

Probabilistic Recognition of Complex Event

Rim Romdhane, Bernard Boulay, Francois Bremond, and Monique Thonnat

INRIA Sophia Antipolis,
France

{Rim.Romdhane, Bernard.Boulay, Francois.Bremond, Monique.Thonnat}@inria.fr,

Abstract. This paper describes a complex event recognition approach with probabilistic reasoning for handling uncertainty. The first advantage of the proposed approach is the flexibility of the modeling of composite events with complex temporal constraints. The second advantage is the use of probability theory providing a consistent framework for dealing with uncertain knowledge for the recognition of complex events. The experimental results show that our system can successfully improve the event recognition rate. We conclude by comparing our algorithm with the state of the art and showing how the definition of event models and the probabilistic reasoning can influence the results of the real-time event recognition.

Keywords: Complex event recognition, uncertainty, event description.

1 Introduction

In the literature, many video event recognition systems have been described. However, many challenging problems still remain to obtain a robust recognition because of noise, illumination changes, segmentation issues and occlusions. We propose a constraint-based approach for real-world video interpretation based on probabilistic reasoning for composite event recognition. The main goal is to improve the techniques of video data interpretation taking into account the imprecision and uncertainty of low level data. To reach this goal, we address uncertainty in event modeling and event recognition processes by a combination of logical and probabilistic methods. In summary, the contributions of this paper are: 1. A general framework for video complex event recognition based on a constraint-based approach for video event recognition and a probabilistic reasoning for handling uncertainty. We propose a dynamic linear model for attributes filtering. 2. New event modeling specification: we improve the event description language proposed by [1] and introduce a new probabilistic description based approach to gain in flexibility for event modeling by adding the notion of utility. Utility expresses the importance of sub-events to the recognition of the whole event. The paper is organized as follows: In section 2, we review the related work. In section 3 and 4 we describe the proposed video interpretation framework for complex event recognition. The experiments realized to evaluate the proposed method are shown in section 4. Finally, we present the conclusion in section 6.

2 Related work

Many approaches for event representation and recognition have been proposed during the last decade [2, 3]. These approaches can be classified into two main categories: probabilistic approaches and symbolic approaches.

The main probabilistic approaches that have been used to recognize video events include Bayesian classifiers [4] and Hidden Markov Models [5, 6]. Bayesian classifiers are well adapted to combine observations at one time point, but they have not a specific mechanism to represent the time and temporal constraints between visual observations. For instance, Dynamic Bayesian Networks (DBN) have been used successfully to recognize short temporal actions [7], but the recognition process depends on time segmentation: when the frame-rate or the activity duration changes, the DBN has to be re-trained. Many probabilistic event recognition approaches can handle uncertainty using a probabilistic framework. For instance, in [8] the authors introduce the switching Hidden Semi-Markov Model (S-HSMM) to deal with time duration modeling. This extension attempts to introduce more semantic in the formalism at the cost of tractability.

Symbolic approaches have been largely used to recognize activities. The main trend consists in designing symbolic networks whose nodes or predicates correspond to the boolean recognition of simpler events. Stochastic grammars have been proposed to parse simple actions recognized by vision modules [9]. Logic and Prolog programming have also been used to recognize activities defined as predicates [10]. Constraint Satisfaction Problem (CSP) has been applied to model activities as constraint networks [11]. The symbolic approaches have shown their efficiency in term of complex event recognition. However, these approaches do not handle the uncertainty of the recognition process leading to recognition errors in complex situations. Thus, in this paper, we propose a new constraint based approach for complex event recognition with probabilistic reasoning to improve the recognition performance.

