Attentional synchrony in films: A window to visuospatial characterization of events

Vipul Nair
University of Skövde, Sweden
vipul.nair@his.se

Jakob Suchan
German Aerospace Center (DLR), Germany
jakob.suchan@dlr.de

Mehul Bhatt
Örebro University, Sweden
mehul.bhatt@oru.se

Paul Hemeren
University of Skövde, Sweden
paul.hemeren@his.se

ABSTRACT

The study of event perception emphasizes the importance of visuospatial attributes in everyday human activities and how they influence event segmentation, prediction and retrieval. Attending to these visuospatial attributes is the first step toward event understanding, and therefore correlating attentional measures to such attributes would help to further our understanding of event comprehension. In this study, we focus on attentional synchrony amongst other attentional measures and analyze select film scenes through the lens of a visuospatial event model. Here we present the first results of an in-depth multimodal (such as head-turn, hand-action etc.) visuospatial analysis of 10 movie scenes correlated with visual attention (eye-tracking 32 participants per scene). With the results, we tease apart event segments of high and low attentional synchrony and describe the distribution of attention in relation to the visuospatial features. This analysis gives us an indirect measure of attentional saliency for a scene with a particular visuospatial complexity, ultimately directing the attentional selection of the observers in a given context.

CCS CONCEPTS

• **Computing methodologies** → **Modeling methodologies**; **Cognitive science**; **Scene understanding**.

KEYWORDS

Visuoauditory cues, Human-interaction, Eye-tracking, Attention

ACM Reference Format:
Vipul Nair, Jakob Suchan, Mehul Bhatt, and Paul Hemeren. 2022. Attentional synchrony in films: A window to visuospatial characterization of events. In *ACM Symposium on Applied Perception 2022 (SAP '22), September 22–23, 2022, Virtual Event, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3548814.3551466

1 INTRODUCTION

A human observer is always in the middle of a continuous stream of dynamic multimodal information, and from this rich plethora of information, the observer picks up specific cues, attends to what the cues lead to and makes sense of the world. Evidence from decades of research on human cognition has indicated that specific visuospatial attributes (cues) get picked over top-down semantic processing[Grèzes 1998; Hochstein and Ahissar 2002], such as the quick and autonomous processing of biological motion[Johansson 1973], face[Farah et al. 1995], gesture[Kang and Tversky 2016], and goal-directed actions[Flanagan and Johansson 2003]. In this study, we pick on those visuospatial cues and demonstrate how everyday interactions can be formally characterized (represented) in terms of visuospatial description. Furthermore, as an example use case, we deploy these visual descriptions to analyse the observer's attentional mechanism while viewing select film scenes.

Films as a case study: We focus on the case of attention in the context of moving images (particularly, visuo-auditory narrative film) to demonstrate the characterization of events from an observer's perspective [Bhatt 2018a,b]. We utilise a visuospatial model for the automated processing of low-level features (e.g., motion), as well as high-level features (e.g., referential gaze) of given human activity [Suchan and Bhatt 2016a]. Furthermore, we take attentional synchrony (multiple viewers looking at the same region) coupled with the event characterization as an investigative window to the human observer − specifically characterizing the human activity (in-scene) with respect to what the viewers attended to. Finally, we present the first results of our event analysis in relation to a semantic interpretation of the multimodal human behaviour data (in-scene) in terms of our visuospatial model. Particular consideration has been given to the multimodality of naturalistic human activity and towards computational requirements, such as ground truth for everyday activities in a context agnostic structure that could have implications for knowledge representation, visual sensemaking, and declarative reasoning within AI systems [Kondyli et al. 2022].

2 VISUOSPATIAL MODEL

We developed a visuospatial model with the aim of providing a semantic interpretation (ground truth) to explicate the visuospatial attributes of an event and how the human observer interprets those attributes. For a human observer the scene is broken down to its objective elements to provide the ground truth, taking into account the various modalities of the human interaction that play out in a typical human-centered interaction scenario. Table.1 lists out the various visuospatial attributes categorized into a taxonomy of elemental relations encompassing the cognitive and multimodal

SAP '22, September 22–23, 2022, Virtual Event, USA

ACM ISBN 978-1-4503-9455-0/22/09.
https://doi.org/10.1145/3548814.3551466

Table 1: A cognitive characterisation of the human interactions and the modalities involved.

