The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Visual Explanation by High-Level Abduction: On Answer-Set Programming Driven Reasoning about Moving Objects

Jakob Suchan,[1] **Mehul Bhatt,**[1,2] **Przemysław Wałęga,**[3] **Carl Schultz**[4]

Cognitive Vision – www.cognitive-vision.org

[1]HCC Lab., University of Bremen, Germany, [2]MPI Lab., Örebro University, Sweden
[3]University of Warsaw, Poland, and [4]Aarhus University, Denmark

Abstract

We propose a hybrid architecture for systematically computing robust visual explanation(s) encompassing hypothesis formation, belief revision, and default reasoning with video data. The architecture consists of two tightly integrated synergistic components: **(1)** (functional) answer set programming based abductive reasoning with SPACE-TIME TRACKLETS as native entities; and **(2)** a visual processing pipeline for detection based object tracking and motion analysis.

We present the formal framework, its general implementation as a (declarative) method in answer set programming, and an example application and evaluation based on two diverse video datasets: the MOTChallenge benchmark developed by the vision community, and a recently developed Movie Dataset.

Introduction

A range of empirical research areas such as cognitive psychology and visual perception articulate human visual sensemaking as an inherently abductive (reasoning) process (Moriarty 1996; Magnani 2015) involving tight linkages between low-level sub-symbolic processes on the one hand, and high-level object and event-based segmentation and inference involving concepts and relations on the other. In spite of the state of the art in artificial intelligence and computer vision, and most recent advances in neural visual processing, generalised *explainable visual perception* with conceptual categories in the context of dynamic visuo-spatial imagery remains an exceptionally challenging problem presenting many research opportunities at the interface of Logic, Language, and Computer Vision.

Explainable Visual Perception We define explainable visual perception from a human-centred, and commonsense reasoning viewpoint. In this paper, it denotes the ability to declaratively:

$\mathcal{VXP}_1$: hypothesise spatio-temporal belief (states) and events; events may be both primitive or temporally-ordered aggregates; from a more foundational viewpoint, what is alluded to here is a robust mechanism for *counterfactual* reasoning.

$\mathcal{VXP}_2$: revise spatio-temporal beliefs, e.g., by non-monotonically updating conflicting knowledge, to fix inherently incompatible configurations in space-time defying geometric constraints and commonsense laws of naive physics, e.g., pertaining to physical (un)realisability, spatio-temporal continuity.

$\mathcal{VXP}_3$: make default assumptions, e.g., about spatio-temporal property persistence concerning occupancy or position of objects; identity of tracked objects in space-time.

Explanatory reasoning in general is one of the hallmarks of general human reasoning ability; robust explainable visual perception particularly stands out as a foundational functional capability within the human visuo-spatial perception faculty. In this respect, the following considerations — establishing the scope of this paper— are important wrt. $\mathcal{VXP}_{1-3}$:

- our notion of explainability is driven by the ability to support commonsense, semantic question-answering over dynamic visuo-spatial imagery within a declarative KR setting;
- the features alluded to in $\mathcal{VXP}_{1-3}$ are not exhaustive; we focus on those aspects that we deem most essential for the particular case of *movement tracking*.

A Hybrid Architecture for Visual Explanation This paper is driven by the development of a *visual explanation component* within a large-scale computational vision & perception system targeted at a range of cognitive interaction technologies and autonomous systems where dynamic visuo-spatial imagery is inherent.

The key contribution is a hybrid visual explanation method based on the integration of high-level abductive reasoning within Answer Set Programming (ASP) ((Brewka, Eiter, and Truszczyński 2011)) on the one hand, and low-level visual processing for object tracking on the other. The core focus of the paper is on the theory, implementation, and applied evaluation of the visual explanation method. We particularly emphasise the closely-knit nature of two key subcomponents representing abductive explanation (Σ_{abd}) and low-level motion tracking (Σ_{trk}) modules respectively:

Σ_{abd}. ASP-based abductive reasoning with abstract visuo-spatial concepts —such as OBJECTS, EVENTS, SPACE-TIME TRACKLETS— as native objects within ASP

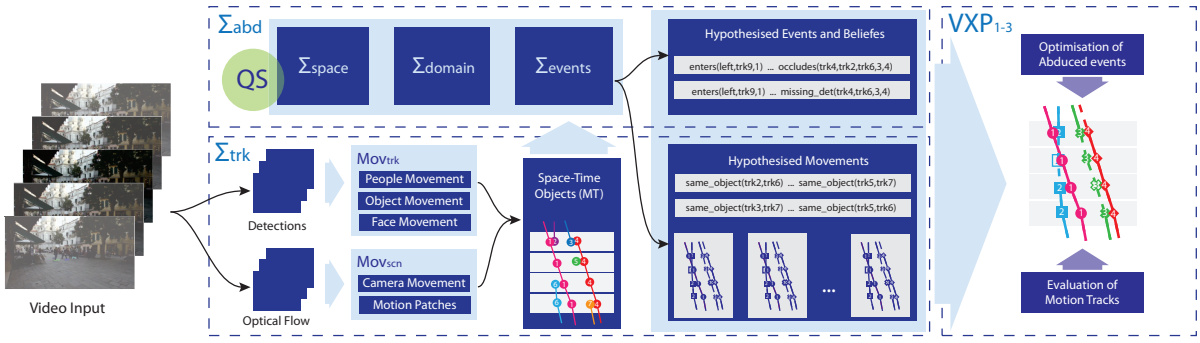

Figure 1: *Visual Explanation* – A Hybrid Architecture Integrating Low-Level Visual Tracking and High-Level Abduction

