The IMMED Project: Wearable Video Monitoring of People with Age Dementia

Rémi Mégret[1], Vladislavs Dovgalecs[1], Hazem Wannous[1], Svebor Karaman[2], Jenny Benois-Pineau[2], Elie El Khoury[3], Julien Pinquier[3], Philippe Joly[3], Régine André-Obrecht[3], Yann Gaëstel[4], Jean-François Dartigues[4]

[1]IMS, UMR 5218 CNRS, University of Bordeaux
{remi.megret, vladislavs.dovgalecs, hazem.wannous}@ims-bordeaux.fr

[2]LaBRI, UMR 5800 CNRS, University of Bordeaux
{jenny.benois, svebor.karaman}@labri.fr

[3]IRIT, UMR 5505 CNRS, University Paul Sabatier
{khoury, pinquier, joly, obrecht}@irit.fr

[4]ISPED, INSERM U593, University of Bordeaux CHU de Bordeaux
yann.gaestel@chu-bordeaux.fr, jean-francois.dartigues@isped.u-bordeaux2.fr

ABSTRACT

In this paper, we describe a new application for multimedia indexing, using a system that monitors the instrumental activities of daily living to assess the cognitive decline caused by dementia. The system is composed of a wearable camera device designed to capture audio and video data of the instrumental activities of a patient, which is leveraged with multimedia indexing techniques in order to allow medical specialists to analyze several hour long observation shots efficiently.

Categories and Subject Descriptors

H.3.1 [Content Analysis and Indexing]: Indexing methods.
J.3 [Life and Medical Sciences]: Health.

General Terms

Algorithms, Measurement, Experimentation.

Keywords

Wearable camera, patient monitoring, audio and video indexing.

1. INTRODUCTION

This paper presents a new application of multimedia indexing methods in healthcare. The monitoring of various diseases by means of a (semi) automatic analysis of recorded video makes its first steps in the medical sector due to the maturity of video acquisition and storage technologies. Nevertheless, it is illusory to think that on-the-shelf solutions both in acquisition set –up and in video indexing are possible. In this work we propose a new application driven system for indexing of activities in wearable video recordings for the monitoring of the dementia disease.

With the ageing of population, dementia cases consequently increases, but the process leading to diagnosis does not allow identifying all people suffering from dementia in the general population. Hence, an early diagnosis of dementia is still a great challenge to prevent insecurity and health worsening in aged people living at home.

Generally, dementia is diagnosed in gathering clues of pathological changes in people life. To assess those changes, physicians use neurological examination, neuropsychological testing, and brain imaging technologies. Finally, diagnosis of possible dementia is asserted by comparing evidences, added to an autonomy decline. This autonomy decline is frequently assessed by Activities of Daily Living interviews [1] in which impairments are known to be related to cognitive decline caused by dementia [2].

Measuring autonomy decline is thus part of the diagnosis process, but it depends on subjective tools and on the patients ability to clearly analyze situations, both challenged by deny or anosognosia in patient, or his caregiver [3].

Our intention is to develop a new system to assess cognitive decline in the patients' life in the most comprehensive manner. By the way, we aim at highlighting early instrumental activities or cognitive processes impairments. This would help in people rehabilitation, by offering strategies to maintain their security and autonomy at home.

Therefore, we develop a wearable video device that allows a capture of the patients' daily activities (see section 2). Wearable image acquisition systems have been already used in a medical context: as a memory aid using the Sensecam device [5] coupled with automatic summarization techniques [4] or for helping in the diagnosis of autism in the Wearcam project [6]. Our system differs in the design of the acquisition device and the indexing tools developed, since observing at video rate the instrumental activities of a patient is a specific objective of our project.

By the use of this new device, we collect video data that can be analyzed to extract meaningful events occurring in patient's everyday life. The device had to follow a double constraint: being wearable in an ergonomic manner and focusing the recording to the inter-personal sphere (i.e. front of person, where instrumental activities and interaction within environment take place). To assist physician in the analysis of video data, we developed a video indexing assistance (see sections 3 and 4). This indexing process uses video and audio media to automatically guide the navigation in data flow straight up to the meaningful recorded situations.

