# Help Me to Help You: how to Learn Intentions, Actions and Plans

Harmish Khambhaita, Geert-Jan Kruijff, Matei Mancas, Mario Gianni, Panagiotis Papadakis, Fiora Pirri, Matia Pizzoli

# Help Me to Help You: how to Learn Intentions, Actions and Plans

**H. Khambhaita** and **G-J. Kruijff**
Language Technology Lab,
DFKI GmbH,
D-66123 Saarbruecken, Germany

**M. Mancas**
F.P.Ms/IT Research Center/TCTS Lab
University of Mons,
31, Bd. Dolez, 7000 Mons, Belgium

**M. Gianni** and **P. Papadakis** and **F. Pirri** and **M. Pizzoli**
ALCOR, Cognitive Robotics Lab,
Sapienza, University of Rome, DIS
via Ariosto 25, 00185 Rome, Italy

**Abstract**

The collaboration between a human and a robot is here understood as a learning process mediated by the instructor prompt behaviours and the apprentice collecting information from them to learn a plan. The instructor wears the Gaze Machine, a wearable device gathering and conveying visual and audio input from the instructor while executing a task. The robot, on the other hand, is eager to learn both the best sequence of actions, their timing and how they interlace. The cross relation among actions is specified both in terms of time intervals for their execution, and in terms of location in space to cope with the instruction interaction with people and objects in the scene. We outline this process: how to transform the rich information delivered by the Gaze Machine into a plan. Specifically, how to obtain a map of the instructor positions and his gaze position, via visual slam and gaze fixations; further, how to obtain an action map from the running commentaries and the topological maps and, finally, how to obtain a temporal net of the relevant actions that have been extracted. The learned structure is then managed by the flexible time paradigm of flexible planning in the Situation Calculus for execution monitoring and plan generation.

## 1 Introduction

In this paper we outline a collaboration model between human-robot in which the final goal is to learn the best actions needed to achieve the required goals (in this case, reporting hazards due to a crash accident in a tunnel, identifying the status of victims and, possibly, rescuing them). The collaboration is here viewed as a learning process involving the extraction of the correct information from the instructor behaviours. The instructor communicate his actions both visually and with the aid of his comments delivered while executing the actions.

In particular, actions and intentions are obtained by elaborating on the instructor path, while inspecting the accident place, what he[1] looks at, together with his running commentaries recorded via the Gaze Machine (GM). The GM (early described in (Marra and Pirri 2008) and in (Belardinelli, Pirri, and Carbone 2007)), worn by the instructor, is a complex wearable device, illustrated in Figure 1 and Figure 3, that allows to gather several perceptual data from a subject executing a task.

Figure 1: The instructor Salvo Candela with the Gaze Machine

The extraordinary vantage point obtained by the Gaze Machine enables an agent to observe, at any time step, not only what effectively the tutor is doing and communicating it but also how the tutor adapts his behaviours, by instantiating with common sense the prescribed laws, that is, those usually regulating his conduct in similar circumstances. It allows to get his intentions, tracking the relationship between saccades and motion towards a direction, namely something interesting in the scene. Finally, by the joint localisation of the instructor's gaze, his current position and his running commentaries and the noise in the scene, it is possible to infer affordances, namely a well defined sequence of the preferred interactions between the instructor and the surroundings.

From these extremely rich source of information an agent is in the condition of learning a well temporised sequence of actions and thus, to generate a suitable plan, in order to correctly operate in a difficult and hazardous environment.