Σ_{trk}. Low-level visual processing pipeline for motion tracking, consisting of detection-based object tracking and optical-flow based (scene-level) movement tracking

The abductive component Σ_{abd} is suited for a wide-range of dynamic visuo-spatial imagery; however, we only focus on video in this paper. As an application, we focus on scene interpretation from video with two datasets: a *Movie Dataset* (Suchan and Bhatt 2016a) and the *MOT16 Dataset*, a benchmark dataset released as part of The Multiple Object Tracking Challenge (Milan et al. 2016).

Visual Explanation: A Hybrid Architecture

We present a general theory for explaining visuo-spatial observations by integrating low-level visual processing and high-level abductive reasoning (Fig. 1). As such, we consider visual abduction as reasoning from visual observations to explanations consisting high-level events grounded in low-level motion tracks. The resulting set of hypotheses is optimised based on the abduced events and the corresponding object movement.

Ontology: Space, Time, Objects, Events

The framework for abducing visual explanations is based on visuo-spatial domain objects representing the visual elements in the scene. The domain objects are associated with spatio-temporal objects describing motion tracks obtained from Σ_{trk}, which form the basis for qualitative spatio-temporal abstractions facilitating high-level reasoning about visuo-spatial dynamics.

The **Qualitative Spatio-Temporal Domain** ($\mathcal{QS}$) is characterised by the basic spatial and temporal entities ($\mathcal{E}$) that can be used as abstract representations of domain-objects and the relational spatio-temporal structure ($\mathcal{R}$) that characterises the qualitative spatio-temporal relationships amongst the supported entities in ($\mathcal{E}$). For this paper, we restrict the basic spatial entities to:

- *points* are a pair of reals x, y,
- *axis-alined rectangles* are a point p and its width and height w, h,

and the temporal entities to:

- *time-points* are a real t

Visuo-spatial **domain objects** $\mathcal{O} = \{o_1, o_2, ..., o_n\}$ are described as spatio-temporal objects by a set of spatial entities, i.e., *points*, and *axis-aligned rectangles*, in time. Towards this, $\mathcal{MT}$ contains all object tracks obtained form Σ_{trk}. The track of a single object o_i is represented by $\mathcal{MT}_{o_i} = (\varepsilon_{t_s}, ..., \varepsilon_{t_e})$, where t_s and t_e denote the start and end frame of the track and ε_{t_s} to ε_{t_e} denotes a spatial primitive representing the object o_i at the time points t_s to t_e, e.g., the axis aligned bounding box of the object.

For reasoning about visuo-spatial phenomena of object tracks, spatio-temporal relationships ($\mathcal{R}$) between the basic entities in $\mathcal{E}$ may be characterised with respect to arbitrary spatial and spatio-temporal domains such as *mereotopology, orientation, distance, size, motion*. From the viewpoint of the examples of this paper, it suffices to focus on the language of the mereotopological system of the Region Connection Calculus (RCC8) (Randell, Cui, and Cohn 1992) consisting of the following jointly exhaustive and pair-wise disjoint relations: disconnected (dc), externally connnected (ec), partially overlapping (po), equal (eq), (non-) tangential proper part ((n)tpp), and their inverse ((n)tppi).

Abducable **events** (Θ) and **beliefs** (Φ) are defined by their (spatio-temporal) *preconditions* and *observer effects*, i.e., for each event $\theta \in \Theta$ we define which properties of the scene have to be true for the event to be possible, and what the (visible) effects of the event are. In the case of visual abduction, properties of the scene are determined by the visually observed object tracks and represent qualitative relations between tracks, i.e., spatial relation $r \in \mathcal{R}$ holding between basic spatial entities ε of a motion track. Complex events are defined by combining multiple events and beliefs, e.g., an event of an object o_i passing behind another object o_j can be defined based on the events of o_i being occluded by o_j and o_i and o_j changing sides.

Abducing Visual Explanations

We implement the theory for visual explanations combining visual processing for object detection and tracking, and estimating movements in the scene, with ASP based reasoning about events, objects, and spatial-dynamics (Fig. 1). The main components of the overall tightly-integrated system comprising of low-level motion tracking with high-level explanation is as follows:

I. *Visuo-Spatial Observations* ($\mathcal{VO}$) – low-level visual processing consisting of *detection based tracking* of object and people movements.

II. *Hypotheses* ($\mathcal{H}$) – abducing hypotheses including belief states, events, and default assumptions given a set of visuo-spatial observations ($\mathcal{VO}$).

III. *Hypotheses to Explanations* — as encompassed in $\mathcal{VXP}_{1-3}$— are generated by evaluating abduced hypothesis ($\mathcal{H}$) based on high-level optimisation of event sequences and low-level cost minimisation of corresponding motion tracks.

