Rule-Based High-Level Situation Recognition from Incomplete Tracking Data

David Münch[1], Joris IJsselmuiden[1], Ann-Kristin Grosselfinger[1], Michael Arens[1], and Rainer Stiefelhagen[12]

[1] Fraunhofer IOSB, Germany, {david.muench|joris.ijsselmuiden|ann-kristin.grosselfinger|michael.arens}@iosb.fraunhofer.de
[2] Karlsruhe Institute of Technology, Germany, rainer.stiefelhagen@kit.edu

Abstract. Fuzzy metric temporal logic (FMTL) and situation graph trees (SGTs) have been shown to be promising tools in high-level situation recognition. They generate semantic descriptions from numeric perceptual data. FMTL and SGTs allow for sophisticated and universally applicable rule-based expert systems. Dealing with incomplete data is still a challenging task for rule-based systems. The FMTL/SGT system is extended by interpolation and hallucination to become capable of incomplete data. Therefore, one analysis to the robustness of the FMTL/SGT system in situation recognition is removing parts of the ground truth input tracks. The recognition results are compared to ground truth for situations such as "load object into car". The results show that the presented approach is robust against incomplete data. The contribution of this work is, first, an extension to the FMTL/SGT system to handle incomplete data via interpolation and hallucination, second, a knowledge base for recognizing vehicle-centered situations.

Keywords: rule-based expert system, fuzzy metric temporal logic, situation graph trees, semantic video understanding

1 Introduction

High-level situation recognition is the process of generating semantic descriptions from a scene observed through machine perception. First, video data needs to be processed by computer vision to obtain corresponding tracks for people, vehicles, and other objects of interest. Second, these tracks need to be processed by high-level situation recognition to detect the occurrence of interesting situations. For this contribution, we concentrate on deducing high-level situations as "loading an object into a car" from tracking data in a surveillance context, as e.g. Figure 1 depicts.

High-level situation recognition should be able to handle any form of uncertainty. This can be partial knowledge of the current state of the world, phenomena which are not observed by our model, and noisy observations. One kind of noisy observations is incomplete data from machine perception, in this case

Fig. 1. Two scenes from the VIRAT video dataset [10]. Car park *VIRAT_S_000002* (left) and car park *VIRAT_S_000200* (right). Typical situations in this context are getting into or getting out of a vehicle and loading or unloading an object.

video-based tracking. Data gaps can occur when objects are occluded, when objects move through areas without sensor coverage, or when machine perception experiences technical problems. For this contribution, existing methods were extended to handle incomplete data.

This article is structured as follows: Section 2 provides a short overview of related work in high-level situation recognition. Our own approach, the methodological improvements to handle incomplete data, and the particular SGT and FMTL rules for a prototypical surveillance scenario are presented in Section 3. Section 4 describes the evaluation and results. Section 5 provides a conclusion.

2 Related Work

A broad overview in situation recognition is given in the survey papers [1, 5, 11]. The whole field can be roughly divided into two main architectural strategies. On the one hand, there are direct approaches working directly on videos. On the other hand, there are the hierarchical approaches built of several layers. The basic idea of using several layers is splitting up the whole recognition process into specialized recognition methods. Usually, there are some methods performing object detection and tracking, others are combining the gathered information in a temporally and spatially limited context, and finally upon this information the high-level situation recognition is performed. Hierarchical approaches are divided into statistical methods often based on probabilistic graphical models such as Bayesian networks or Markov models, syntactic approaches representing actions trough symbols and combining them to situations with grammar-like structures, and description-based approaches using formal languages such as logic to describe situations. Usually, the latter rely on temporal and spatial properties to describe situations [4].

SGTs were presented as knowledge representation for situation recognition based on FMTL in [9]. [6, 7] extended the situation recognition framework to concurrent multi-hypothesis inference and optimized the runtime performance for real-time operation in several domains.

3 Methods

The general framework that underlies the high-level situation recognition is the layered model for cognitive vision systems initially described in [8]. FMTL is a powerful logic which can deal with notions of fuzziness and time. Handling fuzziness allows for handling both uncertainty and inherently vague concepts in the inference process itself. In [2] FMTL and SGTs are applied to the traffic domain and [3] applies them to human behavior. The advantages of using SGTs are the integrated modeling of knowledge, defining the rules and the inference algorithm in a precise formalism and the consolidation in one powerful framework – the SGT-Editor. The internal representation of SGTs is in FMTL rules. The inference algorithm for SGTs is programmed in FMTL, too. Thus, the whole situation recognition is built upon formal FMTL. This allows precise and fast inference about complex information of a particular scene.

