Event Recognition System for Older People Monitoring Using an RGB-D Camera

Baptiste Fosty *, Carlos Fernando Crispim-Junior †, Julien Badie ‡, François Bremond § and Monique Thonnat ¶

INRIA Sophia Antipolis
2004, Route des Lucioles
06560 Sophia Antipolis, FRANCE

* baptiste.fosty@inria.fr
† carlos-fernando.crispim_junior@inria.fr
‡ julien.badie@inria.fr
§ francois.bremond@inria.fr
¶ monique.thonnat@inria.fr

***Abstract*— In many domains such as health monitoring, the semantic information provided by automatic monitoring systems has become essential. These systems should be as robust, as easy to deploy and as affordable as possible. This paper presents a monitoring system for mid to long-term event recognition based on RGB-D (Red Green Blue + Depth) standard algorithms and on additional algorithms in order to address a real world application. Using a hierarchical model-based approach, the robustness of this system is evaluated on the recognition of physical tasks (*e.g.*, balance test) undertaken by older people ($N = 30$) during a clinical protocol devoted to dementia study. The performance of the system is demonstrated at recognizing, first, human postures, and second, complex events based on posture and 3D contextual information of the scene.**

Keywords : health monitoring system, RGB-D camera, complex event recognition, performance evaluation

I. INTRODUCTION

Event recognition remains an open topic of computer vision whereas the needs in terms of data processing and analysis increase with the growing number of cameras. Event recognition approaches can be categorized according to the input features they use and the reasoning methods they apply for [1]. The two main approaches for input features to construct events are pixel-based (low-level) and object-based (high-level). Pixel based approaches are using for instance colors or textures like in [2] while the other category builds an abstraction of the low-level data as objects including inherent properties (*e.g.*, speed, trajectory) [3].

The reasoning methods applied on event recognition can be classified into three main categories : classification methods (*e.g.*, SVM) [4], probabilistic graphical models (*e.g.*, HMM) [5] and semantic models (*e.g.*, description based models) [6]. *Sadanand et al.* [7] have proposed a classification method for activity recognition, where each action (*e.g.*, boxing, diving) is represented by a set of examples on different scales, viewpoint and time-resolution. A set of detectors is used for each action detection, and the output of all action detectors is then combined using a Support-Vector Machine approach. This method outperforms most of state of the art methods on benchmarking datasets. However, classification methods and probabilistic graphical models are generally based on low-level data (*e.g.*, pixel-based, feature-based) and on a training procedure involving a large dataset to be able to generalize among the activities performed on different scenarios. It is difficult to foresee the behavior/performance of this algorithm when applied for a different environment from the one of the training.

The use of semantic models is an alternative approach as it does not require learning but a set of event models provided by domain expert, for instance based on logics or grammars rules. An example of such approach has been described by [8]. They evaluate the detection of complex events by constructing a tree composed of the related sub events. The major limitation relies on how to assign, for example, a sub event to one of two complex events when it cannot be part of both, in the presence of high level noise.

Most of the work presented previously uses RGB cameras. Nevertheless, the use of RGB-D cameras is growing in the domain of event recognition as recently they have become more affordable, they can provide real 3D information of the scene and ease the deployment of the system to new environment. *Banerjee et al.* [9] have developed a system for falling down detection in hospital rooms using RGB-D camera and a fuzzy inference system. The system infers facts using approximative descriptions of the world.

