ECAI 2014
T. Schaub et al. (Eds.)

doi:10.3233/978-1-61499-419-0-1031

Video Event Recognition by Dempster-Shafer Theory

Xin Hong, Yan Huang, Wenjun Ma, Paul Miller, Weiru Liu and **Huiyu Zhou** [1]

Abstract. This paper presents an event recognition framework, based on Dempster-Shafer theory, that combines evidence of events from low-level computer vision analytics. The proposed method employing evidential network modelling of composite events, is able to represent uncertainty of event output from low level video analysis and infer high-level events with semantic meaning along with degrees of belief. The method has been evaluated on videos taken of subjects entering and leaving a seated area. This has relevance to a number of transport scenarios, such as onboard buses and trains, and also in train stations and airports. Recognition results of 78% and 100% for four composite events are encouraging.

1 Introduction

A key to the success of active CCTV surveillance is the use of video analytics. Through video analysis process at the low layer, visual features that a computer vision system can extract and handle without human intervention, are mapped onto concepts (so called events) as perceived by humans, e.g. a male walking. Conceptual events can then be used to infer complex events that have significant semantics to humans with respect to decision making. Dynamic environments such as changing illumination and moving platforms make event recognition very challenging in video based applications. Main concerns include unreliable video equipment, imprecise output of low-level computer vision analytics, varying renditions of events of the same type, and similar appearance of events of different types [2].

Based on the Dempster-Shafer theory of Evidence (DS theory) [1] [3], we propose an evidential reasoning framework for event recognition in video surveillance. The proposed event network model can hierarchically represent structural relationships between composite events, atomic events, contexts and sensor evidence (output of low-level computer vision analytics). An embedded evidential reasoning mechanism provides the ability to numerically represent uncertainty in relation to event recognition, infer the occurrence of complex events with belief values, and make a decision on the most possible complex event having taken place based on the presented visual evidence.

2 Evidential Event Recognition

The proposed system is composed of components at two levels. At the low level, subjects are detected and video features are extracted, using computer vision techniques, to give low level semantic concepts such as a female face has been detected and a person has moved from the door towards the gang-way on bus. The high level modules of the system are designed to maintain the semantic hierarchy of events obtained from domain knowledge and human experts. At this level events of interest are recognised based on information given from the low-level modules with a degree of belief. In this paper, we focus on the investigation that occurs on the high level.

In the context of video surveillance, an *event* is an observation (or collection of observations) of significance to the end-user. An event can be simple or complex, which is composed of simpler events. To distinguish these two different concepts, we call the former as an atomic event and the latter as a composite event. An atomic event can be directly detected from sensors or video analytics, or derived from observations through video cameras. Atomic events can be aggregated to generate composite events which are more meaningful. Sensors and video analysis algorithms may provide the contexts with semantic meanings but may not be of users' interest. To avoid any confusion, we call those as context events.

To reflect the hierarchical structure of the relationships between composite events and atomic events, atomic events and context events, atomic events (also contexts) and outputs of video analytics, we propose an evidential network model for event recognition.

Definition 1 *An evidential event network (EEN) is a graph of upside-down tree $EEN = (ND, EG, MM)$, where:*

- ***ND** = $\{n_1, ..., n_N\}$ is a set of nodes representing events,*
- ***EG** is a set of edges over ND, each of which represents a close relation between the nodes at two consecutive layers,*
- ***MM** is a set of multi-valued mappings, which describe the compatibility relations between the node at the layer where an edge starts and the node at the layer where the edge ends.*

Definition 2 *In an evidential event network, an event node is a tuple:*

n = (nType, Level, Date, Time, source, r, Θ, m)

where *nType* is the descriptor of an event node, such as *Female boards the bus*. *Level* informs whether the event is *Context*, *Atomic* or *Composite*. *Date* and *Time* are related to the event that occurs on-site. *Source* denotes the unique identifier of a source such as a seat sensor and a gender classification algorithm. Here, we use numerical numbers, e.g. 1, 2, to indicate a source. *r* is the degree of reliability of a source. Θ is the frame of discernment that holds all its values. *m* is the mass function of the node. It is worth mentioning that for an inferred atomic or composite event, sources and *r* are not required.

Event inference starts from obtaining inputs from computer vision analysis modules and goes through three stages over an event network.

Stage 1 - At the beginning, evidence casted by low-level computer vision analysis on the context/atomic nodes, *i.e.* leaf nodes is represented by a mass function. Here, the reliability of computer vision analysis is evaluated by applying the *discounting* operation on the original mass functions as shown in Eq. 1.

$$m^r_{\Theta_l}(A) = \begin{cases} (1-r)m_{\Theta_l}(A), & A \subset \Theta_l \\ r + (1-r)m_{\Theta_l}(\Theta_l), & A = \Theta_l \end{cases} \quad (1)$$

[1] CSIT, School of EEECS, Queen's University Belfast, UK, email: {x.hong; w.ma; y.huang; p.miller; w.liu; h.zhou}@qub.ac.uk.

where l is a leaf node, $0 \leq r \leq 1$.

