The original publication is available at www.springerlink.com
KI 2012: Advances in Artificial Intelligence, Lecture Notes in Artificial Intelligence, 2012, Volume 7526/2012, 233-236, DOI: 10.1007/978-3-642-33347-7

Supporting Fuzzy Metric Temporal Logic based Situation Recognition by Mean Shift Clustering

David Münch, Eckart Michaelsen, and Michael Arens

Fraunhofer IOSB, Gutleuthausstraße 1, 76275 Ettlingen, Germany
{david.muench|eckart.michaelsen|michael.arens}@iosb.fraunhofer.de

Abstract. This contribution aims at assisting video surveillance operators with automatic understanding of situations in videos. The situations comprise many different agents interacting in groups. To this end we extended an existing situation recognition framework based on Situation Graph Trees and Fuzzy Metric Temporal Logic. Non-parametric mean-shift clustering is utilized to support the logic-based inference process for such group-based situations, namely to improve efficiency. Additionally, the underlying knowledge base was augmented to also handle multi-agent queries and the situation inference was adapted to also handle inference for group-based situations. For evaluation the publicly available BEHAVE video dataset was used consisting of partially annotated real video data of persons. The results show that the proposed system is capable of correctly and efficiently understanding such group-based situations.

Keywords: Situation Recognition, Situation Graph Trees (SGT), Fuzzy Metric Temporal Logic (FMTL), Mean-Shift Clustering

1 Introduction

Automatic video understanding is an important and challenging task. Frequent queries in surveillance for security issues consider not primarily the actions of individuals but instead situations where a couple of humans act as a group. A knowledge-based logic understanding approach can handle such reasoning by introducing a group concept. This will be instantiated from data containing individuals based on predicates such as proximity. However, such concept may cause considerable computational effort. In such situations logical systems – in their emphasis of soundness – tend to lead to deep exponentially branching search. Here, benign predicates such as proximity – not only in space, but also in time or intention etc. – allow the utilization of machine learning methods to aid the search. In this work we propose an automatic video understanding system for assisting human operators in surveillance applications.

Situation recognition using Fuzzy Metric Temporal Logic and Situation Graph Trees in the domain of traffic is presented by [1], in the domain of human behavior in [3], and in the domain of video surveillance in [5]. [4] presents a way to include a kind of Hough-transform into a knowledge-based representation. This

Fig. 1. Two snapshots from the BEHAVE video dataset [2]. The proposed SGT/FMTL framework recognizes the situations *InGroup* (yellow, thick) and concurrently *Approach* (red, thin) of a person (left). Mean-shift clustering results of frame 5370 (right).

approach demonstrates how the combinatorial limitations of rule-based systems can be supported by prominent non-declarative methods such as clustering. The declarative aspect remains; moreover the declarative approaches become productive systems in real applications.

2 Methods

The SGT/FMTL framework was originally used within the cognitive vision system architecture described in [6]. The framework is extended in [5] to recognize multiple concurrent situations with each situation having an independent Degree of Validity. Basic knowledge is encoded in FMTL rules. On the one hand, basic knowledge is canonical knowledge such as relations like *Distance_is(agent,patient,distance)*, on the other hand these FMTL rules are concepts on a lower level with minor complexity such as *Have_distance(agent2, agent6,small)* which means that the distance of *agent2* and *agent6* is *small*. The knowledge about the expected situations in the domain of video surveillance is encoded in an SGT.

We assume a calibrated camera and a given transformation from the real observed scene to the image plane. The mean-shift clustering is performed in the provided ground plane of the observed scene. The density to be considered is the spatial and temporal proximity of persons. Figure 1 (right) depicts the mean-shift clustering result for an example image sequence where five persons are present. Two groups of two persons each are walking together and one single person is passing by one group. The applied clustering performs well without merging the single person with the group.

When there occur more agents the number of binary and n-ary relations to be examined by the inference process exponentially rises. We overcome this severe limitation and introduce list-based rules in the SGT/FMTL framework. Other languages such as Prolog support list-based computations. Motivated by its pure functionality we extended the knowledge base and inference process of FMTL by so called filters which apply predicates on a whole list, see Equation (1). Internally the *call/N* predicate is called recursively on the whole input list. In [7] the use of *call/N* is discouraged. Thus, we introduce and apply *call/3*

throughout this work. In Equation (1) the proposed *Truefilter* is shown. It is implemented as FMTL rule. The $\Box$ operator is the temporal *always* operator, all the other syntax is standard logic syntax. When trying to satisfy Equation (1), Equation (2) is recursively called until end recursion terminating. The variables of *Truefilter* are defined as follows: *res* contains all elements of *in* with *Fun(agent, elem, parameter)* true.

