A Framework dealing with Uncertainty for Complex Event Recognition

Rim Romdhane, Francois Bremond, Monique Thonnat
INRIA Sophia Antipolis, PULSAR team
2004, route du Lucioles,
BP93 06902 Sophia Antipolis Cedex France

Rim.Romdhane@sophia.inria.fr, Francois.Bremond@sophia.inria.fr, Monique.Thonnat@sophia.inria.fr

Abstract

This paper presents a constraint-based approach for video event recognition with probabilistic reasoning for handling uncertainty. The main advantage of constraint-based approaches is the possibility for human expert to model composite events with complex temporal constraints. But the approaches are usually deterministic and do not enable the convenient mechanism of probability reasoning to handle the uncertainty. The first advantage of the proposed approach is the ability to model and recognize composite events with complex temporal constraints. The second advantage is that probability theory provides a consistent framework for dealing with uncertain knowledge for a robust and reliable recognition of complex event. This approach is evaluated with 4 real healthcare videos and a public video ETISEO'06. The results are compared with state of the art method. The comparison shows that the proposed approach improves significantly the process of recognition and characterizes the likelihood of the recognized events.

1. Introduction

In the literature, many video event recognition systems have been described [11, 16, 17]. However, these systems are not robust enough for coping with computer vision challenges, such as illumination changes, segmentation issues and occlusions [3]. Most of these systems do not handle the uncertainty in the event recognition process. Most of the previous approaches able to recognize events and handling uncertainty, are 2D approaches which model an activity as a set of pixel motion vectors [4, 14]. These 2D approaches can only recognize short and primitive events but cannot address composite events.

We propose a constraint-based approach for real-world video interpretation based on probabilistic reasoning for composite event likelihood computation. The main goal is to improve the techniques of video data interpretation taking into account the imprecision and uncertainty of low level data. To attain our goal, we extend the event recognition approach described in [17] for a robust recognition and we compute the likelihood of the event recognition. Likelihood can be defined as the probability or degree of trust that an event occurs. This approach is tested on healthcare videos and ETISEO'06.

The paper is organized as follow: in section 2, we review the related work. In section 3 and 4 we describe the different stages and the main contribution on the proposed video interpretation framework for composite events recognition. 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 during the last decade. Event recognition approaches can be classified into two main categories: probabilistic approaches and constraint-based approaches. This section describes several of these approaches and a short discussion on the remaining open issues. The main probabilistic approaches that have been used to recognize video events include neural networks [1, 2], Bayesian classifier [13] and Hidden Markov Models (HMM) [6, 9, 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 main advantage of Bayesian networks is that they are able to model the uncertainty of the recognition by using probabilities based on Bayes Theory. However, they have two main drawbacks. First, the a priori

probability needs to be learned and this learning stage is often tiresome: due to the construction of the learning set. Second, they have been often used to recognize elementary actions at the numerical level with only one physical object. The advantage of the HMMs compared to NNs and Bayesian classifier is the ability to recognize sequences of events, but they cannot model easily complex temporal relationships (e.g. Allens interval algebra operators) and they are limited when the recognition involves several mobile objects. The probability of being in a state for a mobile object has to be combined with the probability of being in another state for all other mobile objects. This combination leads to combinatorial explosion of the recognition process.

Many probabilistic event recognition approaches can handle uncertainty using a probabilistic framework. For instance, Chomat and Crowley [4] 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 rules. In [6] the authors introduce the switching Hidden Semi-Markov Model (S-HSMM) to deal with time duration modelling based on the use of discrete coxian distribution. This extension attempts to introduce more semantic in the formalism at the cost of tractability. Constraint-based 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 [10]. Logic and Prolog programming have also been used to recognize activities defined as predicates [5]. Constraint Satisfaction Problem (CSP) has been applied to model activities as constraint networks [15, 16, 18]. For example, Ghallab [7] 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. The approach proposed by Vu et al [17] uses similarly a declarative representation of event defined as a set of spatio-temporal and logic constraints. This approach is very performant in term of composite event recognition with complex temporal constraint and in term of temporal execution time but it made two strong assumptions: 1) good detector (all persons are correctly detected) 2) good tracker (all persons are correctly tracked). The constraint-based 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 reliable and robust complex event recognition with probabilistic reasoning.
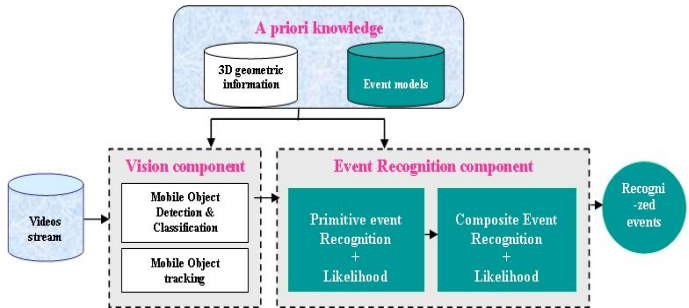