3 Event Description Language

The proposed approach relies on a priori knowledge including the description of the expected objects in the scene, the observed scene, the sensors (*e.g.* fixed video cameras) and the definition of the event models. The expected objects are the physical objects moving in the scene (*e.g.* person, vehicle) which are organized hierarchically (*e.g.* a car is defined as a sub-type of vehicle). We call domain ontology the description of the expected objects and the set of event models which are predefined by human expert. An event model (fig. 1) is composed of five elements:

- Physical objects**: including mobile objects (*e.g.* person, vehicle), contextual objects (equipments, zones).
- Components**: the sub-events composing the event.
- Constraints**: conditions between the physical objects and/or the components including symbolic, logical, spatial and temporal constraints based on [13].

```

CompositeEvent (Up-Go,
  PhysicalObjects ((p: Person), (eq: equipment), (z1: Zone), (z2: Zone), (z3: Zone))
  Components ((c1: CompositeState Person-interacts-with-chair (p, eq, z1) [1]),
    (c2: PrimitiveState Person-walking (p, z2) [0.2])
    (c3: PrimitiveState Person-inside-Stop-zone (p, z3) [1])
    (c4: PrimitiveState Person-inside-zoneUsechair (p, z1) [1])
    (c5: PrimitiveEvent Change-posture-stand-to-sit (p) [1])
    (c6: PrimitiveEvent Change-posture-sit-to-stand (p) [1]))
  Constraints (c1 before c2 ; c2 before c3; c3 before c4; c4 before c5; c5 before c6)
  Alarm (Priority "Normal"))

CompositeEvent (Begin-Guided-test,
  PhysicalObjects ((p1: Person), (p2: Person), (z1: Zone), (z2: Zone), (z3: Zone))
  Components ((c1: PrimitiveState Person-at-Entrance (p1, z1) [1]),
    (c2: PrimitiveState Person-at-Entrance (p2, z1) [1])
    (c3: PrimitiveEvent Person-change-to-ZoneUsechair (p1, z2) [1])
    (c4: PrimitiveEvent Person-change-to-ZoneStop (p2, z3) [1]))
  Constraints (c1 meet c2; c3 meet c4)
  Alarm (Priority "Normal"))

```

Fig. 1. Two Event models: “Up-Go” illustrates a medical exercise for testing the ability of the patient to perform several activities. The model is composed of five steps: (1) the patient is standing at the chair of exercise for a predefined period of time, (2) he/she walks up to a stop zone marked by a red line, (3) returns close to the chair, (4) he/she sits at the chair and (5) gets up. “Begin-Guided-test” describes the beginning of the medical exercise: the nurse and the patient entering together in the room and then going to different places. An utility coefficient was associated to each sub-event.

-**Alarm**: describes the level of importance of an event.

-**Action**: describes a specific treatment to be executed when an instance of an event model is recognized: (e.g. launch a specific vision task such as the monitoring of PTZ cameras (zoom on to get better classification of the mobile object) or provide feedback to vision components to enhance the tracking task).

We propose a notion of utility in the definition of the event model by associating a coefficient to each sub-event. Utility which is defined by a human expert expresses the importance of sub-events for the recognition of the whole event. Its range is in the interval]0,1], higher is the utility value higher is the importance of the sub-event in the recognition of the whole event. The value 1 means that the sub-event is required for the recognition. At least one of the sub-events must have a high utility value otherwise the event model will not be considered during the event recognition process.

4 Event Recognition Process

The proposed event recognition algorithm uses as input the tracked mobile objects (extracted by vision algorithms, segmentation, detection, tracking), a priori knowledge of the scene and predefined event models.

The algorithm operates in 2 stages: (i) at each incoming frame, it computes all possible primitive states related to all mobile objects present in the scene, and (ii) it computes all possible events (*i.e.* primitive events, and then composite states and events).

An event model ω is composed of the set of physical objects $\xi(\omega)$, their associated attributes $A(\xi(\omega))$ and the set of sub-events $Se(\omega)$. The recognition of the event model ω consists of a loop to select a set of physical objects $\xi(\omega)$ then verify the corresponding temporal/spatial/logical constraints $\zeta(\omega)$ until all combinations have been tested. Once a set of physical objects satisfies all constraints we consider that the event is recognized and we generate an event instance p attached to the corresponding event model, the physical object and the recognition time t . The event instance is then stored in the list of recognized events. To prevent from event fragmentation, we consider that if at the previous instant, an event instance p' of the same type (same model, same physical objects) was recognized on a time interval $[t_0, t_1]$ with $|t_1 - t| < \delta$, the two event instances are merged into an instance that is recognized on the time interval $|t_0 - t|$.

