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, bhett]@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 integration 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 With 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

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. 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,

---

2Our 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 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). 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.
Scene Structure  Computer vision (CV) research has resulted in varieties of methods for detecting humans, body structure, interactions [Hoai and Zisserman, 2014; Bojanowski et al., 2013; Laptev and Perez, 2007], as well methods for estimating facing directions [Mammele 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 the visual data fundamental optical flow [Farnebäck, 2003]; estimating the horizontal and vertical camera movement is done by calculating the average movement of all points in the x and 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 (DFM) [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: \( O = \{ o_1, o_2, \ldots, o_n \} \); space-time primitives (regions, points etc): \( E = \{ e_1, e_2, \ldots, e_m \} \); time points: \( T = \{ t_1, t_2, \ldots, t_l \} \); 1D intervals: \( \Delta = \{ \delta_1, \delta_2, \ldots, \delta_k \} \); fluenets: \( \Phi = \{ \phi_1, \phi_2, \ldots, \phi_l \} \); actions and events: \( \Theta = \{ \theta_1, \theta_2, \ldots, \theta_1 \} \). The spatial configuration of objects in the scene is represented using n-ary spatial relations \( R = \{ r_1, r_2, \ldots, r_n \} \) of a particular logic of space / time. \( \Phi = \{ \phi_1, \phi_2, \ldots, \phi_n \} \) is a set of propositional and functional fluenets, e.g. \( \phi_1(e_1, e_2) \) denotes the spatial relationship between \( e_1 \) and \( e_2 \). We use functions that map from the object to the corresponding spatial primitive — extend: \( O \times T \rightarrow E_\phi \) where \( O \) is the object and \( E_\phi \) is the spatial primitive denoting a spatial property of the object at time \( t \). Predicates \( \text{holds-at}(\phi, r, t) \) and \( \text{holds-in}(\phi, r, \delta) \) are used to denote that the fluent \( \phi \) has the value \( r \) at time \( t \), resp. in time interval \( \delta \). Accordingly, we use \( \text{occurs-at}(\theta, t) \), and \( \text{occurs-in}(\theta, \delta) \) to denote that an event or action \( \theta \) occurred at a time point \( t \) or in an interval \( \delta \).
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 $R_{tpp} \equiv \{dc, ec, po, eq, tpp, ntpp, tpp^{-1}, ntpp^{-1}\}$, the positional relations using the rectangle algebra which uses the relations of Allen’s interval algebra [Allen, 1983] $R_{interval} \equiv \{before, after, during, contains, starts, by, finishes, finished, by, overlaps, overlapped, by, meets, met, by, equal\}$, 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 temporal 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 $sth: O \rightarrow E \times T$, which maps the object to its appearance in space and time. $sth(o, \delta) = (\varepsilon_1, \varepsilon_2, \varepsilon_3, ..., \varepsilon_n)$, where $\varepsilon_1$ to $\varepsilon_n$ 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.

1. \begin{align*}
  \text{holds-in}(moving(o, true, \delta) \supset (t_1, \delta) \land (t_2, \delta) \land before(t_1, t_2) \land (\text{position}(o, t_1) \neq \text{position}(o, t_2))).
\end{align*}

2. \begin{align*}
  \text{holds-in}(stationary(o, true, \delta) \supset (t_1, \delta) \land (t_2, \delta) \land before(t_1, t_2) \land (\text{position}(o, t_1) = \text{position}(o, t_2))).
\end{align*}

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

1. \begin{align*}
  \text{holds-in}(growth(o, true, \delta) \supset (t_1, \delta) \land (t_2, \delta) \land before(t_1, t_2) \land (\text{size}(o, t_1) < \text{size}(o, t_2))).
\end{align*}

2. \begin{align*}
  \text{holds-in}(shrinkage(o, true, \delta) \supset (t_1, \delta) \land (t_2, \delta) \land before(t_1, t_2) \land (\text{size}(o, t_1) > \text{size}(o, t_2))).
\end{align*}

Movement Pattern (MP) describe spatio-temporal dynamics, 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. following pattern describes that one object moves inside another object.

