Hierarchical Hidden Markov Model in Detecting Activities of Daily Living in Wearable Videos for Studies of Dementia

Svebor Karaman, Jenny Benois-Pineau, Vladislavs Dovgalecs, Rémi Mégret, Julien Pinquier, Régine André-Obrecht, Yann Gaëstel, Jean-François Dartigues

Hierarchical Hidden Markov Model in Detecting Activities of Daily Living in Wearable Videos for Studies of Dementia

Svebor Karaman[(1)], Jenny Benois-Pineau[(1)], Vladislavs Dovgalecs[(2)], Rémi Mégret[(2)], Julien Pinquier[(3)], Régine André-Obrecht[(3)], Yann Gaëstel[(4)] and Jean-François Dartigues[(4)]

[(1)] LaBRI – University of Bordeaux, 351 Cours de la Libération, 33405 Talence Cedex, France. svebor.karaman@labri.fr, jenny.benois@labri.fr

[(2)] IMS – University of Bordeaux, 351 Cours de la Libération, Talence, France. remi.megret@ims-bordeaux.fr, vladislavs.dovgalecs@ims-bordeaux.fr

[(3)] IRIT – University of Toulouse, 118 route de Narbonne, 31062 Toulouse Cedex 9, France. pinquier@irit.fr, andre-obrecht@irit.fr

[(4)] INSERM U.897, University Victor Ségalen Bordeaux 2, Bordeaux, France. jean-francois.dartigues@isped.u-bordeaux2.fr, yann.gaestel@isped.u-bordeaux2.fr

Abstract

This paper presents a method for indexing activities of daily living in videos obtained from wearable cameras. In the context of dementia diagnosis by doctors, the videos are recorded at patients' houses and later visualized by the medical practitioners. The videos may last up to two hours, therefore a tool for an efficient navigation in terms of activities of interest is crucial for the doctors. The specific recording mode provides video data which are really difficult, being a single sequence shot where strong motion and sharp lighting changes often appear. Our work introduces an automatic motion based segmentation of the video and a video structuring approach in terms of activities by a hierarchical two-level Hidden Markov Model. We define our description space over motion and visual characteristics of video and audio channels. Experiments on real data obtained from the recording at home of several patients show the difficulty of the task and the promising results of our approach.

1. Introduction

The application which drives our research is the vast and urgent area of sustaining capacities of the society to afford increasing costs and ensure convenient

treatment to the rapidly aging population of the planet. The aging of the world population has led to an unprecedented rise in the spread of dementia globally, with significant personal and socio-economic costs. For instance, Europe is currently undergoing a demographic change with the average age of the population increasing significantly. According to EUROPOP2008 [1] (EUROpean POpulation Projections, base year 2008), by 2060 it is projected that almost three times as many people will be aged 80 or more. An unfortunate effect of ageing is the greater likelihood of exposure to chronic conditions like dementia; today 10% of people over the age of 65 have Alzheimer's disease. The situation will be even worse if Science does not make significant progress: 25% of population after the age of 65 will have dementia diseases. Elderly people with strong cognitive impairments such as dementia cannot sustain their life independently, and the placement in nursing homes becomes unavoidable with a high cost for the society. One of the major goals in medical research is the early diagnosis of a dementia disease. This would help proposing appropriate care giving and later entering to the specialized institutions, thus reducing ever growing costs of these diseases for the society. According to results of medical research [2] the first decline in cognitive performances can appear as early as 12 years before dementia phase and up to 10 years before the individuals become slightly dependent in their activities of daily living. To detect the first signs a subject could show and also to assess the progression of the disease with patients in established dementia phase, an objective observation of various activities in an everyday life is required for medical practitioners. This is why the observation with various types of sensors including video cameras is now entering in clinical practice [31]. The amount of generated video data is usually very large: indeed the observation with external cameras in smart homes can last for several hours [3]; the monitoring with wearable sensors can be shorter, but still unexplorable by a medical practitioner in the short time allocated to the preparation of an appointment with a patient. Hence the necessity of automatic recognition of activities of interest is obvious. In literature, a large amount of research is now devoted to the recognition of human activities in video recorded with stationary video sensors installed in buildings, e.g. [4]. Some of them are specifically devoted to the elderly people for observation to assess the degree of autonomy and signs of coming dementia [3]. Nevertheless, the "external observation" does not allow for a medical practitioner

to observe fine actions of the patient, such as instrumental activities of daily leaving. To see how a person uses aliments when cooking or how he/she is knitting or washing dishes would require a very dense installation of various sensors at home for each observed patient. Furthermore, equipping homes with sensors is not always well accepted by elderly persons. Hence, during the last decade various attempts have been made to use a wearable video acquisition set-up. Thus the SenseCam [6] device, which is worn by a person, serves to constitute his life-log and to rememorize the events for a person with memory impairments. The WearCam [7] project uses a camera strapped on the head of young children. This setup allows capturing the field of view of the child together with his gaze in order to monitor the impairments in child's development. All these wearable video sensors somehow give "the point-of-view" of a person in his activity. They are thus precious for the fine studies of behavior as required in particular for dementia diseases. Hence, in this paper we develop the research we first proposed in [5] on activities recognition in videos recorded with cameras worn by patients for diagnosis and monitoring of dementia. We specifically focus on the formalism used for indexing of instrumental activities in such videos, which is a hierarchical two level Hidden Markov Model (HMM). The paper is organized as follows: in section 2 we explicit the origin and the nature of video data, in section 3 we describe the proposed formalism. In section 4, we present our method for partitioning the videos and in section 5 we present our strategy for the extraction of meaningful features which feed the HMM as observations. Experiments with various configurations of the model and description space are reported in section 6. Finally, we conclude and give perspectives of this work in section 7.

2. Origin of the data

a. The visit and recording protocol

The idea of this research from a medical point of view is to use the video recording in the same way as a clinical test such as MRI or radiography and to get this observations in a stress-less and friendly environment for the patient, while at home. In a target usage scenario the doctor will ask the paramedical staff to visit the patient with the recording system. Then, the recorded video is automatically processed and indexed by our method off-line. Finally, the doctor will use the

video and indexes produced by our analysis to navigate in it and search for the activities of interest. Visual analysis of the latter serves to diagnose the disease or assess the evolution of the patient's condition. The typical recording scenario consists of two stages. A small bootstrap video for estimation of patient's localization in his home environment is recorded at the beginning of the recording session. Indeed, when a paramedical assistant comes to visit a patient for the first time, the patient "visits" his house when recording. Then, the patient is asked to perform some of the activities which are a part of clinical evaluation protocols in assessing dementia progress. These activities define the targeted events to be detected by our method.

b. The video recording device

The video acquisition device should be easy to put on, should remain in the same position even when the patient moves hectically. It has to bring as less discomfort as possible to an aged patient. Regarding these constraints, a vest was adapted to be the support of the camera. The camera is fixed approximately on the shoulder of the patient with hook-and-loops fasteners which allow the camera's position to be adapted to the patient's morphology. This position combined with the wide angle lens of the camera offers a large view field similar to the patient's one. With the camera being light and the vest distributing the weight on all the upper body, the acceptance of the device is very good. The volunteers have felt no discomfort while wearing it and were able to perform their activities as if the device was not present. An illustration of the device is given in Figure 1.

c. The video characteristics

The videos obtained from wearable cameras are quite different from the standard edited videos in e.g. cinema, commercials, sports or TV programs of other genres. Indeed, edited videos which are usually a target of video indexing methods have a "clean" motion and are assembled from video shots with discontinuities on the shot borders. In our case, the video is recorded as a long continuous sequence, as in surveillance applications. The latter deals with stationary cameras or with regular motions, such as PTZ. In our wearable setting the camera has a wide angle lens in order to capture a large part of the patient's environment. Hence ego-motion of the patient even of a weak physical magnitude can yield strong changes

FIGURE 1: The recording device (red circle) fixed on the vest adapted to be the support of the camera.

in the recorded scene content in the field of view of the wearable camera as well as a strong blur. Furthermore, when moving in a natural home environment, the patients face strong light sources, such as windows resulting in saturation of luminance in the field of view. Examples of such challenging videos are represented in Figure 2 by key-frames.

(a) Motion blur due to strong motion.

(b) Low lighting while in dark environment.

(c) High lighting while facing a window.

FIGURE 2: Examples of frames presenting challenging data for video analysis.