During the event recognition process, the system estimates the confidence of primitive states and composite events. The confidence measures describe the quality of the analyzed data based on the temporal coherence of the attribute values.

4.1 Probabilistic Primitive State Recognition

The confidence of primitive state is estimated based on Bayes formula (Eq 1).

$$P(w|\zeta(\omega), Id(\xi(\omega))) = P(\zeta(\omega)|w) \times P(Id(\xi(\omega))|w) \times \frac{P(w)}{P(\zeta(\omega), Id(\xi(\omega)))} \quad (1)$$

We compute then the ratio: $\frac{P(w|\zeta(\omega), Id(\xi(\omega)))}{P(\neg w|\zeta(\neg\omega), Id(\xi(\neg\omega)))}$. with $\neg\omega$ is equal to $\omega =$ false. If the ratio value is upper than 1, the primitive state has a high chance to be recognized.

$P(Id(\xi(\omega))|w)$ is the identifier confidence which indicates how well the mobile object $\xi(\omega)$ has been correctly tracked. This probability is obtained by estimating the quality of the tracking process depending on several criteria: the displacement, the appearance and the attribute consistency over the tracking period as described in [18]. The constraint confidence $P(\zeta(\omega)|w)$ is computed depending on the constraint type. There are 2 types of constraint for primitive state: spatial (i.e. a person in a zone) and logical. The confidence of logical constraints (i.e. associate a symbol to a contextual object) is equal to 1 as we consider that the user has a negligible chance to associate a wrong symbol.

The confidence of spatial constraints is obtained by multiplying the confidence of object attributes $P(A(\xi(\omega))|w)$ involved in the constraint with the probability of the constraint to be verified (Eq. 2). For the spatial constraints such as 'inside-zone' or 'close to equipment', we compute the distance $dist$ of the person to the contextual objects (i.e. zone, equipment), more this distance

is small more the probability that this constraint is satisfied is close to 1. .

$$P(\zeta(\omega)|w) = P(A(\xi(\omega))|w) \cdot \frac{1}{\sigma_d \sqrt{2\pi}} \exp\left(-\frac{dist^2}{2\sigma_d^2}\right) \quad (2)$$

The confidence of mobile object attributes $P(A(\xi(\omega))|w)$ can be retrieved from vision algorithms (detection, tracking, posture recognition...). If this confidence is not directly provided, we compute this confidence using a dynamic linear filter such as Kalman filter algorithm.

- Dynamic model for temporal attributes filtering

We propose a dynamic linear model for computing and updating the attributes of the mobile objects to deal with tracking errors. This process works in two steps. The first step consists in computing the expected value α_{exp} of an attribute α at the current instant t_c given the estimated value of α and its velocity at the previous time t_p . The second step is to compute the estimated value α_{est} of the attribute based on the previous one. The final value of the attribute $\bar{\alpha}$ is the mean between the expected and the estimated values of the attribute.

$$\bar{\alpha}(t_c) = \text{mean}(\alpha_{exp}(t_c), \alpha_{est}(t_c)) \quad (3)$$

$$\alpha_{exp}(t_c) = \bar{\alpha}(t_p) + V_\alpha(t_c)(t_c - t_p); \quad (4)$$

$$V_\alpha(t_c) = \frac{V_{\alpha_c} \cdot R_v + e^{-\lambda(t_c - t_p)} \cdot V_\alpha(t_p) S_{V_\alpha}(t_p)}{S_{V_\alpha}(t_c)}; \quad (5)$$

$$S_{V_\alpha}(t_c) = R_v + e^{-\lambda(t_c - t_p)} \cdot S_{V_\alpha}(t_p) \quad (6)$$

V_{α_c} corresponds to the instantaneous velocity of the attribute α at time instants t_{c-1} and t_c , R_v is the instantaneous reliability of the velocity computed as the mean between the reliability of α at time instants t_{c-1} and t_c . $V_\alpha(t_p)$ is the estimated velocity at the previous time t_p . S_{V_α} is the temporal reliability of velocity. The value $e^{-\lambda(t_c - t_p)}$ corresponds to the cooling function of the previously observed attribute values. It can be interpreted as a forgetting factor for reinforcing the newer information.