Pramerdorfer [10] has evaluated RGB-D camera (Kinect, Microsoft) concerning its suitability and robustness for people and fall detection systems, with respect to particular conditions like distance from camera, illuminance or clothing materials and color. For instance, it has been shown that clothing colors can be an issue for people detection and

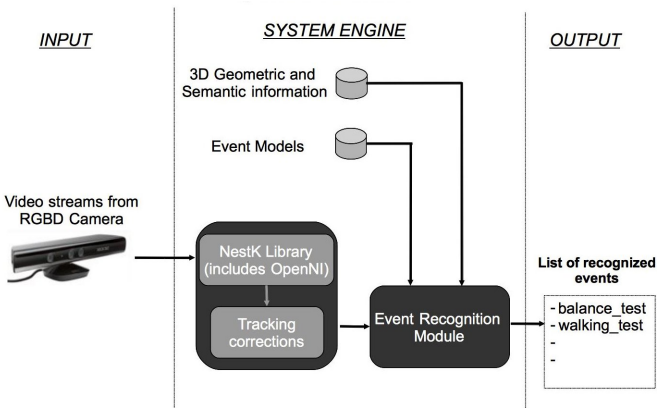

Fig. 1. System architecture

Fig. 2. Black clothes consequences example. This figure shows the consequences of the absorption of the infrared rays by the black clothes on the quality of the person's detection.

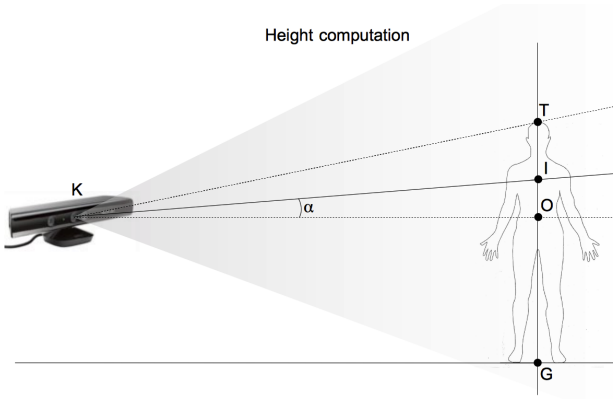

Fig. 3. Height computation. The system is able to recompute the height of a person with the distance between the ground and the RGB-D camera GO, the angle of the camera α and the coordinates of the top point of the person T. I is the intersection between the optical axis of the camera and the person.

tracking using depth data and need to be considered.

This paper's contributions are twofold :

- a set of people detection and tracking techniques for improving the robustness of the system :
  - an alternative method to recompute the height of a person in case of a partial loss of body area detection caused by the issues described by *Pramerdorfer* and also observed in our activity dataset,
  - a reidentification algorithm.
- an evaluation of the robustness of the system for mid to long-term event recognition using hierarchical model-based approach combined with a RGB-D camera,

The proposed approach is presented in section II. The sections III and IV are dedicated respectively to the evaluation and the results of the system, with and without improvements and compared to a 2D camera system. Finally, we conclude and examine the possible extensions of our work in sections V and VI.

II. PROPOSED APPROACH

The proposed system is presented in two main subsections. The first subsection describes the different issues encountered with people detection and tracking and the proposed solutions. The event detection module is presented in the second subsection, with the detailed description of the information needed by the system. For more information on the system architecture, see Fig. 1.

A. Vision Component

The first layer of the vision component performs people detection and tracking based on the open source framework OpenNI, through NestK library. The second layer is composed of the proposed solution to cope with poor estimation of person height due to clothes color, noisy object and people identity maintenance.

Height Computation. In some cases, infrared rays are absorbed when the tracked person is wearing black clothes. The consequence is that some parts of the body (generally lower body parts) are excluded from the 3D bounding box (see Fig. 2), which leads to a smaller height for the tracked person. This is an important issue since the posture is inferred from the height of the person. An alternative method is proposed to compute the height of a person based on the top point of the person (highest point of the person's point cloud), the angle of the camera from the horizontal position and the distance between the base of the device and the ground (see Fig. 3). So, the system computes the new height H as follows :

$$H = GT \tag{1}$$

$$H = GO + OI + IT \tag{2}$$

where,

- GO : the a priori known distance between the sensor and the ground,
- OI : line segment computed from the angle α and the distance between the sensor and the person,

- IT : the distance from the intersection between the optical axis of the camera and the person to the top point of the person (vertical coordinate of this point),

The computed height is used to detect whether the person is sitting or standing based on a thresholding method. We take into account the average height for sitting and the person is considered standing when his/her height value is above the sitting average height.