3.1 Handling Incomplete Data

Interpolation of Input Data For a deduction at time t, data $x(t)$ from interval $[t - \Delta t_1, t + \Delta t_2]$ is used. Δt_1 and Δt_2 are dependent on the temporal range of the applied FMTL rules. If data is missing from t to $t + \Delta t_0$, each $x(t')$ with $t' \in [t, t + \Delta t_0]$ should be calculated as the average over all corresponding values in $T_p = [t - \Delta t_1, t - 1]$ and $T_f = [t + \Delta t_0 + 1, t + \Delta t_0 + \Delta t_2]$ with Δt_1 and Δt_2 chosen freely:

$$x(t') = \frac{1}{|T_p| + |T_f|} \left(\sum_{t_p \in T_p} (w_{x(t_p)} \cdot x(t_p)) + \sum_{t_f \in T_f} (w_{x(t_f)} \cdot x(t_f)) \right). \quad (1)$$

The weights $w_{x(t_p)}$ and $w_{x(t_f)}$ can be used to reflect a larger influence of T_p or T_f when t' is closer to the beginning or the end of $[t, t + \Delta t_0]$ respectively. This procedure works well for linear metric values, and radial values need to be handled differently with radial metrics.

Hallucinating High-Level Evidence. When rule-based systems are getting more complex they consequently have to deal with increasing challenges of noisy and incomplete input data. The general drawbacks of such a system are when trying to instantiate the preconditions of any situation scheme and all of the rules can be satisfied except of a very few ones which leads to a discontinuation of the situation recognition of the current path of inference.

To overcome this drawback we extended the situation recognition inference algorithm presented in [7] to hallucinate missing evidence. If the predicted situation scheme cannot be instantiated due to missing evidence, the missing evidence is hallucinated. That means, the algorithm creates satisfied dummy predicates for the missing evidence so that the situation scheme can be instantiated. It is, of course, internally known which situation schemes are hallucinated. And finally, the situation graph traversal gets continued with the new hallucinated situation scheme.

3.2 Knowledge Base

Figure 2 depicts the SGT representing the knowledge for the situation recognition. The right specialization edge emerging from *Root* is not included in this paper since it specializes for situations that are not evaluated in Section 4 like people approaching each other, standing together, and walking together.

For a detailed description on traversing the SGT to recognize situations refer to [7]. The traversal starts in the *Root* situation scheme and continues along the left specialization edge emerging from *Root*. Then, the start situation scheme *PatientCar* instantiates a car as the patient for the current agent. From there, the situation schemes *CarFar* and *CarNear* can be reached through temporal edges and so on.

The head of an FMTL rules activated by the SGT in Figure 2 is e.g. $\boldsymbol{HaveDistance}(agent, patient, category)$. The body of this rule consists of $DistanceIs(agent, patient, distance) \wedge AssociateDistance(distance, category)$. $DistanceIs(agent, patient, distance)$ calculates the Euclidian distance which is then associated with distance categories using $AssociateDistance(distance, category)$ as described e.g. in Figure 7 in [2].

4 Evaluation

Experimental Setup We implemented and evaluated the proposed method on videos with annotated ground truth data from the VIRAT video dataset, see Figure 1. The VIRAT Video Dataset Release 1.0 was made publicly available in 2011 and is presented in [10]. For three out of six places there exist ground truth annotated files where each object or person of interest is annotated. Additionally, in a second file there are annotated semantically interesting situations and all the participating agents in the environment of a car park. The annotated vehicle-centered situations comprising persons, objects and cars are getting into or getting out of a vehicle, opening or closing trunk, and loading or unloading an object of a vehicle.

The provided ground truth annotated data of the VIRAT video dataset is regarded as complete information. In this evaluation every experiment was performed ten times with a probabilistic unique removal of data.

First, we extended the situation recognition system as mentioned in Section 3.1 to be capable of incomplete data. Second, we developed the lower level basic knowledge which is represented in FMTL rules. This universally valid knowledge is domain independent and does not need to be changed when a different domain is considered. Third, the knowledge about the expected situations is encoded in an SGT. Consequently, the specific SGT, see Figure 2, describes all of the expected situations of a certain domain.