In this paper we describe at a very general level, the following aspects of this learning and generation process:

1. We define two paths, the instructor path in the scene ob-



[1] A effective fire fighter instructor

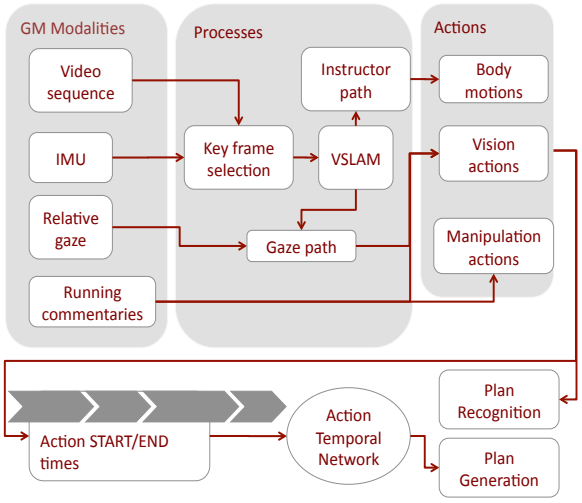


Figure 2: Schema of the flow of information and processing to learn actions from the collaboration instructor-robot, starting from the gaze machine.

tained by visually localising the instructor via the gaze machine (Section 2) and the instructor gaze path, obtained via the stereo pairs mounted on the GM and the cameras staring at the eye pupils.

2. From the two paths and a suitable segmentation and clustering of motion directions (of both body and head), both the motion and vision actions are obtained and labelled. On the other hand, as the instructor actively (and benignly) comments his behaviours, all the manipulation actions are identified by the the running commentaries and the association of head motions and body position (Section 3). Indeed, actions are defined as processes with a start and an end action, and with time varying.
3. A plan library of possible activities and affordances, according to the context, is defined a priori with the contribution of the instructor. In particular, the "what to do in such a situation" can be earlier formulated. According to the prior plan library and the effective sequence accomplished, following the instructor behaviours induced by *common sense* a flexible plan of action processes is generated, where the timelines are settled according to the flexible instantiation provided by the difference between coded rules and common sense (Section 4).

A schema of the model is given in Figure 2.

The problem of inferring a plan from the observations of actions, in the context of knowledge representation, is called *plan recognition*, and it has been earlier introduced by (Schmidt, Sridharan, and Goodson 1978; Kautz and Allen 1986; Kautz 1987). For a review of the consistency based and probabilistic based approaches to plan recognition see (Armentano and Amandi 2007). Geib (Geib 2009) introduced a method of plan-recognition where plan-library is first converted to a lexicon similar to that used in combinatory categorical grammar. By this way author is able to introduce concept of headedness, which avoids early commitments to plan and goal hypothesis in the process of plan-recognition, which eventually results in increased speed of the plan-recognition system. On the other hand in the realm of learning and computer vision the analogous concepts of acting based on observations have been specified as *action recognition, imitation learning* or *affordances learning*, as mainly motivated by the neurophisiological studies of Rizzolatti and colleagues (Pellegrino et al. 1992; Gallese et al. 1996) and by Gibson (Gibson 1977; 1955). Reviews on action recognition are given in (Moeslund, Hilton, and Krüger 2006; Poppe 2010; Aggarwal and Cai 1999) and on learning by imitation in (Argall et al. 2009; Schaal, Ijspeert, and Billard 2009).

The two approaches have, however, evolved in completely different directions. Plan recognition assumed actions to be already given and represented, in so being concerned only in the technical problems of generating a plan, taking into account specific preferences and user choices, and possibly interpreting plan recognition in terms of theory of explanations (Charniak and Goldman 1993). On the other hand action recognition and imitation learning has been more and more concerned with the robot ability to capture the real and effective sequence and to adapt it to changing contexts. As noted by Krüger and colleagues in (Krüger, Kragic, and Geib 2007) the terms action and intent recognition, in plan recognition, often obscure the real task achieved by these approaches. In fact, as far as plan recognition assumes an already defined set of actions the observation process is purely indexical. On the other hand the difficulties with the learning by imitation and action recognition approaches is that they lack important concepts such as execution monitoring, intention recognition and plan generation. The problem of learning a basic theory of actions from observations has been addressed in (Pirri 2010). The author shows how it is possible to automatically derive a model of the Situation Calculus from early vision, thus providing an example of bridging from perception to logical modeling.