I. Visuo-Spatial Observations ($\mathcal{VO}$) Visual explanations are based on observations obtained from visuo-spatial imagery, e.g., video, RGB-D. For the examples in this paper, we focus on *detection and tracking* of people and objects for estimating motion trajectories of semantic entities in the scene. However, the presented approach is also capable of incorporating other kinds of motion, e.g., optical flow based *low-level movement analysis* using long term observations (Ochs, Malik, and Brox 2014), or dense motion-tracklets (Gaidon, Harchaoui, and Schmid 2014) for estimating pixel level motion, corresponding to *camera movement*, or *fine grained object motion*, etc. This may be used to abduce fine grained interactions and the interplay of different movements, e.g. people movement in the presence of camera movement, by combining motion trajectories of semantic entities with pixel movements.

Movement of people and objects is estimated following the *tracking by detection* paradigm, which is based on object detections for each frame and association of the detections across frames. Object detections can in principle be obtained using any state of the art (deep learning based) detector (e.g., *faster RCNN* (Ren et al. 2015), YOLO (Redmon et al. 2016)), or deformable part models (DPM) (Felzenszwalb et al. 2010). For the examples in this paper we are using *faster RCNN* in the movie examples and DPM detections for the MOT dataset (which come as part of the dataset). For association of detections we apply the well established approach of combining *min cost assignment* and *kalman filters* for finding optimal tracklets, where the cost for assigning a detection to a track is calculated by the distance between the prediction for a track and the detection.

- *Prediction* for each track *Kalman filters* are used to predict the next position of the track and the costs for each detection is calculated based on the distance between the prediction and the detection.
- *Assignment* detections are assigned to a track using *min cost assignment* which calculates the best assignment of detection to tracks based on the costs calculated in the prediction step. If no assignment is possible for a detection a new track is started.

The resulting object tracks $\mathcal{MT}$ form the basis for abducing explanations on movement events occurring in the input data.

II. Hypotheses ($\mathcal{H}$) Explanations for visual observations are abduced based on a sequence of visual observations obtained from the video data. For abducing visual explanations from $\mathcal{VO}$, given,

- set $\mathcal{VO}$ consisting of visuo-spatial observations obtained from Σ_{trk},
- domain independent theory of space and time (Σ_{space}) based on the spatio-temporal ontology ($\mathcal{QS}$)
- observable events (Σ_{events})
- domain dependent background knowledge, describing properties of the domain (Σ_{domain})

the task of visual abduction is to find a set of logically consistent hypotheses $\mathcal{H}$ consisting of high-level events and beliefs grounded in low-level motion tracks, such that:

$$\Sigma_{space} \wedge \Sigma_{events} \wedge \Sigma_{domain} \wedge \mathcal{H} \models \mathcal{VO}$$

The computed hypotheses ($\mathcal{H}$) are based on abducibles constituting primitive events and beliefs: $\mathcal{H} \equiv \mathcal{H}_{Events} \wedge \mathcal{H}_{Belief}$; these hypotheses in turn are directly usable for inducing motion tracks:

$$\mathcal{MT}_{\mathcal{VXP}} \longleftarrow \mathcal{H}_{event} \wedge \mathcal{H}_{belief} \wedge \mathcal{MT}$$

The resulting motion tracks $\mathcal{MT}_{\mathcal{VXP}}$ represent the low-level instantiation of the abduced high-level event sequence.

III. Hypotheses to Explanations Hypotheses for visual observations ($\mathcal{VO}$) may be ranked based on the abduced event sequences and cost minimisation of corresponding motion trajectories, i.e., the costs for connecting motion tracks in the hypothesised movements, e.g. considering changes in *velocity*, *size*, and *length* of missing detections. As such, hypothesised explanations are ranked using the built in optimisation functionality of ASP[1]. In particular, we use minimisation by assigning preferences to the abducables events and beliefs and optimise towards minimising the costs of events and beliefs in the answer. E.g., by minimising the duration of missing detections for a particular object, or minimising assigning the property noise to a track to explain its observation.

▷ *High-level event sequences* the cost for high-level events is estimated by assigning a cost for each event. Additionally, for events having a duration there is also a cost assigned to the length of the event, e.g., to abduce that a track is noise is more likely, when it is a very short track. These costs are weighted based on the abduced event that caused the missing detections, e.g., missing detections caused by an occlusion are more likely to be longer (and therefore have a lower cost), than missing detections caused by the detector.

[1]For optimisation we use ASP with the so-called *weak constraints* (Gebser et al. 2012), i.e., constraints whose violation has a predefined cost. When solving an ASP program with weak constraints, a search for an answer set with a minimal cost of violated constraints is performed. Each such minimal-cost answer set is called optimal. The mechanism involving weak constraints enables us to set preferences among hypothesised explanations and search for the ones that are most preferred (optimal). Importantly, the approach enables us to exhaustively search for all optimal explanations. As a result, we can subsequently use other (more fine-graded) evaluation techniques to choose the most preferred explanations.