1. \begin{align*}
  \text{holds-in}(\text{inside}(o_1, o_2, true, \delta) \supset \text{holds-in}(\text{inside}(o_1, o_2, true, \delta) \land \text{holds-in}(\text{inside}(o_1, o_2, true, \delta) \land before(t_1, t_2) \land (\text{distance}(o_1, o_2, t_1) > \text{distance}(o_1, o_2, t_2))).
\end{align*}

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:

1. \begin{align*}
  \text{holds-in}(\text{approaching}(o_1, o_2, true, \delta) \supset \text{holds-in}(\text{approaching}(o_1, o_2, true, \delta) \land before(t_1, t_2) \land (\text{distance}(o_1, o_2, t_1) > \text{distance}(o_1, o_2, t_2))).
\end{align*}

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 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](image)
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 \( \delta \); 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 movement-patterns 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.

\[
\text{occurs-in(appearance}(o, \delta)) \supset \ \text{starts}(t_i, \delta) \land \text{finishes}(t_j, \delta) \land \text{meets}(t_i, t_j) \\
\text{holds-at(exists}(o), false, t_i) \land \text{holds-at(exists}(o), true, t_j).
\]

- **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.

\[
\text{occurs-in(crossing}(a, o_j)) \supset \ \text{holds-at}(\phi_{\text{cross}}(a, o_j), left, t_i) \land \text{holds-at}(\phi_{\text{cross}}(a, o_j), right, t_j) \lor \\
\text{holds-at}(\phi_{\text{cross}}(o_j, a), left, t_i) \land \text{holds-at}(\phi_{\text{cross}}(o_j, a), right, t_j) \land \\
\text{starts}(t_i, \delta) \land \text{finishes}(t_j, \delta) \land \text{meets}(t_i, t_j).
\]

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

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.

\[
\begin{align*}
\text{attn_on(Obj, Int):} & \quad \text{sth(Obj, ST_Obj),} \\
& \quad \text{sth(aggregate_aoa(spectator_set(gp_list)), ST_AOA),} \\
& \quad \text{holds_in(inside(ST_Obj, ST_AOA), Int),} \\
& \quad \text{holds_in(overlapping(ST_AOA, ST_Obj), Int).}
\end{align*}
\]
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:

\[-\text{Int} = \text{interval}(5, 30), \text{attn\_on}(\text{jack}, \text{Int})\]

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

\[-\text{Int} = \text{interval}(5, 30); \ldots\]

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:

\[-\text{attn\_following}(\text{Att}, \text{Obj}, \text{Int}) -\ \text{sth}(\text{Obj}, \text{ST\_Obj}),
\text{sth}(\text{aggregate\_aoa}(\text{spectator\_set}(\text{obj\_list})), \text{Att}),
\text{occurs\_in}(\text{following}(\text{Att}, \text{ST\_Obj}), \text{Int}).\]

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

\[-\text{Int} = \text{interval}(5, 30), \text{attn\_following}(\text{Obj}, \text{Int})\]

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

\[-\text{Obj} = \text{jack}, \text{Int} = \text{interval}(5, 30); \text{Obj} = \text{francis}, \text{Int} = \text{interval}(13, 30); \ldots\]

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

\[-\text{Int} = \text{interval}(5, 30), \text{TP} = \text{timepoint}(5),
\text{sth}(\text{jack}, \text{st\_jack}), \text{sth}(\text{francis}, \text{st\_francis}),
\text{occurs\_in}(\text{following}(\text{ST\_AOA\_1}, \text{st\_jack}), \text{TP}),
\text{occurs\_in}(\text{following}(\text{ST\_AOA\_2}, \text{st\_francis}), \text{TP}),
\text{occurs\_at}(\text{merge}(\text{ST\_AOA\_1}, \text{ST\_AOA\_2}), \text{time}(\text{TP}, \text{Int}, \text{during})).\]

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

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. 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.

---

3NLG [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