Figure 3: The instructor while rescuing a victim.

Our contribution fosters a more tight integration between the plan recognition and learning approaches wherein the actions are segmented via the Gaze Machine and the instructor running commentary and the consequent plan recognition that is based on these actions.

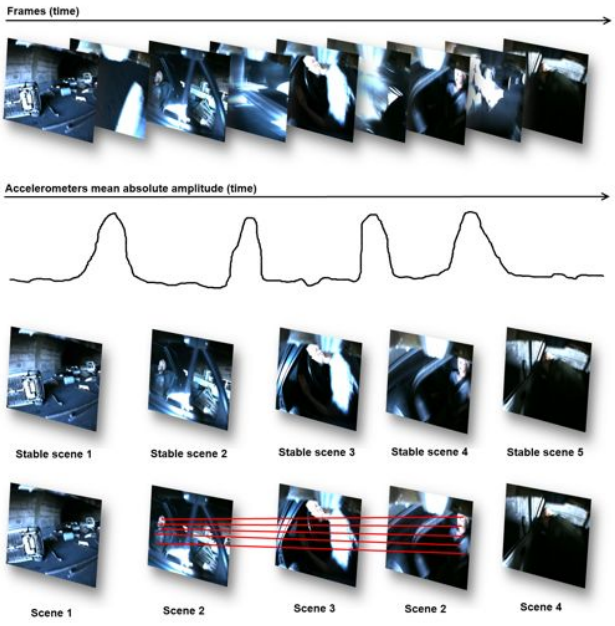


Figure 4: Key frames extraction. First row: acquired frames through time. Second row: acquired accelerometer absolute mean amplitude through time. Third row: frames corresponding to accelerometers peaks (movement from a LOI to another) are discarded. Forth row: features are extracted by the different key frames. If a lot of features match between the key frames of different scenes, this means that those scenes are the same and thus they are grouped together under the same label.

## 2 Visual Localisation

Two paths can be obtained by the instructor running in the disaster theatre. The first concerns the position of his body and the second the position of his gaze, not mentioning the position and direction of his head obtained via the inertial sensor placed on the GM.

Many popular approaches to real-time visual localisation and mapping rely on Extended Kalman Filter (EKF) (Davison et al. 2007) or Particle Filter (Pupilli and Calway 2006). The challenges that we face in this procedure stem from the inherent scene peculiarities of rescue environments as well as the loosely constrained movement of the camera setup which follows the movement of the instructor's head. In detail, the scene characteristics of a rescue environment include a wide range of lighting conditions and a plurality of solid but also non-solid obstacles (such as smoke). The position-orientation of the camera setup is also highly variable as the instructor rushes within the accident area due to looming hazards. As a consequence, we designed the Gaze Machine localisation to cope with the specific head motions and consequent change blindness (Simons and Levin 1998), and take advantage of the calibrated stereo pairs.

We can note that, since most of the computational effort is carried out off-line, we can take advantage of the techniques developed in the context of *Structure and Motion* recovery in order to deal with the higher variance in the camera motion and particular lighting conditions. The selection of a stable sequence of frames, that turn out to be key frames not only for the localisation and mapping process but also for action segmentation is a crucial step. This is more deeply discussed in Section 3, see Figure 4. Furthermore, we rely on well known methods for feature extraction and optical flow to predict the displacement of the tracked features and bundle adjustment between pairs of stereo images for motion estimation. In SAM problems, 3D structure is used to estimate the camera pose by resectioning. Thus, the computation of the motion is also complemented by the usage of dense disparity maps. The process goes through three main steps:

1. for each key frame build a Viewing Graph (Fig. 6);
2. given the estimated position at time $t$, compute the Essential Matrices and estimate the motion from time step $t$ to $t'$ (Hartley and Zisserman 2004);
3. bundle adjust among the estimated sequence of 3D structure and camera motion (Pollefeys et al. 2004).