$$\alpha_{est}(t_c) = \frac{\alpha_c \cdot R_{\alpha_c} + e^{-\lambda(t_c - t_p)} \cdot \alpha_{est}(t_p) \cdot S_\alpha(t_p)}{S_\alpha(t_c)} \quad (7)$$

$$S_\alpha(t_c) = R_{\alpha_c} + e^{-\lambda(t_c - t_p)} \cdot S_\alpha(t_p) \quad (8)$$

Where α_c is the value of the attribute given by vision algorithm and R_{α_c} is the reliability of this attribute α_c at time t_c . The reliability estimation of the attribute changes according to its type. For 2D attributes, the reliability is estimated inversely proportional to the distance to the camera accounting that the segmentation errors increase when the object is farther from the camera. For 3D attributes such as 3D position, we create a history H for the attribute values. Based on this temporal history we compute the confidence of the current attribute value using the Gaussian function. The Gaussian parameters (μ , σ) are computed dynamically using the temporal history.

4.2 Hierarchical Uncertainty Propagation

The recognition of a given complex event is triggered only if its last sub-event (called event terminaison) is recognized which avoids an exponential computation.

Thus the algorithm runs in real time since only the events which their terminaison is recognized are processed.

We compute the confidence of the complex event at time t given its probability at previous time $t-i$ and the probability of its sub-events $Se(w)$ (Eq 9). The probability of the event at previous time is weighted by the coefficient $\gamma \in [0, 1]$ which decrease when the last recognized instance of the event is far in time.

$$P^t(w) = P^t(Se(w), w^{t-i}) = P(Se(w)) \cdot \gamma P(w^{t-i}). \quad (9)$$

w^{t-i} is the last recognized instance of the event at the previous instant $t-i$.

To improve the temporal constraints verification process, we add the notion of tolerance when processing the temporal intervals comparison. For example to improve the verification of the temporal constraint 'A before B' we need to find a time t' such that event A has started and ended at time t' and an event B has started after A at time $t' + \beta$. β is the tolerance coefficient.

After calculating the probability associated to an event, the system can make a recognition decision by accepting events with a probability above a threshold and rejecting others. That is, only the events with high confidence probability are recognized.

5 Experimental Results

We show the effectiveness of using an ontology by applying our algorithm to three different applications: two health care and one airport activity monitoring applications (Fig. 2). Airport application consist in monitoring aircraft and vehicle behaviours whereas health care application consist in monitoring elderly persons observed in an experimental laboratory/hospital room during one hour. The video sequences are challenging in term of illumination changes and shadow. An ontology for airport activity monitoring was built. It is composed of 4 physical object type (person, aircraft, vehicle and zone) and 81 event models: 8 primitive states, 3 primitive events, 24 composite states and 45 composite events. We enhance this ontology for the health care applications by adding new physical objects such as equipment and by modifying some existing primitive events (*e.g.* adding posture attribute to the person). We have reused simple events defined for the airport activity monitoring application and define new event models adapted to the health care. We have tested the event recognition accuracy of our algorithm on health care applications and have compared our results with the approach proposed in [1].

The vision chain algorithms (segmentation, classification, detection and tracking) fails sometimes to provide correct outputs (misclassification, misdetection,...) due to changes of luminosity and noise from video acquisition.

Events	#videos	#actor	% R	FP	FN
Deterministic algo					
Up-Go	27	1	59.25	3	11
Begin-Guided-test	9	2	88.9	1	1
Interacts-with-chair	10	1	100	0	0
Stay-at-kitchen	15	1	86	1	2
prepare-meal	8	1	75	1	2
Probabilistic algo					
Up-Go	27	1	92.59	5	2
Begin-Guided-test	9	2	100	1	0
Interacts-with-chair	10	1	100	0	0
Stay-at-kitchen	15	1	93.3	1	1
prepare-meal	8	1	87.5	3	1

Table 1. Comparison of recognition rate (% R), the false positive (FP) and the false negative (FN) of our algorithm with probabilistic reasoning (probabilistic) and without probabilistic reasoning (deterministic).