Visuospatial Features	Multimodal Interaction (non-exhaustive list)				**Count**	**Sec**
Scene Elements						
Types (Taxonomy)	object	dynamic	person, animal, ...		46	2148.2
			body-parts	face, head, hands, torso, ...		
			vehicle	car, truck, motorcycle, bicycle, train, ...		
			gaze	gaze-point, scan-path, ...		
		static	phone, bag, table, door, wall, ...			
	region	corridor, elevator, doorway, window sill, train cabin, stairway, ...				
Scene Structure						
Visibility	visible(X)				388	2148.2
Presence	present(X)				87	3278.6
Motion	stationary(X), moving(X), turning(X), moving_towards(X, Y), moving_away(X, Y), moving_together(X, Y), moving_next_to(X, Y), turning_towards(X, Y), turning_away(X, Y)				664	2167.3
Spatial Position	behind(X, Y), front(X, Y), left(X, Y), right(X, Y), above(X, Y), below(X, Y), front_left(X, Y), front_right(X, Y), behind_left(X, Y), behind_right(X, Y)				528	1673.8
Human Action	speaking(X)				262	371.1
Head Movement	steady_head(X), turn_left_head(X), turn_right_head(X), turn_upwards_head(X), turn_downwards_head(X), turn_upwards_left_head(X), turn_upwards_right_head(X), turn_downwards_left_head(X), turn_downwards_right_head(X)				1127	1623.4
Gaze	looking_at(X, Y)				446	711.4
Hand Action	hold(X), pull(X), push(X), reaching_towards(X), grasp(X)				325	761.5
Body Pose	bending(X), crouching(X), kneeling(X), lean_backward(X), lean_forward(X), lean_sideways(X), leaning_against(X), sitting(X), lying_down(X), standing(X)				466	2158.0
Visual Attention						
Low-Level	fixation(ID), saccade(ID)				293663	31878.7
Object-Level	attention_on(face(X)), attention_on(head(X)), attention_on(hands(X)), attention_on(torso(X)), ...				14772	16584.5

nature of interactions. Here we present a brief argument to their role in perception and semantic grounding, for a detail explaining of their definition and usage see Appendix-Table.3.

Scene Elements: All scene elements are broadly classified into its several types. The broader categories such as the region of the scene place a huge role in how an embodied human interacts with the environment as well as the attention of an observer [Smith and Mital 2013]. Similarly attentional strategies vary over watching static and dynamic stimuli [Smith and Mital 2013].

Scene Structure: The primary focus is on the human, thereby the visuospatial features of the interaction is classified into the various measurable modalities of the human behavior. Each factor is carefully chosen to enable a partonomical and hierarchical analysis into how the different modalities play into the observer’s semantics. Furthermore this structure enables a multi-factorial analysis to see how certain factors (or combination) act cohesively to enable observers to predict and segment events. Moreover the schema of this structure is designed to be context and environment independent such that the resulting semantic interpretation can be agnostic to the scene context. The modalities were picked with the human observer in mind:

(1) Visibility: Attention tends to be modulated heavily by the mere visibility of a person [Cutting 2005].
(2) Presence: Being present in scene (may or not be visible) is an influential factor in directing attention [Loschky et al. 2015], also in analysing occlusion scenarios [Suchan et al. 2019].
(3) Motion: Motion sensitivity to human vision is well documented, especially that of biological motion [Hemeren and Rybarczyk 2020; Johansson 1973; Viviani and Stucchi 1992].
(4) Spatial Position: The spatiality of a scene is crucial to the observer in understanding the scene and predicting events.
(5) Human Action: Only the act of speaking is considered.
(6) Head Movement: Observers are cued by agent head movement as a first step towards the agent’s forth coming action.
(7) Gaze: The process of predicting action or ascribing intention begins with the gaze of the actor[Smith et al. 2012].
(8) Hand Action: Numerous studies have highlighted the importance of hand action to Action observation and learning. This is tightly linked to mirror neurons [Flanagan and Johansson 2003], so even kinematic information of an action gives rise to mostly accurate semantic interpretations, and these are widely extrapolated for various classification models [Nair et al. 2020]
(9) Body Pose: Humans easily picks up affection [Clarke et al. 2005], identity [Cutting and Kozlowski 1977] and kinematic information [Koul et al. 2019] from body pose.

Visual Attention: Here we shift the focus to the observer of the event, and characterize their attention according to how and what did they attend to. The attentional data is characterised into:

(1) Low-Level: This is the information (ID) of the fixation and saccade data which can be pointed directly to the output format of the eye-tracker in use.
(2) Object-Level: Attention on the objects (scene elements) at high-level observations, e.g., attention is on person X’s face (attention_on(face(X))). The relations are non-exhaustive and are tightly coupled to the Scene Elements’ taxonomy. Table 1 shows relations with respect to the type of ’person’ and respective ’body-parts’.