2. GENERAL ARCHITECTURE

2.1 Processing flow principle

The general principle of the system is represented in figure 1. The activities of the patient are acquired as audio and video data using a wearable device (see section 2.2) under the supervision of a medical assistant. This data is stored on a SD-card which is transferred to the browsing station. A bootstrap annotation of the data is done on the beginning of the video in order to facilitate the automatic analysis of the rest of the video (see section 2.3.1). The video data is transferred through a secure connection to the computation center that indexes the whole video (see section 3) in order to detect the events of interest. The indexes are sent back to the browsing station to help a medical specialist visualize (see section 2.3.2) and analyze patients' behavior and spot meaningful impairments in their daily activities.

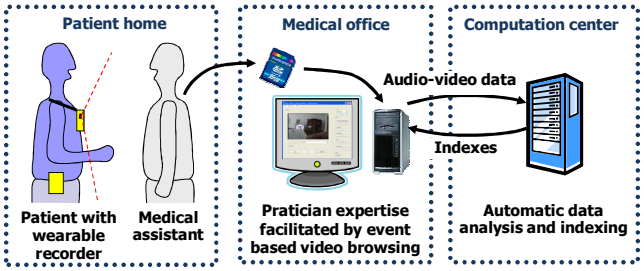

Figure 1. Global processing flow

2.2 Recording device

Regarding the positioning of the camera on the person, the tests showed that the positioning on the shoulder allows both good stability and a good overview [8]. This is a similar conclusion as in [7], and it was favoured over the position on the chest used by the Sensecam device [4] [5] or the position on the head used in the Wearcam project [6] in order to capture the full field of view of instrumental activities with low motion. For the current system, we have therefore designed a fabric jacket adapted to patient comfort, on which velcro straps are sewn at the shoulders to allow easily positioning the camera according to the patient corpulence.

Figure 2. Recording device (left), and camera angle of view when positioned on the shoulder (right)

The current device, represented in Figure 2, includes a HD GoPro Fisheye camera, which integrates the battery and the recorder. This high-resolution 1280x960 digital camera produces very good image quality of the instrumental actions of the patient, with a video rate (25 Hz) that is required to analyze them precisely.

The measured the vertical angle of view of the current device is 98.7°. This measure could be performed precisely using a calibration toolbox dedicated to wide angle cameras [9]. Such a wide angle is a requirement for our application, since it determines the ability to capture both the instrumental activities near the body and the general context of the action as illustrated in Figure 3.

Figure 3. Sample images showing the ability to capture instrumental activities even close to the patient's body

2.3 Video annotation and visualization

2.3.1 Video annotation

Video indexing methods require a learning phase before being able to automatically detect localization (section 4.1) and recognize activities (section 4.2). This data are very much patient-dependent, as home environments do not contain a large amount of invariants. Hence, a protocol has to be defined to annotate a minimum amount of data to allow the learning phase. The difficulty in here is that the annotation interface will be used by a medical practitioner who is not accustomed to advanced ICT. Hence the interface prototype developed comprises the activities and also localization elements. In the protocol, the medical assistant will annotate the first minutes of the video which will contain a tour of the patient's house. Therefore, the video annotation tool (Figure 4) should be easy to use and cross-platform.

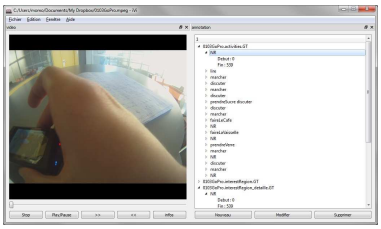

Figure 4. Event based annotation interface

2.3.2 Video visualization

The video footage at the patient's house provides a long sequence shot. The sequence can be of one hour up to a half-a-day duration, which is too long for a medical practitioner to watch entirely. Moreover, activities are of interest only when the autonomy of the patient may be evaluated. Hence the navigation is proposed via pre-defined set of activities, but also in a purely sequential manner to ensure all sequences of interest are viewed. The methods used for the automatic indexing of the activities will now be presented in section 3.