Furthermore, the variability of the data is very strong: the same activities are not performed by different patients in the same environment as this is the case in "smart homes" [3].

d. Activities of Daily Living

Up to now, the medical practitioners were using a paper questionnaire while interviewing the patient and his relatives to determine his ability to correctly perform the following Activities of Daily Living (ADL): "Hoovering",

"Sweeping", "Washing Clothes", "Serving", "Making Coffee", "Making Snack", "Hair Brushing", "Phone", "TV", "Knitting", "Plant Spraying", "Listening to Radio", "Wiping Dishes", "Brushing Teeth", "Washing Dishes". These ADL are the target of automatic recognition in order to provide doctors with an efficient navigation tool throughout the video recorded at the patient's home. Hence, we come now to the problem of recognition of activities sequential in time, on the basis of noisy and variable data with some possible constraints on their time scheduling. We resort to HMMs, which proved to be an excellent model for such types of problems [35].

3. Hidden Markov Models for Video Structuring

a. Classical Hidden Markov Models

The Hidden Markov Model (HMM) is a statistical model which was first introduced in [9] where its application to speech recognition was presented. An HHM is composed of m states: $Q = \{q_1, \dots, q_m\}$. An observation model, which is usually a Gaussian Mixture Model (GMM) for continuous observations, and transition probabilities towards other states and itself are associated to each state q_i. The transition matrix $A = (a_{ij})$ contains all the transition probabilities between all states of the HMM, a_{ij} is the transition probability between state q_i and q_j and the diagonal of the matrix contains all the loop probabilities.

HMMs were later applied to many fields such as handwriting and gesture recognition [12], bioinformatics and video. The video applications of the HMMs have been first designed for low-level temporal structuring like the method for video segmentation using image, audio and motion content presented in [10], were the HMM states represent the camera motion and the transitions between shots. Since the work [9] the research in HMMs has been very intensive and resulted both in more sophisticated observation models and more complex architecture. Hence in [12] a non-Gaussian HMM has been proposed to cope with outliers in observations. As for the structure of HMMs, the richness of application contexts and constraints imposed in various spatio-temporal scenarios to model yielded a wide range of HMMs. An HMM can be fully connected or partially connected. In the later case some transitions probabilities in matrix A are forced to be null. Amongst the variety of HMMs hierarchical and segmental HMMs

turned to be the most popular for modeling of activities in video streams. In the following sub-sections we analyze their advantages (and draw backs) and propose our solution.

b. Hierarchical Hidden Markov Models

With regards to the complexity and inherent hierarchical structure in video scenes coming from numerous application domains the classical "flat" HMM, as presented above, is limited. Indeed the structure of video scenes in e.g. sports video can be mapped to more than one HMM. Thus; in tennis videos [11] one HMM can describe a match as a set of "sets", each set can be represented as a set of "games" and each game can be represented as set of "points" etc.., up to an elementary events such as "rally", "first missed serve". In our case of modeling of activities of daily leaving each activity, such as "washing dishes", "making coffee", "watching TV" etc can also be decomposed into some elementary events. In both cases a vertical link of hierarchy exists between states of more "global" HMM and "more detailed" HMM. This can be represented by a specific case of HMM, the so-called "hierarchical HMMs (HHMMs)" modeling both the hierarchy of events (states) and transitions between them. Usually [14] the hierarchical structure is defined using the bottom-level states as emitting sates (where the observation distribution has to be learnt) and high-level states as "internal" states to model the structure of the events. In the formulation of [14] a state is denoted by q_i^d $(d \in \{1, \dots, D\})$ where i is the state index and d is the hierarchy index i.e. the state level, with $d = 1$ the top-level HMM. The possible transitions are both horizontal (between the states of the same level) and vertical between states of neighboring levels d and $d+1$. A transition matrix $A^{q_i^d} = \left(a_{jk}^{q_i^d}\right)$ is defined for the sub-states of each internal state q_i^d, where $a_{jk}^{q_i^d} = P(q_j^{d+1}, q_k^{d+1})$ is the probability of making a horizontal transition from sub-states j and k of q_i^d. The vertical transitions $\Pi^{q_i^d} = \left\{\pi^{q_i^d}(q_j^{d+1})\right\} = \left\{P(q_j^{d+1}|q_i^d)\right\}$ can be seen as the probability of entering state q_j^{d+1} from its "parent" state q_i^d. This probability is related to the initial probabilities of the classical HMM as the probability that state q_i^d will initially activate the state q_j^{d+1}. Each production state (state at lower level) is parameterized by an observation model B^{q^D}. A two-level

approach was proposed in [13] where the bottom level is composed of HMMs for features analysis, and the top level is a stochastic context-free parsing. This model was applied to gesture recognition and video surveillance. The results presented in [13] show improvement in recognition performance of the proposed HHMM over a flat HMM. However, one of the main drawbacks of these fully hierarchical models is the higher number of parameters to train (such as complementary "vertical" transition probabilities $\Pi^{q_i^d}$) which induces the need of a large amount of learning data.

c. Segmental Hidden Markov Models

The major limitation of classical HMMs and derived HHMMs is the invalidity of the "Independence Assumption" [16]. This means the hypothesis on independence of subsequent observations. In video this is clearly not the case. Hence the Segmental Hidden Markov Model (SHMM), introduced in [16], addresses the problem of variable length sequences of observation vectors as presented in [17]. The application to video has been for example shown for tennis video parsing [18] where thanks to SHMM different modalities can be processed with their native sampling rates and models. Once again, despite the gain in performance, these models have a much higher computational cost and number of parameters than the flat HMM.

d. Design of an activities recognition model: a two-level hierarchical HMM

In order to take into account both the complexity of our data and the lack of large amount of training data for learning purposes, we propose the following model. If we abstract our problem of recognition of daily activities in the video to its simplest core, we can draw an equivalence between an activity and a hidden state of an HMM. The connectivity of the HMM can, at this level, be defined by the spatial constraints of the patient's environment when it is known. The easiest way is to design a fully connected HMM and train the inherent state-transition probabilities from the labeled data. Unfortunately, the ADL we consider are very much heterogeneous and often very complex, therefore the suggested equivalence between an activity and a hidden state cannot hold together.

Hence, we propose a two-level Hierarchical HMM (HHMM). The activities that are meaningful to the medical practitioners are encoded in the top-level HMM, the set of possible states is thus defined accordingly. We also introduce a reject state "None" to model non-meaningful observations from doctors' point of view. Thus defined, the top-level HMM contains the transitions between "semantic" activities including the reject class. A bottom-level HHM models an activity with m non-semantic states, as in [19]. The number of states at bottom level m is fixed to 3, 5 or 7 for ADL states and to 1, 3, 5 or 9 for the reject class "None" in our experiments. The overall structure of the HHMM is presented in Figure 3, with 3 states at the bottom level.

i. Top-level HMM

The top-level HMM represents the relations between the actions of interest, which are the ADL defined by the medical practitioners. In this work, the actions of interest are ADL such as "Meds Management", "Hand Cleaning" , "Brushing Teeth", "Plant Spraying", "Washing Dishes", "Sweeping", "Making Coffee", "Making Snack", "Hair Brushing", "Phone", "TV" etc, and another activity for all the rest which is not relevant to the ADL of interest named "None". We denote the set of states at this level as $Q^1 = \{q_1^1, \dots, q_{n_1}^1\}$ and transition matrix $A^1 = (a_{ij}^1)$, where n_1 is the number of activities. In this work, no constraints were specified over the transitions between activities as such restrictions are very difficult to know a priori when addressing a larger set of activities and when analyzing a large set of videos where the physical constraints of each patient's house are different. Moreover, the ADL a patient is asked to fulfill depend very much on his condition and their sequencing cannot be fixed for all patients in the same way. Hence, we design the top-level HMM as a fully connected one. We consider equiprobable transitions from activities states to one another, hence $\forall i, j: a_{ij}^1 = \frac{1}{n_1}$. The states of the top-level HMM modeling activities are denoted in Figure 3 as "Act" for the sake of simplicity. In our model, $a_{start,j}^2$, for $j = 1, .., 3$ in example in Figure 3, correspond to vertical transitions $\pi^{Act_i}(q_j^2)$ in the HHMM formalism of [14].

