Handling Uncertainty for Video Event Recognition

Rim Romdhane, François Bremond, Monique Thonnat

Handling Uncertainty for Video Event Recognition

RIM ROMDHANE, FRANCOIS BREMOND, MONIQUE THONNAT

INRIA, Sophia Antipolis research unit, PULSAR project
rromdham@sophia.inria.fr, Francois.Bremond@sophia.inria.fr, Monique.Thonnat@sophia.inria.fr

Abstract

This paper presents a cognitive vision approach for video event recognition able of handling the uncertainty of the recognition process. The recognition task is complex because of image noise, of segmentation and classification issues. In this work, we extend the event recognition algorithm (crisp algorithm) proposed in [1] by proposing a geometric method which handles the uncertainty of the recognition process. This method consists in computing the precision of the 3D information of the mobile objects evolving in the scene for each frame of the video sequence. We use the computed information to calculate the probability of the event. The proposed method is tested with videos of everyday activities of elderly people. Events of interest have been modeled with the help of medical experts (i.e. gerontologists). The experimental results show that the proposed approach improves significantly the process of recognition and can characterize the likelihood of the recognized events.

1 Introduction

In the literature, many video event recognition systems have been described [2]. However, these systems are not robust enough for coping with computer vision challenges, such as illumination changes, segmentation issues and occlusions. Most of these systems do not address the issue of handling uncertainty in the event recognition process. The various sources generating uncertainty in vision systems are described in [3].

Current event recognition systems have usually a low performance. Most previous approaches, able to recognize events and handling uncertainty, are 2D approaches which model an activity as a set of pixel motion vectors [4][5]. These 2D approaches can only recognize short and primitive events but cannot address composite events.

We propose a video interpretation approach based on uncertainty handling. The main goal is to improve the techniques of automatic video data interpretation taking into account the imprecision of the recognition. To attain our goal, we extended the event recognition described in [1] (figure 1) by modelling the uncertainty and computing the precision of the 3D information characterising the mobile objects evolving in the scene. We use the 3D information to compute the spatial probability of the event. We also compute the temporal probability of the event based on its spatial probability at previous instant.

This approach is validated using a homecare application which tracks elderly people living at home and recognizes events of interest specified by gerontologists. The paper is organized as follow: in section 2, we review the related work. In section 3, we describe the different stages of the video interpretation framework that our work is based on. In the section 4, we propose the main contribution remaining on the uncertainty handling in the event recognition algorithm. The experiments realized to evaluate the proposed method are shown in the section 5. Finally, we present the conclusion in the section 6.

2 State of the art

The research field of event representation and recognition has been very active for the last decades. Event recognition approaches are classified into two main categories: probabilistic approaches and constraint-based approaches. This section describes several of these works and a short discussion on the remaining open issues.

The main probabilistic approaches that have been used to recognize video events include neural networks [6], [7], [8]. Bayesian classifier ([9]) and Hidden Markov Models (HMM) ([10], [11], [12])

The two first approaches (i.e. neural networks and 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 [8], but the recognition process depends on time segmentation: when the frame-rate or the activity duration changes, the DBN has to be re-trained. The approaches based on HMM are very popular and have been successfully used to recognize temporal sequences of simple events, but cannot model easily complex temporal relationships (e.g. Allen's interval algebra operators).

Many probabilistic event recognition approaches can handle the uncertainty using a probabilistic framework. For instance, Chomat and Crowley [5] address the problem of probabilistic recognition of activities (such as a person is walking) and hand gestures using local spatio-temporal appearance and the

Bayes rule. Also HMM approaches recognize temporal sequences of simple events by taking into account the uncertainty of the visual observations. For instance, a typical algorithm as the one presented by Hongeng et al. [12] uses HMM to recognize multi-state activities in dynamic scenes. Despite that the probabilistic event can model the uncertainty in the event recognition, they cannot recognize complex events.

Constraint-based approaches have been largely used to recognize activities for few decades. The main trend consists in designing symbolic networks whose nodes or predicates correspond to the boolean recognition of simpler events. Stochastic grammar has been proposed to parse simple actions recognized by vision modules [13]. Logic and Prolog programming have also been used to recognize activities defined as predicates [14]. Constraint Satisfaction Problem (CSP) has been applied to model activities as constraint networks [15], [16]. For example, Ghallab [17] represents an event as a set of temporal constraints on time-stamped events. The event recognition algorithm implements the propagation of temporal constraints based on the RETE algorithm. Vu et al. in [1] use similarly a declarative representation of event defined as a set of spatio-temporal and logic constraints. The constraint-based approaches have shown their efficiency in term of primitive and 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 geometric method which handles the uncertainty of the recognition process. This method consists in computing the precision of the 3D information of the mobile objects evolving in the scene for each frame of video sequence.