Results We choose the following six videos to evaluate on, due to the availability of annotations and the occurrence of different situations. From scene 00 we

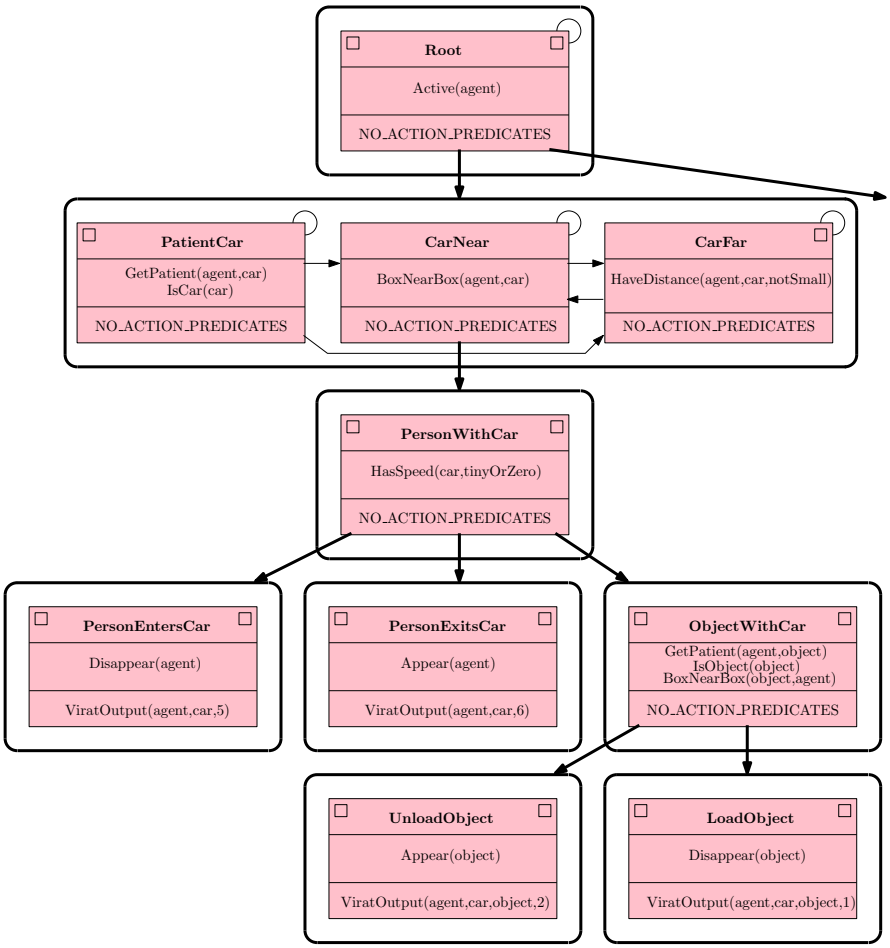

Fig. 2. Part of the SGT representing the knowledge to detect the expected vehicle-centered situations used in the evaluation. The basic structural element of an SGT is a situation scheme which is identified by a unique name, a precondition, and a postcondition both out of one or more FMTL predicates. An example is the "Root" situation scheme with the precondition *Active(agent)* and without any postcondition. A situation scheme can be a start resp. end situation which is marked with a small box on the upper left resp. right of the situation scheme. Thin edges represent the temporal structure of the situation schemes within a unit visualized with a thick box called situation graph. Thick edges from a single situation scheme to a situation graph model the conceptional refinement of a situation scheme. The resulting structure is a hypergraph and is called SGT. In this figure the situation schemes *PersonEntersCar*, *PersonExitsCar*, *UnloadObject*, and *LoadObject* raise as postcondition a message that they could be instantiated with a distinct configuration of the variables.