Object Detection Filtering. Objects, furnitures and walls are sometimes detected as a person by the OpenNI tracking algorithm and therefore generate unreliable events. To avoid it, we propose to erase noisy objects which are not inside the expected range size for a human being.

Reidentification. In some cases, the tracking algorithm may have difficulties to keep track of a person. For example, it happens when a person leaves the scene and re-enters or when an element of the scene occludes the person for a significant time. In this case, the tracking algorithm considers the single person as two different persons and labels him/her with a different IDs. This leads to event misdetections since the event history of the person is lost.
In order to tackle this issue, we use a method called re-acquisition [11] that tries to connect the current IDs with the previously lost IDs in a predefined time window, based on the appearance of each person. This step can be considered as an extension of the tracking algorithm at a larger scale. The first step of the re-acquisition method is to extract relevant data from each tracklet to compute a visual signature. To compute this visual signature, a descriptor based on covariance matrices is used [12]. This descriptor has shown very good results in the case of multi-camera re-identification of people. The second step is to arrange the tracklet into several clusters depending on the distance between their visual signature. The tracklets belonging to the same cluster are then considered as representing the same person because their visual signature is nearly identical. The main advantage of the re-acquisition method is to reduce the tracking errors due to occlusions by merging the IDs of tracklets representing the same person.

B. Event Recognition Framework.

The description of event models is defined using a declarative language [13] [6]. This language is affordable by expert since it uses a proper structure and explicit key words. Event Models are composed of six components :

- *Physical Objects* : objects involved in the recognition of the event modeled (*e.g.*, person or spatial zone),
- *Components* : sub-events that the model is composed of,
- *Forbidden Components* : events that should not occur in case of the event model is recognized,
- *Constraints* : conditions that the physical objects and/or the components should hold,
- *Alert* : importance level of the scenario model in terms of priority,
- *Action* : in association with the Alert type, specific action which would be performed when an event of the model is recognized (*e.g.*, send a SMS to a carer).

The physical objects refers to objects detected dynamically at previous steps of the computer vision chain (people detection and tracking algorithms) or related to a priori knowledge of the scene (*e.g.*, spatial zones of interest which contains semantic information in regard to the activity to detect). For instance, a person refers to a mobile object in the scene containing the following attributes : x-y-z 3D coordinates, width, height, and depth. Constraints define conditions that physical object property(ies) should meet, or components should hold. They could be atemporal, such as spatial and appearance constraints, or temporal, such as, Person_in_zone1 before Person_in_zone2. Temporal constraints are defined using Allens interval algebra (*e.g.*, BEFORE, MEET, and AND) [14]. This implies to know a priori to the processing the event we want to recognize and the contextual information on the scene.

Event models are hierarchically categorized according to their complexity (ascending order) :

- *Primitive State* : instantaneous value of a property of a physical object (*e.g.*, Sitting or Inside_zone_couch),
- *Composite State* : composition of two or more primitive states,
- *Primitive Event* : change in a value of a physical object property (*e.g.*, Person changes from Sitting to Standing posture),
- *Composite Event* : composition of two previous models.

Here is an example of the definition of a complex event model with its sub events :

```
CompositeEvent (Sitting_in_couch,
  PhysicalObjects (
    (p1 : Person),
    (z1 : Zone)
  )
  Components (
    (c1 : Person_sitting (p1))
    (c2 : Person_in_zone_couch (p1, z1))
  )
  Constraints ((c1 and c2))
  Alarm (URGENT)
)

PrimitiveState (Person_in_zone_couch,
  PhysicalObjects (
    (p1 : Person),
    (z1 : Zone)
  )
  Constraints (
    (p1 -> Position in z1 -> Vertices)
    (z1 -> name = zone_couch)
  )
  Alarm (NOTURGENT))
```

To summarize, the extraction of complex events from video sequences is performed by a combination of the RGB-D data stream, the corresponding tracking information (delivered mainly by the libraries NestK and OpenNI), the contextual objects (zones or equipments) and the event models.