3 Overview of the proposed Video Interpretation framework

The proposed event recognition approach is based on a video interpretation framework described in (figure 1). This framework contains a vision component (e.g detection and tracking task) and an event recognition component. The video event framework takes as input video streams and a priori knowledge composed of 3D geometric information and pre-defined event models.

- **A priori knowledge**: The a priori knowledge is all the information used by the event recognition process to infer high semantical representation of the scene. This knowledge is composed of 3D geometric information (i.e. empty scene model, camera calibration) and pre-defined event models. The 3D geometric information includes in particular a decomposition of the 3D scene ground-floor into a set of zones of interest which are the main contextual elements for the event recognition process used for homecare applications (figure 2). The event models are pre-defined by human experts using a description language explained in the next section. These event models correspond to activities of interest ("eating", "preparing meal",…) which characterize people behavior living at home.

- Detection & tracking task: This task consists in detecting and tracking mobile objects in the scene. First, it consists in detecting for each frame the mobile objects in the scene and in classifying them with labels such as PERSON, corresponding to their type based on their 3D size and their shape. A mobile object is described by 3D numerical features (center of gravity, position, height, width, length) and by a semantic class (Person, Occluded Person, group pf persons, noise or unknown). The 3D point is computed by projecting the middle bottom point of the 2D bounding box into the 3D ground plane of scene.

The tracking task associates to each new mobile object an identifier and maintains it globally throughout the whole video.

- **Event Representation**: The goal of event representation is to formalize the a priori knowledge for the scene understanding process. This knowledge corresponds to a 3D empty scene model of the observed environment and a set of event models specified by human experts. The description of a priori knowledge has to be declarative and in natural terms, so that the experts of the application domain can easily define and modify it. There are four types of perceptual activities: primitive states, composite states, primitive events and composite events. A state describes a stable situation in time characterizing one or several physical objects. A primitive state (e.g. a person is located inside a zone) corresponds to a spatio-temporal property directly computed by the vision component. A composite state is the combination of primitive states. An event is an activity containing at least a change of state value between two consecutive times (e.g. a person enters a zone of interest (kitchen): he/she is outside the zone and then inside). A primitive event corresponds to a change of primitive state value and a composite event is a combination of primitive events. An event (and more generally any activity) is composed on five elements:

-**Physical objects**: including mobile objects (e.g. individuals), equipment or zones of interest.

Components: corresponding to the sub-events composing the event.

-**Forbidden components**: corresponding to the events which should not occur during the main event.

-**Constraints**: conditions between the physical objects and/or the components (constraints can be temporal, spatial or logical).

We represent an event model with the list of the physical objects involved in the event and a set of constraints on these physical objects. The recognition process consists in verifying the set of constraints to recognize in real time event occurrences. We illustrate below an example of primitive state inside_zone (person P is inside a zone Z) and primitive event enters_zone (corresponds to the change of person position and moving from outside the zone to the inside the zone).when the constraint 'p in z' is verified the primitive state inside zone is recognized.

PrimitiveState (Inside_zone,
      PhysicalObjects ((p : Person), (z : Zone))
      Constraints((p in z))
PrimitiveEvent(enters_zone,
      PhysicalObjects((p : Person), (z : Zone))

Components((s_outside : PrimitiveState outside_zone(p, z))
      (s_inside : PrimitiveState inside_zone(p, z)))

Constraints(((s_inside 's Duration) <= 1)
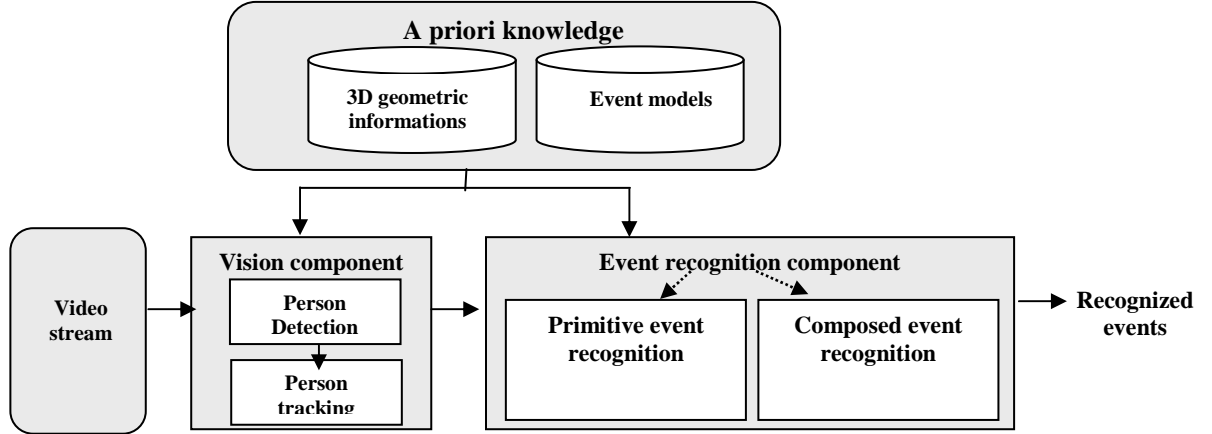   *(s_outside   before_meet   s_inside))*