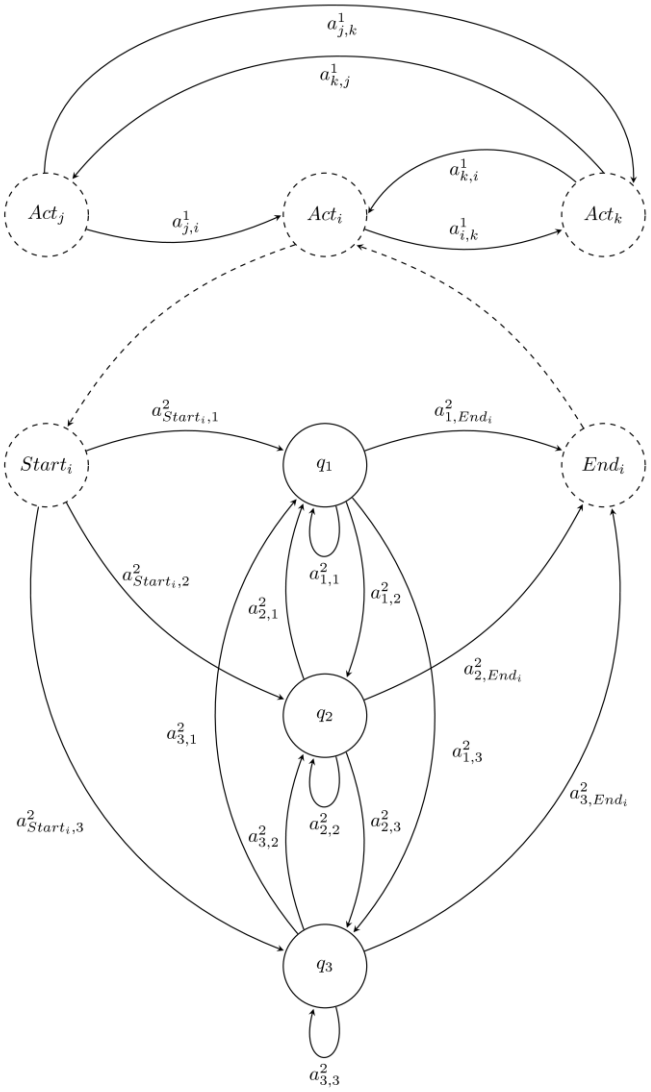


FIGURE 3: The two-level HMM structure for modeling activities of a patient. Act. Activities, q: emitting states. Dashed circled states are non emitting states.

ii. Bottom-level HMM

Most of the activities defined in the above section are complex and could not easily be modeled by one state. For each activity q_i^1 in the top-level HMM a bottom-level HMM is defined with the set of states $Q_i^2 = \left\{q_{i1}^2, \dots, q_{n_i^2}^2\right\}$ with $n_i^2 = 3,5 \text{ or } 7$ for ADL states and $n_i^2 = 3,5,7 \text{ or } 9$ states for the reject class "none" in our experiments. The state transition matrices A_i^2, for $i = 1, \dots, n_1$ also correspond to a fully connected HMM: $a_{ikl}^2 \neq 0$, at initialization, for $k = 1, \dots, n_i^2$ and $l = 1, \dots, n_i^2$. For the video stream not to be over-segmented the loop probabilities a_{ikk}^2 have to be initialized with greater values than other transition probabilities: $a_{ikk}^2 > a_{ikl}^2, \forall\, k \neq l$, this will be explicitly defined in our

experimental study. Activities are more likely to involve several successive observations rather than just one, this explains the choice for such a higher loop probability.

At the bottom level, each non semantic state models the observation vector o by a Gaussian Mixture Model (GMM). In our model we consider diagonal covariance matrix. The GMM and the transitions matrix of all the bottom-level HMMs are learned using the classical Baum Welsh algorithm [9] with labeled data corresponding to each activity.

iii. Implementation, training and recognition

HMM is a well-studied subject for today, and a lot of implementations of HMMs are available in open source software. In our implementation of the designed two-level HHMM, we used the HTK library [32]. This probably is the mostly used software for HMMs [8].

We consider a continuous HMM that models observations probability with GMM. The nature of the observations will be detailed in section 5. For training the bottom-level HMMs we use the Baum-Welsh algorithm. In the Baum-Welsh algorithm, an initialization is needed. The number of states at the bottom level m is fixed and will not be changed during the learning process. The transition probabilities are initialized with greater values for loop probabilities as stated in previous section, the exact values are precised in each experiment presented in section 6. We used a fixed number of Gaussian components for the observation model. The HTK Baum-Welsh training implementation may discard low-weight Gaussian components in a mixture. Precisely, the component l of the GMM is discarded if the re-estimated weight is lower than a minimal "threshold" weight. The initialization of the GMM can be done as a "flat-start" i.e. setting all means and variances to be equal to the global mean and variance. However, since the Baum-Welsh would only find a local optimum and that the amount of learning data in our context is not very large, a more detailed initialization is possible by using iterative Viterbi alignments.

For the recognition, the Viterbi algorithm is used. The HTK implementation makes efficient use of the "token passing" paradigm to implement a beam pruned Viterbi search. Details on the HTK library can be found in the HTK Book [33].

4. Partitioning into analysis units

The video structuring will rely on an analysis unit. We want to establish a minimal unit of analysis which is more relevant than the video frames. The objective is to segment the video into the different viewpoints that the patient provides by moving throughout his home. In contrast to the work in [6] where the description space is based on a fixed key-framing of the video, our goal is to use the motion of the patient as one of the features. This choice corresponds to the need to distinguish between various activities of a patient which are naturally static (e.g. reading) and dynamic (e.g. hoovering). This viewpoint segmentation of our long uninterrupted video sequences may be considered as an equivalent to shots in edited video sequences. We now detail the designed motion-based segmentation of the video.

a. Global Motion Estimation

Since the camera is worn by the person, the global motion observed in an image plane can be called the "ego-motion". We model the ego-motion by the first order complete affine model and estimate it with a robust weighted least squares by the method we reported in [20]. The parameters of (1) are computed from the motion vectors extracted from the compressed video stream (H.264 in the current recording device) where one motion vector $\vec{d}_i = (dx_i, dy_i)$ is extracted per i-th image block and is supposed to follow the model

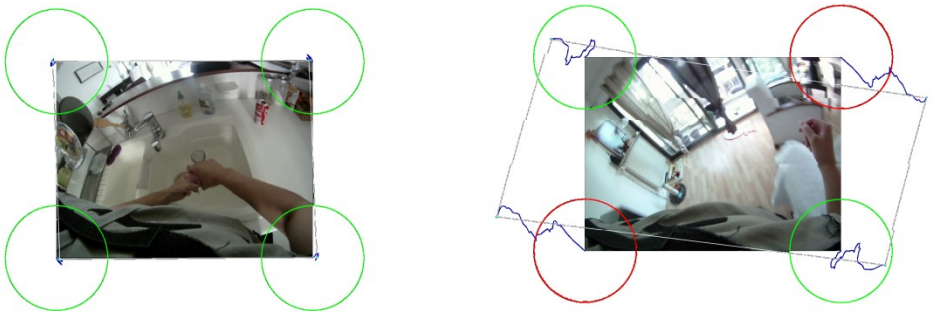
$$\begin{pmatrix} dx_i \\ dy_i \end{pmatrix} = \begin{pmatrix} a_1 \\ a_4 \end{pmatrix} + \begin{pmatrix} a_2 & a_3 \\ a_5 & a_6 \end{pmatrix} \begin{pmatrix} x_i \\ y_i \end{pmatrix} \quad (1)$$

with (x_i, y_i) being the coordinates of a block center.

b. Corners Trajectories

To split the video stream into segments we compute the trajectories of each corner using the global motion estimation previously presented. For each frame the distance between the initial and the current position of a corner is calculated. We denote w as the image width and s as a threshold on the frame overlap rate. A corner is considered as having reached an outbound position once it has had a distance greater than $s * w$ from its initial position in the current segment. These

boundaries are represented by green and red (when the corner has reached an outbound position) circles in Figure 4.



(a) Corner trajectories while the person is static.

(b) Corner trajectories while the person is moving.

FIGURE 4: Example of corners trajectories.

c. Definition of Segments

Each segment S_k corresponds to a temporal interval $S_k = [t_k^{min}, t_k^{max}]$ which aims to represent a single "viewpoint". The notion of viewpoint is clearly linked to the threshold s, which defines the minimal proportion of the first frame of a segment, which should be contained in all its frames. We define the following rules: a segment should contain a minimum of 5 frames and a maximum of 1000 frames. These boundaries on segment duration are defined to avoid an over-segmentation of the video by setting a minimal duration corresponding to a sixth of second, and to avoid having a long static activity represented by a single segment [34]. The end of the segment is the frame corresponding to the time when at least 3 corners have reached at least once an outbound position. The key frame is then chosen as the temporal center of the segment, see examples in Figure 5. Hence the estimated motion model serves for two goals: i) estimated motion parameters are used for the computation of dynamic features in the global description space, and ii) the key frames extracted from motion-segmented "viewpoints" are the basis for

FIGURE 5: An example of key frame (center) with the beginning (left) and ending (right) frames of the segment.

extraction of spatial features. We will now focus on the definition of all features and the design of the global description space candidates.