III. EVALUATION

The evaluation of the event description based approach using RGB-D camera is divided into four main parts :

- a posture recognition evaluation to assess the improvement brought by the proposed techniques on the height computation,
- a performance comparison between the proposed event recognition system and the system using only NestK people tracking functionalities (without the proposed improvements),
- a performance comparison with a system following the same description based approach but using a RGB camera (AXIS, Model P1346). The reference system uses image segmentation algorithm proposed by [15] and the people algorithm tracking proposed by [16].
- a complementary evaluation of the assessed event duration compared to the real event duration provided by the ground truth.

A. Dataset

The proposed system has been evaluated at monitoring the physical tasks of participants of a medical protocol for Alzheimer disease (AD) study. Participants aged more than 65 years were recruited by the Memory Center of the Nice Hospital. Inclusion criteria of the AD group were: diagnosis of AD according to NINCDS-ADRDA criteria and a Mini-Mental State Exam (MMSE) score above 15. AD participants with significant motor disturbances as per the Unified Parkinsons Disease Rating Scale were excluded. Participants are asked to perform a set of physical tasks and daily living activities as a basis to a clinical evaluation of their executive functions. The protocol is divided into three scenarios : directed activities, semi-directed activities and undirected activities. In this paper, we focus on the analysis of the directed activities.

Scenario 01 (S1) or Directed activities is intended to assess kinematic parameters about the participant's gait profile (*e.g.*, static and dynamic balance test, walking test). During this scenario an assessor stays with the participant inside the room (see figure 4) and asks him/her to perform mainly four physical activities within 10 minutes (divided in sub activities).

The RGB-D camera recordings are acquired at a framerate of 10 frames per seconds with an angle of view of 57 degrees horizontally and 43 degrees vertically ($\pm$ 27 because of the motorized tilt). These activities are briefly described as follows :

- *Balance test* : the participant should keep balance while performing exercises (*e.g.*, standing with feet side by side or standing on one or the other foot)

Fig. 4. Room where the patients evaluations take place.

TABLE I
POSTURE RECOGNITION PERFORMANCE

	Recall (%)	Precision (%)	F-Score (%)
Sitting	100	75.5	86.0
Standing	100	89.2	94.3
Total	**100**	**82.6**	**90.5**

- *Walking test* : the assessor asks the participant to walk through the room, following a straight path from one side of the room to another (go attempt, four meters), and then to return (return attempt, four meters);
- *Repeated transfer test* : The assessor asks the participant to make the first posture transfer (from sitting to standing posture) without using help of his/her arms. The examiner will then ask the participant to repeat the same action five times in a row;
- *Up & go test* : participants start from the sitting position, and at the assessor's signal he/she needs to stand up, to walk a three meters path, to make a U-turn in the center of the room, return and sit down again.

B. Performance Evaluation

The system is evaluated compared to event annotation provided by domain expert. The following indices are computed :

$$precision = \frac{TP}{TP + FP}, \quad recall = \frac{TP}{TP + FN}$$

$$F - Score = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

where : TP : true positive events, FP : false positive events, FN : false negative events. The evaluation is performed by taking into account the number of detected events.

IV. RESULTS AND DISCUSSION

The following results refers to 30 videos with average time length of 6.9 minutes. Table I presents the event recognition system performance at posture recognition. A recall of 100% is achieved (out of 190 events, 107 for standing posture and 83 for sitting posture).