3 SEMANTIC EVALUATION

This section describes the process of annotation of the film scenes and corresponding eye-tracking data by Human experts.

3.1 Scenes

We choose ten film scenes (see Table 2) from a larger dataset focused on qualitative spatio-temporal analysis and the semantic interpretation of films [Suchan and Bhatt 2016a,b]. The dataset also has eye-tracking data from 32 participants (per scene). The eye-tracking data was collected using using a Tobii X2-60 Eye Tracker at a rate of

60 Hz. These selected scenes were used for our high-level semantic analysis.

3.2 Procedure

ELAN[1] tool; a non web application where users can add textual descriptions (manual annotations) to video and audio recordings [Sloetjes and Wittenburg 2008], was used for annotating the visuospatial features for the chosen scenes. Expert human evaluators annotated the scenes and their corresponding eye-tracking data in order to ensure high-quality data. Furthermore to ensure uniformity in the annotation language, the schema of our visuospatial features is transposed to the ELAN's annotation structure. Such that an evaluator needs only go to a modality (e.g, bodypose) and pick the appropriate description (referred to as controlled vocabularies) (e.g., lean_forward(X)) for what might be happening in the scene.

3.2.1 Annotation on scene structure: For each scene we specified certain number of entities that were characters/objects of interest, and the evaluators annotated what these entities were doing in terms of our controlled vocabularies. See Fig.1, where *motion* feature is annotated for three characters(entities) in S10, with rough sketch based on stills from the the scene (credits[2]).

3.2.2 Annotation on eye-tracking: The evaluators annotated the attention attributes for all the eye-tracking participants for the chosen scenes. They were guided by low-level information (fixation and saccade) both visually (video export from eye-tracker) and in form of timeline data (automated annotation of low-level data). See Fig.1, where *attention* feature is annotated for one of the eye-tracking participant for S10.

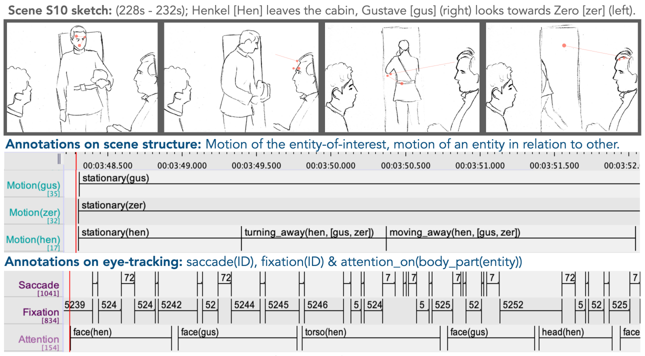
Scene S10 sketch: (228s - 232s); Henkel [Hen] leaves the cabin, Gustave [gus] (right) looks towards Zero [zer] (left).
Annotations on scene structure: Motion of the entity-of-interest, motion of an entity in relation to other.
00:03:48.500
00:03:49.000
00:03:49.500
00:03:50.000
00:03:50.500
00:03:51.000
00:03:51.500
00:03:52.
Motion(gus) [35]
stationary(gus)
Motion(zer) [32]
stationary(zer)
Motion(hen) [17]
stationary(hen)
turning_away(hen, [gus, zer])
moving_away(hen, [gus, zer])
Annotations on eye-tracking: saccade(ID), fixation(ID) & attention_on(body_part(entity))
Saccade [1041]
72
72
7
7
7
7
72
7
Fixation [834]
5239
524
524
5242
52
5244
5245
5246
5
524
5
525
52
5252
5
52
525
Attention [154]
face(hen)
face(gus)
torso(hen)
face(gus)
head(hen)
face

Figure 1: Annotation (ELAN) example for scene S10. Credits[2].

3.2.3 Annotation summary: Table 1 (right column) shows the summary of the annotations with respect to the model's taxonomy. Similarly Table 2 (right columns) show the summary of the annotations for each scene with respect to scene structure (event data) and visual attention. For a more detailed distribution of the annotation summary, see Appendix- Table.3(eye-tracking), Table.4(event-data) and Table.5(event-data at the elemental level).