5. Audiovisual Information Extraction

The description space aims to describe the different modalities that can be extracted from the video stream. We first introduce descriptors that characterize the motion within the video recorded, then define the audio analysis and finally present static descriptors that gather the context of the patient's environment. The fusion of all these features will be presented in the last sub-section of this section.

a. Motion description

The motion contains interesting information that can be used to characterize an activity. The camera being worn by the patient, the global motion corresponds to the ego-motion. Thus, the parameters of the global motion model are directly linked to the instantaneous displacement of the patient and can help to distinguish between static or dynamic activities. However, the instantaneous motion may be limited to describe highly dynamic activities such as hoovering. We therefore seek a description of motion history defining the dynamics on a longer term. Finally, the local motion is also important and may characterize a moving object or an interaction with an object. Therefore, a set of descriptors for these several properties of the motion will be defined.

i. Global and Instant Motion

The ego-motion is estimated by the global motion analysis presented in section 4a. The parameters a_1 and a_4 are the translation parameters. We limit our analysis to these parameters, since as in the case of wearable cameras, they better express the dynamics of the behavior, and pure affine deformation without any translation is practically never observed.

The instant motion histogram is defined as the histogram of the log-energy of each translation parameter H_{tpe}, as expressed in (2), defining a step s_h and using a log scale. Since this histogram characterizes the instant motion it is computed for each frame. This feature is designed to distinguish between "static" activities e.g. "knitting" and dynamic activities, such as "sweeping".

$$H_{tpe,j}[i] = \begin{cases} 1 \text{ iff } a_j(t) \in B_i \\ 0 \text{ otherwise} \end{cases} \quad \text{with the bins } B_i \text{ defined as}$$

$$\begin{aligned} a_j \in B_1 \quad & \text{iff} \quad \log\left(a_j{}^2 \left| a_j \right.\right) < i * s_h && \text{for } i = 1 \\ a_j \in B_i \quad & \text{iff} \quad (i-1) * s_h \leq \log\left(a_j{}^2 \left| a_j \right.\right) < i * s_h && \text{for } i = 2..N_e - 1 \\ a_j \in B_{N_e} \quad & \text{iff} \quad \log\left(a_j{}^2 \left| a_j \right.\right) \geq (i-1) * s_h && \text{for } i = N_e \end{aligned}$$

Eq. (2): Translation parameter histogram, associated to a segment S_k, a_j is either a_1 or a_4.

The feature for a video segment S_k is an averaged histogram on all its frames: $\overline{H_{tpe,j}}$, j=1,4 for horizontal and vertical translations parameters, respectively a_1 and a_4. The global instant motion feature is the concatenation of both: $\overline{H_{tpe}} = \left(\overline{H_{tpe,1}}, \overline{H_{tpe,4}}\right)$.

We denote $H_{tpe}(x) = H_{tpe,1}$ the histogram of the log-energy of horizontal translation, and $H_{tpe}(y) = H_{tpe,4}$ the histogram of the log energy of vertical translation observed in image plane. The number of bins is chosen empirically and equally with regards to x and y, $N_e = 5$, the threshold s_h is chosen in such a way that the last bin corresponds to the translation of the image width or height respectively.

ii. History of Global Motion

Another element to distinguish static and dynamic activities is the motion history. On the contrary to the instant motion, we design it to characterize long-term dynamic activities, such as walking ahead, vacuum cleaning, etc. The estimation of this is done by computing a "cut histogram" H_c. The i-th bin of this histogram contains the number $H_c(i)$ of cuts (according to the motion based segmentation presented in section 5a) that happened in the last 2^i frames, see Figure 6. The number of bins N_c is defined as 8 in our experiments providing a history horizon of 256 frames. This represents almost 9 seconds of our 30 fps videos. The history horizon was chosen to be the highest power of two lower than the minimal average duration of an activity. Such a definition is a good trade-off between long term history and potential overlapping of activities. Thus defined, the cut histogram is associated to each frame in the video. The descriptor associated to a segment is the average of the cut histograms of the frames belonging to the segment.

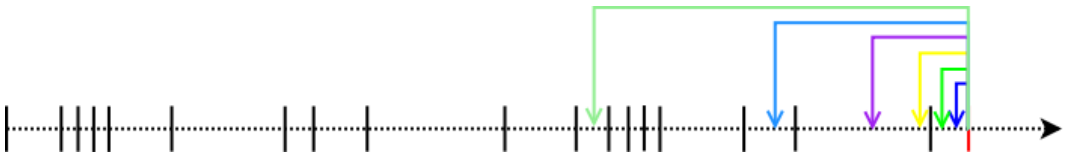


FIGURE 6: The number of cuts (black lines) is summed to define the value of each bin. In this example: $H_c[1]=0$, $H_c[2]=0$, $H_c[3]=1$, $H_c[4]=1$, $H_c[5]=2$, $H_c[6]=7$.

iii. Local Motion

All the previous motion descriptors focus on the global motion which is very important as it provides a characterization of the ego-motion. However, the residual motion may reveal additional information, such as the occurrence of a manual activity or the presence of a moving object or a person in the visual field of the patient. We introduce a descriptor which is computed on each block of a 4x4 grid partitioning of an image. The value representing each block (3) is computed as the Root Mean Square (RMS) of the difference $\overrightarrow{\Delta \boldsymbol{d}_{k,l}} = \left(\Delta \boldsymbol{dx}_{k,l}, \Delta \boldsymbol{dy}_{k,l}\right)^{\boldsymbol{T}}$ between motion vector extracted from compressed stream and the one obtained from the estimated model (1). The residual motion descriptor RM of the whole has therefore a dimensionality of 16.

$$RM_b = \sqrt{\frac{\sum_{k=1,l=1}^{k=N,l=M} (\Delta \boldsymbol{dx}_{k,l}{}^2 + \Delta \boldsymbol{dy}_{k,l}{}^2)}{N * M}} \tag{3}$$

Eq. (3): Residual Motion value for block b of width N and height M.

b. Audio

The particularity of our contribution in the design of a description space consists in the use of low-level audio descriptors. Indeed, in the home environment with ambient TV audio track, noise produced by different objects that the patient is manipulating, conversations with the persons, etc. All are good indicators of activity and its location. In order to characterize the audio environment, different sets of features are extracted. Each set is characteristic of a particular sound: speech, music, noise and silence [28]. Energy is used for silence detection. 4 Hertz energy modulation and entropy modulation give voicing information, being specific to the presence of speech. The number of segments per second and the segment duration, resulting from a "Forward-Backward" divergence algorithm [27], are used to find harmonic sound, like music.

Spectral coefficients are proposed to detect noise: percussion and periodic sounds (examples: footstep, home appliance, flowing water, vacuum cleaner, etc.). An original low level descriptor called "spectral cover" is used and allows recognizing two specific sound events: water flow and vacuum cleaner [35]. Finally, the complete set of audio descriptors is composed of 7 possible events: speech, music, noise, silence, periodic sounds, water flow and vacuum cleaner.

c. Static descriptors

Static descriptors aim to characterize the instantaneous state of the patient within his/her environment. The first static descriptor is the localization estimation. As many activities are linked to a specific location e.g. cooking in the kitchen, it can be helpful to have a estimation of the localization. The second defines the local spatial and color environment using the MPEG-7 descriptor "Color Layout". This descriptor aims at capturing the spatial and color organization of local pattern when facing a sink or a gas cooker for example.

i. Localization

We use the method of Bag of Visual Words [21] for representing an image as a histogram of visual words. Low level visual information contained within an image is captured using local features SURF [24] descriptors. Descriptors are quantized into visual words using a pre-built vocabulary which is constructed in a hierarchical manner [22]. The Bag of Words vector is built by counting the occurrence of each visual word. Due to rich visual content, the dimensionality of such histograms is very high (we used a 1111 word dictionary in our context). A kernel based approach based on the SVM classifier [26] was therefore chosen to obtain location estimates. The histograms were compared with the intersection kernel, which is adapted to such features. In practice, the feature extraction step can be done without annotation, and can be run as a preprocessing routine. Dimensionality reduction through non-linear Kernel PCA [23] with intersection kernel was included in this routine to reduce the size of the stored descriptors to several hundred linear dimensions [25]. Classification was then applied directly on these simplified descriptors. A one-vs-all approach was used to address the multi-class classification problem. The final location was represented as a vector, containing a 1 for the detected class, and 0 for other classes.