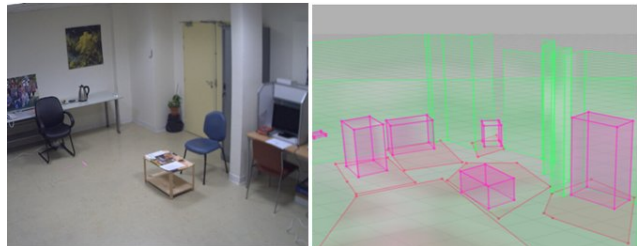Figure 1. Overview of the proposed Event Recognition Framework.

Figure 2. Contextual elements of the scene.

3. Overview of the proposed Video Interpretation Framework

The proposed event recognition approach is based on a video interpretation framework described in figure 1. The video event framework takes as input video streams and a priori knowledge. The a priori knowledge is all the information used by the event recognition process to infer high semantically 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 framework contains a vision component (e.g. detection, classification and tracking tasks) and an event recognition component. In this section, we detail the a priori Knowledge of the framework. The event recognition component is detailed in the section 4.

3.1. 3D Geometric Information

The 3D geometric information includes in particular a decomposition of the 3D scene ground-floor into a set of zones of interest corresponding to the equipment 3D projection on the ground floor which are the main contextual elements for the event recognition process (Figure 2). The green colour represents the walls, the pink coulor is used for the equipment and the red is for zones of interest.

```
PrimitiveState (Inside-zone,
  PhysicalObjects ((p : Person), (z : Zone))
  Constraints ((p in z))
  Action (Priority "Normal"))


PrimitiveEvent (moves-close-to,
  PhysicalObjects ((p: Person), (eq: Equipment))
  Components ((s-far: PrimitiveState far-from (p, eq))
              (s-close: PrimitiveState close-to (p, eq)))
  Constraints ((s-close's Duration<= Threshold)
               (s-far before-meet s-close))
  Action (Priority "Normal"))


CompositeEvent (person-interacts-with-TV,
    PhysicalObjects ((p: Person), (eq: equipment), (z: Zone))
    Components (c1: PrimitiveEvent stay-at-TV (p, eq),
               (c2: PrimitiveState inside-zone-use-TV (p, z)),
    Constraints ((c1 meet c2)
                 (c1's Duration >=threshold_{useTV})
                 (c2's Duration <= threshold_{useTV})
                 (z's Name=zone-use-of-TV)
                 (eq's Name=TV)
    Action (Priority "Normal"))
```

Figure 3. Events Models

3.2. Event Models

The event models correspond to the modelling of all the knowledge used by the system to detect event occurring in the scene. The description of this knowledge has to be declarative and intuitive (in natural terms). In this work, we propose to represent the activities of interest into a formal model that satisfies a number of spatial and temporal constraints by using the event description language proposed by Vu et al. [17]. We have used this language to address complex activity recognition involving several physical objects of different types (e.g. person, equipment) in a scene observed by video cameras over an extended period of time.

There are four types of 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 using the vision component results. A composite state is a 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: 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 of five elements:
- Physical objects: including mobile objects (e.g. persons, 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 including symbolic, logical, spatial and temporal constraints.
- Action: describes the actions to be taken when the event is recognized.
Composite events are also called in video understanding community complex events, behaviours and scenarios. We illustrate in figure 3 as examples the primitive state inside-zone (person P is inside a zone Z), the primitive event moves-close-to (corresponding to the change of person position and moving close to an equipment) and a composite event person interacts with an equipment (TV). This model consists of a person being close to the equipment, inside the defined zone of use of the equipment and during a long enough time (duration of use of the equipment). This composite event contains three physical objects, two components, and three constraints.

