Evaluation of a Monitoring System for Event Recognition of Older People

Carlos Fernando Crispim-Junior, Vasanth Bathrinarayanan, Baptiste Fosty, Rim Romdhane, Alexandra Konig, Monique Thonnat, François Bremond

Evaluation of a Monitoring System for Event Recognition of Older People

Carlos Fernando Crispim-Junior[1], Vasanth Bathrinarayanan[1], Baptiste Fosty[1], Alexandra Konig[2], Rim Romdhane[1], Monique Thonnat[1], Francois Bremond[1,2]
[1] INRIA – Sophia Antipolis, [2] CMRR-Nice
[1] 2004 Route de Lucioles, Sophia Antipolis
[2] 4 avenue Reine Victoria, Nice.

Abstract

Population aging has been motivating academic research and industry to develop technologies for the improvement of older people's quality of life, medical diagnosis, and support on frailty cases. Most of available research prototypes for older people monitoring focus on fall detection or gait analysis and rely on wearable, environmental, or video sensors. We present an evaluation of a research prototype of a video monitoring system for event recognition of older people. The prototype accuracy is evaluated for the recognition of physical tasks (e.g., Up and Go test) and instrumental activities of daily living (e.g., watching TV, writing a check) of participants of a clinical protocol for Alzheimer's disease study (29 participants). The prototype uses as input a 2D RGB camera, and its performance is compared to the use of a RGB-D camera. The experimentation results show the proposed approach has a competitive performance to the use of a RGB-D camera, even outperforming it on event recognition precision. The use of a 2D-camera is advantageous, as the camera field of view can be much larger and cover an entire room where at least a couple of RGB-D cameras would be necessary.

1. Introduction

Population aging has been motivating academic research and the industry to develop technologies which would improve the quality of life of older people, and provide support on daily living activities, especially in cases of frailty and disablement.

Monitoring systems have been proposed to support doctors during objective assessment of health status of older people. Further examples are: gait analysis and the identification of motor disturbances; assessment of physical task performance; health status monitoring (*e.g.*, rising patterns of frailty) and the identification of harmful situations (*e.g.*, fall).

Gao *et al*. [1] have demonstrated the fusion of data from inertial sensors worn at the waist, chest, thigh, and side of a person body to action recognition (sitting, standing, and walking). A Naïve Bayes Classifiers is used for inertial sensor fusion. Similarly, Rong and Ming [2] have used C4.5 algorithm and a Naïve Bayes classifier for inertial sensors fusion on activity recognition. Disadvantages of these approaches are motion noise; inter sensor-calibration, and sometimes the assumption that the sensors are always placed at the same body position among patients, which can add noise in large scale studies.

Fleury *et al*. [3] have presented a multi-modal system for activity recognition of older people on Smart-Homes. The approach uses sensors such as actimeter, microphones, passive infrared sensors, and door contacts whose outputs are fused by an SVM classifier. Medjahed *et al*. [4] have proposed a similar approach relying on infrared, change state, audio, and physiological sensors; but the combination of sensor events is performed using a Fuzzy inference approach.

Cao *et al*. [5] have proposed a description-based approach for older people monitoring whereas the human body context (*e.g.*, sitting, standing, walking) and the environment context are described in function of event models. The body context data is provided by a set of cameras, while the environmental context is obtained of accelerometers attached to objects of daily living (*e.g.*, TV remote control or doors use). A rule-based reasoning engine is used for processing and combining both context types.

Zouba *et al*. [6] have evaluated a video monitoring system at the identification of activities of daily living of older people on a model apartment equipped with home appliances. A set of environmental sensors (pressure, contact) is attached to the home appliances, and their change of state is modeled using a hierarchical description- based approach. A 2D-RGB camera is used to track the people over the environment and estimate their posture. A long term evaluation is performed (4 hours per patient), but the results are demonstrated only for four participants.

Joumier *et al*. [7] have evaluated motor disturbances among older people using a video monitoring system based on a hierarchical description-based approach. They extract attributes (*e.g.*, duration, walking speed) of automatically recognized physical task events in order to identify differences between Alzheimer and healthy participants groups.

Banerjee *et al* [8] have presented a video monitoring for fall detection in Hospital Rooms using a RGB-D camera as input. A Fuzzy inference approach is proposed to reason over features extracted from depth information provided by the camera.