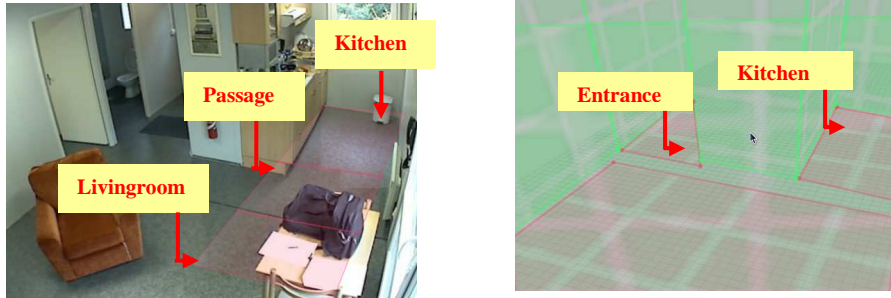Figure 1:Architecture of the proposed Video Interpretation framework.

Figure 2: Defining zones of interest in the 3D scene.

4 Event recognition based uncertainty handling

The video event recognition algorithm recognizes which activities are occurring in a stream of tracked mobile objects. The algorithm performs a loop of: (1) selection of a set of physical objects then (2) verification of the corresponding atemporal constraints until all combination of physical objects have been checked. Once a set of physical objects satisfies all the constraints, the event is said to be recognized. Due to the vision algorithms errors (segmentation, tracking, …) the performance of the recognition can be poor and the recognition of event can be missed. As the video event recognition is highly affected by noise and errors, we propose to improve the performance of the recognition taking into account its imprecision and uncertainty. We propose to measure the imprecision of the computation of the 3D distance between the 3D position of the tracked mobile objects and different static object of the scene: the distance of

the tracked object to an equipment (table, chair, …), the distance between the different tracked object (i.e. the different persons) and the distance between the tracked object and a

zone. For the case of equipment, we consider the plan projection of the equipment to define an equipment zone. We compute the distance of the tracked object to this zone. Based on this distance, we are able to recognize the primitive state "close_to" and "far_from" equipment. In the following paragraph, we detail the recognition process of the primitive state "inside_zone" we compute the distance of the tracked object to the zones of interest. The computation of this 3D distance is the main element to recognize primitive states such as "inside_zone". The mobile object in the scene (i.e. person) is represented by the position of its feet corresponding to the middle of the bounding box bottom segment. For distance calculation, we compute the distance of the person to the different borders of the zone and the

distance that we take into account is the minimum of all these distances (the closest one). This distance is zero when the person is inside the zone and the distance is high if the person is far from the zone. Based on this distance, we compute the Gaussian probability that a primitive state occurs within the zone (1.a).

$$P(person \in zone / d_{p \to z}) = f(x); \qquad (1.a)$$

With $d_{p \to z}$ is the distance of the tracked object to a zone. The Gaussian probability technique is defined by:

- If the tracked person is near the zone ($d_{P \to Z} \to 0$), then the probability $f(x) \to 1$.

- If yhe tracked object is far from the zone ($d_{P \to Z} \to \infty$), the probability $f(x) \to 0$.

- Otherwise, the probability is in [0, 1].

Based on this definition the probability follows a normal law as it has a density of probability f:

$$f(x) = \frac{1}{\sigma \sqrt{2 \Pi}} \exp(\frac{-1}{2} (\frac{x - \mu}{\sigma}))^2 \qquad (1.b)$$

In our method, μ equal to 0 and we compute experimentally the value of σ.

We compute also the temporal probability of primitive state occurrence based on the current spatial probability and the previous ones. The temporal probability is the weighted sum of the n previous spatial probabilities of the event

$$P_{temporal}(e_{f_c}) = \frac{\sum_{f < f_c} w_{f_i} P_{spatial}(e_i)}{n_{f_i}} \qquad (2)$$

With e_{f_c}, the specific event to be recognized at frame f_c, w_i the weights and $P_{spatial}(e_{f_i})$ the spatial probabilities of the same event computed at previous frames f_i. n_{f_i}, the number of considered previous frames.