EVENTS	Description
enters(Border, Trk, T)	The object corresponding to track Trk enters the scene at time point T.
exits(Border, Trk, T)	The object corresponding to track Trk exits the scene at time point T.
occludes(Trk_1, Trk_2, Trk_3, T_1, T_2)	The object corresponding to track Trk_1 and track Trk_2 is occluded by the object corresponding to track Trk_3 between time points T_1 and T_2.
missing_det(Trk_1, Trk_2, T_1, T_2)	Missing detections for the object corresponding to the tracks Trk_1 and Trk_2 between time points T_1 and T_2.
COMPLEX EVENTS	**Description**
passing_behind(O_1, O_2, T_1, T_2)	Object O_1 is passing behind object O_2 between time points T_1 and T_2.
moving_together(O_1, O_2, T_1, T_2)	Objects O_1 and O_2 are moving together between time points T_1 and T_2.
BELIEFS	**Description**
same_object(Trk_1, Trk_2)	The tracks Trk_1 and Trk_2 belong to the same object.
belongs_to(Trk_1, Trk_2)	The object corresponding to track Trk_1 is a part of the object corresponding to track Trk_2.
noise(Trk)	Track Trk is a faulty detection.

Table 1: *Abducibles*: Events and Beliefs for Explaining Observed Object Tracks.

▷ *Low-level motion characteristics* the cost of the motion tracks $\mathcal{MT}$ is estimated based on the characteristics of the abduced movement. Towards this we consider changes in *velocity*, for each abduced event that connects two object tracks. For the examples in this paper we use a constant velocity model to minimize changes in velocity of an abduced object track.

The best explanation is selected by *minimising* the costs of the hypothesised answer set based on the motion and the high-level event sequence. The final movement tracks for the optimal explanation $\mathcal{MT}_{\mathcal{VXP}}$ are then generated by predicting the motion of the object for each hypothesised event associating two tracks, using linear interpolation.

Visuo-Spatial Phenomena

The framework may be used for abducing explanations by modelling visuo-spatial phenomena including but not limited to:

- *Object Persistence* objects can not appear and disappear without a cause, e.g. getting occluded, leaving the field of view of the camera, etc.
- *Occlusion* objects may disappear or re-appear as a result of occlusion between two non-opaque objects.
- *Linkage* objects linked to each other, such that movement of one object influences movement of the other object, e.g. a face belonging to a person.
- *Sensor Noise* observations that are based on faulty data, e.g. missing information, miss-detections, etc.

Event Semantics as Spatial Constraints For explaining perceived visuo-spatial dynamics of objects in the scene, we define the basic events listed in Table 1 to assure spatio-temporal consistency, e.g. object persistence, or occlusion. The focus is on explaining appearance and disappearance of objects in the scene.[2]

► *Entering and Leaving* Objects can only enter or exit the scene by leaving the screen at one of its borders. For these events to happen the object has to be overlapping with the border of the screen while appearing or disappearing.

enters:

```
topology(po,TRbox,left_border) :-
    enters(from_left,TR,T), track(TR,TRbox,T).

topology(po,TRbox,right_border) :-
    enters(from_right,TR,T), track(TR,TRbox,T).
```

exits:

```
topology(po,TRbox,left_border) :-
    exits(to_left,TR,T), track(TR,TRbox,T).

topology(po,TRbox,right_border) :-
    exits(to_right,TR,T), track(TR,TRbox,T).
```

► *Missing Detections and Occlusion* Appearance and disappearance of tracks in the middle of the screen can be either caused by a missing detection or by an occlusion from some other object. The event that an object gets occluded by some other object may be possible, when the object disappears while overlapping with the other object.

occludes:

```
topology(po,TRbox1,TRbox2) :-
    occludes(TR1,TR2a,TR2b,T1,T2),
    track(TR1,TRbox1,T1), track(TR2a,TRbox2,T1).

topology(po,TRbox1,TRbox2) :-
    occludes(TR1,TR2a,TR2b,T1,T2),
    track(TR1,TRbox1,T2), track(TR2b,TRbox2,T2).
```

Generating Hypotheses on Events We generate hypotheses explaining the observation of a track starting and ending based on the defined events, such that the spatial constraints defined above are satisfied.

starts:

```
1{
noise(TR);
enters(from_left,TR,T);
enters(from_right,TR,T);
missing_det(TR1,TR,T1,T) : ends(TR1,T1), T1<T;
occludes(TR1,TR2a,TR,T1,T): ends(TR2a,T1), T1<T,
    type(TR1,Type), Type!=border, starts(TR1,T11),
    ends(TR1,T12), T11<=T1, T<=T12
}1
:- starts(TR,T).
```

[2]The semantics of the underlying spatial and temporal relations with ($\mathcal{QS}$) is founded on the geometric and spatial reasoning capability provided by the ASPMT(QS) spatial reasoning system (Wałęga, Bhatt, and Schultz 2015); the system, implemented within ASPMT (Lee and Meng 2013), is directly available to be used as a black-box within our visual explanation framework.