Steps 1-2 provide a local consistency between different temporal frames. On the other hand they do not take into account sudden movements, which are filtered out in the key frame selection. In order to maintain a global consistency a bundle adjustment step is required where the re-projection error is minimised. Using the above described visual-based SLAM we are able to obtain an estimate of the instructor's path which, in turn, is used to derive the gaze path within the scene. It is interesting to note that due to inhibition of return, typical of the gaze when a salient feature come up hiding previous saliency levels, often a large amount of images are required in order to effectively track features. The viewing graph will tell on which configurations it is possible to rely in so avoiding the constraints induced by a sequence of pairs of images. Given the position of the instructor the localisation of his gaze is immediately obtained by the stereo pair.

In the following section the two paths are going to be segmented according to the recognition of actions from (i) the running commentary and (ii) the video sequence. The recognition of the actions will in turn enable to infer the spatiotemporal information of an action, Indeed, the two path prove to be essential for action segmentation as they can correctly specify where and when an action is performed as well as the corresponding spatiotemporal information of the instructor's gaze: what and when the instructor is gazing at, during a particular action.

## 3 Segmentation and Action Maps

In this section we discuss how we can segment the data acquired using the Gaze Machine to obtain a sequence of per-

formed actions. We shall also discuss the intention recognition via the *coup d'oeil*, i.e. how it is possible to extract the instructor's intention on the basis of his fixations and the spoken running commentaries.

A library of possible activities and affordances has been compiled in advance with the contribution of the instructor but, due to the high changeability of the scenario, the instructor will not follow a predefined, prioritised sequence of actions. The decision on what to do next is taken on the run, according to the task (i.e. plan the rescue) and the affordances characterising the scenario. The current instructor's intention involves what he is actually able to capture via attention. A saccade that is directed toward a location that is not involved in the current action may indicate a shift in the instructor's attention; depending on the associated saliency, this may or may not fire a head movement. However, also the information provided by the peripheral vision is enough to increase the situation awareness and take decisions. We, thus, introduce the concept of *coup d'oeil* to refer to those time instants in which something in the peripheral view fires a running commentary reporting something relevant in the scene.


Figure 5: Fixations from the tunnel sequence labelled by the instructor running commentaries: these are examples of key frames used for action segmentation and to define compatibilities; the third figure above induce the constraint *lookingAt(victim,t)* **during** *openingDoor(car,t')*.

The Gaze Machine records the instructor's saccade sequence by tracking his gaze in space. This is accomplished by projecting in the 3D scene the estimated point of regard. Scene structure is recovered via the Gaze Machine stereo rig while both pupils are tracked to extract visual axes (see Figure 1). A first kernel-based segmentation is performed to extract the fixation scan path from the acquired sequence of 3D points of regard. The main problem we address in this step is taking into account the instructor 3D position, as the 3D fixated points changes if the instructor moves.

The segmentation of the image flow acquired from the experienced firefighter is needed as a prior to further analysis of his actions. Key frame selection has been thoroughly investigated in the context of SAM recovery (Torr, Fitzgibbon, and Zisserman 1998; Pollefeys et al. 2004). In this paper we face the problem in the case of wearable cameras and unpredictable human motions. When performing some activity, a person is acquiring information (by gazing) in some important location in order to perform actions and then he moves to another location of interest (LOI). During the movement between two LOIs the acquired images are of little interest as they are most of the time fuzzy and very unstable. Moreover, the extensive visual disruptions caused by the firefighter fast motion imply a high probability of change blindness (Simons and Levin 1998), which decreases again the usability of the gaze data acquired during those periods. It is thus important to discard the frames which are recorded during the LOI change in order to extract the more stable scenes (Figure 4, third row). Finally, the firefighter can move from one LOI to another and then come back to the first LOI, or he can also be disturbed by some important bottom-up distractor which makes him turn his head and then he can look again to the previous scene. This shows that two stable scenes are not necessarily different scenes or LOIs (Figure 4, forth row).