4. Event Recognition Process with uncertainty handling

We propose two extensions of the event recognition algorithm [17]. The first extension consists in the dealing of the crisp aspect in the process of spatial primitive state detection for a robust recognition. The first algorithm (algorithm 1) deals with mis-detection and allows a large recognition of events. The second extension consists in the computation of the likelihood of the event recognition based on the reliability computation of its components (algorithm 2). The reliability measures describe the visual quality of the analysed data and the temporal coherence of the obtained mobile object attribute values. In the next section we detail the two proposed algorithms.

4.1. Robust Recognition

The algorithm [17] is based on a crisp method for the recognition of states and events. The first extension consists in a non binary interpretation of the spatial primitive states based on [0, 1]-valued Gaussian functions.
The proposed algorithm 1 (figure 4) takes as input the mobile object (i.e. person) coordinates at each time instant t and the 3D coordinates of contextual objects (i.e. 3D projection coordinates of zones and equipment on the ground). The algorithm computes the distance of the person (represented by the position of its feet corresponding to the middle of the bounding box bottom segment) to the different borders of the contextual zone and the distance that it takes 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
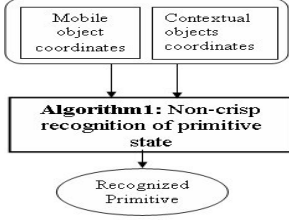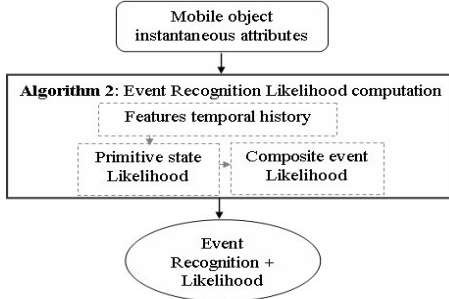
Figure 4. Inputs and outputs of the algorithm 1.

Figure 5. Event recognition likelihood computation algorithm.

the zone. Based on this distance, it computes the Gaussian probability P(S) that a primitive state occurs within the zone as shown in (Eq.1). The event is recognized when the computed probability is over a predefined threshold. The variance σ is learned experimentally.

$$P(S) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{d^2}{2\sigma^2}} \qquad (1)$$

4.2. Likelihood of the Event Recognition process

The second extension (algorithm 2) consists in the computation of the likelihood of the recognition of an event based on the reliability computation of its components (see section 3.2). The algorithm 2 takes as input the mobile object (i.e. person) instantaneous features values, creates a temporal history of these features values, estimates how much the mobile object features vector deviates from an estimated distribution model and then computes the likelihood of a recognized event (figure 5). We believe that the temporal and spatial coherency of the mobile object features is the key to evaluate the event recognition process. In the following, we detail the process of the computation of the likelihood for the primitive state and composite events.

4.2.1 Likelihood of Primitive State Recognition

A primtive state (see section 3.2) is modelled based on the physical objects involved in the state (mobile objects (i.e. person), contextual objects (i.e. zones, equipment,...)) and spatio-temporal constraints.
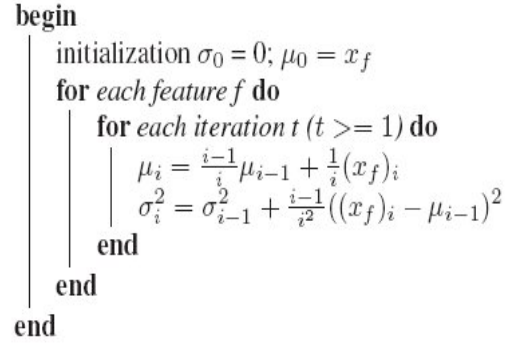

Figure 6. algorithm for parameters learning.

The primitive state likelihood $P_t(S)$ is computed based on the combination of the mobile object (Mo) reliability.

The computation of mobile object reliability aims at evaluating how well the classification, detection and tracking of mobile objects work.
The reliability computation method consists mainly in estimating how much the mobile object features vector deviate from an estimated distribution model.
For that, a vector of seven features (motion step, velocity, direction, 3D height, 3D width and temporal trajectory coherency) $F = \{f_1, \cdots, f_7\}$ (figure 7) that best characterize the mobile object is extracted. These features are detailed in the next section.