3. INTERPERSONAL INTERACTIONS

We are interested in defining the audiovisual environment in which the patient is interacting, as summarized in Figure 5. First, the audio stream is split into homogeneous segments where the audio classification is based on speaker diarization and sound event segmentation. Second, the video stream is divided into shots. On each shot, person detection is achieved thanks to face and clothing descriptors.

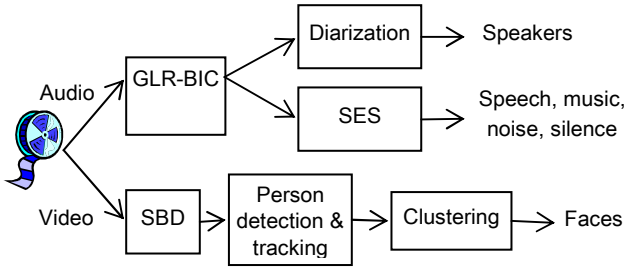

Figure 5. Audiovisual environmental indexing architecture.

3.1 Audio classification

The objective of speaker diarization is to segment the audio stream according to the current speaker. A first step partitions the audio content into segments (every segment must be as long as possible and the most acoustically homogeneous). A second step gathers segments corresponding to the same speaker. In the best case, a cluster corresponds to only one speaker. The first step is based on the GLR (Generalized Likelihood Ratio) and the BIC (Bayesian Information Criterion) methods. The parameters used are MFCC (Mel Frequency Cepstral Coefficients). This helps minimizing the tuning phase of the system and makes it robust to any type of audio content without any specific tuning. For more details, see [9].

Additionally, sound event segmentation (SES) is processed, that discriminate speech, music, noise and silence. This system [9] uses 4 parameters: 4 Hz energy modulation and entropy modulation for speech detection, number of segments per second and segment duration for music detection. Segments are resulting from a "Forward-Backward" divergence algorithm. A fusion with the previous system (speaker diarization) improves the results. Audio indexing results are good for the person wearing the device, but much more complicated for others (when the sound level becomes low).

3.2 Person detection

The audiovisual environment analysis is completed by visual person detection. First, we perform Shot Boundaries Detection (SBD), which relies on the same GLR-BIC algorithm as speaker diarization where the parameters used are the RGB color histograms. Then, face (using OpenCV [11]) and clothing are detected. Face tracking (using skin color) and costume tracking (in HSV space) improve the performance. The diarization phase mainly uses SIFT parameters [12] and 3D color histograms of the costume. Such an approach was shown to give very good results on heterogeneous TV corpus [13] and seems quite robust to the hard constraints of our project recordings.

4. ACTIVITY RECOGNITION

4.1 Location recognition

The location of the person in its environment is important information which provides contextual constraints for activity recognition. Our approach uses a Bag of Features approach [15] coupled with non-linear dimensionality reduction [17]. We extract SURF [16] descriptors and then quantize them to "visual words" using hierarchical k-means tree [14]. Each image is represented by a signature containing "visual word" frequencies weighted using a tf-idf scheme [15].

As the amount of supervised annotation to be expected from the medical practitioner is low, we adopted a semi-supervised classification approach by applying a dimensionality reduction step to the signatures using Laplacian Eigenmaps [18] in a transductive setup, which was shown to improve the classification performances [17]. Location is estimated by a simple nearest-neighbor classifier on the reduced signatures.

4.2 Activities indexing

4.2.1 Temporal Segmentation

The first step towards indexing the activities is to first segment the video into coherent segments. Since the camera captures continuously, the video corresponds to one long shot sequence. The segmentation in shots usually defined in video indexing [25] can not be used. Since the camera is worn on the shoulder, close to the patient's view, we propose a temporal segmentation based on the coherence of the viewpoint location, thanks to motion analysis. The camera motion in image plane is robustly estimated according to the complete 1st order affine model [22]. Then the trajectories of each corner of video frames are computed. When their distances from the initial position are greater than a predefined threshold the current "viewpoint" segment ends and a new one starts at the next frame. This set of segments forms the initial temporal partition of the recorded video to be labelled.

