THE GEOMETRY OF A SCENE

On Deep Semantics for Visual Perception Driven Cognitive Film Studies

Jakob Suchan   Mehul Bhatt

Human-Centred Cognitive Assistance Lab. | hcc.uni-bremen.de/
University of Bremen, Germany
{jsuchan, bhatt}@uni-bremen.de

Abstract

We present a general computational narrative model encompassing primitives of space, time, and motion from the viewpoint of deep knowledge representation and reasoning about visuo-spatial dynamics, and (eye-tracking based) visual perception of the moving image. The declarative model, implemented within constraint logic programming, integrates knowledge-based qualitative reasoning (e.g., about object / character placement, scene structure) with state of the art computer vision methods for detecting, tracking, and recognition of people, objects, and cinematographic devices such as cuts, shot types, types of camera movement. A key feature is that primitives of the theory —things, time, space and motion predicates, actions and events, perceptual objects (e.g., eye-tracking / gaze points, regions of attention etc)— are available as first-class objects with deep semantics suited for inference and query from the viewpoint of analytical Q&A or studies in visual perception.

We present the formal framework and its implementation in the context of a large-scale experiment concerned with analysis of visual perception and reception of the moving image in the context of cognitive film studies.

1. INTRODUCTION

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],[2] Within cognitive film theory, the role of *mental activity of observers* (e.g., subjects / spectators, analysts / critics) has been regarded as one of the most central objects of inquiry [2, 29]. Principal research questions that emerge in the context of cognitive film theory pertain to the systematic study and generation of evidence that can characterise and establish strong correlates between principles for the *synthesis of the moving image* (Listing L1), and its cognitive (e.g., embodied visual, auditory, aesthetic, emotional) recipient effects and influences on observers.

DRIVE (2011) | QUADRANT SYSTEM. VISUAL ATTENTION.
Director. Nicolas Winding Refn

This scene, involving The Driver (*Ryan Gosling*) and Irene (*Carey Mulligan*), adopts a TOP-BOTTOM and LEFT-RIGHT quadrant system that is executed in a SINGLE TAKE / without any CUTS

The CAMERA MOVES BACKWARD tracking the movement of The Driver and Irene; DURING MOVEMENT_1, Irene OCCUPIES the right half, WHILE The Driver OCCUPIES the LEFT half

Spectator eye-tracking data suggests that the audience is repeatedly switching their attention between the LEFT and RIGHT half, with a majority of the audience fixating visual attention on Irene as she MOVES into an extreme CLOSE-UP SHOT

Credit. Quadrant system method based on study by film analyst Tony Zhou. L1

Visual Semantics of the Moving Image Driven by cognitive studies of cinema, and cognitive film theory in particular, we interpret *the moving image* in a broad sense to encompass: *multi-modal* visuo-auditory perceptual signals (also including depth sensing, haptics, and empirical observational data) where basic concepts of semantic or content level coherence, and spatio-temporal continuity and narrativity are applicable. With this as a basis, this paper focusses on methods for investigating the visuo-spatial semantics of the moving image at the interface of artificial intelligence based *spatial representation and reasoning*, *visuo-spatial cognition*, and *computational models of narrative*. In particular, we develop and demonstrate foundational methods focussing on cognitively-driven qualitative analysis of dynamic visuo-spatial imagery encompassing:

- **geometry of a scene**. content-level deep semantic analysis of scene structure and semantics —object / character identity and placement, cuts, shot types, categories of camera movement— pertaining to the moving image
- **perception & reception**. visual perception analysis of spectator behaviour and engagement with the medium, e.g., visual fixation on film characters, gaze patterns

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

[2] PROJECTIONS: The Journal for Movies and Mind. www.berghahnbooks.com/journals/proj/

Film / Director	Scenes	Duration (minutes)
Jaws (1975) Steven Spielberg	1	4:36
The Untouchables (1987) Brian De Palma	1	9:44
Paprika (2006) Satoshi Kon	1	1:49
Grand Budapest Hotel (2014) Wes Anderson	2	{1:41, 4:17}
Moonrise Kingdom (2012) Wes Anderson	1	1:56
Darjeeling Limited (2007) Wes Anderson	1	1:25
The Hunger Games (2012) Gary Ross	1	2:48
Solaris (1972) Andrei Tarkovsky	1	7:46
The Shining (1980) Stanley Kubrick	2	{2:26, 0:38}
Drive (2011) Nicolas Winding Refn	3	{2:59, 0:51, 1:59}
The Bad Sleep Well (1960) Akira Kurosawa	1	2:46
Goodfellas (1990) Martin Scorsese	1	3:03
Total (per subject)	16	50:44

Table 1: Experiments in Deep Semantics and Eye-Tracking Based Visual Perception. (case-study developed in this paper is part of this experiment with 31 subjects)

co-related with influence of cinematographic aids such as *cuts*, *long takes*, *symmetry* on attention and whilst watching a film

Our research addresses *space* and *spatio-temporal dynamics* from the viewpoint formal representation and computational reasoning about space, events, actions, and change, especially focussing on space and motion as interpreted within artificial intelligence and knowledge representation and reasoning (KR) in general, and *declarative spatial reasoning* [5, 37] in particular.