4 HIGH-LEVEL SEMANTIC ANALYSIS

Attentional synchrony: Fig.2.(a) shows the attentional synchrony (%) for scene S2 with rough sketch based on stills from the scene (credits[3]). The synchrony for the same region in a frame is computed based on the annotation: same body part of the same entity (e.g., attention_on(hand(X)) at time t). The segment shown is an excerpt from 285s -323s of S2. The static frames depict the corresponding scene events overlapped with the information of whom the participants attended(%). The example showcases a viewing trend of high synchrony when characters are alone in the frame or do a specific behaviour compared – an interesting case for investigating reactive and anticipatory gaze scenarios.

Segments based on high-low attentional synchrony: Fig.2.(b) shows the attentional distribution –participants (% of total viewers) whose gaze is synchronous and duration (% of overall synchrony period) of gaze – for high- attentional synchrony segments for scene S2 segregated in terms of high (>50%) and low (<50%) synchrony measure. Low synchrony segments have low attentional distribution, hence not shown. The arrows from Fig.2.(a) point to the corresponding segment number. We further take this segmentation process to tease apart the visuospatial structure of the scenes.

Feature analysis on high-low synchrony segments: Fig.2.(c) shows the distribution of the scene structure (event data) for the high-low synchrony segments cumulated for all the scenes. Note that the example case S2 is amongst the halves(S1, S2, S3, S9) where low-synchrony has more event data (scene structure) than high-synchrony. Again this presents a case to study cognitive films – how directional style, symmetry and narrative styles, among other cinematic practices, affect synchronous gaze behaviour.

Low-level event instances: Fig.2.(d) shows the distribution for the low-level visuospatial features (i.e., scene structure modalities from Table.1) for S2. In comparison, Fig.2.(e) shows the distribution for the same low-level visuospatial features, but only when a change of state occurs (e.g., X is moving(t1,t2):X is stationary(t2, t3)). Change in the state of modalities is a higher-level abstraction of scene structure. Note that this is a simple case of change situations; more complex abstraction could be similarly achieved by cross-modal feature analysis.

High-level event instances: Fig.2.(f) shows a much higher-level of the presented case of visibility change (to occlusion) and gaze change (to gaze-transition). Here attentional distribution is in focus to showcase how many (and how much) viewers attended these instances. Occlusion here is abstracted as someone moving or stationary, is visible and gets occluded for a brief time(<5s) and becomes visible again. Similarly gaze-transition is abstracted as someone switches gaze from one person to another (object-of-interest), while the visibility information of the pre-switch and post-switch object-of interest should be clear. Finally, Fig.2.(g) shows the attentional

[1] ELAN Computer software. (2020). Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive. https://archive.mpi.nl/tla/elan)

[2] Credits: "The Grand Budapest Hotel", directed by Wes Anderson, produced by Wes Anderson, Scott Rudin, Steven Rales, and Jeremy Dawson, Fox Searchlight Pictures, TSG Entertainment, Indian Paintbrush, Studio Babelsberg, American Empirical Pictures, USA and Germany, 2014

[3] Credits: "Solaris", directed by Andrei Tarkovsky, produced by Vyacheslav Tarasov, Mosfilm, Russia, 1972

Table 2: Selected scenes, length, description, ID, total count and duration of respective annotated features.

Film, Director	Year	Scene ID	Scene	Mins.	Event Data		Visual Attention Data *(average)*			
					Scene-Level		Object-Level		Low-Level	
					freq	*sec*	*freq*	*sec*	*freq*	*sec*
The Bad Sleep Well, Akira Kurosawa	1960	S1	Triangle scene	2:46	287	2421.8	59.0	54.5	1447.3	159.2
Solaris, Andrei Tarkovsky	1972	S2	Opening scene	7:46	644	3024.5	59.0	91.2	3541.8	398.7
Goodfellas, Martin Scorsese	1990	S3	Copacabana scene	3:03	570	2403.6	110.2	99.9	1706.6	165.6
Paprika, Satoshi Kon	2006	S4	Opening scene	1:48	178	581.6	34.8	43.1	954.6	109.1
The Drive, Nicolas Winding Refn	2011	S5	Irene's flat scene	2:58	394	1685.2	71.5	118.2	1523.4	163.4
		S6	First meet scene	0:50	143	546.4	33.8	35.3	547.7	56.4
		S7	Corridor scene	1:59	282	1419.3	57.3	76.5	924.8	102.7
The Hunger Games, Gary Ross	2012	S8	Selection scene	2:48	316	1424.1	86.5	95.8	1585.5	155.8
The Grand Budapest Hotel, Wes Anderson	2014	S9	Lobby scene	1:41	474	1368.5	70.0	74.3	914.7	104.7
		S10	Train scene	4:17	1005	3947.4	167.6	163.6	1989.1	224.9
TOTAL				**29m 56s**	**4293**	**18841.2**	**749.7**	**852.4**	**15135.5**	**1640.5**