Activity Recognition in the context of older people monitoring has been mainly presented using ambient sensors (ambient intelligence) and wearable sensors (*e.g.*, inertial), with a few cases where they are combined with video-cameras [5-8]. In this sense, video monitoring can incorporate scene semantics while performing people localization and extracting body features (as posture) at the same time. The use of video monitoring avoids missing data in cases of a participant considers uncomfortable or refuses to use wearable sensors, or cannot use it due to a medical condition (*e.g.*, pacemaker).

A description-based approach can be used to provide flexible way of adding and changing event models to domain experts, as presented in [6,7]. Comparatively to Classification methods and Probabilistic Graphical models approaches, such as [8], it does not require as much data as no training phase is needed, and its hierarchical nature allows the explicit modeling of composite events.

We herein present the evaluation of a video monitoring system for event recognition of older people using a hierarchical description-based approach and a 2D-RGB camera as input. The evaluation is performed on the recognition of physical tasks and instrumental activities of daily living (IADL) of participants of a clinical protocol for an Alzheimer disease study. To assess the improvement brought by the use of an RGB-D camera, as the prototype proposed by Barnejee *et al*. [8], we compare our research prototype with a RGB-D camera.

The presented evaluation is as large as the one found in Joumier *et al*. [6] in terms of number of patients, but we also evaluate complex activities of longer duration, such as Instrumental Activities of Daily Living. Although we do not evaluate patients on hourly-basis period as Zouba *et al*. [5], the video recordings herein used have a higher number of events.

The paper is organized as follows: the Video Monitoring system is presented in section 2, the Evaluation procedure is described in section 3, Results and Discussion in section 4, followed by the Conclusions in section 5.

2. Video Monitoring

The video monitoring system is divided into two main modules: the vision and the event recognition module. The vision module is responsible for detecting and tracking people on the scene. The event recognition module uses a generic constraint-based ontology language proposed by [5] for event modeling, and the reasoning algorithm proposed by Vu *et al*. [8]. Figure 1 presents the architecture of the proposed video monitoring system.

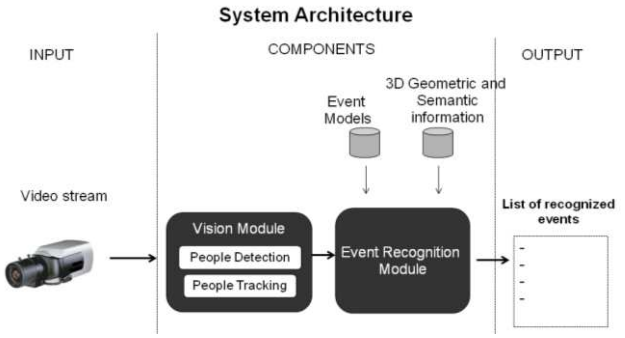

Figure1. Overall Architecture of the Video Monitoring System

2.1. Vision component

This component is a modular platform, locally developed, that allows testing of different algorithms for each step of a computer vision chain (*e.g.*, image acquisition, image segmentation, physical objects detection, physical objects tracking, actor identification, and event recognition). It extracts foreground objects in the current frame using an extension of the Gaussian Mixture Model algorithm for background subtraction proposed by Nghiem *et al*. [12]. Object tracking is performed by a multi-feature algorithm proposed by Chau *et al*. [13] using features such as 2D size, 3D displacement, color histogram, and dominant color.

The vision component is also responsible for classifying objects according to a set of a priori defined objects of interest (called scene actors, *e.g.*, a person, a vehicle). The detected scene actors are then passed to the event recognition module which assesses whether the actions/activities of these actors match the event models defined by the domain experts. Figure 2 illustrates a detected person been tracked on the scene. Blue dots represent previous positions of the person.

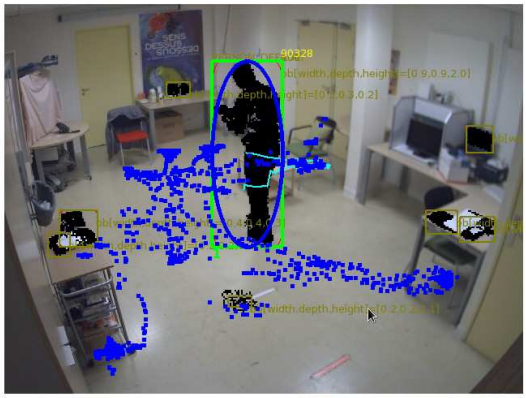

Figure 2. Scene Image of a detected Person been tracked by the vision component. Blue dots represent the person previous positions on the scene.