Declarative Narrativisation and Deep Semantics With respect to a broad-based understanding of the moving image (as aforediscussed), we define dynamic visuo-spatial *perceptual narratives* as declarative models of visual, auditory, haptic and other (e.g., qualitative, analytical) observations in the real world that are obtained via artificial sensors and / or human input. Deep semantics denotes the existence of declaratively grounded models (e.g., for spatial and temporal knowledge) and systematic formalisation that can be used to perform reasoning and query answering, relational learning, or more broadly, even embodied simulation.[3] Deep semantics, founded on declarative representation and inference, serves as basis to externalise explicit inferred knowledge, e.g., using modalities such as diagrammatic representations (e.g., Fig. 3), natural language (e.g., Listing L1), complex (dynamic) data visualisation (e.g., Fig. 2) etc.

Evaluation: An Experimental Case-Study We demonstrate the model by its application to the domain of cognitive film studies for analysing visual experience combining deep visual analysis of the "geometry of a scene" with analysis of eye movement behaviour. Examples and empirical evaluation are presented in the context of large-scale experiment with a total of 31 subjects, and involving 16 scenes (per subject) from 12 films, with each scene ranging between $0:38$ minute to max. of $9:44$ minutes in duration) (Table 1).[4]

Core Contributions

We present a computational narrative model for performing Q/A centered deep semantic analysis of the geometry —structure and semantics— of the moving image and its visual perception and reception by the audience:

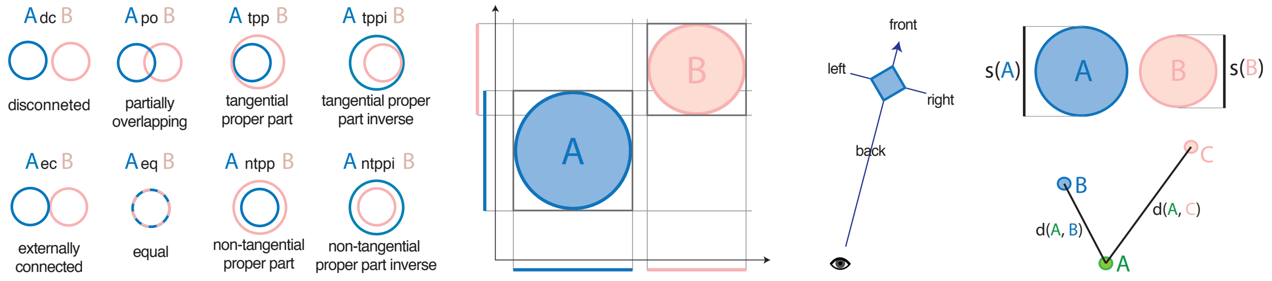
(1). Space & Motion a domain-independent formal framework encompassing primitives of space, time, and motion for commonsense representing and reasoning about dynamic visuo-spatial imagery. The framework is founded in logic programming such that a corresponding implementation is seamlessly usable as a generic library of space & motion via declarative programming frameworks based on logic programming (e.g., Prolog based CLP(QS) [5]) and answer-set programming (e.g., ASPMT(QS) [37]).

(2). Commonsense Cognitive Vision integration of the formal KR-based commonsense theory of space, time and motion with state of the art computer vision methods that have been customised herein for the film domain. This encompasses detection, tracking, and recognition of people, objects, cinematographic devices such as (camera) motion, cuts, shot types, object / character placement & scene structure. Whereas our application of state of the art computer vision is film domain specific, the integration with KR methods serves as a model for other areas in AI, e.g., vision & robotics, where commonsense reasoning about space and motion is crucial.

(3). Implementation The proposed framework has been fully modelled and implemented declaratively within constraint logic programming (CLP). We emphasize that the level of declarativeness within logic programming is such that each aspect pertaining to the overall framework can be

[3] Whereas this paper alludes to logic programming, the broader agenda of "deep semantics" indeed relates to "deep KR" also encompassing other declarative KR frameworks such as description logic based (spatio-terminological) reasoning, answer-set programming based non-monotonic (spatial) reasoning, or even other specialised commonsense reasoners based on expressive action description languages for handling *space, events, action, and change*.

[4] We conducted the experiments with the stationary Tobii X2-60 Eye Tracker, collecting eye movement data with a rate of 60 Hz.



(a) (b) (c) (d)

Figure 1: Spatial Relations: (a) Region Connection Calculus (RCC-8), (b) 2-Dimensional Position using Rectangle Algebra (RA), (c) 2-Dimensional Intrinsic Orientation, and (d) Size and Distance

seamlessly customised and elaborated, and that question-answering & query can be performed using the primitives of the theory —*things, space and motion, actions and events, perceptual objects* (e.g., eye-tracking / gaze points, regions of attention etc)— as first class objects within the CLP environment.