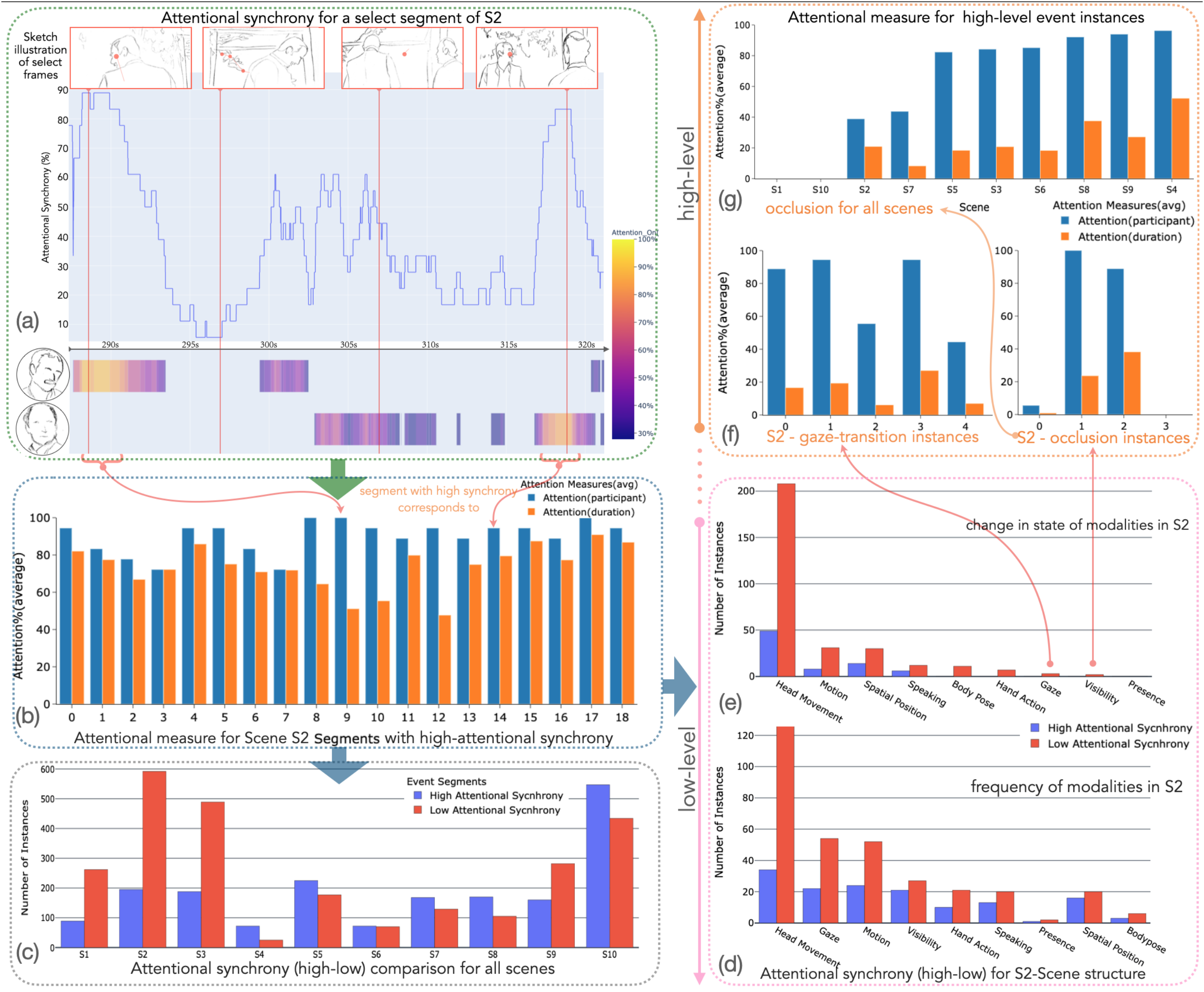

Figure 2: High-level semantic analysis process and flow, with example case of scene S2 and cumulative observations. Credits[3]

distribution for all observed occlusion cases for all the scenes (note that S1 and S10 did not have any occlusion cases).