A two-step approach can be used to extract meaningful scenes or key frames from the video flow: first the data from the accelerometer can provide cues on the head stability and then computer vision techniques are able to recognise already seen scenes or novel scenes. Figure 4 illustrates the process. The top row shows the successive frames through time. The shape of the absolute mean amplitude of the accelerometers located in the gaze machine is presented on the second row and shows picks during the firefighter movements between two LOIs and valleys during his stay in the same LOI. By discarding the frames which correspond with the accelerometer peaks, it is possible to keep only the stable scenes. Feature extraction and matching between those different scenes provide information to group together the scenes which are the same. If the features extracted from some key frames of one scene match a number of features above a given threshold on some key frames from another scene, this means that the two scenes are the same as it can be seen on Figure 4, forth row. In that case the two scenes are labelled with the same label.

Along with the instructor changes in position, pose and the running commentaries, 3D fixations are used to detect the starting/ending of an action. We are interested in producing a Map of basic actions, divided in 1) body motion, 2) vision actions and 3) manipulation actions, labelled with the correspondent starting/ending time.

Actions related to body motions are segmented on the basis of the instructor position in the 3D scene. Vision ac-

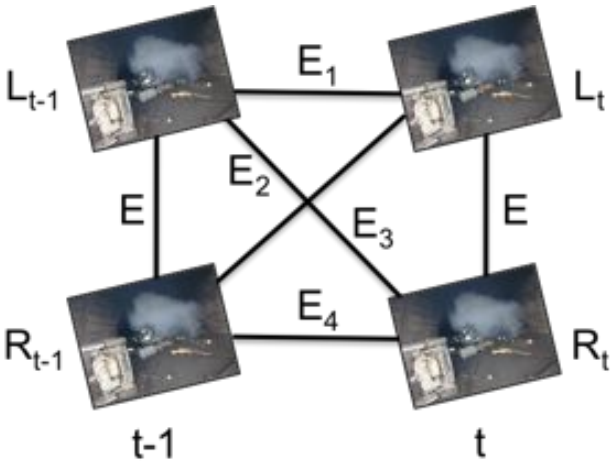


Figure 6: The Viewing Graph (Levi and Werman 2003; Rudi, Pizzoli, and Pirri 2010) used for the visual localisation. $L_t$ and $R_t$ are respectively the left and the right scene cameras at time t. $E_i$ is the Essential Matrix between different views.

tions involve the generation of a sequence of fixations and are detected by clustering in time and space and recognising special sequences in the running commentaries (i.e. *I see...*). A coupe d'oeil belongs to the vision actions category. It is detected making use of the special sequences in the running commentary and, when significant, sudden changes in inertial measurements. Indeed, the coupe d'oeil involves saccades followed by head movements and changes in body directions. For the detection of the the manipulation actions we completely rely on the running commentary, as the scene cameras on the GM don't provide a good point of view for gesture recognition.

From the Action Map we can define the compatibility conditions for generating a flexible plan. We assume we are given a plan library from the usual instruction on behaviours, and that this plan library includes affordances, given a specific rescue situation. Our aim is to show within the map the common sense raising from the choice of an action according to the urgency of a decision.

## 4 From Actions to flexible plan and plan recognition

The Actions map is constituted by a timeline indicating the time stamp of each action, the temporal relations among actions and the spatial cluster they belong to. The spatial cluster is obtained by the instructor path (see Section 3). Using the rules specified in the plan library and the Action Map the instructor plan execution can be suitably labelled for planning.