2. 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 [6]. 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*: architecture design [7], geographic information systems [10], cognitive vision and robotics [8, 32, 34]. [5]

The high-level semantic interpretation and qualitative analysis of visual attention in the context of visual perception studies requires the representational and inferential mediation of (declarative) qualitative abstractions of the visuo-spatial dynamics, encompassing *space, time, motion, and interaction*. We use a first-order typed language ($\mathcal{L}$) with the following alphabet: $\{\neg, \wedge, \vee, \forall, \exists, \supset, \equiv\}$ (respectively meaning negation, conjunction, disjuction, universal quantification, existential quantification, implication, and equivalence).

Notation: Spatial and temporal objects may be abstracted with *primitives* such as *regions*, *points*, *oriented points*, *line segments* as per needs. We use a first-order language with sorts for: objects: $\mathcal{O} = \{o_1, o_2, ..., o_i\}$; spacetime 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, t) are used to denote that the fluent ϕ has the value r at time t. 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.

Space and Time

Spatial and temporal relations (Fig. 1) are used to represent the perceived dynamics in a scene. The spatio-temporal domain is modelled using the topological relations of the RCC8 fragment of the RCC calculus [30] (Fig. 1a), which consists of the eight base relations $\mathcal{R}_{\mathsf{top}} \equiv \{$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 [3] $\mathcal{R}_{\mathsf{interval}} \equiv \{$before, after, during, contains, starts, started_by, finishes, finished_by, overlaps, overlapped_by, meets, met_by, equal$\}$, for representing position for each dimension (horizontal and vertical) separately (Fig. 1b). We use ordering relations $\mathcal{R}_{\mathsf{ord}} \equiv \{<, =, >\}$ to compare properties of spatial objects, i.e. size and distance. Further, we also use Allen’s intervals 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, i.e. $t = interval(t, t)$. [6]

[5] Knowledge representation and reasoning about *space* may be formally interpreted within diverse frameworks such as: (a) geometric reasoning & constructive (solid) geometry [22]; (b) relational algebraic semantics of ‘qualitative spatial calculi’ [25]; and (c) axiomatically constructed formal systems of mereotopology and mereogeometry [1]. Within a mereotopological theory such as the Region Connection Calculus (RCC) [30] (Fig. 1a), primitive entities are arbitrarily dimensioned extended regions in space, time, space-time, e.g., 4D spatio-temporal histories [18, 28].

[6] These characterisations are sufficient for the examples of this paper; an elaborate set of spatio-temporal relations to represent and reason about

Space-Time Histories

These are defined as regions in space-time. The space-time history sth of an object o is given by the function $sth : \mathcal{O} \mapsto \mathcal{S} \times \mathcal{T}$, which maps the object to its appearance in space and time. For representing connectedness of space-time histories, we appeal to spatial and temporal connectedness (s-connected and t-connected) [18, 28]. If two space-time histories are connected in space and time we say they are st-connected. 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.

Movement Pattern (MP) describe spatio-temporal dynamic, by combining relations, $MP = r_1 \times r_2 \times ... \times r_i$ where $r_i \in \mathcal{R}$ for 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.

$$\begin{aligned}&\mathsf{holds\text{-}in}(\mathsf{inside}(o_1, o_2), true, \delta_1) \supset \\ &\quad \mathsf{holds\text{-}in}(\phi_{\mathsf{top}}(o_1, o_2), \{tpp, ntpp, eq\}, \delta_1).\end{aligned} \tag{1}$$

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:

$$\begin{aligned}&\mathsf{holds\text{-}in}(\mathsf{approaching}(o_1, o_2), true, \delta_1) \supset (\forall t_1, t_2 \in \delta_1 \wedge t_1 < t_2) \\ &\quad \mathsf{holds}(\phi_{\mathsf{ord}}(at(dist(o_1, o_2), t_1), at(dist(o_1, o_2), t_2)), >).\end{aligned} \tag{2}$$

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

$$\begin{aligned}&\mathsf{holds\text{-}in}(\mathsf{parallel}(o_1, o_2), true, \delta_1) \supset (\forall t_1, t_2 \in \delta_1 \wedge t_1 < t_2) \\ &\quad \mathsf{holds}(\phi_{\mathsf{ord}}(at(dist(o_1, o_2), t_1), at(dist(o_1, o_2), t_2)), =) \wedge \\ &\quad \mathsf{holds\text{-}in}(\phi_{\mathsf{top}}(o_1, o_2), dc, \delta_1).\end{aligned} \tag{3}$$

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, ends to exist.