5 DISCUSSION

This study took inspiration from the film domain, where the directors use their know-how of visuospatial cues to direct viewers' attention. In that sense, these cues are a working prototype in the hands of filmmakers, and we use that to understand human perception and set criterias for building human-centric applications. Additionally, with high-low attentional synchrony, which is a simple case of high-low gaze clustering of multiple viewers towards a common point in a scene, we bring forth a novel way of analysing and investigating visual and event perception. Use cases of the showcased approach are many, specifically in areas of human-centred design, social-robotics, autonomous driving, AI methods on human events and benchmarking datasets. Finally, we put forth this study in support of the need to study human behaviour in ecologically valid natural settings in order to facilitate uniform and replicable studies.

ACKNOWLEDGMENTS

We would like to thank all the participants who performed the semantic evaluations for all the scenes in this article and the anonymous reviewers for their helpful comments and suggestions.

REFERENCES

Mehul Bhatt. 2018a. Cognitive media studies : Potentials for spatial cognition and AI research. *Cognitive Processing* 19, Suppl. 1 (2018), S6–S6. https://doi.org/10.1007/s10339-018-0884-3

Mehul Bhatt. 2018b. Minds. Movement. Moving Image. *Cognitive Processing* 19, Suppl. 1 (2018), S5–S5. https://doi.org/10.1007/s10339-018-0884-3

Tanya J Clarke, Mark F Bradshaw, David T Field, Sarah E Hampson, and David Rose. 2005. The perception of emotion from body movement in point-light displays of interpersonal dialogue. *Perception* 34, 10 (2005), 1171–1180.

James E Cutting. 2005. Perceiving scenes in film and in the world. *Moving image theory: Ecological considerations* (2005), 9–27.

James E Cutting and Lynn T Kozlowski. 1977. Recognizing friends by their walk: Gait perception without familiarity cues. *Bulletin of the psychonomic society* 9, 5 (1977), 353–356.

Martha J Farah, James W Tanaka, and H Maxwell Drain. 1995. What causes the face inversion effect? *Journal of Experimental Psychology: Human perception and performance* 21, 3 (1995), 628.

J Randall Flanagan and Roland S Johansson. 2003. Action plans used in action observation. *Nature* 424, 6950 (2003), 769–771.

Julie Grèzes. 1998. Top down effect of strategy on the perception of human biological motion: A PET investigation. *Cognitive Neuropsychology* 15, 6-8 (1998), 553–582.

Paul Hemeren and Yves Rybarczyk. 2020. The Visual Perception of Biological Motion in Adults. In *Modelling Human Motion*. Springer, 53–71.

Shaul Hochstein and Merav Ahissar. 2002. View from the top: Hierarchies and reverse hierarchies in the visual system. *Neuron* 36, 5 (2002), 791–804.

Gunnar Johansson. 1973. Visual perception of biological motion and a model for its analysis. *Perception & psychophysics* 14, 2 (1973), 201–211.

Seokmin Kang and Barbara Tversky. 2016. From hands to minds: Gestures promote understanding. *Cognitive Research: Principles and Implications* 1, 1 (2016), 1–15.

Vasiliki Kondyli, Jakob Suchan, and Mehul Bhatt. 2022. Grounding Embodied Multimodal Interaction: Towards Behaviourally Established Semantic Foundations for Human-Centered AI. In *First International Workshop on Knowledge Representation for Hybrid Intelligence (KR4HI 2022)., part of International Conference on Hybrid Human-Artificial Intelligence (HHAI 2022), Amsterdam, The Netherlands.*

Atesh Koul, Marco Soriano, Barbara Tversky, Cristina Becchio, and Andrea Cavallo. 2019. The kinematics that you do not expect: Integrating prior information and kinematics to understand intentions. *Cognition* 182 (2019), 213–219.

Lester C Loschky, Adam M Larson, Joseph P Magliano, and Tim J Smith. 2015. What would Jaws do? The tyranny of film and the relationship between gaze and higher-level narrative film comprehension. *PloS one* 10, 11 (2015), e0142474.

Vipul Nair, Paul Hemeren, Alessia Vignolo, Nicoletta Noceti, Elena Nicora, Alessandra Sciutti, Francesco Rea, Erik Billing, Francesca Odone, and Giulio Sandini. 2020. Action similarity judgment based on kinematic primitives. In *2020 Joint IEEE 10th International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*. IEEE, 1–8.

Han Sloetjes and Peter Wittenburg. 2008. Annotation by category-ELAN and ISO DCR. In *6th international Conference on Language Resources and Evaluation (LREC 2008).*

Tim J Smith, Daniel Levin, and James E Cutting. 2012. A window on reality: Perceiving edited moving images. *Current Directions in Psychological Science* 21, 2 (2012), 107–113.