For example, according to the plan recognition algorithm of (Geib 2009), and using the specified plan library, we fist generate a combinatory categorical grammar (CCG) type plan-lexicon which maps observations to CCG categories. The algorithm results in a set of $explanations$, mentioning a goal and an ordered sequence of actions. The choice of assigning categories to observations is made according to specified $head$ value. $Headedness$ is a powerful method of controlling the space of possible explanations to be considered during the plan-recognition procedure.

In any case we mainly base the mapping from the Action Map to a possible plans via a temporal network compiling constraints and compatibilities within the Situation Calculus. Temporal relations specify how activities, such as $looking\ at\ a\ victim$ and of $opening\ the\ car\ door$ are correlate along time. For modelling both temporal constraints and cause-effect relations between activities we adopt, in fact, Temporal Flexible Situation Calculus(Finzi and Pirri 2005), accommodating Allen temporal intervals, multiple timelines among actions and concurrent situations. It intermediates between Situation Calculus formulae and temporal constraint networks. For example, the temporal relations illustrated in Figure 5, first row, last image, can be expressed by the compatibilities

$$T_c = [comp(lookingAt(victim, t), \\ [[(during, openingDoor(car))]])]$$

Here the compatibility states that the activities *look a victim*, involving vision actions, and *opening the car door* have to be performed according to the $during$ temporal relation. The temporal network associated with the compatibilities $T_c$ is represented in Figure 7. Therefore a way to generate a plan is to exploit the obtained temporal network and the flexible plan in the Situation Calculus.

## 5 Conclusion

In this work we have described a new framework for the collaboration between a human and a robot based on a wearable device, the Gaze Machine. This device creates a strong communication between the human, in this case an instructor, and the robot, by allowing the agent to look straightly into the perceptual flux of the companion. We have described how to process this perceptual information in order to obtain an Action Map. The Action Map is a rich labelled graph, starting from which it is possible to use specific methods, such as the transformation from a temporal network to a flexible plan and plan recognition, to generate a plan for the robot to correctly explore the environment.

## Acknowledgements

This paper describes research done under the EU-FP7 ICT 247870 *NIFTI* project.

## References

Aggarwal, J. K., and Cai, Q. 1999. Human motion analysis: A review. *Computer Vision and Image Understanding* 73:428–440.

Argall, B. D.; Chernova, S.; Veloso, M.; and Browning, B. 2009. A survey of robot learning from demonstration. *Robot. Auton. Syst.* 57(5):469–483.

Armentano, M. G., and Amandi, A. 2007. Plan recognition for interface agents. *Artif. Intell. Rev.* 28(2):131–162.

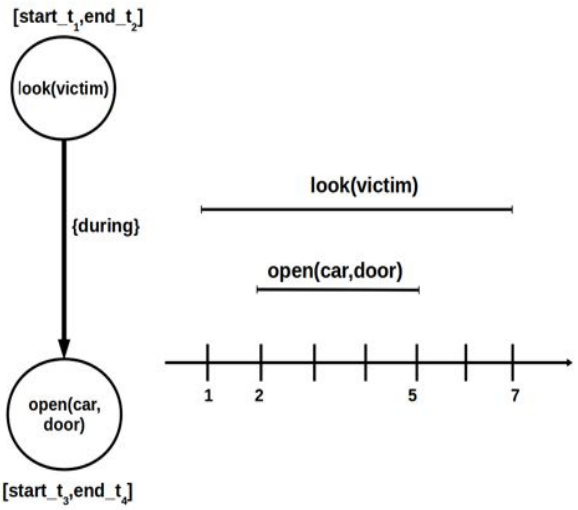


Figure 7: Temporal constraint network represented by $T_c$ along the timelines $do([start_{look}(victim, t_1), end_{look}(victim, t_2), S_0])$ and $do([start_{open}(car, door, t_3), end_{open}(car, door, t_4)], S_0)$


Belardinelli, A.; Pirri, F.; and Carbone, A. 2007. Bottom-up gaze shifts and fixations learning by imitation. *IEEE Transactions on Systems, Man, and Cybernetics, Part B* 37(2):256–271.