$$\begin{aligned}&\mathsf{occurs\text{-}at}(\mathsf{appearance}(o), true, t) \supset \mathsf{holds\text{-}at}(\mathsf{exists}(o), false, t_{prev}) \wedge \\ &\quad \mathsf{holds\text{-}at}(\mathsf{exists}(o), true, t) \wedge \mathsf{holds}(\mathsf{meets}(t_{prev}, t), true)\end{aligned} \tag{4}$$

$$\begin{aligned}&\mathsf{occurs\text{-}at}(\mathsf{disappearance}(o), true, t) \supset \mathsf{holds\text{-}at}(\mathsf{exists}(o), true, t_{prev}) \wedge \\ &\quad \mathsf{holds\text{-}at}(\mathsf{exists}(o), false, t) \wedge \mathsf{holds}(\mathsf{meets}(t_{prev}, t), true)\end{aligned} \tag{5}$$

▸ **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.

$$\begin{aligned}&\mathsf{occurs\text{-}at}(\mathsf{crossing}(o_1, o_2), true, t) \supset \\ &\quad (\mathsf{holds\text{-}at}(\phi_{\mathsf{orient}}(o_1, o_2), left, t_{prev}) \wedge \mathsf{holds\text{-}at}(\phi_{\mathsf{orient}}(o_1, o_2), right, t)) \vee \\ &\quad (\mathsf{holds\text{-}at}(\phi_{\mathsf{orient}}(o_1, o_2), right, t_{prev}) \wedge \mathsf{holds\text{-}at}(\phi_{\mathsf{orient}}(o_1, o_2), left, t))\end{aligned} \tag{6}$$

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. E.g. consider the cinematographic device of a *Tracking Shot* describes the action, that the camera is tracking the movement of some objects in the scene.

$$\begin{aligned}&\mathsf{occures\text{-}in}(tracking(\mathsf{cam}_1, [\mathsf{o}_1, \mathsf{o}_2, ..., \mathsf{o}_i]), true, \delta_1) \supset \\ &\quad \bar{\mathsf{o}} = [\mathsf{o}_1, \mathsf{o}_2, ..., \mathsf{o}_i] \wedge \mathsf{holds\text{-}in}(parallel(\mathsf{cam}_1, \bar{\mathsf{o}}), true, \delta_1) \wedge \\ &\quad \mathsf{occures\text{-}in}(\mathsf{move}(\mathsf{cam}_1), true, \delta_1).\end{aligned} \tag{7}$$

3. VISUAL PROCESSING: PERCEPTION AND 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, shot types, cuts and scene structure), and *perceptual objects* (eye-tracking / gaze points, areas of attention). The obtained individuals are represented as space-time objects in the context of the presented visuo-spatial narrative model (see Alg. 1).

Scene Structure

Detecting visual elements in movies is a key focus in computer vision research and resulted in a variety of methods for detecting humans (including body structure), and their interactions[11, 19, 24], as well methods for estimating facing directions [27] or recognising the identity of characters in movies [35]. The low-level visual processing algorithms that we utilise for high-level semantic analysis are founded in state-of-the-art outcomes from the computer vision community for detection and tracking of *people, objects, and motion* in the context of film analysis.

▸ **Identifying Cuts**. Analysing the structure of the scene, includes, identifying cuts [4], i.e. segmenting the scene into

space, time, and motion in the context of dynamic visuo-spatial imagery can be utilised as per [34].

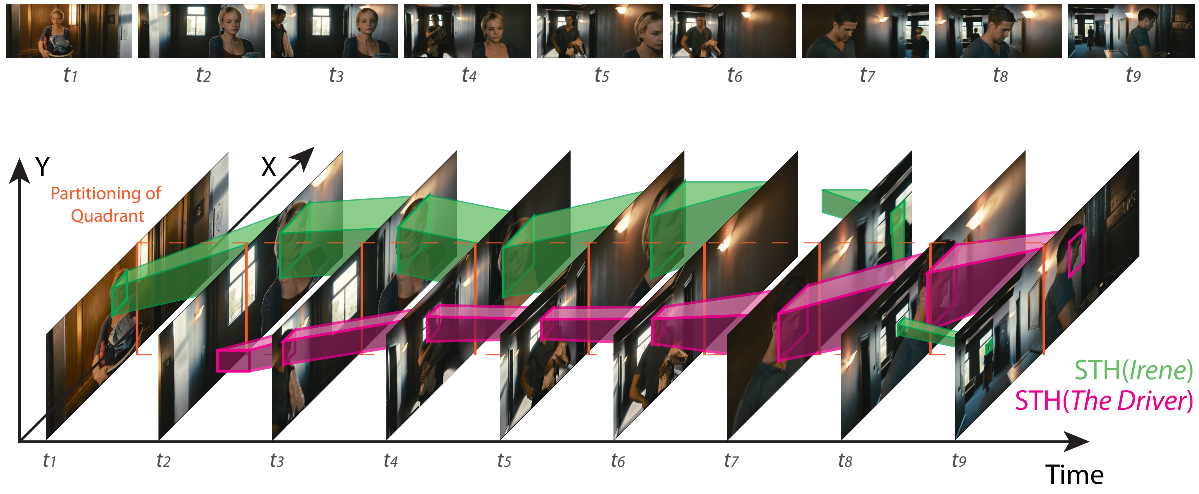


Figure 2: Space-Time Histories (2D + Time). [$STH(Irene)$, $STH(The\ Driver)$] generated from people tracking in a scene from the movie “Drive (2011)”.

its basic elements. In this way, we obtain single shots, that are used for further analysis of the semantics of the scene.