Tim J Smith and Parag K Mital. 2013. Attentional synchrony and the influence of viewing task on gaze behavior in static and dynamic scenes. *Journal of vision* 13, 8 (2013), 16–16.

Jakob Suchan and Mehul Bhatt. 2016a. The geometry of a scene: On deep semantics for visual perception driven cognitive film, studies. In *2016 IEEE Winter Conference on Applications of Computer Vision, WACV 2016, Lake Placid, NY, USA, March 7-10, 2016*. IEEE Computer Society, 1–9. https://doi.org/10.1109/WACV.2016.7477712

Jakob Suchan and Mehul Bhatt. 2016b. Semantic question-answering with video and eye-tracking data: AI foundations for human visual perception driven cognitive film studies. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*. 2633–2639.

Jakob Suchan, Mehul Bhatt, and Srikrishna Varadarajan. 2019. Out of sight but not out of mind: an answer set programming based online abduction framework for visual sensemaking in autonomous driving. *arXiv preprint arXiv:1906.00107* (2019).

Paolo Viviani and Natale Stucchi. 1992. Biological movements look uniform: evidence of motor-perceptual interactions. *Journal of experimental psychology: Human perception and performance* 18, 3 (1992), 603.

APPENDIX

Table 3: A detailed description of the different visuospatial features shown in Table.1

Visuospatial Features	Multimodal Interaction (non-exhaustive list)	Description	Example
Scene Elements			
Types (Taxonomy)	object – dynamic – person, animal, ... body-parts – face, head, hands, torso, ... vehicle – car, truck, motorcycle, train, ... gaze – gaze-point, scan-path, ... static – phone, bag, table, door, wall, ... region – corridor, elevator, doorway, window sill, train cabin, ...	Non-exhaustive list of elements –	expand based on context
Scene Structure			
Visibility	visible(X)	The entity is visible in scene	X is visible
Presence	present(X)	The entity is present (may or may not be visible) in scene	X is present
Motion	stationary(X), moving(X), turning(X), moving_towards(X, Y), moving_away(X, Y), moving_together(X, Y), moving_next_to(X, Y), turning_towards(X, Y), turning_away(X, Y)	Relative displacement of the entity with respect to (or irrespective) to other visible entity(s)	X is moving X moves towards Y
Spatial Position	behind(X, Y), front(X, Y), left(X, Y), right(X, Y), above(X, Y), below(X, Y), front_left(X, Y), front_right(X, Y), behind_left(X, Y), behind_right(X, Y)	Relative position of the entity with respect to another (or more than one entity)	X is in front of Y X is behind of Y, Z, W
Human Action	speaking(X)	Entity that is speaking	X is speaking
Head Movement	steady_head(X), turn_left_head(X), turn_right_head(X), turn_upwards_head(X), turn_downwards_head(X), turn_upwards_left_head(X), turn_upwards_right_head(X), turn_downwards_left_head(X),turn_downwards_right_head(X)	The different head movement types are described as a spatial motion with respect to the agent (X turned head towards own left)	X turns his/her/its head towards his/her/its left.
Gaze	looking_at(X, Y)	Entity looking at another entity or object-of-interest	X is looking at Y
Hand Action	hold(X), pull(X), push(X), reaching_towards(X), grasp(X)	Hand action that aggregates towards one of these actions	X is pulling something
Body Pose	bending(X), crouching(X), kneeling(X), lean_backward(X), lean_forward(X), lean_sideways(X), leaning_against(X), sitting(X), lying_down(X), standing(X)	Posture of the Entity	X is in standing posture
Visual Attention			
Low-Level	fixation(ID), saccade(ID)	The data points(or taken from) to the eye-tracker	–
Object-Level	attention_on(face(X)), attention_on(head(X)), attention_on(hands(X)), attention_on(torso(X)), ...	Gaze of the viewer on which part of which entity pointing to the body/object parts under fixation	hands(kar)

Table 4: Summary of attention annotation for S1.

Scene S1	Body-Parts											
19 participants	Face			Head			Hands			Torso		
Entities	*freq*	*%*	*sec*	*freq*	*%*	*sec*	*freq*	*%*	*sec*	*freq*	*%*	*sec*
moriyama	132	11.8	121.8	51	4.5	40.9	51	4.5	54.5	49	4.4	25.1
shirai	225	20.1	259.1	111	9.9	95.2	45	4.0	33.6	60	5.4	37.1
iwanbuchi	14	1.2	7.7	5	0.4	2.4	2	0.2	0.9	1	0.1	0.8
nishi	194	17.3	231.5	44	3.9	31.8	102	9.1	78.3	35	3.1	16.2
TOTAL	565	50.4	620.2	211	18.8	170.3	200	17.8	167.2	145	12.9	79.1

Table 5: Summary of event (scene structure) annotation for S1.