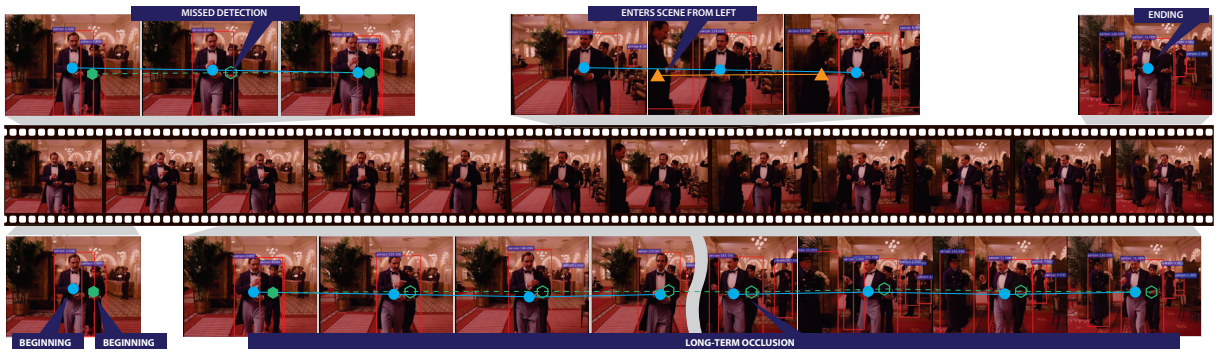

Figure 2: *People Movement* – Scene from the movie The Grand Budapest Hotel (2014) by Wes Anderson

ends:

```
1{
noise(TR);
exits(to_left,TR,T);
exits(to_right,TR,T);
missing_det(TR,TR2,T,T2) : starts(TR2,T2), T<T2;
occludes(TR1,TR,TR2b,T,T1): starts(TR2b,T1), T<T1,
    type(TR1,Type), Type!=border, starts(TR1,T11),
    ends(TR1,T12), T11<=T, T1<=T12
}1
:- ends(TR,T).
```

Beliefs as (Spatial) Constraints Beliefs about objects in the scene are stated as constraints in ASP.

► *Part-Whole Relations* E.g. the fact that every face belongs to exactly one person is stated as follows.

```
:- belongs_to(Face1,Person),
       belongs_to(Face2,Person), Face1!=Face2.
```

Further we define that the face of a person has to stay together with the person it belongs to, using spatial constraints, i.e. the face track is a non-tangential proper part of the person track.

```
topology(ntpp,Fbox,Pbox) :-
       belongs_to(Face, Person),
       track(Person, Pbox,N), track(Face, Fbox,N).
```

Generating Hypotheses on Beliefs Hypotheses on faces belonging to persons are generated by stating that for each detected face, there has to be a corresponding person, such that the spatial constraint is satisfied.

```
1{ belongs_to(Face, Person) : type(Person,person) }1
:- type(Face,face).
```

Costs of Hypotheses using Optimization Costs for abduced visual explanations are minimized using ASP based optimization, e.g., the cost for missing detections are based on their length.

```
#minimize {(T2_start-T1_end)*ALPHA, TR1, TR2 :
       missing_det(TR1,TR2,T1_end,T2_start),
       weight(missing_det, ALPHA)}.
```

Further, the characteristics of the underlying motion is taken into account, assuming constant velocity, by taking differences in velocity between the two tracks and the interpolated segment in between.

```
#minimize {((X_vel_prev-2*X_vel_during+X_vel_next)**2+
    (Y_vel_prev-2*Y_vel_during+Y_vel_next)**2)*ALPHA,
    TR1, TR2 :
    missing_det(TR1,TR2,T1,T2),
    track(TR1, Box_T1, T1), box(Box_T1, X_e, Y_e, _, _),
    [... long ...]
    X_vel_during = (X_e - X_s) / (T2-T1),
    Y_vel_during = (Y_e - Y_s) / (T2-T1),
    [... long ...]
    weight(missing_det_vel, ALPHA)}.
```

Application and Evaluation: Scene Interpretation with Moving Objects

We demonstrate the proposed theory of visual abduction by applying it in the context of scene interpretation focussing on generating visual explanations on perceived motion. In particular, the emphasis is on spatio-temporal consistency of abduced explanations with respect to the underlying motion tracks.

Movie Dataset (Suchan and Bhatt 2016a; 2016b). We use the video part of the Movie Dataset consisting of 16 select scenes from 12 films, with each scene ranging between $0:38$ minute to max. of $9:44$ minutes in duration. Most of the scenes involve multiple moving objects, and moving camera(s). Object detection with the movie dataset is performed using faster RCNN (Ren et al. 2015) with the pretrained VGG16 model for detection of people and objects in the scene.

Visual Explanation of Object Movement As an example consider the scene from the movie The Grand Budapest Hotel (2014) by Wes Anderson (Figure 2). Here we abduce the movement of the two main characters walking down the hallway of the hotel. The set of visual observations consist of 11 tracks for the detected people in the scene. The abduced events explain occuring missing detections, occlusion and re-appearance, as well as entering, and leaving the scene.

```
exits(to_right,trk10,1511) enters(from_left,trk9,1500)
occludes(trk4,trk2,trk6,1490,1495)
occludes(trk1,trk0,trk7,1490,1496)
...
noise(trk4) noise(trk8)  ending(trk11,1512)
```

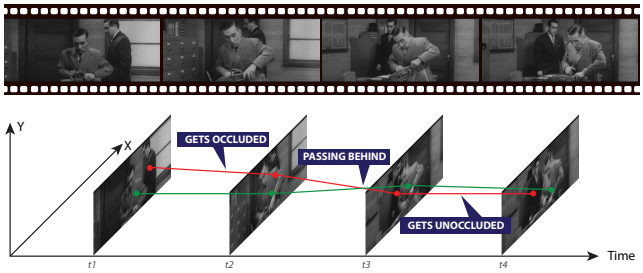

Figure 3: *Occlusion, while passing behind* – Scene from "The Bad Sleep Well" by Akira Kurosawa (1960)