TABLE II

EVENT RECOGNITION PERFORMANCE WITH THE PROPOSED VISION COMPONENT IMPROVEMENTS

	Only NestK		NestK + Improvements	
Event category	Recall (%)	Precision (%)	Recall (%)	Precision (%)
Balance test	90.0	96.4	100	100
Walking test (go attempt)	93.3	93.3	100	90.9
Walking test (return attempt)	**73.3**	95.7	**90**	100
Repeated transfer test	96.7	**60.4**	100	**90.9**
Up & go test	80.0	85.7	93.3	90.3
Total	**86.7**	**82.8**	**96.6**	**94.2**
Global F-Score	**84.7**		**95.4**	

Total number of events to be detected : 150 (1 event of each category per video)

TABLE III

COMPARISON BETWEEN THE EVENT RECOGNITION PERFORMANCES OF THE SYSTEM USING RGB AND RGB-D CAMERAS

Camera	RGB camera		RGB-D camera	
Event category	Recall (%)	Precision (%)	Recall (%)	Precision (%)
Balance test	95.8	95.8	100	100
Walking test (go attempt)	91.7	100	100	90.9
Walking test (return attempt)	87.5	95.5	90	100
Repeated transfer test	75.0	100	100	90.9
Up & go test	91.7	100	93.3	90.3
Total	**88.3**	**98.3**	**96.6**	**94.2**
Global F-Score	**93.0**		**95.4**	

Total number of events to be detected : 150 (1 event of each category per video)

Table II and III evaluate the system for the same videos but with respect to complex event recognition (5 complex events per video, 150 in total). Table II shows the differences obtained for complex event recognition with and without the proposed improvements done in the vision component (see section II-A). Results obtained directly with NestK people detection output are presented on the left, while the results obtained from the proposed system are on the right (with improvements).

The observed gain of performance of the proposed approach is approximately of 10% for precision, recall and F-Score. On improved version, a recall of approximately 97% is obtained on the overall activities (true positive rate) while the precision is close to 94%. This fact means that the system recognizes most of the activities from the video sequence (around 3% missed) with an acceptable amount of false positive events. For the repeated transfer test, we highlight that the improvement of the height computation of the person has improved the precision of the detection of this event, directly related to posture (from 60.4% to 90.9%). Concerning the return attempt of the Walking test, its detection is mainly improved by the use of the reidentification algorithm and the improvement of the recognition of the go attempt.

Table III compares the results obtained with a RGB-D camera (on the right) and with a RGB camera (on the left). RGB-D camera results refer to the improved version on Table II. The proposed approach has a higher recall (less false negatives) than RGB camera one, but a lower precision (more false positives). This means that the system using real 3D information obtains a lower rate of misdetected events. In total, for the system using the RGB-D camera, this trade-off between the numbers of false positive and false negative detected events leads to an improvement of the recognition of around 2.5% (F-Score).

TABLE IV

EVALUATION OF THE ASSESSED DURATION OF EVENT

	Recall (%)	Precision (%)
Balance test	99.9	70.8
Walking test (go attempt)	55.2	94.1
Walking test (return attempt)	60.1	62.6
Repeated transfer test	86.6	94.6
Up & go test	79.8	86.2
Total	**94,5**	**73,6**
Global F-Score	**82.8**	

The previous tables have presented the evaluation of the system in terms of event frequency. Table IV shows an evaluation of the proposed approach in terms of assessed duration of a given event compared to the real event duration annotated as the ground truth.

The recall value of the assessed duration is close to the one obtained in terms of event frequency, matching the real duration of event. On the other hand, the precision is lower. This lower value is due to the fact that RGB-D camera (Kinect) field of view do not cover the whole scene where the physical tests have been undertaken. Therefore, in the current system, the time spent by the person performing a physical test outside the field of view (Walking tests and Up & Go test) is not taken into account.

V. CONCLUSION

This paper has presented an event recognition system using RGB-D camera based on hierarchical descriptive models. While RGB cameras has to be calibrated to obtain a 3D estimation of the scene, the use of RGB-D camera provide real 3D information which tends to be more reliable. Besides the affordability of the nowadays RGB-D cameras and the robustness of their 3D information, the

use of a description based language allows us to easily adapt the event models to new environments. Moreover, the proposed improvements for the vision components such as the reidentification algorithm enable the proposed approach to achieve the desired robustness. Finally, while most of the computer vision algorithms are developed to work on short video clips and for short terms activities, we are more focused on detecting mid to long term activities (*e.g.*, walking test) for long term monitoring (*e.g.*, weeks or months) in order to track people habits directly at home and therefore detect any behavioral change that can lead to an increasing frailty level.