▸ **3D Camera Movement**. Estimation of camera movement is done using Fernaback’s dense optical flow [16] for two consecutive frames where we reject samples in homogeneous regions, based on the eigen values of the sample points. Estimateing 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.

▸ **People Detection and Tracking**. We are using *histograms of oriented gradients (HOG)* [12] for face detection and *deformable part models* (DPM) [17, 31] to detect people, and upper bodys. To associate detections over time and to generate tracks of people movement, we use particle filters for each potential track in the scene. We use optical flow [26] and color histograms to track the movement of the detected entities. In this way we obtain space-time histories for all detected entities in the scene (see Figure 2).

▸ **Character Identification by Deep Learning**. We use *Convolutional Neural Networks (CNN)* based deep learning as implemented and made available in the Caffe framework [21]; 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.

Perceptual Artefacts

Visual attention may be estimated based on the dynamics and distribution of eye movement data [14, 20]. 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.

Algorithm 1: $\mathcal{STH}(\mathcal{GP}, \mathcal{V})$

```
Data: Gaze data ($\mathcal{GP}$) given as gaze point at time point $t$, and
      Video sequence ($\mathcal{V}$) given as frame at time point $t$
Result: A set of Space-Time Histories ($\mathcal{STH}$) where each
        $sth \in \mathcal{STH}$ is a set of detected regions / points for
        consecutive time points $t$
$\mathcal{STH} \leftarrow \emptyset$
for $t \in \mathcal{GP}$ do
    $Att\_regions \leftarrow detect\_attention(\mathcal{GP}, t)$
$\mathcal{STH}_{att} \leftarrow associate(Att\_regions)$
$Shots \leftarrow detect\_shots(\mathcal{V})$
for $shot \in Shots$ do
    for $t \in Shots$ do
        $Faces \leftarrow DPM\_detect(\mathcal{V})$
        $Upper\_Bodys \leftarrow DPM\_detect(\mathcal{V})$
        $People \leftarrow DPM\_detect(\mathcal{V})$
        $Individuals \leftarrow Individuals \cup$
          $\{(Faces, t), (Upper\_Bodys, t), (People, t)\}$
    for $individual \in \{Faces, Upper\_Bodys, People\}$ do
        $Tracks \leftarrow pf\_tracking(individual)$
    for $track \in Tracks$ do
        $ID \leftarrow cnn\_identify(track)$
        $\mathcal{STH}_{vis} \leftarrow \mathcal{STH}_{vis} \cup \{(track, ID_{track})\}$
$\mathcal{STH} \leftarrow \mathcal{STH}_{vis} \cup \mathcal{STH}_{att}$
return $\mathcal{STH}$
```

▸ **Spatial Clustering of Gaze Points**. We estimate regions of high attention for a group of people using density based clustering (DBSCAN) [15] on the gaze points of all participants at a single time point.

▸ **Regular and Dynamic Heat Maps**. 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* [26]; (2) afterwards the heat map is generated by weighted addition of the gaussian for the motion compensated gaze points for n consecutive time points.

4. DEEP SEMANTICS FOR 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?

Our space-time history model and its integration with low-level visual processing supports such Q/A based on the visual analysis of the scene and the eye movement data. Looking at the space-time history of the aggregated area of attention of all participants, the system is able to answer queries concerning the focus of the attention of the spectators and also involving people and objects in the scene, e.g. at which time(s) was the attention fixated on a certain character.

Drive (2011). Dir: Nicolas Winding Refn As a use-case, consider the scene of the movie Drive (2011) (see Fig. 3). using our framework, it is possible to define (manually, or using other UI means)[7] high-level rules and execute queries in the logic programming language PROLOG to reason about spectator attention;

The domain-specific input data for this scene is as follows:

```
% people tracks and camera movement
...
at(639, person(1), pos(300, 220), size(106, 253)).
at(639, person(1), pos(300, 221), size(105, 252)).
at(744, person(2), pos(514, 103), size(93, 371)).
...
identity(person(1), ’Irene’).
identity(person(2), ’The Driver’).
...
at(658, camera_movement(0,0,-16)).
at(659, camera_movement(0,1, -20)).
...

%Basic scene structure (scene, shots, and cuts)
at(scene(scene1), true, interval(0, 1214)).
at(cut(cut1), true, timepoint(602)).
in(shot(shot1), true, interval(0, 602)).
in(shot(shot2), true, interval(603, 1214)).
```