$$\Box\{\boldsymbol{Truefilter}(in, Fun, agent, parameter, res) \leftarrow \\ Truefilter\_(in, Fun, agent, parameter, res) \wedge\ res <> []\} \tag{1}$$

$$\Box\{\boldsymbol{Truefilter\_}([elem|in], Fun, agent, parameter, res) \leftarrow \\ functor = ..[Fun, agent, elem, parameter] \\ \wedge\ [(call(functor) \wedge\ res = [elem|new] \wedge\ !) \vee\ res = new] \\ \wedge\ Truefilter\_(in, Fun, agent, parameter, new)\} \tag{2}$$

Thus, the introduction of list-based rules in the SGT/FMTL framework allows easily recognizing situations where more than two agents are involved. However this combinatorial explosion of satisfying instances leads to a decreasing runtime of the FMTL inference engine.

3 Evaluation

The proposed methods were evaluated on the BEHAVE video dataset [2]. Not for every frame but for some parts of the video there exists annotated ground-truth. The situations of interest are: *InGroup, Approach, WalkTogether, Meet, Split, Ignore, Chase, Fight, RunTogether, and Following*. It has to be said that *Meet* occurs only once in the ground-truth and *Ignore* twice. Thus, both situations cannot be evaluated properly.

Table 2 (left) depicts the confusion matrix of frames $18000 - 23700$ of the BEHAVE video dataset. The true positive rate of the seven situations to be recognized is almost 1 when using an interval based measure as e.g. proposed in [8]. Thus, practically no situation is missed. But there do arise some false positives. The confusion matrix gives a short overview of these. In Figure 2 (right) the runtime of the presented approach without and with the mean-shift clustering from Section 2 is shown. It can be seen that applying mean-shift clustering on the whole BEHAVE video dataset reduces the runtime of the situation recognition significantly. Thus, in this case real-time processing is reached.

4 Conclusion

In this article the SGT/FMTL situation recognition framework was extended by the concept "group". Thus, situations in which groups of individuals interact can be described more naturally. In order to do so we made use of certain higher-order logic programming mechanisms processing logical queries on possibly large

	WalkTogether	RunTogether	InGroup	Approach	Split	Chase	Following
WalkToget.	.45	.1	.3	.1	0	0	.05
InGroup	0	0	.9	.05	.05	0	0
Approach	.05	0	.3	.6	.05	0	0
Split	0	0	.25	.05	.7	0	0

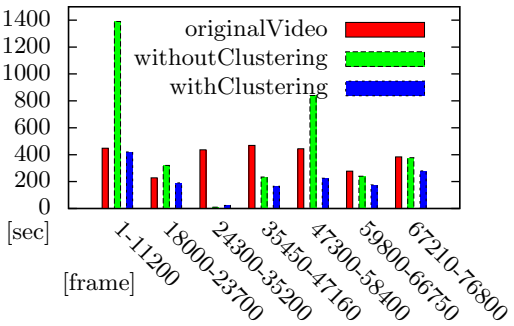

Fig. 2. The confusion matrix of situation recognition applied to frames $18000 - 23700$ (left). On the left the actual situation; above the recognized situation. The duration of the original video file and the runtime with and without the clustering (right).

data bases. Thus, the declarative SGT model describing the interaction of groups of people on surveillance videos turns out to be tractable in real-time on contemporary standard hardware. For verification we used the publicly available BEHAVE video dataset as representative example.

The introduced clustering concept needs further comparative evaluation on larger datasets. As real-time performance is achieved with this improvement the next steps are the integration into a multi-camera network. We have shown that the clustering performs well on basic "group" concepts; therefore we will investigate in how far such methods can be applied on higher levels such as how the same behavior of agents leads to groups.

References

1. Arens, M., Gerber, R., Nagel, H.H.: Conceptual representations between video signals and natural language descriptions. IVC 26(1), 53–66 (2008)
2. Blunsden, S., Fisher, R.: The behave video dataset: ground truthed video for multi-person behavior classification. Annals of the BMVA 2010(4), 1–12 (2010)
3. González, J., Rowe, D., Varona, J., Roca, F.X.: Understanding dynamic scenes based on human sequence evaluation. IVC 27(10), 1433 – 1444 (2009)
4. Michaelsen, E., Doktorski, L., Arens, M.: Shortcuts in production-systems. In: PRIA, vol. 2, pp. 30–38 (2008)
5. 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)
6. Nagel, H.H.: Steps toward a cognitive vision system. AI Mag. 25(2), 31–50 (2004)
7. Naish, L.: Higher-order logic programming in prolog. Tech. rep., In Workshop on MultiParadigm Logic Programming (1996)
8. Oh, S., et. al.: A large-scale benchmark dataset for event recognition in surveillance video. In: CVPR. pp. 3153 –3160 (2011)