Scene S1	Visuospatial features of events																	
	Visibility		Presence		Motion		Spatial Position		Gaze		Human Action		Head Movement		Hand Action		Body Pose	
Entities	*freq*	*sec*	*freq*	*sec*	*freq*	*sec*	*freq*	*sec*	*freq*	*sec*	*freq*	*sec*	*freq*	*sec*	*freq*	*sec*	*freq*	*sec*
shirai	2	136.5	2	136.5	29	136.5	8	136.5	10	64.6	6	10.3	25	128.1	3	90.6	6	136.5
moriyama	2	87.0	2	87.0	18	87.0	3	87.0	5	60.1	9	22.6	11	87.0	8	15.5	4	87.0
nishi	5	117.7	1	163.7	17	119.1	8	108.2	18	24.8	3	1.7	35	118.7	26	37.1	10	117.7
iwanbuchi	1	2.5	1	2.5	1	2.5	1	2.5	1	0.7	1	1.5	1	0.7	2	1.1	2	2.5
Total	10	343.8	6	389.8	65	345.1	20	334.2	34	150.3	19	36.2	72	334.5	39	144.3	22	343.7

Table 6: Summary of event (scene structure) annotation at the elemental level for S1.

Scene S1		**Entities**									
Visuospatial Features	**Relations**	**shirai**		**moriyama**		**nishi**		**iwanbuchi**		**TOTAL**	
		freq	*sec*	*freq*	*sec*	*freq*	*sec*	*freq*	*sec*	*freq*	*sec*
Visibility	visible	2	136.5	2	87.0	5	117.7	1	2.5	10	343.8
Presence	present	2	136.5	2	87.0	1	163.7	1	2.5	6	389.8
Motion	stationary	12	85.2	7	61.0	9	104.0	1	2.5	29	252.8
	moving	1	1.3	2	7.5	3	5.5			6	14.3
	turning	2	4.0	1	0.8	2	3.3			5	8.0
	moving_towards	2	2.5	4	12.1	1	0.8			7	15.5
	moving_away	5	31.6	1	1.0					6	32.7
	moving_together										
	moving_next_to	1	2.6			2	5.4			3	8.0
	turning_towards	3	5.4	2	3.0					5	8.4
	turning_away	3	3.7	1	1.5					4	5.3
SpatialPosition	behind	1	5.2			2	10.2			3	15.5
	front	5	101.5	1	79.1	1	2.8	1	2.5	8	186.6
	left			1	5.2	4	68.0			5	73.3
	right										
	above										
	below										
	front_left			1	2.5					1	2.5
	front_right	2	29.6							2	29.6
	behind_left										
	behind_right					1	26.9			1	26.9
HumanAction	speaking	6	10.3	9	22.6	3	1.7	1	1.5	19	36.2
HeadMovement	steady_head	8	83.1	6	81.8	11	73.9	1	0.7	26	239.5
	turn_left_head	6	10.3			4	4.2			10	14.5
	turn_right_head	3	2.1	1	0.4	3	2.7			7	5.2
	turn_upwards_head	3	9.3	1	1.3	2	2.3			6	13.0
	turn_downwards_head	4	18.7	1	0.9	2	1.1			8	23.8
	turn_upwards_left_head										
	turn_upwards_right_head	1	4.6	1	0.8	4	9.4			7	20.4
	turn_downwards_left_head			1	1.9	4	7.4			5	9.3
	turn_downwards_right_head										
Gaze	looking_at	10	64.6	5	60.1	18	24.8	1	0.6	34	150.2
HandAction	hold	1	89.9	2	6.6	5	13.0			8	109.5
	pull			1	1.0	2	1.4			3	2.4
	push			1	0.6	4	3.0	1	0.5	6	4.1
	reaching_towards	1	0.3	2	1.8	10	11.8	1	0.6	14	14.5
	grasp	1	0.4	2	5.5	5	7.9			8	13.7
BodyPose	bending			1	8.0					1	8.0
	crouching										
	kneeling										
	lean_backward										
	lean_forward	2	2.0			1	6.7	1	1.9	4	10.6
	lean_sideways										
	leaning_against										
	sitting					7	77.3	1	0.7	8	78.0
	lying_down										
	standing	4	134.5	3	79.1	2	33.7			9	247.3
TOTAL		91	975.7	62	620.0	118	790.6	11	16.5	284	2413.0