Rule	Description
attn_on(Obj, Int)	aggregate subject attention is overlapping or covering object Obj during time interval Int
attn_following(Att, Obj, Int)	s-t history of attention Att is following the movement of object Obj during time interval Int
attn_shift(Att, T)	aggregated attention of all subjects shifts to the space-time region Att at time point T
attn_focusing(Att, Int)	aggregated attention of all subjects Att becomes more focused during the time interval Int

Table 2: Visual Attention Predicates (select)

Given this data we calculate different kinds of geometric representations, e.g. points, regions, line-segments, etc., wich serve as a basis for analysing the spatio-temporal dynamics of the scene.

```
at(exists(person(P)), true, timepoint(T)) :-
   at(T, person(P), _, _).
at(position(person(P)), point(X, Y), timepoint(T)) :-
       at(T, person(P), pos(X, Y), _).
at(region(Obj), polygon(Poly), timepoint(T)) :- (long).
at(movement_dir(Obj), dir(X, Y), timepoint(T)) :- (long).
```

Sample Predicates and Queries. The set of *rules* characterising different kinds of attention and fixation behaviours vis-a-vis deep video analysis is in principle extensive, and open-ended. Here, we illustrate some select sample encodings (Table 2) 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. [8]

```
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).
```

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

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

The query results in all time intervals in which the spectators attention is on the character *Irene*:

```
...
Int = interval(643, 741);
...
```

To semantically analyse the cinematographic characteristics of a scene as in Listing L1 using film analysis techniques,

[7] Within a *usable* product, it is expected to have UI modalities that will facilitate the creation of user / domain specific rules (i.e., rules need not be predefined, and may be created easily).

[8] Within PROLOG, ‘ **,** ’ corresponds to conjunction, ‘ **;** ’ to a disjunction, and ‘a **:-** b, c.’ denotes a *rule* where ‘a’ is true if both ‘b’ and ‘c’ are true; capitals are used to denote variables, whereas lower-case refers to constants; ‘_’ (i.e., the underscore) is a “dont care” variable, i.e., denoting placeholders for variable in cases where one doesn’t care for a resulting value.

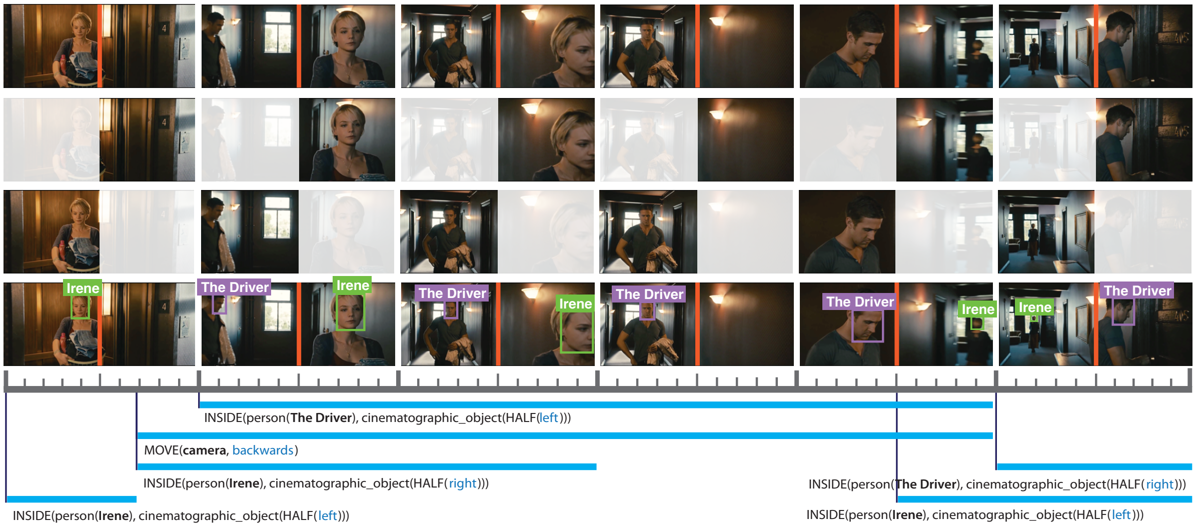

Irene
The Driver
Irene
The Driver
Irene
The Driver
The Driver
Irene
Irene
The Driver
INSIDE(person(The Driver), cinematographic_object(HALF(left)))
MOVE(camera, backwards)
INSIDE(person(Irene), cinematographic_object(HALF(right)))
INSIDE(person(The Driver), cinematographic_object(HALF(right)))
INSIDE(person(Irene), cinematographic_object(HALF(left)))
INSIDE(person(Irene), cinematographic_object(HALF(left)))

Figure 3: Visuo-Spatial Narrative: Diagrammatic representation of deep analysis of a scene from ("Drive 2011"). Analysis based on the Quadrant system method. Eye-tracking data and gaze patterns have been excluded for clarity.