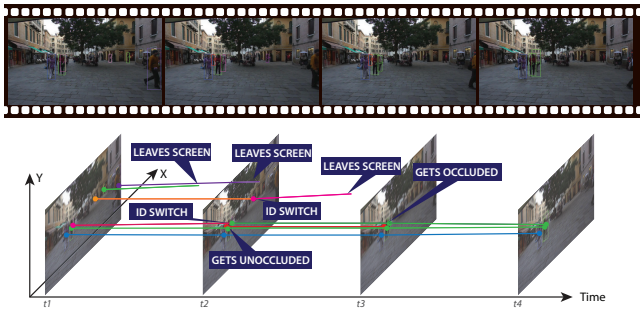

Figure 4: *Detection Errors* – Scene from the MOT2016 dataset

Similarly we can abduce complex events based on movement events and belifes, e.g., based on the following sequence of movement events from the movie "The Bad Sleep Well" (1960) by Akira Kurosawa (depicted in Figure 3), we can abduce the occurrence of the complex event *passing behind* between two objects in the scene.

```
...
missing_det(trk1,trk3,556,561) noise(trk2)
occlusion(trk3,trk4,trk0,561,629)
missing_det(trk4,trk5,629,634)
...
...
passing_behind(obj1,obj2,561,629)
...
```

Hypotheses on People and their Faces As an example for abducing properties of objects, we use face detections and abduce which face belongs to which person, using the part-whole relations defined in Section *Visuo-Spatial Phenomena*. The spatial constraint that a face track has to be inside a person track, is used to improve abduced object tracks.

```
belongs_to(trk2,trk1) belongs_to(trk4,trk3)
```

MOT16 Benchmark Video Dataset We use the MOT16 (Milan et al. 2016) dataset consisting of highly accurate and consistent annotation protocols. MOT16 is a benchmark dataset released as part of The Multiple Object Tracking Challenge (MOTChallenge). It consists of 14 complex video sequences in highly unconstrained environments filmed with both static and moving cameras. We use the detections (provided by the dataset) based on deformable part models (DPM): these are noisy and include numerous miss detections, i.e. false positives and false negatives. We focus on abducing people motion and on generating concise explanations for the perceived movements, i.e. under consideration of occlusion and appearance / disappearance of characters as per the abducible events in Table 1. As a result of the noisy detections and the complexity of the movements in the dataset the obtained motion tracks include a high amount of errors, e.g. identity switches, missing detections, etc. (Figure 4). For the sample scene we abduced the following events:

```
...
missing_det(trk19,trk23,45,49) exits(to_right,trk25,47)
exits(to_right,trk6,46) occludes(trk1,trk23,trk28,50,54)
...
```

Evaluating Visual Explanations

We evaluate the generated visual explanations based on their ability to generate low-level object tracks. Towards this we compare the accuracy and precision of the movement tracks the hypothesised event sequences are grounded in.

Multi-Object Tracking For evaluating the precision and accuracy of the abduced object tracks we follow the ClearMOT evaluation schema for evaluating multi-object tracking performance as described in (Bernardin and Stiefelhagen 2008).

- *MOTA* describes the accuracy of the tracking, taking into account the number of missed objects / false negatives (FN), the number of false positives (FP), and the number of miss-matches (MM).
- *MOTP* describes the precision of the tracking based on the distance of the hypothesised track to the ground truth of the object it is associated to.

These metrics are used to assess how well the generated visual explanations describe the low-level motion in the scene.

Results & Discussion

We present results of the presented approach for abducing visual explanations ($\mathcal{VXP}$) on improving multi-object tracking performance using selected scenes from the Movie Dataset and the MOT 2016 Dataset. Overall the results show that using our proposed method can increase accuracy (*MOTA*) of the tracking. However, the precision (*MOTP*) of the tracking is dropping a little, which is a result of the interpolation, which is not as precise as the detections.

Movie Dataset The scenes in the Movie Dataset contain relatively controlled scenes with few targets. Results on these scenes show that the presented approach can abduce correct event sequences and is capable of correcting many of the errors normally occurring in multi-object tracking tasks, e.g., fragmented object tracks, id-switches, etc. I.e., the object tracks obtained from the high-level event sequences improve the accuracy (*MOTA*) of the tracking (see Table 2).

