] T. Blascheck, K. Kurzhals, M. Raschke, M. Burch, D. Weiskopf, and T. Ertl. State-of-the-Art of Visualization for Eye Tracking Data. pages 63–82, Swansea, UK, 2014. Eurographics Association.

[Bojanowski *et al.*, 2013] Piotr Bojanowski, Francis Bach, Ivan Laptev, Jean Ponce, Cordelia Schmid, and Josef Sivic. Finding actors and actions in movies. In *Proc. ICCV*, 2013.

[Cohn *et al.*, 2003] Anthony G. Cohn, Derek R. Magee, Aphrodite Galata, David C. Hogg, and Shyamanta M. Hazarika. Towards an architecture for cognitive vision using qualitative spatio-temporal representations and abduction. In *Spatial Cognition III, Routes and Navigation, Human Memory and Learning, Spatial Representation and Spatial Learning*, pages 232–248, 2003.

[Dalal and Triggs, 2005] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE (CVPR 2005), 20-26 June 2005, San Diego, CA, USA*, pages 886–893. IEEE Computer Society, 2005.

[Farnebäck, 2003] Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Proceedings of the 13th Scandinavian Conference on Image Analysis*, SCIA'03, pages 363–370, Berlin, Heidelberg, 2003. Springer-Verlag.

[Felzenszwalb *et al.*, 2010] Pedro F. Felzenszwalb, Ross B. Girshick, David A. McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(9):1627–1645, 2010.

[Hoai and Zisserman, 2014] M. Hoai and A. Zisserman. Talking heads: Detecting humans and recognizing their interactions. In *IEEE CVPR*, 2014.

[Holmqvist *et al.*, 2011] Kenneth Holmqvist, Marcus Nystrom, Richard Andersson, Richard Dewhurst, Halszka Jarodzka, and Joost van de Weijer. *Eye Tracking. A comprehensive guide to methods and measures*. Oxford University Press, 2011.

[Jia *et al.*, 2014] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[Laptev and Pérez, 2007] Ivan Laptev and Patrick Pérez. Retrieving actions in movies. In *IEEE 11th ICCV 2007, Rio de Janeiro, Brazil, October 14-20, 2007*, pages 1–8. IEEE, 2007.

[Lucas and Kanade, 1981] Bruce D. Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. pages 674–679, 1981.

[Marin-Jimenez *et al.*, 2014] M. Marin-Jimenez, A. Zisserman, and V. Ferrari. Detecting people looking at each other in videos. *IJCV*, 106(3):282–296, feb 2014.

[Muller, 1998] Philippe Muller. A qualitative theory of motion based on spatio-temporal primitives. In: *KR 98, Trento, Italy, June 2-5, 1998*, pages 131–143. Morgan Kaufmann, 1998.

[Nannicelli and Taberham, 2014] Ted Nannicelli and Paul Taberham. Contemporary cognitive media theory. In, *Cognitive Media Theory*, AFI Film Readers. Routledge, 2014.

[Randell *et al.*, 1992] David A. Randell, Zhan Cui, and Anthony Cohn. A spatial logic based on regions and connection. In *KR'92.*, pages 165–176. Morgan Kaufmann, California, 1992.

[Reiter and Dale, 2000] Ehud Reiter and Robert Dale. *Building Natural Language Generation Systems*. Cambridge University Press, Cambridge, U.K., 2000.

[Rodriguez-Molina and Marin-Jimenez, 2011] Daniel Rodriguez-Molina and Manuel J. Marin-Jimenez. LibPaBOD: A library for part-based object detection in C++, 2011. Software available at http://www.uco.es/ in1majim/.

[Sobchack, 2004] Vivian Sobchack. *Carnal Thoughts: Embodiment and Moving Image Culture*. University of California Press, November 2004.

[Suchan and Bhatt, 2016] Jakob Suchan and Mehul Bhatt. The Geometry of a Scene: On Deep Semantics for Visual Perception Driven Cognitive Film Studies. In *WACV 2016: IEEE Winter Conference on Applications of Computer Vision.* IEEE, 2016.

[Suchan *et al.*, 2014] Jakob Suchan, Mehul Bhatt, and Paulo E. Santos. Perceptual narratives of space and motion for semantic interpretation of visual data. In: *Computer Vision - ECCV 2014 Workshops - Zurich, Switzerland, September 6-7 and 12, 2014, Proceedings, Part II*, volume 8926 of *LNCS*, pages 339–354. Springer, 2014.

[Suchan *et al.*, 2015] Jakob Suchan, Mehul Bhatt, and Harshita Jhavar. Talking about the moving image: A declarative model for image schema based embodied perception grounding and language generation. *CoRR*, abs/1508.03276, 2015. http://arxiv.org/abs/1508.03276.

[Tapaswi *et al.*, 2012] Makarand Tapaswi, Martin Bäuml, and Rainer Stiefelhagen. "Knock! Knock! Who is it?" Probabilistic Person Identification in TV Series. In *IEEE CVPR*, Jun. 2012.

[Vernon, 2008] David Vernon. Cognitive vision: The case for embodied perception. *Image Vision Comput.*, 26(1):127–140, 2008.