ii. Spatial and color description

Using the extracted key frames representing each segment, a simple description of the local spatial and color environment is expected. In this choice we seek for the global descriptors which characterize the color of frames while still preserving some spatial information. The MPEG-7 Color Layout Descriptor (CLD) proved to be a good compromise for both [29]. It is a vector of DCT coefficients computed on a roughly low-passed filtered and sub-sampled image. We compute it on each key frame and retain 6 parameters for the luminance and 3 for each chrominance as in [30]. This descriptor gives a coarse but yet discriminative visual summary of the local environment.

d. Descriptors fusion

Hence, for description of the content recorded with wearable cameras we designed three description subspaces: the "dynamic" subspace has 34 dimensions, and contains the descriptors $D = (H_{tpe}(x), H_{tpe}(y),\ H_c, RM)$; the "audio" subspace contains the $k = 7$ audio descriptors $p = (p_1, \ldots, p_k)$; the "static" subspace contains 19 coefficients, more precisely $l = 12$ CLD coefficients $C = (c_1, \ldots, c_l)$ and m = *7* localization coefficients $L = (l_1, \ldots, l_m)$.

We design the global description space in an "early fusion" manner concatenating all descriptors in an observation vector o in R^n space with n = 60 dimensions when all descriptors are used, thus, the designed description space is inhomogeneous. We will study the completeness and redundancy of this space in a pure experimental way with regard to the indexing of activities in Section 6, by building all the possible partial fusions.

6. Experiments

a. Corpus

The experiments are conducted on corpora of videos recorded with our wearable device by patients in their own houses. A video recording is of an average duration of 40 minutes and contains approximately 10 activities; not all activities are present in each video. Each video represents an amount of 50 to 70 thousands frames, which induces hundreds to a thousand segments according to our motion-based temporal segmentation, see section 4. The description spaces are built using

each descriptor separately and with all possible combinations of descriptors where order is not considered. Therefore, a total of 63 different descriptions spaces are considered.

The experiments are conducted in two stages. First, on a corpus of 5 videos recorded with 5 different patients. The aim of this first experiment is to analyze the overall performances of all the descriptors combinations and of the HMM configurations. The influence of the proposed motion-based temporal segmentation is also discussed in the first experiment. The second experiment use a subset of all the descriptors combinations, selected as the 13 best performances. The HMM configurations are also limited to those who have shown the best performances on the first experiment. In this experiment the corpus is larger as it contains 26 videos. The latter constitutes a unique corpus which has been recorded on healthy volunteers and patients during two years since the beginning of the research. The analysis of the performances is also two-fold: we evaluate in terms of global accuracy and for singular activities.

b. Evaluation metrics

To evaluate the *overall performance* of the proposed model we used the global accuracy metric, which is a ratio between the number of correct estimations and the total number of observations. Any misclassification of an activity, which will correspond to a *false negative* with regard to the ground truth activity and to a *false positive* with regard to the detected activity, will decrease the global accuracy metric.

Table 1: Evaluation metrics.

$precision = \frac{TP}{TP + FP}$	$recall = \frac{TP}{TP + FN}$
$accuracy = \frac{TP + TN}{TP + FP + TN + FN}$	$F - score = \frac{1}{1/precision + 1/recall}$

When the analysis is aiming for detailed performances for particular activity recognition the precision, recall and F-score metrics are used. *True positives* (TP), *true negatives* (TN), *false positives* (FP) and *false negatives* (FN) values

corresponds to the correct detection, correct absence, misdetection and missed detection respectively for a given activity. According to these definitions, we can now evaluate the metrics as presented in Table 1. These metrics can only be used when computed for each activity separately. For the global performance evaluation a detection is either true or false, the notions of *false positive* or *true negative* are irrelevant.

c. Learning and testing protocol

The experiments are conducted in a leave-one-out cross validation scheme, i.e. the HMMs are learned using all videos except one which is used for testing. The results are presented in terms of global accuracy of recognition averaged over the cross validation process. The learning is performed over a sub sampling of smoothed data extracted from frames. The smoothing substitute the value of each frame descriptor by the average value on the 10 surrounding frames, then one of ten samples is selected to build ten times more learning sequences. The testing has been done on frames or segments of the last video.

In the first experiment presented here, the bottom level HMM of each activity has 3 or 5 states. For one evaluation all activities have the same number of states, except the "None" which may be modeled with more or fewer states, here 9 or only one. All HMMs observation models are 5 Gaussians mixtures except the "None" one state-HMM which has only one Gaussian. The activities of interest for the first experiment are the ADL: "Plant Spraying", "Remove Dishes", "Wipe Dishes", "Meds Management", "Hand Cleaning" , "Brushing Teeth", "Washing Dishes", "Sweeping", "Making Coffee", "Making Snack", "Picking Up Dust", "Put Sweep Back", "Hair Brushing", "Serving", "Phone" and "TV". In the following we report the results of the evaluation for all activities, i.e. "global" evaluation and for some activities of interest. In figures the results per descriptor are sorted in decreasing order.

d. Evaluation of the influence of temporal segmentation

The proposed temporal segmentation reveals three main advantages. First, the amount of data to process in the recognition process is divided by a factor between 50 and 80 since one observation is defined for a segment and not for a frame. Second, the key frames may be used as a summary of the whole video