2.2. Event Recognition Module

This module is mainly composed of an event modeling framework, and a temporal scenario recognition algorithm which assesses whether the constraints of event models are satisfied [10]. The event models are built taking into account a priori knowledge of the experimental scene, and attributes of dynamic objects (herein called Physical Objects, *e.g.*, a person, a car) detected and tracked by the vision component. The event modeling follows a declarative and intuitive ontology-based language which uses natural terminology to allow end users (*e.g.*, medical experts) to easily add and change models. The a priori knowledge consists of the decomposition of a 3D projection of the scene floor plan in a set of spatial zones which have semantic information regarding the events (*e.g.*, TV zone, Armchair zone, Office Desk, Coffee machine object).

An event model is composed of six components [5]:

- Physical Objects refers to real objects involved in the recognition of the event modeled. Examples of physical object types are: mobile objects (*e.g.* person herein, or vehicle in another application), contextual objects (equipments) and contextual zones (chair zone);
- Components refer to sub-events that the model is composed of;
- Forbidden Components refer to events that should not occur in case of the event model is recognized;
- Constraints are conditions that the physical objects and/or the components should hold. These constraints could be logical, spatial and temporal;
- Alert describes the importance of a detection of the scenario model for a given specific treatment; and
- Action in association with the Alert type describes a specific action which would be performed when an event of the described model is detected (*e.g.* send a SMS to a caregiver responsible to check a patient over a possible falling down).

Three types of Physical Object are defined for this prototype: Person, Contextual Zones and Contextual Objects. The Person class is an extension of a generic class named mobile, which contains information of mobile objects (*e.g.*, 3D position, width, height). The Person class model has attributes like body posture, appearance, etc. Contextual Zone and Object classes refer to static objects a priori defined in which the Person interaction with is of particularly interest for an event modeling.

Constraints define conditions that physical object property (ies) and/or components should satisfy. They can be a-temporal, such as spatial and appearance constraints; or they could be temporal and specify two instances ordering which should generate a third event, for example, Person_crossing_from_Zone1toZone2 is defined as Person_in_zone1 before Person_in_zone2. Temporal constraints are expressed using Allen's interval algebra (*e.g.*, BEFORE, MEET, and AND) [11].

The ontology hierarchically categorizes event models according to their complexity (in ascending order):

- Primitive State models an instantaneous value of a property of a physical object (Person posture, or Person inside a semantic zone).
- Composite State refers to a composition of two or more primitive states.
- Primitive Event models a change in a value of physical object property (*e.g.*, Person changes from Sitting to Standing posture).
- Composite Event refers to the composition of two previous event models which should hold a temporal relationship (Person changes from Sitting to standing posture before Person in Corridor Zone).

Figure 3 presents an example of a primitive state model. This model checks for the state of the attribute Posture of a detected and tracked Person whether it fits the desired posture value.

```
PrimitiveState(Person_ sitting,
  PhysicalObjects( (p1:Person) )
  Constraints ( (P1->Posture = sitting) )
)
```

Figure 3. Primitive State of Person sitting

Figure 4 presents the Composite Event "Person using OfficeDesk". The model has two components which must be detected together (expressed by the use of AND operator) to allow the model recognition. The second constraint specifies that the component 1 (Person inside OfficeDesk zone) has to be already recognized on the scene for at least 2 seconds.

```
CompositeEvent(Person_using_OfficeDesk,
  PhysicalObjects(
      (p1:Person), (z1:Zone) )
  Components(
    (c1:CompositeEvent
  P_insideOfficeDeskZone(p1,z1))
    (c2:PrimitiveState  P_sitting (p1)))
  Constraints(  ( duration(c1) > 2 )
                    (c1 and c2)
  )
 )
```

Figure 4. Description of the event model "Person using office desk". P_sitting states for Person sitting.