Fig. 2. Two health care and one airport activities monitoring applications.

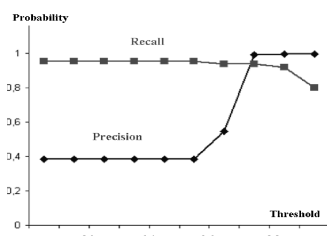


Fig. 3. The performance of primitive states detection was measured depending on the threshold defining the level of likelihood to decide that an event is recognized. With the threshold equal to 0.8, the performance of our system is 0.96 for precision and 0.93 for recall.

We tested the recognition performance of the primitive state of the proposed system by varying the decision threshold value. The precision and recall rates of the primitive states detection are shown in figure 3. The primitive states are sometimes wrongly recognized due to video noise and vision errors. However, by fixing for all experiments the threshold of detection of primitive states to 0.75

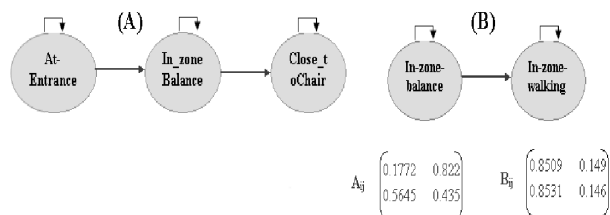


Fig. 4. HMM model for the event (A) ‘begin balance exercise’: the person enters the room, go to the zone of balance and get close to the chair to begin balance exercise. The event (B) ‘change-zone’ with the learned transition and observation matrices.

we manage to successfully decrease the false detection of primitive states. By avoiding miss detections of primitive states and using a flexible event description, the proposed system recognizes the complex events with a recognition rate about 92.59 % for the ‘Up-Go’ event and 100% for the event Begin-Guided-test (Tab. 1). The low rate of false alarms in the case of complex events can be explained by the fact that the event models are very constrained and they are unlikely to be recognized by error.

The comparison (Table 1) shows that the complex event ‘Up-Go’ in the case of the probabilistic algorithm (92.59 %) is higher than the recognition rate of the deterministic algorithm [1] (59.25%). This can be explained by the fact that the deterministic algorithm fails to recognize the primitive state Person-inside-Stop-zone because the person was not correctly detected. However, the probabilistic algorithm manages to recognize this primitive state and as a consequence the complex event.

Comparison with probabilistic method

For comparison with probabilistic method, Bayesian Network models were developed. In our case, the structure of the network is derived from the knowledge about the application domain. For example, logical constraints of sub-events that represent the recognition of a particular event indicate the direct causal link between them. The conditional probabilities were learned using the expectation maximization (EM) algorithm.

In addition, the proposed algorithm was compared with HMMs. We use a left-right HMM for representing the temporal constraints (Fig 4). We model different event such as change zone and change-posture. In the phase of training, we use the sequences of health care database manually classified as belonging in an event. For each event, a HMM is trained. For training, we use the expectation maximization algorithm to estimate the parameters of the HMM model. We use the Forwards-Backwards algorithm for the probability computation.

Table 2 shows the confusion matrices for the proposed algorithm (PA), BNs and HMMs experiments. The proposed algorithm outperform the HMMs and BNS for the event recognition rate. It can be explained by the lack of training

PA			BNs			HMMs		
C	T	L	C	T	L	C	T	L
1	0	0	.88	0	.12	.78	0	.22
0	.78	.22	0	.78	.22	0	.67	.33
.11	0	.89	.33	0	.67	.33	0	.67
average Pcc=89%			average Pcc=77%			average Pcc=70%		

Table 2. Confusion matrix for the proposed algorithm (PA), the BNs and HMMs. C: Person-sit-at-chair, T: Person-watch-TV, C: Person-interacts-with-Library.

data. To have a good recognition rate for the probabilistic approaches like HMMs and BNs we need to have a good parameter estimation. The learning stage need a large and pertinent amount of data.