i.e. the quadrant system, and to validate the claim about shifting attention while the camera is tracking Irene and the Driver, one could formulate as follows: [9]

```
?- I = interval(_, _), I1 = interval(_, _),
|    tracking(cam, [’Irene’, ’The Driver’], I1),
|    sth(gaze(Spectator), ST_GP),
|    inside(ST_GP, quadrant(half(Half)), I),
|    time(during, I, I1).
```

The answer to this is a sequence of attention shifts between the two right and the left half of the quadrant system.

```
...
Spectator = subject(s1),
I = interval(639, 646),
Half = left;
Spectator = subject(s1),
I = interval(647, 746),
Half = right;
Spectator = subject(s1),
I = interval(747, 833),
Half = left;
Spectator = subject(s1),
I = interval(834, 894),
Half = right;
...
false.
```

This way, semantic Q/A becomes possible with spatio-temporal entities of visual attention as well as domain-specific perceptual elements within the scene; both categories exist as native entities within our CLP based framework.

A Note on Software Integration. Whereas the sample queries in this section have been exemplified using the interactive capabilities of PROLOG, note that it is not necessary to manually use the framework as such; the complete logical reasoning engine of PROLOG (i.e., also our space-time history extensions implemented in PROLOG) can be embedded as a reasoning component within larger software frameworks / middleware etc for online processing, or reasoning results may be serialised within a database for offline / processing of sets of experiments .

5. SUMMARY AND OUTLOOK

Cognitive vision as an area of research has already gained prominence, with several recent initiatives addressing the topic from the perspectives of language, logic, and artificial intelligence [9, 13, 32, 36]. There has also been an increased interest from the computer vision community to synergise with cognitively motivated methods for perceptual grounding and inference with visual imagery [23, 38]. We posit that knowledge representation and reasoning can serve a crucial role for the development of next-generation methods and tools for large-scale experiments in visual perception in cognitive science and psychology. Driven by this, our research has laid out the conceptual, formal, and computational foundations for a general, declarative model of representing and reasoning with deep semantics about visuo-spatial narrative primitives identifiable with respect to a broad-based interpretation of "the moving image". Our narrative model and approach can directly provide the foundations that are needed for the development of novel assistive technologies in areas where high-level qualitative analysis and perceptual sensemaking of dynamic visuo-spatial imagery are central.

ACKNOWLEDGEMENTS. Tobii Pro (Sweden) donated eye-trackers ┃ J. Sutton (Cal Poly, USA) helped with visual processing ┃ S. Goyal (Amity University, IN) supported film experiments and data collection, and programming of clustering algorithms.

[9] The visuo-spatial narrative model can be used as a basis for automatic generation of natural language descriptions using the declarative Prolog based natural language generator provided by [33].

References

[1] M. Aiello, I. E. Pratt-Hartmann, and J. F. v. Benthem. *Handbook of Spatial Logics*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2007.

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

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

[4] E. E. Apostolidis and V. 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.

[5] M. Bhatt, J. H. Lee, and C. P. L. Schultz. CLP(QS): A declarative spatial reasoning framework. In *Spatial Information Theory - 10th International Conference, COSIT 2011, Belfast, ME, USA, September 12-16, 2011. Proceedings*, volume 6899 of *Lecture Notes in Computer Science*, pages 210–230. Springer, 2011.

[6] M. Bhatt, C. Schultz, and C. Freksa. The 'Space' in Spatial Assistance Systems: Conception, Formalisation and Computation. In T. Tenbrink, J. Wiener, and C. Claramunt, editors, *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.

[7] M. Bhatt, C. P. L. Schultz, and M. Thosar. Computing narratives of cognitive user experience for building design analysis: KR for industry scale computer-aided architecture design. In C. Baral, G. D. Giacomo, and T. Eiter, editors, *Principles of Knowledge Representation and Reasoning: Proceedings of the Fourteenth International Conference, KR 2014, Vienna, Austria, July 20-24, 2014*. AAAI Press, 2014.

[8] M. Bhatt, J. Suchan, and C. Schultz. Cognitive Interpretation of Everyday Activities – Toward Perceptual Narrative Based Visuo-Spatial Scene Interpretation. In M. Finlayson, B. Fisseni, B. Loewe, and J. C. Meister, editors, *Computational Models of Narrative (CMN) 2013., a satellite workshop of CogSci 2013: The 35th meeting of the Cognitive Science Society.*, Dagstuhl, Germany, 2013. OpenAccess Series in Informatics (OASIcs).

[9] M. Bhatt, J. Suchan, and C. P. L. Schultz. Cognitive interpretation of everyday activities - toward perceptual narrative based visuo-spatial scene interpretation. In M. A. Finlayson, B. Fisseni, B. Löwe, and J. C. Meister, editors, *2013 Workshop on Computational Models of Narrative, CMN 2013, August 4-6, 2013, Hamburg, Germany*, volume 32 of *OASICS*, pages 24–29. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, 2013.