For each feature value x_f we compute respectively the instantaneous reliability P_{inst}^f at time instant T as a Gaussian probability density function with parameters (μ_f, σ_f)(Eq 2). The Gaussian parameters (μ_f, σ_f) are learned based on the algorithm detailed below (6). This algorithm allows to learn iteratively and fastly the parameters (μ_f, σ_f) values for a large learning database. As the value of i increases, the ratio $\frac{i-1}{i^2}$ becomes very low and has less impact on the calculated value of σ_i. For the 3D height (H), 3D width (W) and shape ratio (Sr) features (see next section), we learn experimentally a Gaussian model for a person with ($\mu_H = 160$, $\sigma_H = 56$) ($\mu_W = 50$,$\sigma_W = 35$) and ($\mu_{Sr} = \mu_W$ / μ_H , $\sigma_{Sr} = \sigma_W$ / σ_H).

We believe that to have a global idea about the feature reliability value at time instant t, it is important to consider not only the instantaneous value but also the previous ones. Thus we compute the temporal reliability (Eq 3) in a temporal window based on previous reliability values $P_{temp}^f(t <= T)$. The value $e^{-(T-t)}$ corresponds to the cooling function of the previous computed reliability values. It can be interpreted as cooling factor for reinforcing the newer values and giving less importance to the previous ones.
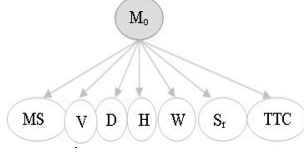
Figure 7. Mobile object features.

$$P_{inst}^f(T) = e^{-\frac{(x_f - \mu_f)^2}{2\sigma_f^2}} \quad (2)$$

$$P_{temp}^f(T) = \frac{P_{inst}^f(T) + \sum_{t<T} e^{-(T-t)} P_{temp}^f(t)}{\sum_{t<T} e^{-(T-t)}} \quad (3)$$

The mobile object reliability is then computed based on its features reliability according to the total probability (see Eq 4)

$$
\begin{aligned}
P_t(S) = P_t(Mo/F) &= \sum_{f \in F} P(Mo/f).P(f/F) \\
&= \sum_{f \in F} P_{temp}(f).P(f/F)
\end{aligned}
\quad (4)
$$

(1) Motion Step (MS)

The motion step is defined as the 3D coordinate difference at two consecutive time instants t and t-1. If the mobile object is well tracked, this motion step is rather small. As many event models are based on the mobile object displacement, this feature is a key to deal with the recognition of the events.

(2) Velocity (V)

When a mobile object is wrongly identified and tracked, the velocity value $V = \sqrt{V_x^2 + V_y^2}$ could increase abnormally. Thus, this feature has an important impact on the quality of the recognition therefore we propose to compute the reliability of this feature.
We compute, first, the value of mobile object velocity based on its n (n= 10) previous positions over time and then we calculate the velocity difference of the tracked object at instant t and t-1.

(3) Direction (D)
In general the direction of a mobile object changes smoothly, thus an important change of the mobile object direction is usually the consequence of tracking errors. For this reason we consider that direction is also an important

feature to evaluate the mobile object reliability. We calculate the angle value of the movement direction based on 2D coordinates of the mobile object at instant t and t-1. The value of this angle is in the interval $[0, 2\pi]$.

(4) 3D Height (H) and **(5) 3D Width (W) features**

The value of the 3D height and 3D width is very important for the classification of a mobile objects as a person. If a mobile object is wrongly classified, it could be not detected and tracked and no event could be recognized.

(6) Shape Ratio(Sr)
The shape ratio of a mobile object at time instant t is computed as follow

$$SR_t = W_t/H_t \quad (5)$$

Errors in vision algorithms can cause a large change in the value of the shape ratio.

(7) Temporal Trajectory Coherency (TTC)

The mobile object trajectory is defined as the localization of the mobile object during an interval of time. Characterize the trajectory coherency of a mobile object over time is an important feature for the evaluation of the likelihood of the event recognition process. In order to estimate correctly this feature, we need to observe the mobile movement in a long enough temporal interval δ_t. We consider that ten frames are long enough for this estimation.
To evaluate the trajectory coherency, we use linear regression formalism. Given a sample of the coordinates (xi, yi), i=1,...n of a mobile object, we aim at determining if this sample fits a linear model: $Yi = aXi + b$;
Based on the least squares method, our goal is to find the equation of the affine function that minimizes the sum of the squared deviations of the points to this affine function. First, we compute the empirical average of the xi ($\bar{x}$) and yi ($\bar{y}$) (Eq 6), the empirical variance of the xi (S_X^2), the empirical covariance (S_{XY}) (Eq 7) and the affine function parameter (a, b) are then calculated. The vertical distance (ε_i) between the affine function and a trajectory point (xi, yi) is called the error (Eq 8).