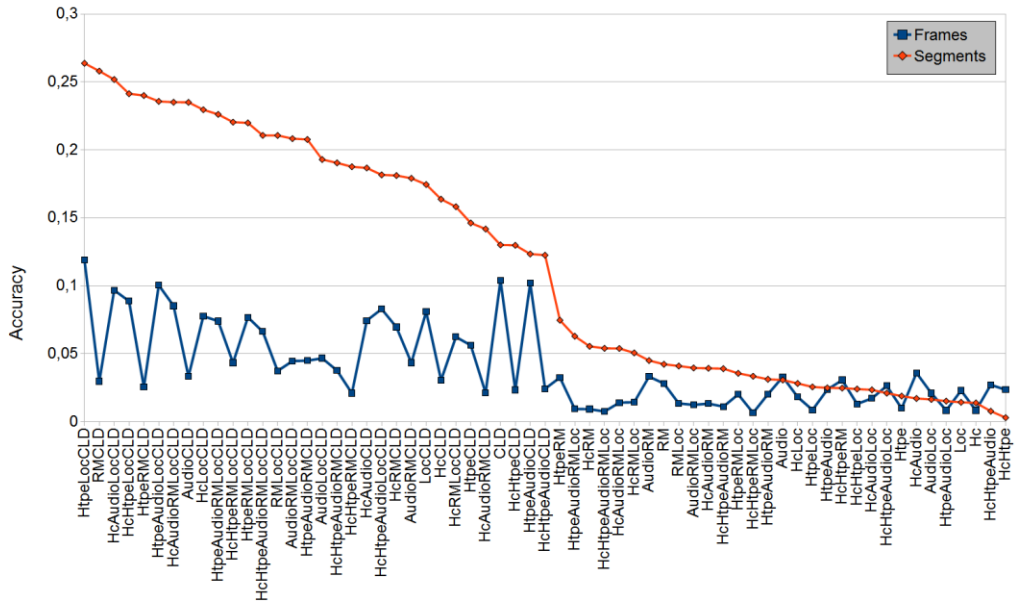


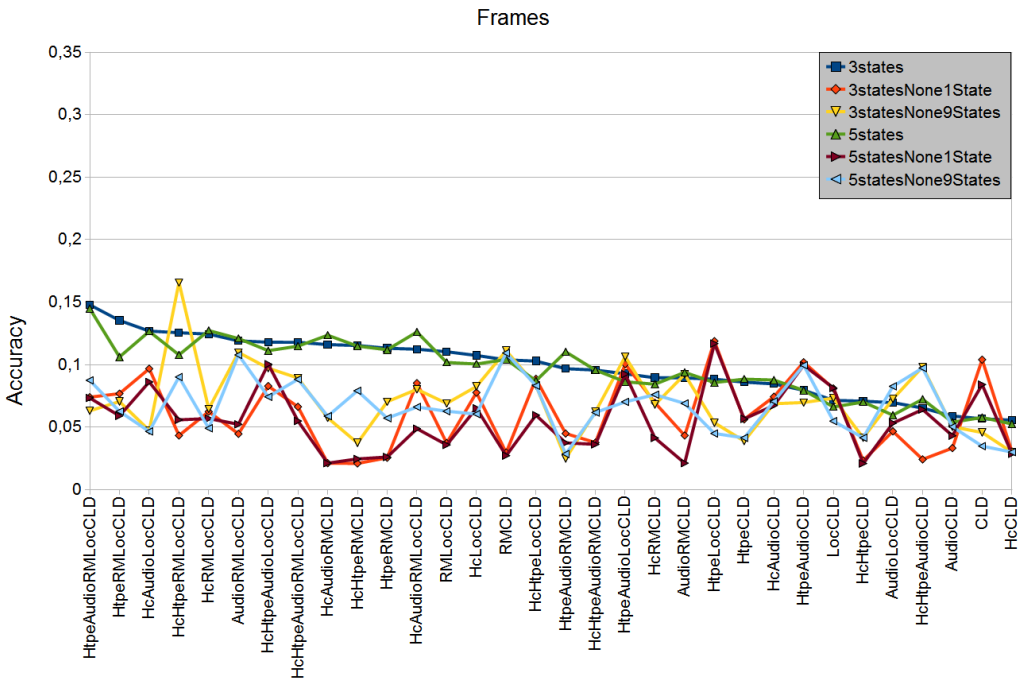
FIGURE 7: Global accuracy evaluation of recognition using frames (blue curve and square points) and segments (red curve and diamond points) over all the description spaces fusion tested (sorted by decreasing accuracy with respect to segments approach).

NB: For a better readability of the figure, results are shown for a selected configuration (3statesNone1State) of the HMMs but are similar for other configurations.

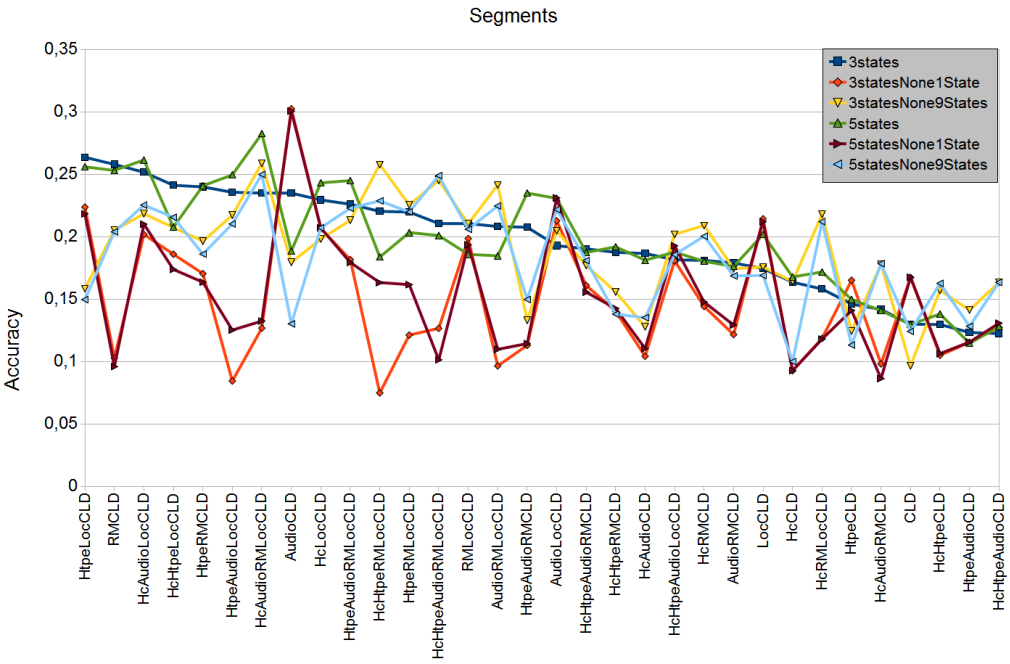
which is relevant as it gathers the evolution of the patient in successive places. Finally, the evaluation of recognition performance presented in Figure 7 shows that the results are better when the recognition process is run on segments. In this figure the results are sorted in decreasing order. The best results are always obtained with segments as observations and other results are similar using frames or segments.

e. Global evaluation of the description space

Figure 7 also shows which configurations are the most successful for the task. All the 33 best configurations are actually all the configurations including the CLD descriptor. We will therefore in the following only consider configurations which include CLD, and evaluate all possible combination of it with the other descriptors. The results are presented in Figure 8. Once again, a significant gain in performance can be observed when using segments instead of frames observations, the best accuracy for segments is 0.31 while the best accuracy for frames is 0.17. Here, the best global performance is obtained for the fusion $AudioCLD$ and good performances are also obtained for description spaces $RMCLD$, $H_cLocCLD$, $H_cAudioLocCLD$ and



(a) Results using frames as observations



(b) Results using segments as observations

FIGURE 8: Global accuracy evaluation of recognition using segments over CLD and all possible fusion with CLD description spaces using frames (a) or segments (b) as observations. The curves represent 6 different HMM configuration: 3 states (blue curve and square points), 3 states with "None" class being modeled with only one state (red curve with circle points), 3 states with "None" class being modeled with 9 states (yellow curve with triangle pointing down points), 5 states (green curve and triangle pointing up points), 5 states with "None" class being modeled with only one state (purple curve with triangle pointing right points), 5 states with "None" class being modeled with 9 states (pale blue curve with triangle pointing left points).

$H_c AudioRMLocCLD$. The $Audio$ descriptor seems efficient to capture some of the characteristic noises of activities which may occur for "Washing Dishes" or "Brushing Teeth" for example.

f. Global evaluation of the reject class model

We have also investigated the influence of modeling the reject class "None" in a different way than all the ADL classes. We have performed experiments when modeling this "None" class by a single state HMM or by a much more complex 9 states HMM. From the same Figure 8, we can see that performances with the reject class being modeled as a single state are clearly poorer and using 9 states does not significantly improve or degrade the performance. However, this configuration with 9 states for the "None" class shows good performances in high dimensionality description spaces built upon video segments.

g. Evaluation recognition of activities on the whole corpus

Finally, the second experiment is run on a corpus of 26 videos following the same leave-one-out cross validation scheme. The description space candidates are the 13 best configurations from the first experiment. The number of states at the bottom-level m is fixed to 3. In this experiment, the 23 different activities are "Food manual preparation", "Displacement free", "Hoovering", "Sweeping", "Cleaning", "Making a bed", "Using dustpan", "Throwing into the dustbin", "Cleaning dishes by hand", "Body hygiene", "Hygiene beauty", "Getting dressed", "Gardening", "Reading", "Watching TV", "Working on computer", "Making coffee", "Cooking", "Using washing machine", "Using microwave", "Taking medicines", "Using phone", "Making home visit".

An overview of the results in this larger scale experiments are given in Figure 9. The best median accuracy (0.42) is obtained for the complete description space $AudioH_c H_{tpe} RMLocCLD$. The gain of performance compared to the first experiment can be explained by the larger amount of training data. However, it is important to state the large variance of accuracy between 0.1 and 0.9. This shows the difficulty of our task.

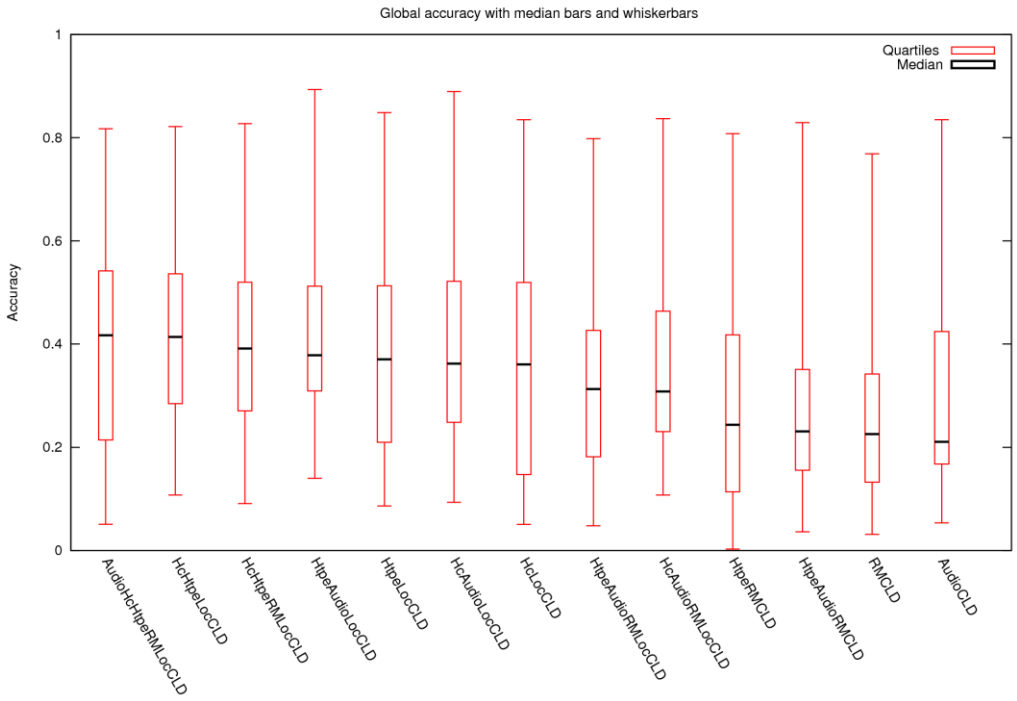


FIGURE 9: Global accuracy with regards to all description space candidates. The results are sorted by decreasing median accuracy.