MOT Dataset The results for the visual tracking on the *Venice-2* file from the MOT2016 dataset (see Table 2) show, that our approach is capable of dealing with complex data in challenging settings. For comparability we use the DPM

Sequence	Tracking	**MOTA**	**MOTP**	FP	M	MM	non-r. MM	r. MM	TP	TR
The Bad Sleep Well	*without* $\mathcal{VXP}$	58.5 %	80.8 %	1	86	5	0	5	0.875	1.0
(107 frames, 2 targets)	*with* $\mathcal{VXP}$	100.0 %	69.1 %	0	0	0	0	0	1.0	1.0
The Drive	*without* $\mathcal{VXP}$	59.8 %	76.7 %	0	345	18	0	18	1.0	1.0
(627 frames, 2 targets)	*with* $\mathcal{VXP}$	79.7 %	76.6 %	0	182	1	0	1	1.0	1.0
MOT2016 - Venice-2	*without* $\mathcal{VXP}$	6.4 %	69.9 %	47	27137	153	4	150	0.987	0.241
(600 frames, 74 targets)	*with* $\mathcal{VXP}$	8.1 %	65.2 %	216	26535	86	27	77	0.946	0.486

Table 2: *Evaluation of Tracking Performance*: false positives (FP), misses (M), miss-matches (MM), non-recoverable miss-matches (non-r. MM), recoverable miss-matches (r. MM), track precision (TP), track recall (TR)

based detections provided with the dataset for our evaluation. These detections suffer from a large number of false positives and negatives. Due to this our underling tracking method is only capable of tracking a small part of the overall targets in the data, resulting in a low *MOTA* score. Even so our results demonstrate that high-level abduction can be used to improve tracking performance, i.e., improved *MOTA* by 1.7%, and number of miss-matches reduced by 43.8%.

Based on the promising results presented in this paper, using basic tracking by detection, we suppose that ASP-based visual explanations can also be used to improve multi-object tracking using more elaborate tracking approaches, e.g., based on continuous energy minimization (Milan, Schindler, and Roth 2016) or minimum cost multi-cuts (Tang et al. 2017).

For the examples presented in this paper, optimal answer-sets are computed rather fast, e.g., for the scene depicted in Figure 2, 457 optimal models are abduced in 0.973s, of which the first model is found after 0.04s and the last one after 0.93s.[3] For longer scenes or in online situations, visual explanations would naturally have to be computed incrementally, as the number of abduced hypothesis grows exponentially with the number of tracks.

Related Work

Answer Set Programming (ASP) has become a widely used tool for abductive reasoning and non-monotonic reasoning in general. The work presented in this paper aims at bridging the gap between high-level formalisms for logical abduction and low level visual processing, by tightly integrating qualitative abstractions of space and time with the underlying numerical representations of spatial change. The significance of abducing high-level explanations in a range of contexts has been well established in AI and KR, e.g. in planning and process recognition (Kautz and Allen 1986; Kautz 1991), vision and abduction (Shanahan 2005), probabilistic abduction (Blythe et al. 2011) etc. Within KR, reasoning about spatio-temporal dynamics on the basis of an integrated theory of space, time, objects, and position (Galton 2000) or defined continuous change using 4-dimensional regions in space-time has also received significant theoretical interest (Muller 1998; Hazarika and Cohn 2002). Dubba et al. (2015) uses abductive reasoning for improving learning of events in an inductive-abductive loop, using inductive logic programming (ILP). The role of visual commonsense in general, and answer set programming in particular, has been used in conjunction with computer vision to formalise general rules for image interpretation in the recent works of Aditya et al. (2015). From the viewpoint of computer vision research there has been an interest to synergise with cognitively motivated methods (Aloimonos and Fermüller 2015); in particular the research on semantic interpretation of visual imagery is relevant to this paper, e.g., for combining information from video analysis with textual information for understanding events and answering queries about video data (Tu et al. 2014), and perceptual grounding and inference (Yu et al. 2015).

[3]We computed hypotheses using a Intel Core i5-4210M 2.60GHz CPU with 12 GB RAM running Ubuntu 16.04

Summary and Outlook

The paper presents a robust, declarative, and generally usable hybrid architecture for computing visual explanations with video data. With a focus on abductive reasoning in the context of motion tracking, the architecture has been formalised, fully implemented, evaluated with two diverse datasets: firstly, the benchmark MOTChallenge (evaluation focus), and secondly a Movie Dataset (demonstration focus).

The overall agenda of the work in this paper is driven by a tighter integration of methods in KR and Computer Vision on the one hand, and the twin concepts of "*deep semantics*" & "*explainability*" on the other. $\mathcal{VXP}$ is rooted in state of the art methods in knowledge representation and reasoning (i.e., answer set programming), and computer vision (detection based object tracking, optical flows, RCNN). The overall system is designed to be a part of a larger perception module within autonomous systems, and cognitive interaction systems. The scope of $\mathcal{VXP}$ may be further expanded, e.g., for visuo-spatial learning (with *inductive logic programming*), ontological reasoning (with *description logics*), are achievable depending on the scope and complexity of the low-level visual signal processing pipeline, and chosen high-level commonsense knowledge representation and reasoning method(s) at hand.

Acknowledgements

We acknowledge funding as part of the German Research Foundation (DFG) CRC 1320 "EASE – Everyday Activity Science and Engineering" (http://www.ease-crc.org/) Project P3: "Spatial Reasoning in Everyday Activity". We also acknowledge the Polish National Science Centre project 2016/23/N/HS1/02168.

References

Aditya, S.; Yang, Y.; Baral, C.; Fermuller, C.; and Aloimonos, Y. 2015. Visual commonsense for scene understanding using perception, semantic parsing and reasoning. In *2015 AAAI Spring Symposium Series*.