The reliability of the trajectory coherency is calculated based on the error ε_i (Eq 9). s_{min} and s_{max} are predefined thresholds.

$$\bar{x} = \frac{1}{n}.\sum_{i=1}^n (x_i); \quad \bar{y} = \frac{1}{n}.\sum_{i=1}^n (y_i) \quad (6)$$

$$S_X^2 = \frac{1}{n}.\sum_{i=1}^n (x_i - \bar{x})^2; \quad S_{XY} = \frac{1}{n}.\sum_{i=1}^n ((x_i - \bar{x})(y_i - \bar{y})) \quad (7)$$

$$a = \frac{S_{XY}}{S_X^2}; \quad b = \bar{y} - a.\bar{x}; \quad \varepsilon_i = y_i - ax_i - b \quad (8)$$

$$\text{TTC}Likelihood = \begin{cases} 1 & \text{if } |\varepsilon_i| <= s_{min} \\ \frac{(\varepsilon_i + s_{max})}{(s_{max} - s_{min}} & \text{otherwise} \end{cases} \quad (9)$$

4.2.2 Likelihood of Composite Event Recognition

A composite event (see section 3.2) is modelled based on the description of the involved physical objects (mobile objects (i.e. person), contextual objects (i.e. zones, equipment,...)) and a combination of states and events with spatio/temporal constraints.

The first idea to handle the uncertainty is to assign at the event modelling step a coefficient of importance level or priority α_i for each component c_i of the composite event model (see figure 9). This coefficient value which is defined by the expert expresses which components are more important for the recognition of the event and allow a more flexible event modelling. At the event recognition step, the recognition of composite events is based on the recognition of its sub-events (i.e. primitive states and primitive events that compose it). The composite event likelihood C_E is computed with Bayes theorem (Eq 10). $P(C_E)$ is the prior probability of C_E, it is the probability that C_E is correct before the sub-event $C = \{c_1, \cdots, c_n\}$ were detected. $P_t(C|C_E)$ is the conditional probability of the sub-event given that the hypothesis C_E is recognized. $P_t(C|C_E)$ is called the likelihood and $P(C)$ is the marginal probability. These probabilities are learned based on statistical method from a set of representative training sequences. Finally we compute the ratio $P_t(C_E|C)/P_t(\neg C_E|C)$ (see Eq 11) with $\neg C_E$ is equal to C_E = false. If the ratio value is upper than 1, the recognition of the composite event has a high chance to be correct. The figure 8 shows a Bayesian representation of the composite event 'interaction-With-chair' composed of two sub-events 'inside-zone-Balance' and 'close-to-chair' and the corresponding probability table.

$$P_t(C_E|C) = \frac{P_t(C|C_E).P(C_E)}{P(C)} \\ = \frac{\prod_{i=1}^{n} P(c_i|C_E).P(C_E)}{P(C)} \quad (10)$$

$$\frac{P_t(C_E|C)}{P_t(\neg C_E|C)} = \frac{P_t(C|C_E).P(C_E)}{P_t(C|\neg C_E).P(\neg C_E)} \\ = \frac{\prod_{i=1}^{n} P(c_i|C_E).P(C_E)}{\prod_{i=1}^{n} P(c_i|\neg C_E).P(\neg C_E)} \quad (11)$$

	Inside-z-Balance		Close-to-chair	
	true	false	true	false
C_E	0.96	0.03	0.99	0.001
$\neg C_E$	0.59	0.37	0.42	0.57

Figure 8. Bayes network and its probability table.

Figure 9. A priority coefficient is assigned to each of the event components.