6 Conclusion

We have proposed a flexible event modeling language and a novel event recognition algorithm to describe and recognize complex video events with probabilistic reasoning to handle the uncertainty. We have proposed a dynamic model for computing and updating the attributes of the mobile objects to deal with tracking errors. We have detailed the estimation of primitive state probability as a Bayesian process and we have computed the confidence of complex event as Markov process taking into account the probability of the event at previous time. A future work consists at deeply studying the uncertainty due to occlusions. Studying more techniques to handle the tracking errors and comparison with those different techniques is also planned. Moreover, a learning stage is still required to learn the algorithm parameters.

7 Acknowledgment

This work was supported partly by the PACA region, the Sweet-home, and CIU projects. However, the views and opinions expressed herein do not necessarily reflect those of the financing institutions.

References

1. Vu, Thinh. and Bremond, Francois. and Thonnat, Monique.: Automatic Video Interpretation: A Novel Algorithm for Temporal Scenario Recognition. The Eighteenth International Joint Conference on Artificial Intelligence, Mexico (2003)
2. Ryoo, M.S and Aggarwal. J.K.: Semantic Representation and Recognition of Continued and Recursive Human Activities. International Journal of Computer Vision, 2009.
3. Chen L., Nugent C.: Ontology-based recognition in intelligent pervasive environments. International journal of Web Information Systems, Vol.5, pp.410-430, 2009.

4. Oliver, N and Horvitz, E.:A comparison of HMMs and dynamic bayesian networks for recognizing office activities. International conference on user modeling, vol. 3538, pp. 199–209., 2005.
5. Hoey, J. and Bertoldi, P.P. and Mihailidis:Assisting persons with dementia during handwashing using a partially observable markov decision process. International Conference on Computer Vision Systems (ICVS), 2007.
6. Dan Kuettel and Breitenstein M. and Van Gool L. and Ferrari V.:Whats going on? Discovering Spatio-Temporal Dependencies in Dynamic Scenes.CVPR, 2010.
7. Gong and Xiang, T.: Recognition of group activities using dynamic probabilistic networks. The 9th International Conference on Computer Vision, 2003.
8. Duong, T.V. and Bui, H.H. and Phung, D.Q. and Venkatesh, S.:Activity recognition and abnormality detection with the switching hidden semi-markov model.CVPR, 2005.
9. Ivanov, Y. and Bobick,A. and Mihailidis:Recognition of visual activities interactions by stochastic parsing. IEEE Trans. Patt. Anal. Mach. Intel, vol. 1, pp. 838–845.1, 2005.
10. Davis,L. and Harwood,D. and Vidmap,D.:Video monitoring of activity with prolog. AVSS, 2005.
11. Reddy,S. and Gal,Y. and Shieber,S.:Recognition of Users Activities Using Constraint Satisfaction. Springer Berlin / Heidelberg, vol. 5535, pp. 415-421.1, 2009.
12. Nevatia, R. and Hongeng,S. and Bremond, F.:Video-based event recognition:activity representation and probabilistic recognition methods. CVIU, vol. 2, pp. 129–162., 2004.
13. Allen, J.F.:Maintaining knowledge about temporal intervals. In Communications of the ACM, 1983.
14. Getoor, L. and Taskar B:Introduction to Statistical Relational Learning. MIT Press, 2007.
15. Alberto Avanzi and Francois Bremond and Christophe Tornieri and Monique Thonnat.: Design and Assessment of an Intelligent Activity Monitoring Platform. EURASIP, 2005.
16. Liao, L. and Fox D. and Kautz, H:Location-based activity recognition using Relational Markov Networks. IJCAI, 2005.
17. Pentney, W. and Popescu A. and Wang S. and Kautz H. and Philipose M.:Sensor-based understanding of daily life via large-scale use of common sense. AAAI, 2006.
18. Chau D.P, Bremond F., Thonnat M.:Robust Mobile Object Tracking Based on Multiple Feature Similarity and Trajectory Filtering, VISSAP, 2010.