[10] M. Bhatt and J. O. Wallgruen. Geospatial narratives and their spatio-temporal dynamics: Commonsense reasoning for high-level analyses in geographic information systems. *ISPRS International Journal of Geo-Information., Special Issue on: Geospatial Monitoring and Modelling of Environmental Change, ISPRS International Journal of Geo-Information*, 3(1):166–205, 2014.

[11] P. Bojanowski, F. Bach, I. Laptev, J. Ponce, C. Schmid, and J. Sivic. Finding actors and actions in movies. In *Proc. ICCV*, 2013.

[12] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893 vol. 1, June 2005.

[13] K. S. R. Dubba, A. G. Cohn, D. C. Hogg, M. Bhatt, and F. Dylla. Learning relational event models from video. *J. Artif. Intell. Res. (JAIR)*, 53:41–90, 2015.

[14] A. T. Duchowski. *Eye Tracking Methodology: Theory and Practice*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2007.

[15] M. Ester, H. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In E. Simoudis, J. Han, and U. M. Fayyad, editors, *KDD-96, Portland, Oregon, USA*, pages 226–231. AAAI Press, 1996.

[16] G. 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.

[17] P. F. Felzenszwalb, R. B. Girshick, D. A. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(9):1627–1645, 2010.

[18] S. M. Hazarika and A. G. Cohn. Abducing qualitative spatio-temporal histories from partial observations. In *KR*, pages 14–25, 2002.

[19] M. Hoai and A. Zisserman. Talking heads: Detecting humans and recognizing their interactions. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[20] K. Holmqvist, M. Nystrom, R. Andersson, R. Dewhurst, H. Jarodzka, and J. van de Weijer. *Eye Tracking. A comprehensive guide to methods and measures*. Oxford University Press, 2011.

[21] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[22] D. Kapur and J. L. Mundy, editors. *Geometric Reasoning*. MIT Press, Cambridge, MA, USA, 1988.

[23] A. Karpathy and L. Fei-Fei. Deep Visual-Semantic Alignments for Generating Image Descriptions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Columbus, Boston, USA*. IEEE, 2015.

[24] I. Laptev and P. Pérez. Retrieving actions in movies. In *IEEE 11th International Conference on Computer Vision, ICCV 2007, Rio de Janeiro, Brazil, October 14-20, 2007*, pages 1–8. IEEE, 2007.

[25] G. Ligozat. *Qualitative Spatial and Temporal Reasoning*. Wiley-ISTE, 2011.

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

[27] M. Marin-Jimenez, A. Zisserman, and V. Ferrari. Detecting people looking at each other in videos. *International Journal of Computer Vision*, 106(3):282–296, feb 2014.

[28] P. Muller. A qualitative theory of motion based on spatio-temporal primitives. In A. G. Cohn, L. K. Schubert, and S. C.

Shapiro, editors, *Proceedings of the Sixth International Conference on Principles of Knowledge Representation and Reasoning (KR'98), Trento, Italy, June 2-5, 1998*, pages 131–143. Morgan Kaufmann, 1998.

[29] T. Nannicelli and P. Taberham. Contemporary cognitive media theory. In T. Nannicelli and P. Taberham, editors, *Cognitive Media Theory*, AFI Film Readers. Routledge, 2014.

[30] D. A. Randell, Z. Cui, and A. Cohn. A spatial logic based on regions and connection. In *KR'92. Principles of Knowledge Representation and Reasoning*, pages 165–176. Morgan Kaufmann, San Mateo, California, 1992.

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

[32] M. Spranger, J. Suchan, M. Bhatt, and M. Eppe. Grounding dynamic spatial relations for embodied (robot) interaction. In *PRICAI 2014: Trends in Artificial Intelligence - 13th Pacific Rim International Conference on Artificial Intelligence, Gold Coast, QLD, Australia, December 1-5, 2014. Proceedings*, volume 8862, pages 958–971. Springer, 2014.

[33] J. Suchan, M. Bhatt, and H. 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.

[34] J. Suchan, M. Bhatt, and P. E. Santos. Perceptual narratives of space and motion for semantic interpretation of visual data. In L. de Agapito, M. M. Bronstein, and C. Rother, editors, *Computer Vision - ECCV 2014 Workshops - Zurich, Switzerland, September 6-7 and 12, 2014, Proceedings, Part II*, volume 8926 of *Lecture Notes in Computer Science*, pages 339–354. Springer, 2014.

[35] M. Tapaswi, M. Bäuml, and R. Stiefelhagen. "Knock! Knock! Who is it?" Probabilistic Person Identification in TV Series. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2012.

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

[37] P. Walega, M. Bhatt, and C. Schultz. ASPMT(QS): Non-Monotonic Spatial Reasoning with Answer Set Programming Modulo Theories. In *LPNMR: Logic Programming and Nonmonotonic Reasoning - 13th International Conference*, 2015.

[38] H. Yu, N. Siddharth, A. Barbu, and J. M. Siskind. A Compositional Framework for Grounding Language Inference, Generation, and Acquisition in Video. *J. Artif. Intell. Res. (JAIR)*, 52:601–713, 2015.