VI. FUTURE WORK

In a very near future, this system will be installed in volunteers home and in a hospital for the evaluation of older people performance to help with the early design of Alzheimer disease. We also plan to evaluate the fusion of multiple RGB-D cameras to cope with the restricted field of view of a single RGB-D camera when compared to an ambient camera.

REFERENCES

[1] Lavee, G., Rivlin, E., Rudzsky, M. : Understanding Video Events: A Survey of Methods for Automatic Interpretation of Semantic Occurrences in Video. IEEE Transactions on Systems, Man, and Cybernetics. **39** (5). (2009) 489–504

[2] Kellokumpu, V., Zhao, G., PietikŁinen, M. : Human Activity Recognition Using a Dynamic Texture Based Method. The British Machine Vision Conference (BMVC 2008), Leeds, UK. (2008)

[3] Chen, L., Nugent, C. D., Wang, H. : A Knowledge-Driven Approach to Activity Recognition in Smart Homes. IEEE Transactions on Knowledge and Data Engineering. **24** (6). (2012) 961–974

[4] Xu, D., Chang, S.-F. : Visual Event Recognition in News Video using Kernel Methods with Multi-Level Temporal Alignment. IEEE Transactions on Pattern Analysis and Machine Intelligence. **30** (11). (2008) 1985–1997

[5] Ogale, A. S., Karapurkar, A., Guerra-Filho, G., Aloimonos, Y. : View Invariant Identification of Pose Sequences for Action Recognition. Presented ate the Video Analysis Contect Extraction Workshop (VACE). (2004)

[6] Zaidenberg, S. and Boulay, B. and Bremond, F. : A generic framework for video understanding applied to group behavior recognition. The 9th IEEE International Conference On Advanced Video and Signal Based Surveillance (AVSS 12). (2012)

[7] Sadanand, S., Corso, J. J. : Action Bank : A High-Level Representation of Activity in Video. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2012) 1234–1241

[8] Summers-Stay, D., Teo, C. L., Yang, Y., Fermller, C., Aloimonos, Y. : Using a Minimal Action Grammar for Activity Understanding in the Real World. IEEE Conference on Intelligent Robots and Systems. 4104–4111 (2012)

[9] Banerjee, T. and Rantz, M. and Li, M. and Popescu, M. and Stone, E. and Skubic, M. and Scott, S. : Monitoring Hospital Rooms for Safety Using Depth Images. Gerontechnology. AI for Gerontechnology. (2012)

[10] Pramerdorfer, C. : Evaluation of Kinect Sensors for Fall Detection. IASTED International Conference. Signal Processing, Pattern Recognition and Applications (SPPRA 2013).

[11] Bak, S., Corvee, E., Bremond, F., Thonnat, M. : Multiple shot Human Re-Identification by Mean Riemannian Covariance Grid. AVSS, Klagenfurt, Austria. (2011)

[12] Badie, J., Bak, S., Serban, S.-T., Bremond, F. : Recovering people tracking errors using enhanced covariance-based signatures. PETS 2012 workshop, associated with AVSS 2012 conference. (2012)

[13] Vu, T., Bremond, 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) 9-15

[14] Allen J. F. : Maintaining knowledge about temporal intervals. Communications of the ACM. **26** (11). (1983) 832–843

[15] Nghiem,A. T., Bremond, F., Thonnat, M. : Controlling Background Subtraction Algorithms for Robust Object Detection. The Proceedings of the 3rd International Conference on Imaging for Crime Detection and Prevention, ICDP 09, Kingston University, London, UK, (2009)

[16] Chau, D. P., Bremond, F., Thonnat, M. : A multi-feature tracking algorithm enabling adaptation to context variations. In the Imaging for Crime Detection and Prevention Conference (ICDP 2011), Kingston University, London, UK, (2011)