Aloimonos, Y., and Fermüller, C. 2015. The cognitive dialogue: A new model for vision implementing common sense reasoning. *Image and Vision Computing* 34:42 – 44.

Bernardin, K., and Stiefelhagen, R. 2008. Evaluating multiple object tracking performance: The clear mot metrics. *EURASIP Journal on Image and Video Processing* 2008(1).

Bhatt, M., and Loke, S. W. 2008. Modelling dynamic spatial systems in the situation calculus. *Spatial Cognition & Computation* 8(1-2):86–130.

Bhatt, M.; Guesgen, H. W.; Wölfl, S.; and Hazarika, S. M. 2011. Qualitative spatial and temporal reasoning: Emerging applications, trends, and directions. *Spatial Cognition & Computation* 11(1).

Bhatt, M. 2012. 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.

Blythe, J.; Hobbs, J. R.; Domingos, P.; Kate, R. J.; and Mooney, R. J. 2011. Implementing weighted abduction in markov logic. In *Proceedings of the Ninth International Conference on Computational Semantics*, IWCS '11. USA: Association for Computational Linguistics.

Brewka, G.; Eiter, T.; and Truszczyński, M. 2011. Answer set programming at a glance. *Commun. ACM* 54(12):92–103.

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

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

Gaidon, A.; Harchaoui, Z.; and Schmid, C. 2014. Activity representation with motion hierarchies. *International Journal of Computer Vision* 107(3):219–238.

Galton, A. 2000. *Qualitative Spatial Change*. Oxford University Press, Oxford, UK.

Gebser, M.; Kaminski, R.; Kaufmann, B.; and Schaub, T. 2012. *Answer Set Solving in Practice*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan and Claypool Publishers.

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

Kautz, H. A., and Allen, J. F. 1986. Generalized plan recognition. In Kehler, T., ed., *Proceedings of the 5th National Conference on Artificial Intelligence. Philadelphia, 1986. Volume 1: Science.*, 32–37. Morgan Kaufmann.

Kautz, H. A. 1991. Reasoning about plans. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. chapter A Formal Theory of Plan Recognition and Its Implementation, 69–124.

Lee, J., and Meng, Y. 2013. Answer set programming modulo theories and reasoning about continuous changes. In *IJCAI 2013, Proceedings of the 23rd International Joint Conference on Artificial Intelligence, Beijing, China, 2013*.

Magnani, L. 2015. *Philosophy and Cognitive Science II: Western & Eastern Studies*. Cham: Springer International Publishing. chapter Understanding Visual Abduction, 117–139.

Milan, A.; Leal-Taixé, L.; Reid, I. D.; Roth, S.; and Schindler, K. 2016. MOT16: A benchmark for multi-object tracking. *CoRR* abs/1603.00831.

Milan, A.; Schindler, K.; and Roth, S. 2016. Multi-target tracking by discrete-continuous energy minimization. *IEEE Trans. Pattern Anal. Mach. Intell.* 38(10):2054–2068.

Moriarty, S. E. 1996. Abduction: A theory of visual interpretation. *Communication Theory* 6(2):167–187.

Muller, P. 1998. A qualitative theory of motion based on spatio-temporal primitives. In Cohn, A. G.; Schubert, L. K.; and Shapiro, S. C., eds., *Proceedings of the Sixth International Conference on Principles of Knowledge Representation and Reasoning (KR'98), Italy*. Morgan Kaufmann.

Ochs, P.; Malik, J.; and Brox, T. 2014. Segmentation of moving objects by long term video analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* 36(6):1187–1200.

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

Redmon, J.; Divvala, S. K.; Girshick, R. B.; and Farhadi, A. 2016. You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 779–788. IEEE Computer Society.

Ren, S.; He, K.; Girshick, R. B.; and Sun, J. 2015. Faster R-CNN: towards real-time object detection with region proposal networks. In *Annual Conference on Neural Information Processing Systems 2015, Canada*.

Shanahan, M. 2005. Perception as abduction: Turning sensor data into meaningful representation. *Cognitive Science* 29(1):103–134.

Suchan, J., and Bhatt, M. 2016a. The geometry of a scene: On deep semantics for visual perception driven cognitive film, studies. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2016, USA*. IEEE Computer Society.

Suchan, J., and Bhatt, M. 2016b. Semantic question-answering with video and eye-tracking data: AI foundations for human visual perception driven cognitive film studies. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, USA*. IJCAI/AAAI Press.

Suchan, J., and Bhatt, M. 2017. Deep semantic abstractions of everyday human activities - on commonsense representations of human interactions. In *ROBOT 2017: Third Iberian Robotics Conference, Seville, Spain*. Springer.

Tang, S.; Andriluka, M.; Andres, B.; and Schiele, B. 2017. Multiple people tracking by lifted multicut and person re-identification. In *CVPR* 2017.

Tu, K.; Meng, M.; Lee, M. W.; Choe, T. E.; and Zhu, S. C. 2014. Joint video and text parsing for understanding events and answering queries. *IEEE MultiMedia*.

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

Yu, H.; Siddharth, N.; Barbu, A.; and Siskind, J. M. 2015. A Compositional Framework for Grounding Language Inference, Generation, and Acquisition in Video. *Journal of Artificial Intelligence Research (JAIR)* 52:601–713.