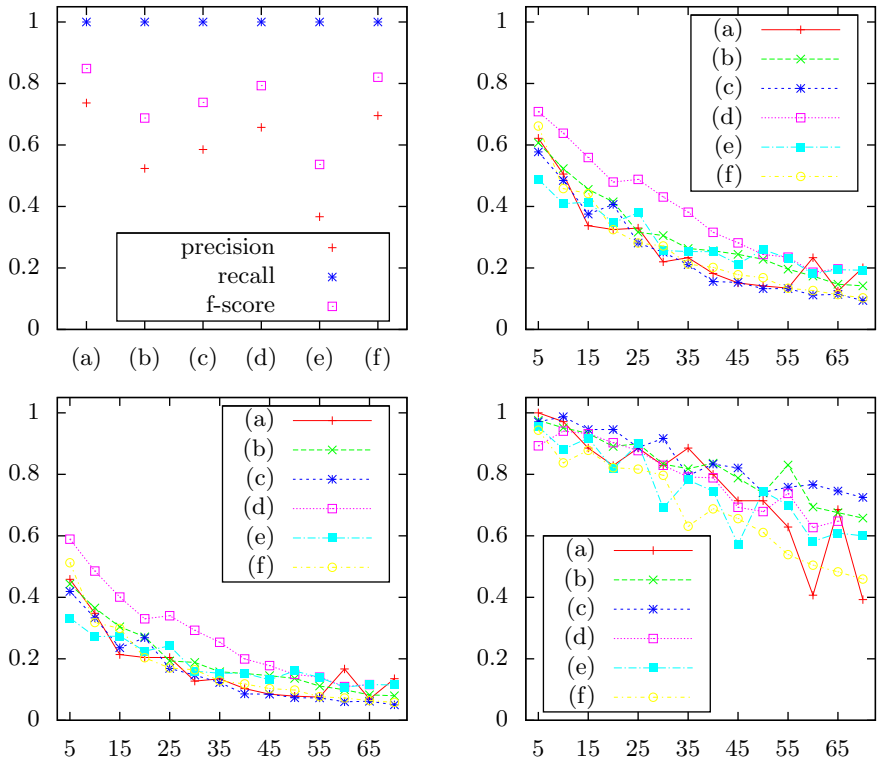

Fig. 3. The results of the unmodified, original scenarios of the six different videos of the VIRAT dataset are visualized in terms of precision, recall, and f-score (upper left). F-score of all evaluated videos with gap size of 5 seconds (upper right), precision (lower left), and recall (lower right). The horizontal axis equals the removed data in percent.

selected sequence 02 (a), 03 (b), 04 (c), and 06 (d); from scene 02 we selected segment 06 (e) of sequence 00 and segment 00 (f) of sequence 02.

The classification rates of the performed experiments on the six different video sequences are shown in Figure 3 (upper left). The recall is throughout all the six sequences equal to 1.0, which means, that the proposed method never misses any interesting situation in the testset. The average precision is far from 1, the f-score, of course, is slightly better. Some false positive classification results cause the bad precision, but we argue that this is not as disappointing because every single occurring situation was recognized.

Figure 3 depicts f-score (upper right), precision (lower left), and recall (lower right) of all evaluated videos. The figures show that the proposed approach is capable of handling incomplete data even if more than half of the data is missing.

Figure 4 shows the ROC-curves of video (d) for a gap size of 5 seconds (left). The false positive rate slightly increases for larger amounts of missing data and

the larger the gaps, the true positive rate decreases slightly. The same evaluation without data interpolation and hallucinating performs worse (right).

Figure 5 (left) consists of three different test configurations: gap sizes of 1, 3, and 5 seconds in video (d). Gap sizes of 1 and 3 perform slightly similar; larger gaps of size 5 result in a roughly worse result. Figure 5 (right) shows f-score of video (d) with gap size 5 including the error bars of the three standard deviations.

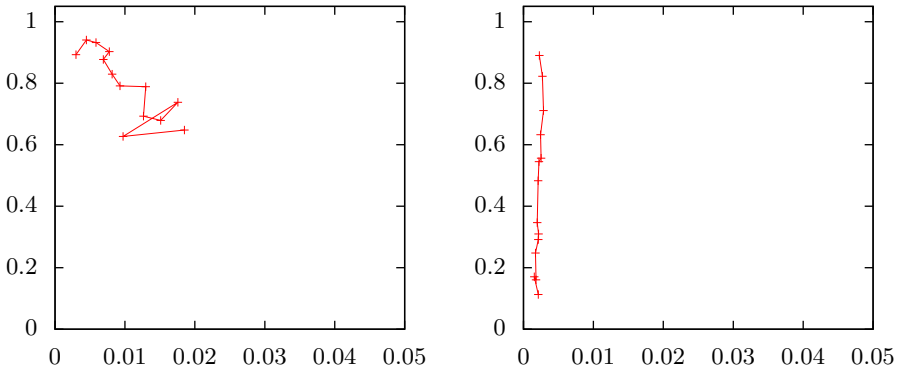

Fig. 4. For video (d) with gap size five seconds the ROC-curves are shown (left). True positive rate on vertical axis; false positive rate on horizontal axis. Without data interpolation and hallucinating (right).