3. Evaluation

The present evaluation assesses the accuracy of the prototype during the event recognition of physical tasks and IADLs. The recognition of physical tasks is performed on 29 videos of ~6 min each. The results are compared to a monitoring system using the same hierarchical description-based approach, but using a RGB-D camera. The RGB-D camera provides real measurements of 3D data, in opposition to the 2D RGB camera used by the proposed prototype, which needs to be calibrated to obtain 3D information.

The 2D RGB camera records video data from the top of one of the corners of the experimentation room, while the RGB-D camera records them from a lateral view of the scene. The lateral view is chosen as this camera placement on a position similar to the RGB camera would decrease its system performance on people detection. The decrease is due to the depth measurement become less reliable when the person moves farther than 4-5 m of the camera (*e.g.*, on events at the back of the room). Figure 5 illustrates the differences in point of view of both cameras. The second part of the evaluation focuses on the evaluation of mid-term duration activities (IADLs).

Figure 5. Scene point of the view in respect to the 2D RGB (A) and RGB-D (B) cameras.

The prototype accuracy is evaluated using the indices of sensitivity, precision, and F-score described in Equations 1, 2, and 3, respectively, in comparison to events annotated by domain experts.

$$Sensitivity = \frac{TP}{TP + FN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

where: TP: True Positive rate, FP: False Positive rate, FN: False Negative rate.

$$F - Score = 2 * \frac{Sensitivity * Precision}{Sensitivity + Precision} \quad (3)$$

All evaluations are performed on video recordings of participants of a clinical protocol devoted to the study of early markers of Alzheimer disease. The dataset is described on the next section.

3.1. Dataset

Participants aged more than 65 years are recruited by the Memory Center (MC) of a collaborating Hospital. Inclusion criteria of the Alzheimer Disease (AD) group are: diagnosis of AD according to NINCDS-ADRDA criteria and a Mini-Mental State Exam (MMSE) [14] score above 15. AD participants who have significant motor disturbances (per the Unified Parkinson's Disease Rating Scale) are excluded. Control participants are healthy in the sense of behavioral and cognitive disturbances.

The clinical protocol asks the participants to undertake a set of physical tasks and Instrumental Activities of Daily Living in a Hospital observation room furnished with home appliances. Experimental recordings use a 2D-RGB video camera (AXIS®, Model P1346, 8 frames per second), a RGB-D camera (Kinect®, Microsoft©). The activities of the clinical protocol are divided into three scenarios: Guided, Semi-guided, and Free activities.

Guided activities (10 minutes) intend to assess kinematic parameters of the participant gait profile (*e.g.*, static and dynamic balance test, walking test):

- Balance testing: the participant should keep balance while performing actions such as keeping both feet side by side stand, standing with the side of the heel of one foot touching the big toe of the other foot, among others;
- Walking Speed test (WS): the assessor asks the participant to walk through the room, following a straight path from one side of the room to the other (chair side to video camera side, outward attempt, 4 meters), and then to return (return attempt, 4 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 assessor will then ask the participant to repeat the same action five times in a row; and
- Time Up & Go test (TUG): participant start from the sitting position, and at the assessor's signal he/she needs to stand up, to walk a 3 meters path, to make a U-turn in the center of the room, return and sit down again.

Semi-guided activities (15 minutes) aim to evaluate the level of autonomy of the participant by organizing and carrying out a list of instrumental activities of daily living (IADL) within 15 minutes. The participant is alone in the room with the list of activities to perform, and he/she is advised to leave the room only when he/she has felt the required tasks are completed.

- Watch TV,
- Make tea/coffee,
- Write the shopping list of the lunch ingredients,
- Answer the Phone,

- Read the newspaper/magazine,
- Water the plant,
- Organize the prescribed drugs inside the drug box according to the daily/weekly intake schedule,
- Write a check to pay the electricity bill,
- Call a taxi,
- Get out of the room.

Free activities (5 minutes) aims to assess how the participant spontaneously initiates activities and organize his/her time.

The present paper focuses on the use of an ambient camera and the recognition of guided- and semi-guided scenarios of the dataset. For these accuracy evaluations, no difference is considered with respect to the doctor diagnosis of the participant condition (*e.g.*, healthy, Alzheimer disease).