4.2.2 Description space and activities modeling

For each segment from the initial video partition we compute content descriptors. At the present state of the project we consider three kinds of global descriptors: colour-based, motion-based and location-based. The colour descriptor is the MPEG7 Colour Layout Descriptor [23] of a key-frame chosen as the temporal centre of the segment. The motion descriptors are "instant motion" and "historic of motion" (averaged). Both of them are computed on the basis of global camera motion. The location descriptor is the result of the localization process (section 4.1) and hence is more semantic than the first two descriptors. The descriptors are merged in an early fusion way to create a feature vector for each segment. A more detailed presentation can be found in [1]. Furthermore, the audio classification results (section 3.1) will shortly complete the feature space as the second "semantic" component of feature vector. The integration of the person detection (section 3.2) features is also promising.

The activities are modelled by a two level Hidden Markov Model (HMM) where the top level, called Activity HMM, contains states corresponding to the semantic activities such as "making coffee" or "cooking". The bottom level, called Elementary HMM, contains *m* non semantic states which are sub-states of an Activity HMM. This structure is motivated by the complexity of the daily activities to recognize. Each state of the Elementary HMM is modelling the descriptors feature space using Gaussian Mixtures Models (GMM). The parameters of the GMM and the state transition matrix of the Elementary HMM are learned using annotated data through the Baum-Welsh algorithm. The HMM are built using the HTK Library [27].

First results of activities recognition presented in [1] are promising. The new descriptors face and person detection and audio classification should enhance the recognition performance.

5. CONCLUSION

In this paper, we have presented the application of multimedia indexing to the monitoring of instrumental daily activities for the diagnosis of dementia. The principle of the system has been presented, and we have shown audio and video indexing approaches paving the way towards the efficient browsing of the video by medical practitioners. The study of the impact of the proposed system to the clinical diagnostic of dementia disease by medical experts is planned in the future of this work. The feedback of medical researchers we have today encourages us to go further; first of all to improve the automatic analysis, and second to establish larger scale experiments on already existing cohort of patients.

6. ACKNOWLEDGMENTS

This work is supported by a grant from Agence Nationale de la Recherche with reference ANR-09-BLAN-0165-02.
Project web site : `http://immed.labri.fr`
We also thank Christian Faurens for video acquisition and editing, and ENSEIRB-Matméca/IPB and SCRIME/Labri for their support.

7. REFERENCES

[1] American Psychiatric Association. Diagnosis and statistical manual of Mental Disorders. DSM III-R. Washington DC: Amer Psychiatr Ass, 1987.

[2] J.-F. Dartigues, “Methodological problems in clinical and epidemiological research on ageing”, *Revue d’épidémiologie et de santé publique*, 53(3): 243-9, 2005.

[3] K. Perez, C. Helmer, H.Amieva, J.-M. Orgogozo, I. Rouch, J.-F. Dartigues and P. Barberger-Gateau, “Natural history of decline in instrumental activities of daily living performance over the 10 years preceding the clinical diagnosis of dementia: a prospective population-based study”. *Journal American Geriatrics Society*, vol. 56, n°1, pp.37-44, 2008.

[4] S. Hodges, L. Williams, E. Berry, S. Izadi, J. Srinivasan, A. Butler, G. Smyth, N. Kapur and K. Wood, "SenseCam: a Retrospective Memory Aid". *UBICOMP*, pp. 177-193, 2006.

[5] E. Berry, N. Kapur, L. Willliams, S. Hodges, P. Watson, G. Smyth, J. Srinivasan, R. Smith, B. Wilson and K. Wood. "The Use of a Wearable Camera, SenseCam, as a Pictorial Diary to Improve Autobiographical Memory in a Patient with Limbic Encephalitis". *Neuropsychological Rehabilitation*, pp. 582-601, 2007.