We propose also to handle the probability of occurrence of the events. As already mentioned in section 3, an event is a combination of primitive states or other events. The recognition of an event is based on the recognition of its components. Thus the estimation of the probability of any event is based on the probability of its sub-components. We propose to compute its probability as the weighted sum of the probability of its sub-components c_i (3).

$$P(e) = \frac{\sum_i w_i P(c_i)}{n_{f_i}} \qquad (3)$$

With, w_i the weight; $P(c_i)$ the probability of occurrence of the sub-component c_i of the event e and n_{fi} is the number of the event components.

2 Experiments

This section describes and discusses the experimental results. First, we describe the experimental site used to validate our approach. Then we show and discuss the results of the event recognition process. To validate the proposed approach, we have performed a set of experiments in a homecare laboratory [18]. For these experiments, we have acquired and processed a video sequence with one actor with 8 frames per second. The video contains about 500 frames.

- Experimental Site

Developing and testing the impact of the activity monitoring solutions requires a realistic near real-life environment in which training and evaluation can be performed. We have performed a set of experiments in the homecare laboratory which is a realistic site reproducing the environment of a typical apartment: 41m2 with entrance, livingroom, bedroom, bathroom, and kitchen. The kitchen includes an electric stove, microwave oven, fridge, cupboards, and drawers. 4 video cameras are installed in the laboratory. One video camera is installed in the kitchen, two video cameras are installed in the livingroom and the last one is installed in the bedroom.

- Experimental Results

In this section, we present the different results obtained. We compare the different results to the ground truth data (GT). The ground truth data is defined at the event recognition level. It contains only the event occurring in the scene. For performance evaluation, we use classical metrics. When the system correctly recognizes an event a true positive (TP) is scored. A false positive (FP) is scored when an incorrect event is recognized. If an event occurs and the system does not report it, a false negative (FN) is scored. We use two standards metric: the precision (4) and the sensitivity (5). The precision is the ratio between the numbers of true positive (correct recognition) and the sum of the numbers of true and false positive. The sensitivity is the ratio between the number of true positive and the sum of the numbers of true positive and false negative. The event is recognized if the computed probability < threshold, this threshold is chosen manually. The value of the threshold enables to control the precision and the sensitivity. For experimentation, we have choose threshold = 0.4. we have set up experimentally the value of σ and for this experiments the value of the weight is set to 1. The results are described in the table 1.

$$\Pr ecision \ = \frac{TP}{TP + FP} \qquad (4)$$

$$Sensitivit\ y = \frac{TP}{TP + FN} \qquad (5)$$

Event	GT	TP	FP	FN	Precision	Sensitivity
"inside_Entrance"	137	126	5	11	0.96	0.91
inside_Kitchen	67	54	1	13	0.98	0.80

Table 1: Recognition results of Crisp algorithm.

Event	GT	TP	FP	FN	Precision	Sensitivity
inside_Entrance	137	130	12	7	0.91	0.95
inside_Kitchen	67	58	5	9	0.92	0.86

Table 2: Recognition results of the proposed algorithm.

The results of table (2) compared to the results of table (1) show that for the proposed algorithm, we obtain a slightly better sensitivity respectively 0.95 instead of 0.91 for the state inside_entrance and 0.86 instead of 0.80 for the state inside_kithcen compared to crisp algorithm.

We can control the recognition performance (sensitivity with respect to precision) by more or less strict value for the threshold. In figure 3, because of segmentation and tracking errors and noise, the crisp algorithm did not manage to recognize any event. All these vision problems affect deeply the event recognition process. Handling uncertainty, the proposed algorithm manages to detect the event with a probability of recognition Pr greater than the recognition threshold.

(a). Inside_kitchen (Pr=0.54)   **(b).** Inside_kitchen (Pr=0.54)   **(c).** Inside_kitchen (Pr=0.37)

Figure3: Illustration of the recognition performance of Crisp with the proposed algorithm (Pr: probability of recognition).

6 Conclusion

In this paper, we have presented an event recognition method for handling uncertainty. The state-of-the-art constraint-based approaches have shown their efficiency in term of primitive and complex event recognition. However, they lack of mechanisms for handling uncertainty can lead to recognition errors of event. Thus, we propose to handle the uncertainty of the recognition process. Our proposed method consists in computing the precision of the 3D information of the mobile objects observed in the scene for each frame of the video sequence. This approach is tested using a homecare application which tracks elderly people living at home and recognizes events of interest specified by gerontologists. The experimental results show that in many cases where the crisp event recognition process scores a false negative, the proposed method manages the event recognition.