4. Results and Discussion

Table 1 presents the results of the research prototype for the event recognition of physical tasks (guided-scenario). The activities in this scenario follow a time order and their recognition generally depends on the People Tracking (*e.g.*, WS return attempt should happen after WS outward), Posture detection (*e.g.*, Repeated transfer is characterized by transfers between sitting and standing posture), and the identification of a person presence in one or more zones. A few events are a combination of the detection of the participant presence in a specific zone with a particular posture, for example, the difference between WS and Up and Go tests rely on the fact Up and Go starts from sitting posture while WS from standing. The reference approach refers to a monitoring system following the same hierarchical description-based approach, with similar event models, but using a RGB-D camera as input. From the medical perspective, these events refer to an evaluation of the motor abilities of a participant, consisting of short-duration events which take place in a predefined area of the observation room.

Table 1. Prototype Accuracy on Physical Tasks

Approach	Reference		Proposed	
IADL	Sens.	Precision	Sens.	Precision
Repeated Transf.	100.00	90.90	75.00	100.00
Up and Go	93.30	90.30	91.66	100.00
Balance Test	100.0	100.00	95.83	95.83
WS.Outward	100.0	90.90	91.66	100.00
WS. Return	90.00	100.00	87.50	95.45
Average	96.60	94.20	88.33	**98.33**

N: 29 participants, ~ 6 min. each, Total: ~150 events. Sens.: Sensitivity index, eq.1; WS: Walking Speed test

It is seen that the proposed approach has a higher precision than the reference approach, although the reference approach presents a higher sensitivity. Table 2 presents the comparison of both systems using the F-Score. In this performance index a slightly higher performance can be observed in favor of the Reference approach.

Table 2. F-Score comparison at Physical Task Recognition

Approach	Reference	Proposed
F-Score	95.40	93.00

Table 3 presents the performance of the prototype at recognizing IADLs (semi-guided scenario). These activities reflect the state of the cognitive abilities of a person, and therefore constitute an important index of an older person health status for medical experts. The results are presented only for the proposed prototype (using a RGB camera) due to the fact IADLs are performed over the whole observation room, and the smaller field of view of the RGB-D camera is insufficient to capture the entire scene. The IADLs are modeled using 7 (seven) composite event models, which are composed of a Primitive State for the recognition of the person position inside a contextual zone, and another for his/her proximity to a contextual object in this zone (*e.g.*, Phone station, Coffee machine). The activities "writing a check" and "writing a shopping list" are not differentiated, and are referred as "Using office desk". This simplification is adopted as the visual component herein used does not provided information regarding contextual object handling. The label of activity "Organize the prescribed drugs…" is shortened as "Using Pharmacy Basket" to improve table layout.

Table 3. Prototype Accuracy on IADLs

IADL	**Sensitivity**	**Precision**
Using Phone	72.83	85.50
Watching TV	80.00	71.42
Using Office Desk	92.72	58.62
Preparing Tea/ Coffee	90.36	69.44
Using Pharmacy Basket	100.00	88.09
Watering plant	100.00	64.91
Reading	71.42	69.76
Average Performance	86.76	72.53

N: 29 participants, ~ 15 min each, Total: 435 min, ~232 events.

The achieved performance index values demonstrated in Table 3 are lower than the ones found for physical tasks. This is explained by the fact that IADL event generally involve the interaction with objects spread over the room, and they do not follow a specific order in opposition to the physical tasks of the guided scenario. Sensitivity values have been affected by noise on underlying vision components that will be fixed with

future improvements of people detection algorithm. Lower precision values refer to activities in which the contextual object and zone are spatially close, and when not fully performed by the patient does not provide enough evidence for the correct recognition. For example, "making tea" and "watering the plant" are spatially close, but involve different contextual objects. Time to time a participant places in between the objects and just stretches him/herself to reach the objects, providing ambiguous evidence to the event recognition module. Possible solutions are the adoption of a probabilistic framework for noise handling and/or a multi-sensor approach for the cases where a lack of field of view prevents the system from capturing the full activity (*e.g.*, wearable camera).

5. Conclusions

An evaluation has been presented for our research prototype devoted to event recognition based on a hierarchical description-based framework. This is the largest evaluation on event recognition for monitoring of older people using video cameras, both in the number of events and of real patients involved.