h. Evaluation for specific activities

A more in depth analysis of the performances for activities recognition is given in Figure 10. We have selected a subset of four activities ("Hoovering", "Making a bed", "Reading" and "Working on computer") where the performances wary strongly when different description spaces are used. The performances are given in terms of accuracy, recall, precision and F-score, see Table 1. Note, the accuracy when computed by activity can easily be high since true negatives have positive impact on the performance. The results are sorted by decreasing precision as exchanges with the doctors have led to the conclusion that it was better to have less but more accurate detections, which is exactly what good precision metric values represent.

For the activity "Hoovering" the $Audio$ is essential as the 7 best performances contains the $Audio$ descriptor, see Figure 10a. This can be clearly linked with the fact that one of the coefficients of the $Audio$ descriptor corresponds to the detection of a "hoover" sound. The best trade-off between recall and precision, i.e. the best F-score, is obtained for the description space $H_{tpe}AudioRMCLD$ which contains global and local instantaneous motion descriptors in addition to $Audio$

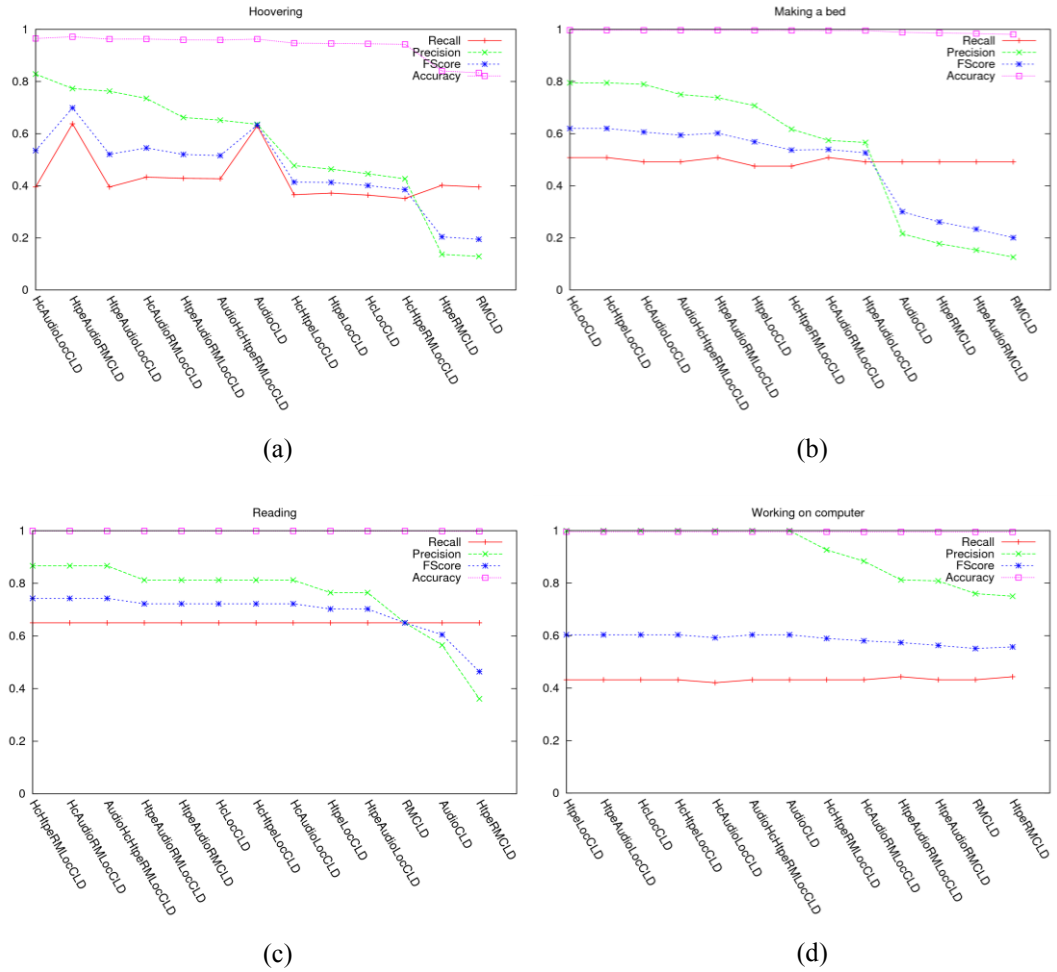


FIGURE 10: Performances according to the four metrics: recall (red curves, "+" points), precision (green curves, "x" points), F-score (blue curves, "*"points) and accuracy (pink curves, square points). The results are sorted by decreasing precision. a) "Hoovering", b) "Making a bed", c) "Reading" and d) "Working on computer".

and CLD. The complete description space $AudioH_cH_{tpe}RMLocCLD$ is also a good trade-off between recall and precision.

The second activity studied is "Making a bed", the results are presented in Figure 10b. For this activity the four top results contains the three descriptors H_c, Loc and CLD, the best results being obtained for the description space $H_cLocCLD$. This activity will always happen in the bedroom, thus the presence of the Loc descriptor in the best description spaces is not surprising. The H_c descriptor is helpful to characterize the fact that a patient moves around the bed while performing this activity.

The third activity "Reading" is more static and involves localized residual motion while turning the pages of the book being read. This is confirmed as the 5 top description spaces incorporate the RM descriptor, see Figure 10c. The static

component of the activity is captured by the motion descriptors, both H_c and H_{tpe} seem efficient for capturing this characteristic. The activity "Reading" being more likely in a limited set of locations, the Loc descriptor is also present in most of the best configurations.

Figure 10d depicts the results for the last activity we studied: "Working on computer". The best description spaces contain the Loc and CLD descriptors combined with at least one motion descriptor. Once again, the complete description space gives one of the best performances.

7. Conclusions and perspectives

Hence in this paper we tackled the problem of recognition of activities in videos acquired with cameras worn by patients for study of dementia disease. These videos are complex, with strong and irregular motion and lighting changes, the presence of activities of interest in the recordings is rare.

We solved the problem using HMM formalism. A Hierarchical two level HMM was proposed modeling both semantic activities from the taxonomy defined by medical doctors and non-semantic intermediate states. In order to define the observations of HMM we introduced a new concept of camera "viewpoint" and proposed a temporal segmentation of video thanks to analysis of apparent motion in it.

For the video frames and viewpoints we defined multimodal description spaces comprising motion features, static visual features and audio descriptors. The observations for training and recognition with HHMM were obtained by combination of proposed features in an early fusion manner with a reasonable dimension of the most complete space.

In the definition of the set of states of upper-level HMM we introduced a "None" state modeling a rejection class, which is necessary for description of our natural content namely transitions between activities and non-relevant actions of patients.

The proposed model was tested on the unique-in-the-world video corpus acquired with healthy volunteers and patients in "ecological" environment, i.e. at their homes. The taxonomy of activities was defined by medical researchers and the proposed framework was tested with cross-validation to recognize them. In these tests the optimal configurations of description space ensure performance which is nearly 8 times better than chance.

Detailed studies of different description spaces; HHMM states configurations and observation collection highlight that:

- (i) temporal segmentation into view-points improves the performances due to the filtering of the descriptors within meaningful unit of time;
- (ii) for the bottom-level HMM in our hierarchical model three states are sufficient to model the internal structure of each semantic activity;
- (iii) as far as the description space is concerned, the complete description space combining all available features performs the best in terms of median accuracy. Even if it is difficult to chose an absolute winner for description space composition, the best overall performances are ensured when static color descriptor of a scene content is present;
- (iv) the optimal description space varies per activity, each descriptor brining more information for a specific activity; the statement which often correlates with a "common sense": e.g. the hovering activity is the best recognized with audio features in description space.