The results are encouraging but still more work is needed to ameliorate the uncertainty estimation for the event recognition process (other type of constraints and states) and more experiments on longer videos and various threshold values.

Future work includes evaluating other methods for computing the 3D information precision and for handling the uncertainty of the recognition process.

A next task consists in studying the impact of the tracking process quality on the estimation of the event recognition uncertainty by taking into account the coherency of the mobile object trajectory. The event recognition uncertainty process can be improved by first estimating the calibration, detection and tracking uncertainty.

These results are preliminary and further experiments with longer videos and various scenarios are planned to test and validate the estimation of event recognition uncertainty.

6 References

[1] Vu, T., Brémond, F. , Thonnat, M. : Automatic Video Interpretation: A Novel Algorithm for Temporal Scenario Recognition. The Eighteenth International Joint Conference on Artificial Intelligence (IJCAI'03), Acapulco, Mexico (2003).

[2] T. B. Moeslund and E. Granum. A survey of computer vision-based human motion capture. Computer Vision and Image Understanding, 81:231--238 (2001).

[3] Cherfaoui V., Burie J., Royere C. and Gruyer D.: Dealing with uncertainty in perception system for the caracterisation of driving situation' 2000 IEEE intelligent Transportation systems. Conference Proceedings Dearborn(MI) USA (2000).

[4] Rapantzikos, K., Avrithis, Y. , Kollias, S.: Handling Uncertainty in Video Analysis with Spatiotemporal Visual Attention. Proc. of IEEE International Conference on Fuzzy Systems (FUZZ-IEEE '05), Reno, Nevada (2005).

[5] Chomat, O. ,Crowley, J.l.: Probabilistic Recognition of Activity using Local Appearance. cvpr, vol. 2, pp.2104, IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'99) - Volume 2, (1999).

[6] A.J. Howell and H. Buxton. Active vision techniques for visually mediated interaction. Image and Vision Computing (2002).

[7] M. Barnard, J. M. Odobez, and S. Bengio. Multi-modal audio-visual event recognition for football analysis. In IEEE Workshop on Neural Networks for Signal Processing (NNSP), France, (2003).

[8] S. Gong and T. Xiang. Recognition of group activities using dynamic probabilistic networks. In The 9th International Conference on Computer Vision, Nice, France (2003).

[9] Oliver, N., Horvitz, E.: A comparison of HMMs and dynamic bayesian networks for recognizing office activities International conference on user modeling No10, vol. 3538, pp. 199--209. Edinburgh , (2005).

[10] J. Hoey, A. V. Bertoldi, P.P., Mihailidis, A.: Assisting persons with dementia during handwashing using a partially observable markov decision process. In: International Conference on Computer Vision Systems (ICVS). Germany (2007).

[11] R. Nevatia, S. Hongeng, and F. Bremond. Video-based event recognition: activity representation and probabilistic recognition methods. CVIU, 96(2):129--162 (2004).

[12] S. Hongeng, F. Bremond, and R. Nevatia. Representation and optimal recognition of human activities. In IEEE Proceedings of Computer Vision and Pattern Recognition, South Carolina, USA (2000).

[13] Y. Ivanov and A. Bobick. Recognition of visual activities and interactions by stochastic parsing. IEEE Trans. Patt. Anal. Mach. Intell., 22(8):852--872 (2000).

[14] L. S Davis, D. Harwood, and V. D. Shet. Vidmap: Video monitoring of activity with prolog. In Advanced Video and Signal-Based Surveillance (AVSS), Como, Italy (2005).

[15] Zouba, N., Bremond, F., Thonnat, N.: Monitoring activities of daily living of elderly based on 3D key human postures. In 4th international cognitive vision workshop, ICVW 2008. Greece (2008).

[16] Vu. T., F. Brémond, G. Davini, M. Thonnat, Q.C. Pham, N. Allezard, P. Sayd, J.L. Rouas, S. Ambellouis and A. Flancquart, Audio Video Event Recognition System for Public Transport Security. The IET conference on Imaging for Crime Detection and Prevention (ICDP 2006),pp.414-419, London, Great Britain (2006).

[17] Ghallab M. On Chronicles: Representation, Online Recognition and Learning. 5th International Conference on Principles of Knowledge Representation and Reasoning (KR'96), Cambridge Temporal (USA), 5-8, pp.597-606. (1996).

[18] removed for anonymity reason.