5 Conclusion

We have presented a cognitive vision system that can deal with incomplete data in the application of situation recognition in a video surveillance setup. The main ideas to deal with incomplete data in a rule-based expert system are on the lower tier the interpolation of input data and its uncertainty and on the upper tier the extension of the situation recognition inference algorithm. These two extensions allows our system both to deal with ordinary incomplete data and to handle *high-level incomplete data* such as occlusions. The contribution of this work is the extension of the SGT-Editor and the formal situation recognition inference algorithm to handle incomplete data. As well as developing a knowledge base for recognizing vehicle-centered situations and the broad evaluation of the VIRAT video dataset on a high semantic level. To the best of our knowledge nobody has evaluated the VIRAT video dataset on a high semantic level before.

Acknowledgements. The authors would like to thank Yvonne Fischer and Wolfgang Hübner for fruitful discussions and for their contributions leading to the success of this work.

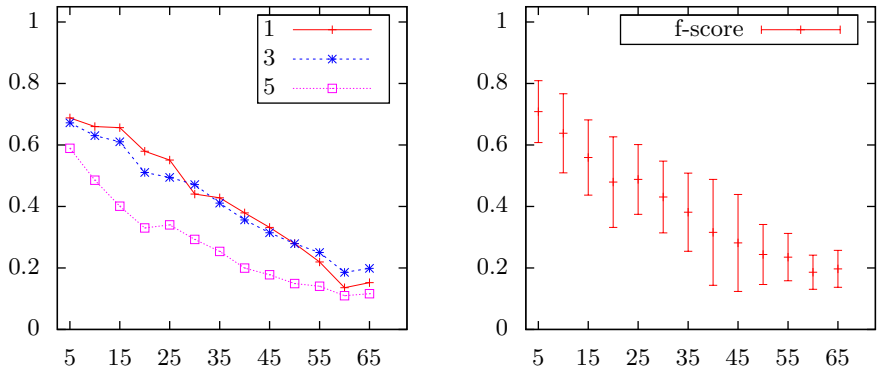

Fig. 5. Video (d) with gap sizes of 1, 3, and 5 seconds (left). F-score (right) of video (d) with gap size 5 including the error bars of the three standard deviations.

References

1. Aggarwal, J.K., Ryoo, M.S.: Human Activity Analysis: A Review. ACM Computing Surveys (2011)
2. Gerber, R., Nagel, H.H.: Representation of occurrences for road vehicle traffic. Artificial Intelligence 172(4-5), 351 – 391 (2008)
3. González, J., Rowe, D., Varona, J., Roca, F.X.: Understanding dynamic scenes based on human sequence evaluation. Image and Vision Computing 27(10), 1433 – 1444 (2009), special Section: Computer Vision Methods for Ambient Intelligence
4. IJsselmuiden, J., Stiefelhagen, R.: Towards high-level human activity recognition through computer vision and temporal logic. In: Proceedings of the 33rd Annual German Conference on Advances in Artificial Intelligence (2010)
5. Lavee, G., Rivlin, E., Rudzsky, M.: Understanding video events: A survey of methods for automatic interpretation of semantic occurrences in video. IEEE Trans Syst Man Cybern C Appl Rev 39(5), 489 –504 (2009)
6. Münch, D., IJsselmuiden, J., Arens, M., Stiefelhagen, R.: High-level situation recognition using fuzzy metric temporal logic, case studies in surveillance and smart environments. In: ICCV Workshops. pp. 882–889 (2011)
7. Münch, D., Jüngling, K., Arens, M.: Towards a Multi-purpose Monocular Vision-based High-Level Situation Awareness System. In: International Workshop on Behaviour Analysis and Video Understanding (ICVS 2011). p. 10 (2011)
8. Nagel, H.H.: Image sequence evaluation: 30 years and still going strong. International Conference on Pattern Recognition 1, 1149 (2000)
9. Nagel, H.H.: Steps toward a cognitive vision system. AI Magazine 25(2), 31–50 (2004)
10. Oh, S., et. al.: A large-scale benchmark dataset for event recognition in surveillance video. In: CVPR. pp. 3153 –3160 (2011)
11. Turaga, P., Chellappa, R., Subrahmanian, V., Udrea, O.: Machine recognition of human activities: A survey. CSVT 18(11), 1473 –1488 (2008)