The last statement makes us think that despite we settled an acceptable framework for this challenging application; the future is in incorporation of more semantic features in the description space. Events from more complete sets of wearable sensors can be used, such as accelerometers and other sensors, the combination of which with wearable video has not been explored yet for medical purposes according to our best knowledge. In video, we think about defining a concept flow related to the recognition of objects the person manipulates. This would leverage the fusion of information from video and other sources. Last, but not least, the acceptability of the service with wearable sensors by patients, let us hope that the proposed approach has a direct clinical perspective.

Acknowledgements. This work is partly supported by a grant from the ANR (Agence Nationale de la Recherche) with reference ANR-09-BLAN-0165-02, within the IMMED project.

8. References

[1] http://epp.eurostat.ec.europa.eu/

[2] H. Amieva et al., "Prodromal Alzheimer's disease: Successive emergence of the clinical symptoms", *Annals of Neurology*, volume 64, issue 5, pages 492–498, November 2008.

[3] N. Zouba, F. Bremond, A. Anfonso, M. Thonnat, E. Pascual, and O. Guerin, "Monitoring elderly activities at home", *Gerontechnology*, volume 9, issue 2, May, 2010.

[4] R. Hamid , S. Maddi, A. Johnson, A. Bobick, I. Essa and Ch. Isbell, "A novel sequence representation for unsupervised analysis of human activities", *Artificial Intelligence* (2009),, volume 173, pages 1221–1244.

[5] R. Megret , D. Szolgay , J. Benois-Pineau, , P. Joly, , J. Pinquier , J.-F. Dartigues and C. Helmer, "Wearable video monitoring of people with age dementia: Video indexing at the service of healthcare", *International Workshop on Content-Based Multimedia Indexing - CBMI* (2008), Conference Proceedings, art. no. 4564934, pages 101-108.

[6] S. Hodges et al., "Sensecam: a retrospective memory aid". *UBICOMP'2006*, pages 177–193, 2006.

[7] L. Piccardi, B. Noris, O. Barbey, A. Billard, G. Schiavone, F. Keller and C. von Hofsten. "Wearcam: A head wireless camera for monitoring gaze attention and for the diagnosis of developmental disorders in young children". *International Symposium on Robot & Human Interactive Communication*, pages 177-193, 2007.

[8] HTK Web-Site: http://htk.eng.cam.ac.uk

[9] L. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition". *Proceedings of the IEEE*, volume 77, number 2, pages 257-286, 1989.

[10] J.S Boreczky and L.D. Wilcox, "A Hidden Markov Model Framework for Video Segmentation Using Audio and Image Features". Proceedings of the IEEE International Conference on *Acoustics, Speech and Signal Processing*, volume 6, pages 3741-3744, 1998.

[11] E. Kijak, G. Gravier, P. Gros, L. Oisel and F. Bimbot, "HMM based structuring of tennis videos using visual and audio cues" *International Conference on Multimedia and Expo* – ICME (2003), volume 3, pages 309-312.

[12] S. P. Chatzis, D. I. Kosmopoulos, and T. A. Varvarigou, "Robust sequential data modeling using an outlier tolerant hidden markov model". *IEEE Transactions on Pattern Analysis and Machine Intelligence*(2009), 31 (9), 1657-1669.

[13] Y. Ivanov, and A. Bobick, "Recognition of visual activities and interactions by stochastic parsing". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22 (8), 852 -872, August 2000.

[14] S. Fine, Y. Singer and N. Tishby "The Hierarchical Hidden Marko Model: Analysis and Applications". *Machine learning* (1998), volume 32, pages 41-62.

[15] Nam T. Nguyen, Dinh Q. Phung, Svetha Venkatesh and Hung Bui, "Learning and Detecting Activities from Movement Trajectories Using the Hierarchical Hidden Markov Models" *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* – CVPR (2005), , volume 2, pages 955-960.

[16] M. Gales and J. Young "The Theory of Segmental Hidden Markov Models". *University of Cambridge*, Department of Engineering, 1993.

[17] M. Ostendorf, V. Digalakis and O. A. Kimball, "From HMMs to Segment Models: A Unified View of Stochastic Modeling for Speech Recognition", *IEEE Transactions on Speech and Audio Processing* (1995), volume 4, pages 360-378.

[18] M. Delakis, G. Gravier and P. Gros, “Audiovisual integration with Segment Models for tennis video parsing” *Computer Vision and Image Understanding* (2008), volume 111, number 2, pages 142-154.

[19] D. Surie, T. Pederson, F. Lagriffoul, L-E. Janlert and D. Sjölie “Activity Recognition using an Egocentric Perspective of Everyday Objects” *Ubiquitous Intelligence and Computing* (2007), Springer, pages 246-257. [20] J. Benois-Pineau and P. Kramer “Camera motion detection in the rough indexing paradigm”. *TREC Video*, 2005.

[21] Lazebnik, C. Schmid and J. Ponce, “Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories”. *IEEE Conference on Computer Vision and Pattern Recognition* - CVPR, (2006) volume 2, pages 2169-2178.

[22] F. Jurie and B. Triggs, “Creating efficient codebooks for visual recognition”, *Tenth IEEE International Conference on Computer Vision* - ICCV (2005), volume 1, pages 604-610.

[23] B. Scholkopf, A. Smola and K.R. Muller, “Nonlinear component analysis as a kernel eigenvalue problem”, *Neural computation* (1998), volume 10 (6), pages 1299-1319.

[24] H. Bay, T. Tuytelaars and Luc Van Gool, “SURF: speeded-up robust features”, *9th European Conference on Computer Vision* (2008), volume 110 (3), pages 346-359.

[25] Y. Bengio, O. Delalleau, N. Le Roux, J.-F. Paiement, P. Vincent and M. Ouimet, “Spectral dimensionality reduction”, *Feature Extraction* (2006), Foundations and Applications, Springer, pages 519-550.

[26] C. Burges, “A tutorial on support vector machines for pattern recognition”, *Data mining and knowledge discovery* (1998), volume 2 (2), pages 121-167.

[27] R. André-Obrecht, “A new statistical approach for automatic speech segmentation”. *IEEE Transactions on Audio, Speech and Signal Processing* (1988), volume 36(1), pages 29–40.

[28] J. Pinquier and R. André-Obrecht, “Audio indexing: Primary components retrieval - robust classification in audio documents”. *Multimedia Tools and Applications* (2006), volume 30(3), pages 313–330.

[29] G. Quenot, J. Benois-Pineau, B. Mansencal, E.Rossi, et al. ., “Rushes summarization by IRIM consortium : redundancy removal and multi-feature fusion”. *VS’08 (Trec Video Summarization)*, 2008.

[30] T. Sikora, B. Manjunath, and P. Salembier, “Introduction to MPEG-7: Multimedia content description interface”. 2002.

[31] Y. Gaëstel, C. Onifade-Fagbemi, F. Trophy, S. Karaman, J. Benois-Pineau, R. Mégret, J. Pinquier, R. André-Obrecht and J.-F. Dartigues “Autonomy at home and early diagnosis in Alzheimer Disease: usefulness of video indexing applied to clinical issues. The IMMED Project”. Alzheimer's Association International Conference on Alzheimer's Disease - AAICAD, 16-21 Juillet, 2011, France

[32] S. J. Young and S. Young, “The HTK hidden Markov model toolkit: Design and philosophy”. *Entropic Cambridge Research Laboratory, Ltd*, 1994.

[33] S. Young, G. Evermann et al., “The HTK book”. 1997.

[34] S. Karaman, J. Benois-Pineau, J.-F. Dartigues, Y. Gaëstel, R. Mégret and J. Pinquier, “Activities of Daily Living Indexing by Hierarchical HMM for Dementia Diagnostics”. Content-

Based Multimédia Indexing and retrieval - CBMI'2011, IEEE Workshop, 13-15 Juin, 2011, Madrid, Espagne.

[35] N. Harte, D. Lennon, and A. Kokaram, "On parsing visual sequences with the hidden Markov model". *EURASIP Journal on Image and Video Processing*, pages 1-13, 2009.

[36] P. Guyot, J. Pinquier and R. André-Obrecht. "Acoustic water flow detection in real life with an original feature: the spectral cover", submitted to IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'2012), 25-30 March 2012.