Stage 2 - Mass functions of an atomic event can be obtained if they have evidence contexts associated. There are two steps involved in this stage. Firstly, mass functions of a context node are translated to a linked atomic event node using Eq. 2. Secondly, if an atomic event node has two or more associated context nodes, the translated mass functions are aggregated following the Dempster's rule of combination [1] as shown in Eq. 3.

$$m^{i}_{\Theta_a}(A) = \sum_{B} m_{\Theta_{l_i}}(B) \tag{2}$$

where $A \subseteq \Theta_a$, $B \subseteq \Theta_{l_i}$, and there exists $A = \Gamma(B)$

$$m_{\Theta_a} = m^{1}_{\Theta_a} \oplus m^{2}_{\Theta_a} \oplus ... \oplus m^{I}_{\Theta_a} \tag{3}$$

Stage 3 - Belief functions of the composite event node are inferred from its associated atomic event nodes. Three steps are involved in this stage. In step 1, mass functions of an atomic event are translated to the composite event using Eq. 4. In step 2, the consensus is obtained by combining the translated mass functions from all the atomic events using Eq. 5. This step may iterate until its reaching the composite event as the root. In step 3, the pignistic probability $BetP$ [4] of the composite event is calculated using Eq. 6.

$$m^{i}_{\Theta_{cp}}(A) = \sum_{B} m_{\Theta_{a_i}}(B) \tag{4}$$

where $A \subseteq \Theta_{cp}$, $B \subseteq \Theta_{a_i}$, and there exists $A = \Gamma(B)$.

$$m_{\Theta_{cp}} = m^{1}_{\Theta_{cp}} \oplus m^{2}_{\Theta_{cp}} \oplus ... \oplus m^{I}_{\Theta_{cp}} \tag{5}$$

$$BetP(w) = \sum_{A} \frac{m(A)}{|A|} \tag{6}$$

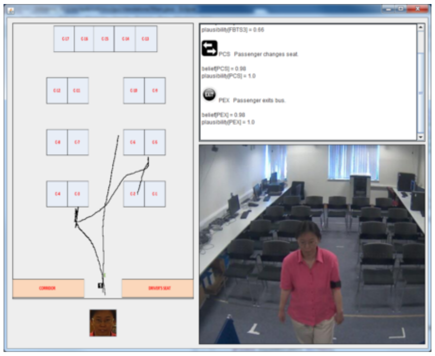
where $A \subseteq \Theta_{cp}$, $w \in A$.

The final decision is made based on the selection of the element which holds the highest pignistic probability.

3 Experimental Work

We have evaluated the performance of our proposed event recognition framework on the dataset collected from a simulated bus scenario. In our simulated bus environment, we use a camera (A) to view the entry, and a camera (B) to view the saloon. Camera A is positioned to capture a passenger's face as (s)he enters the bus. The imagery from camera A is provided as the input to a face detection module with a gender classification tool. The imagery from camera B is provided as the input to a human detection and spatial tracking module. A Vicon sensor is also worn by a passenger to provide ground truth motion.

We start from tracking passengers who board the bus and continuously track them as they move, sit and later alight from the bus. Within this context, there are four broad human activities: boarding, moving, sitting and alighting. We are particularly interested to evaluate the system performance in the detection and recognition of composite events: MBTSt - Male boards and transits to seat X; FBTSt - Female boards and transits to seat X; PEX - Person exits the bus and PCS - Person changes seat. Fig. 1 show the screen shot of the graphical interface of our event recognition system at an instance. The left side is the plan view of the bus area with trajectories. The bottom right shows a captured view. The top right shows the events occurring in the video, with the belief and plausibility of an event that was automatically recognised by the proposed system.



plausibility[FBTS3] = 0.66

PCS Passenger changes seat.

belief[PCS] = 0.98
plausibility[PCS] = 1.0

PEX Passenger exits bus.

belief[PEX] = 0.98
plausibility[PEX] = 1.0

Figure 1: Passenger's event reasoning demonstration

Three female and three male adults participated in the experiment. Video recordings include eighteen trails, each lasts around thirty second for a single passenger taking the bus. Three grouped scenarios were formed. In the first group, a passenger entered the bus, selected a seat and then exited the bus. In the second group, a passenger entered the bus, took a seat, changed to another seat, and then exited the bus. In the final group, a passenger entered the bus, walked towards the back row and turned around, sat down, and finally exited the bus. In total, 78% of 9 MBTS and 9 FBTS, and 100% of 18 PEX and 6 PCS, have been detected and recognised.

4 Conclusions

In this paper we have presented a framework of representing the structural knowledge of events and reasoning complex events based on the outputs of low-level video analytics. our approach has an ability to bridge the semantic gap between the low level video data and the high level human interpretation. With the support of DS theory, our approach is effective and can infer reliable events at a high level. The proposed approach takes into account the uncertainty in the stages of event representation, recognition processing and low-level video analytics. The proposed framework has provided reliable recognition results of complex scenarios using numerical belief measures. The experiments show that the proposed framework is able to recognise complex events not only when the tracking results were perfect but also when tracking process has deficiency.

ACKNOWLEDGEMENTS

This work is partially supported by CSIT project funded by UK EPSRC, Invest NI, and various industrial partners under grant EP/H049606/1.

REFERENCES

[1] A.P. Dempster, 'Upper and lower probabilities induced by a multivalued mapping', *The Annals of Statistics*, **28**, 325–339, (1967).

[2] G. Lavee, E. Rivlin, and M. Rudzsky, 'Understanding video events: A survey of methods for automatic interpretation of semantic occurrences in video', *IEEE Trans. SMC, Part C: Applications and Reviews*, **39**(5), 489–504, (2009).

[3] G. Shafer, *A Mathematical Theory of Evidence*, Princeton University Press, 1976.

[4] P. Smets, 'Constructing the pignistic probability function in a context of uncertainty', in *Proceedings of UAI*, pp. 29–40, (1990).