5. Experimental results

The proposed method was tested in two different applications. First, 4 real world videos of healthcare application with 4 actors were tested. Because of change of luminosity, cameras vibration and noise of video acquisition, these videos have bad quality so that the vision chain algorithms (segmentation, classification, detection and tracking) fails sometimes to provide correct outputs (misclassification, misdetection). We compare the different results with the ground truth data (GT). The ground truth data is defined at the event 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 (Eq 12): the precision (P) and the sensitivity (S) . 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.

$$P = \frac{TP}{TP + FP}; \quad S = \frac{TP}{TP + FN} \quad (12)$$

Algo	Crisp Algo/UH Algo					
Events	**GT**	**TP**	**FP**	**FN**	**S (%)**	**P (%)**
Primitive Event and primitive state						
Person-Walking	5	2/ 5	3/ 10	2/ 0	50/ 100	40/ 33
Inside-zone-Balance	35	14 / 21	3/8	22/ 14	39/ 60	82/72
Close-to-chair	33	16/ 21	1/ 1	17/ 12	48/ 63	94/ 95
Composite event						
interaction-with-chair	21	5/16	0/ 4	16/5	24/ 76	100/ 80
Stay-at-Equipment	8	2/ 5	0/ 0	5/ 3	40/ 62	100/ 100
Move-close-chair	2	1/ 1	0/ 0	1/ 1	50/ 50	100/ 100
interaction-with-TV	10	4/ 8	0/ 0	6/ 2	40/80	100/ 100
Person-reading	3	1/3	0/1	2/0	100/100	50/ 75

Table 1. Recognition results of the Crisp (in black)/ UH algorithms (in bleu).

The method was tested also on a public video sequence ETISEO'06. The proposed method is valid for any type of video. Infact, we test some scenarios with important spatial localization changes on ETISEO'06. The scenario position-changes (figure 10) consists of a person going out of a room, moving to the corridor, stopping near a pillar (figure 11) was recognized correctly with the proposed method. For the crisp algorithm [17], because of the mis-recognition of the primitive state close-Room-Door, the whole scenario was not recognized. In many case when the crisp algorithm fails to recognize an event because of a misdetection, the proposed algorithm with uncertainty handling (UH algorithm) recognizes it correctly. The sensitivity of the proposed algorithm is higher than the crisp one. The precision of the proposed algorithm is high in the case of complex events (Person-reading, interaction-with-chair,...). False positive detection occurs mainly at the borders of zones and equipments as the contextuals objects (i.e. equipments) are very close otherwise less false detection can be scored. The non-detection of false alarms in the case of complex event can be explained by the fact that the scenarios are very constrained and there are unlikely to be recognized by error.

6. Conclusions

The lack of mechanisms for handling uncertainty and the imprecision of low level data can lead to recognition errors or mis-recognition of events for real-world videos. Thus, the interest in the approach of event recognition with uncertainty handling is growing rapidly. We propose in this paper an approach to handle the uncertainty of the event recognition. This approach takes into account the event modelling and the event recognition process. The proposed approach

```
CompositeEvent (position-changes,
    PhysicalObjects((p : Person),(z1 : Zone),(z2 : Zone),(w1 : Wall),(w2 : Wall))
    Components (c1: PrimitiveState inside-Exit-Room (p, z1)),
              (c2: PrimitiveState close-Room-Door (p, w1)),
              (c3: PrimitiveState inside-zone-corridor (p, z2)),
              (c4: PrimitiveState close-to-pillar (p, w2)))
    Constraints ((c1 before-meet c3)
              (c3 before-meet c4)
              (c3's Duration >=threshold)
              (c4's Duration <=threshold)
              (z1's Name=roomExit)
              (z2's Name=corridor)
              (w1's Name=RoomDoor)
              (w2's Name=pillar))
    Action (Priority "Normal"))
```

Figure 10. The composite event "position-changes". This model consists of a person going out of a room, moving to the corridor, stopping near a pillar.

Figure 11. Etiseo'06: Recognition of the composite event "position-changes".

is tested using a health care application and ETISEO'06 database. The experimental results show that in many cases

where the crisp event recognition obtain false negative, the proposed method recognize correctly the events. In the proposed method, the recognition of false alarm in the case of primitive events (Close-to-chair, Inside-zone-Balance, ...) occur mainly on the border of zones and equipments especially when they are very close.

The low number of false alarms in the case of composite event can be explained by the fact that the scenarios are very constrained and there is a little chance to be wrongly detected.

A future work consists in studying how to establish cooperation between the event recognition and the vision modules. It will be useful for the event recognition module to send a feedback to vision modules to track again people. A next task consists in learning automatically event models based on spatio-temporal physical objects relationships and modelling their uncertainty. We plan also to study different types of interaction of the mobile objects with the contextual objects based on facial, gaze, gesture and posture detections. More experimental evaluation and comparison with other state of the art event recognition methods are planned to assess the robustness of the proposed system.