[6] L. Picardi, B. Norris, O. Barbey, A. Billard, G. Schiavone, F. Keller and C. von Hofsten. "WearCam: A Head Mounted Wireless Camera for Monitoring Gaze Attention and for the Diagnosis of Developmental Disorders in Young Children". *International Symposium on Robot & Human Interactive Communication*, 2007.

[7] W. W. Mayol, B. J. Tordoff, and D. W. Murray. “Designing a miniature wearable visual robot.” *ICRA*, Vol. 4, pp. 3725–3730, 2002.

[8] R. Mégret, D. Szolgay, J.Benois-Pineau, Ph. Joly, J. Pinquier, J.-F. Dartigues and C. Helmer “Wearable video monitoring of people with age dementia: Video indexing at the service of healthcare” *CBMI*, pp.101-108, 2008.

[9] D. Scaramuzza, A. Martinelli and R. Siegwart (2006). "A Toolbox for Easy Calibrating Omnidirectional Cameras", *IROS*, Beijing China, 2006.

[10] E. El Khoury, C. Senac and J. Pinquier. “Improved Speaker Diarization System for Meetings”. *ICASSP*, Taipei, Taiwan, pp. 4241-4244, April 2009.

[11] P. Viola and M. Jones. “Robust real-time face detection”. *IJCV,* 57:137-154, 2004.

[12] D. G. Lowe. “Distinctive image features from scale-invariant keypoints”. *IJCV*, 60:91-110, 2004.

[13] E. El Khoury, C. Senac, P. Joly. “Face-and-Clothing Based People Clustering in Video Content”. *ACM Multimedia Information Retrieval*, Philadelphia, pp. 295-304, 2010.

[14] D. Nister and H. Stewenius, “Scalable Recognition with a Vocabulary Tree”. *CVPR*, 2006.

[15] J. Sivic and A. Zisserman, “Video Google: A text retrieval approach to object matching in videos”. *ICCV*, 2003.

[16] H. Bay, A. Ess, E. Tuytelaars and Luc Van Gool, “SURF: Speeded Up Robust Features”. *CVIU*, 110(3):346-359, 2008.

[17] V. Dovgalecs, R. Mégret, H. Wannous and Y. Berthoumieu. "Semi-Supervised Learning for Location Recognition from Wearable Video". *CBMI*, 2010.

[18] M. Belkin and P. Niyogi, “Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering”, *NIPS*, Vancouver, British Columbia, Canada, 2002.

[19] E. Dumont and B. Merialdo, “Rushes video parsing using video sequence alignment” *CBMI*, pp. 44-49, 2009.

[20] J. Borezcky, L. Rowe, "Comparison of Video Shot Boundary Detection Techniques". *Journal of Electronic Imaging*, pp. 170-179, 1996.

[21] E. Kijak, P. Gros and L. Oisel, "Hierarchical Structure Analysis of Sport Videos Using HMMS". *ICIP,* 2003.

[22] J. Benois-Pineau and P. Kramer, "Camera Motion Detection in the Indexation Primaire Paradigm". *TREC Video,* 2005.

[23] T. Sikora, B.S. Manjunath and P. Salembier, "Introduction to MPEG-7". *MCDI,* 2002.

[24] G. Abdollahian, Z. Pizlo and E. J. Delp “A study on the effect of camera motion on human visual attention”. *ICIP*, pp. 693-696, 2008.

[25] W. Dupuy, J. Benois-Pineau and D. Barba “Recovering of Visual Scenarios in Movies by Motion Analysis and Grouping Spatio-temporal Colour Signatures of Video Shots”, *EUSFLAT*, pp. 385-390, 2001.

[26] M. Delakis, G. Gravier and P. Gros “Audiovisual integration with Segment Models for tennis video parsing”. *CVIU*, vol. 111, pp. 142-154, 2008.

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

[28] S. Karaman, J. Benois-Pineau, R. Mégret, V. Dovgalecs, J.-F. Dartigues and Y. Gaëstel, “Human Daily Activities Indexing in Videos from Wearable Cameras for Monitoring of Patients with Dementia Diseases”, *ICPR,* 2010.