Charniak, E., and Goldman, R. P. 1993. A bayesian model of plan recognition. *Artif. Intell.* 64(1):53–79.

Davison, A. J.; Reid, I. D.; Molton, N. D.; and Stasse, O. 2007. Monoslam: Real-time single camera slam. *IEEE Trans. Pattern Analysis and Machine Intelligence* 29:2007.

Finzi, A., and Pirri, F. 2005. Representing flexible temporal behaviors in the situation calculus. In *Proceedings of IJCAI-2005*, 436–441.

Gallese, V.; Fadiga, L.; Fogassi, L.; and Rizzolatti, G. 1996. Action recognition in the premotor cortex. *Brain* 119:593–609.

Geib, C. 2009. Delaying commitment in plan recognition using combinatory categorial grammars. In *Proc. of the IJCAI 2009*, 1702–1707.

Gibson, J. 1955. Perceptual learning: differentiation or enrichment? *Psyc. Rev.* 62:32–41.

Gibson, J. 1977. The theory of affordances. In Shaw, R., and Bransford, J., eds., *Perceiving, Acting, and Knowing: Toward an Ecological Psychology*. Hillsdale, NJ: Lawrence Erlbaum. 67–82.

Hartley, R. I., and Zisserman, A. 2004. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition.

Kautz, H. A., and Allen, J. F. 1986. Generalized plan recognition. In *AAAI*, 32–37.

Kautz, H. A. 1987. *A formal theory of plan recognition*. Ph.D. Dissertation, Department ofComputer Science, University of Rochester.

Krüger, V.; Kragic, D.; and Geib, C. 2007. The meaning of action a review on action recognition and mapping. *Advanced Robotics* 21:1473–1501.

Levi, N., and Werman, M. 2003. The viewing graph. *CVPR*.

Marra, S., and Pirri, F. 2008. Eyes and cameras calibration for 3d world gaze detection. In *ICVS*, 216–227.

Moeslund, T. B.; Hilton, A.; and Krüger, V. 2006. A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding* 104(2-3):90–126.

Pellegrino, G. D.; Gallese, V.; Fadiga, L.; Fogassi, L.; and Rizzolatti, G. 1992. Understanding motor events: a neurophysiological study. *Exp. Brain Research* 91:176–180.

Pirri, F. 2010. The well-designed logical robot: learning and experience from observations to the situation calculus. *Artificial Intelligence* 1–44.

Pollefeys, M.; Van Gool, L.; Vergauwen, M.; Verbiest, F.; Cornelis, K.; Tops, J.; and Koch, R. 2004. Visual modeling with a hand-held camera. *Int. J. Comput. Vision* 59(3):207–232.

Poppe, R. 2010. A survey on vision-based human action recognition. *Image and Vision Computing* 28:976–990.

Pupilli, M., and Calway, A. 2006. Real-time visual slam with resilience to erratic motion. *Proc. CVPR* 1:1244– 1249.

Rudi, A.; Pizzoli, M.; and Pirri, F. 2010. Linear solvability in the viewing graph. In *Proceedings of the Tenth Asian Conference on Computer Vision*.

Schaal, S.; Ijspeert, A.; and Billard, A. 2009. Computational approaches to motor learning by imitation. *Philosophical Trans. of the Royal Soc. B: Biological Sciences* 358(1431):537–547.

Schmidt, C. F.; Sridharan, N. S.; and Goodson, J. L. 1978. The plan recognition problem: An intersection of psychology and artificial intelligence. *Artif. Intell.* 11(1-2):45–83.

Simons, D. J., and Levin, D. T. 1998. Failure to detect changes to people during a real-world interaction. *Psychonomic Bulletin and Review* 5:644–649.

Torr, P.; Fitzgibbon, A.; and Zisserman, A. 1998. Maintaining multiple motion model hypotheses over many views to recover matching and structure. 485 –491.