7. Acknowledgments

References

[1] M. Barnard, , J. Odobez, , and M. Bengio. Multi modal audio-visual event recognition for football analysis. *IEEE Workshop on Neural Networks for Signal Processing (NNSP)*, 2003. 1

[2] X. Chen, , and C. Zhang. An interactive semantic video mining and retrieval platform-application in transportation surveillance video for incident detection. *Sixth IEEE International Conference on Data Mining (ICDM'06)*, 2006. 1

[3] V. Cherfaoui, , J. Burie, , C. Royere, , and D. Gruyer. Dealing with uncertainty in perception system for the caracterisation of driving situation 2000 ieee intelligent transportation systems. *2000 IEEE intelligent Transportation systems. Conference Proceedings Dearborn(MI)*, 2000. 1

[4] O. Chomat, , and J. Crowley. Probabilistic recognition of activity using local appearance. *IEEEComputer Society Conference on Computer Vision and Pattern Recognition (CVPR'99)*, 2:2104, 1999. 1, 2

[5] L. Davis, , D. Harwood, , and D. Vidmap. Video monitoring of activity with prolog. *Advanced Video and Signal-Based Surveillance (AVSS)*, 2005. 2

[6] T. Duong, , H. Bui, , D. Phung, , and S. Venkatesh. Activity recognition and abnormality detection with the switching hidden semi-markov model. *IEEE computer society Conference on Computer Vision and Patern recognition CVPR'05*, 2005. 1, 2

[7] M. Ghallab. On chronicles: Representation, online recognition and learning. *5th International Conference on Principles of Knowledge Representation and Reasoning (KR'96)*, 5(8):597–606, 1996. 2

[8] Gong, , and T. Xiang. Recognition of group activities using dynamic probabilistic networks. *The 9th International Conference on Computer Vision*, 2003. 1

[9] J. Hoey, , P. Bertoldi, , and Mihailidis. Assisting persons with dementia during handwashing using a partially observable markov decision process. *nternational Conference on Computer Vision Systems (ICVS)*, 2007. 1

[10] Y. Ivanov, , and A. Bobick. Recognition of visual activities and interactions by stochastic parsing. *IEEE Trans. Patt. Anal. Mach. Intell.*, 1:838–845, 2005. 2

[11] Y. Ke, , R. Sukthankar, , and M. Hebert. Understanding video event. *IEEE Transactions on Systems Man and Cybernetics*, 2007. 1

[12] R. Nevatia, , S. Hongeng, , and F. Bremond. Video-based event recognition: activity representation and probabilistic recognition methods. *CVIU*, 2(96):129–162, 2004. 1

[13] N. Oliver, , and E. Horvitz. A comparison of hmms and dynamic bayesian networks for recognizing office activities. *International conference on user modeling*, 3538(10):199–209, 2005. 1

[14] K. Rapantzikos, , Y. Avrithis, , and S. Kollias. Handling uncertainty in video analysis with spatiotemporal visual attention. *Proc. of IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 05)*, 2005. 1

[15] S. Reddy, , Y. Gal, , and S. Shieber. Recognition of users activities using constraint satisfaction. *Springer Berlin / Heidelberg*, 5535:415–421, 2009. 2

[16] T. Vu, , F. Bremond, , M. G. Davini, , M. Thonnat, , Q. Pham, , N. Allezard, , P. Sayd, , J. Rouas, , S. Ambellouis, , and F. A. Audio video event recognition system for public transport security. *The IET conference on Imaging for Crime Detection and Prevention (ICDP 2006)*, pages 414–419, 2006. 1, 2

[17] T. Vu, , F. Bremond, , and M. Thonnat. Automatic video interpretation: A novel algorithm for temporal scenario recognition. *The Eighteenth International Joint Conference on Artificial Intelligence (IJCAI'03)*, 2003. 1, 2, 3, 7

[18] N. Zouba, , F. Bremond, , M. Thonnat, , A. Anfonso, , E. Pascual, , P. Mallea, , V. Mailland, , and O. Guerin. A computer system to monitor older adults at home: preliminary results. *In the international journal Gerontechnology, SF-TAG: Gerontechnology-French Issue*, 8(3), 2009. 2