This experimentation shows that by using state of art algorithms for image segmentation and object tracking we can obtain an event recognition accuracy that is competitive to a similar system using a RGB-D camera, if not better for certain cases (*e.g.*, precision). One advantage of RGB cameras is their larger field of view compared to the compared RGB-D cameras. For the same surface, it will be necessary to employ at least two RGB-D cameras to monitor the same area, what will consequently increase the system processing time and complexity.

The prototype successfully recognizes physical tasks with F-Score of 93.00%, and IADLs with a sensitivity of 86.76 % and a precision of 72.53 %. Performance values at IADLs are lower than the ones obtained on physical task recognition due to the fact that they are obtained from less structured and more complex activities.

The presented prototype also follows an easy to deploy approach as it is based on description-based event recognition framework to allow domain experts to add and change models without having to change the system core.

Future work will focus on the development of objective evaluation of differences between healthy and dementia patients by focusing on the attributes of automatically recognized events (*e.g.*, duration), which are free of biases related to human factors, like fatigue, stress.

References

[1] L. Gao, A.K. Bourke, J. Nelson. A system for activity recognition using multi-sensor fusion. In Proceedings of Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pages 7869-7872, August 2011.

[2] L. Rong and L. Ming. Recognizing Human Activities Based on Multi-Sensors Fusion. In Proceedings of 4th International Conference on Bioinformatics and Biomedical Engineering, pages 1-4, June 2010.

[3] A. Fleury, N. Noury, M. Vacher. Introducing knowledge in the process of supervised classification of activities of Daily Living in Health Smart Homes. In Proceedings of 12th IEEE International Conference on e-Health Networking Applications and Services, pages 322-329, July 2010.

[4] H. Medjahed, D. Istrate, J. Boudy, J.-L. Baldinger, B. Dorizzi. A pervasive multi-sensor data fusion for smart home healthcare monitoring. In Proceedings of IEEE International Conference on Fuzzy Systems, pages 1466-1473, June 2011.

[5] Y. Cao, L. Tao, and G. Xu. An event-driven context model in elderly health monitoring. In Proceedings of Symposia and Workshops on Ubiquitous, Autonomic and Trusted Computing, 2009, pages 120-124.

[6] N. Zouba, F. Bremond and M. Thonnat. An Activity Monitoring System for Real Elderly at Home: Validation Study. In the 7th IEEE International Conference on Advanced Video and Signal-Based Surveillance, AVSS 2010, Boston, 29 August 2010.

[7] J. Joumier, R. Romdhane, F. Bremond, M. Thonnat, E. Mulin, P.H. Robert, A. Derreumeaux, J. Piano and L. Lee. Video Activity Recognition Framework for assessing motor behavioural disorders in Alzheimer Disease Patients. In the International Workshop on Behaviour Analysis, Behave 2011, Sophia Antipolis, France.

[8] T. Banerjee, M. Rantz, M. Popescu, E. Stone, M. Li and M. Skubic. Monitoring Hospital Rooms for Safety Using Depth Images. AI for Gerontechnology, Arlington, Virginia, US, November 2012.

[9] H. Izadinia and M. Shah. Recognizing complex events using large margin joint low-level event model. In Proceedings of the 12th European conference on Computer Vision, 4:430-444, Firenze, Italy, October 2012.

[10] T. Vu, F. Brémond and M. Thonnat, Automatic Video Interpretation: A Novel Algorithm for Temporal Scenario Recognition. The Eighteenth International Joint Conference on Artificial Intelligence (IJCAI'03), Acapulco, Mexico, 9-15 August 2003.

[11] J.F. Allen. Maintaining Knowledge about temporal intervals. Communications of the ACM, 26(11):832-843, November 1983.

[12] A.T. Nghiem, F. Bremond and M. Thonnat. Controlling background subtraction algorithms for robust object detection. In Proceedings of 3rd International Conference on Imaging for Crime Detection and Prevention, pages 1-6, London, UK, December 2009.

[13] D.P. Chau, F. Bremond, and M. Thonnat. A multi-feature tracking algorithm enabling adaptation to context variations. In Proceedings of International Conference on Imaging for Crime Detection and Prevention, 2011.

[14] M.F. Folstein, L.M. Robins, and J.E. Helzer; The mini-mental state examination. Arch Gen. Psychiatry, 40(7):812, 1983.