Title: Video monitoring for activities of daily living recognition

Authors: Konstantinos Avgerinakis, Alexia Briassouli, Ioannis Kompatsiaris
Affiliation: Centre for Research and Technology Hellas, Surrey university

Abstract

A novel algorithm for helping patients with dementia, based on computer vision and machine learning technologies, is presented. Static and wearable cameras are used in order to record the activities that an elder performs throughout day. The goal of this task is to recognize daily activities of the patients with dementia in order to develop behavioural profile and be able to track the progress of their condition and detect potential deteriorations

Introduction

The worldwide increase in life expectancy (Fig1) entails age-related health issues, multiplying healthcare costs every year, among which dementia is prominent, with a new case every four seconds. Although currently more common in high-income countries (Fig 2), dementia is expected to increase significantly in developing countries, which are projected (http://www.alz.co.uk/research/statistics) to account for 71% of cases by 2050 (Fig. 3). Technologies that monitor activities of daily living (ADL) can allow a person with dementia to remain independent, reducing the burden on family/friends and decreasing healthcare costs. They also offer an increased sense of safety, since emergencies can be detected and appropriate feedback will be provided to assist the person with dementia in daily life and cognition issues.

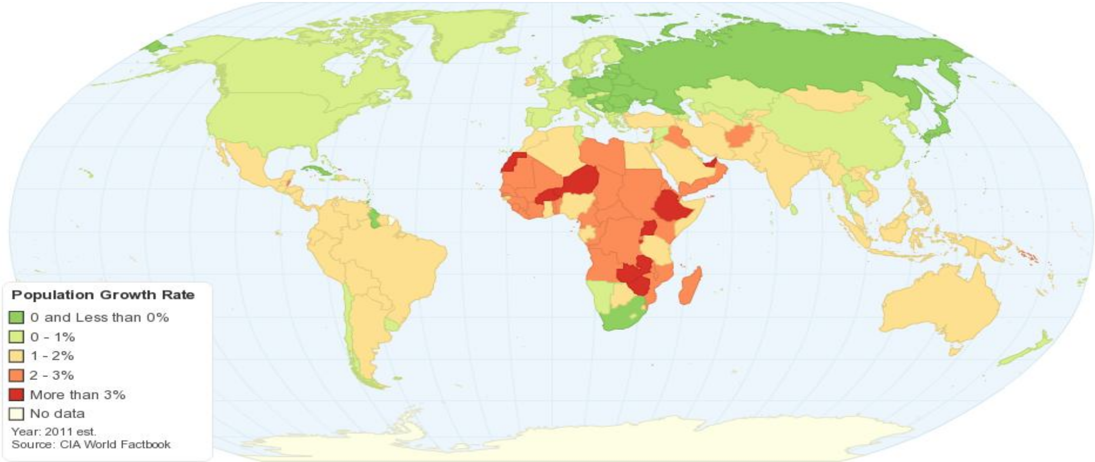
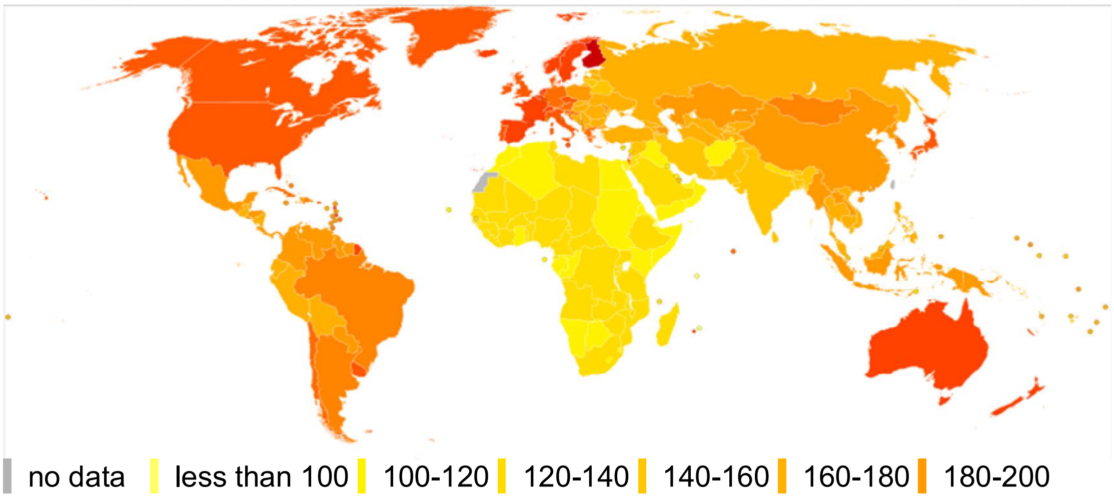


Fig1. Worldwide population growth rate.



🟧 200-220 🟧 220-240 🟧 240-260 🟥 260-280 🟥 280-300 🟥 more than 300

Fig2. Alzheimer’s and other dementia diseases per 100.000 inhabitants in 2002.

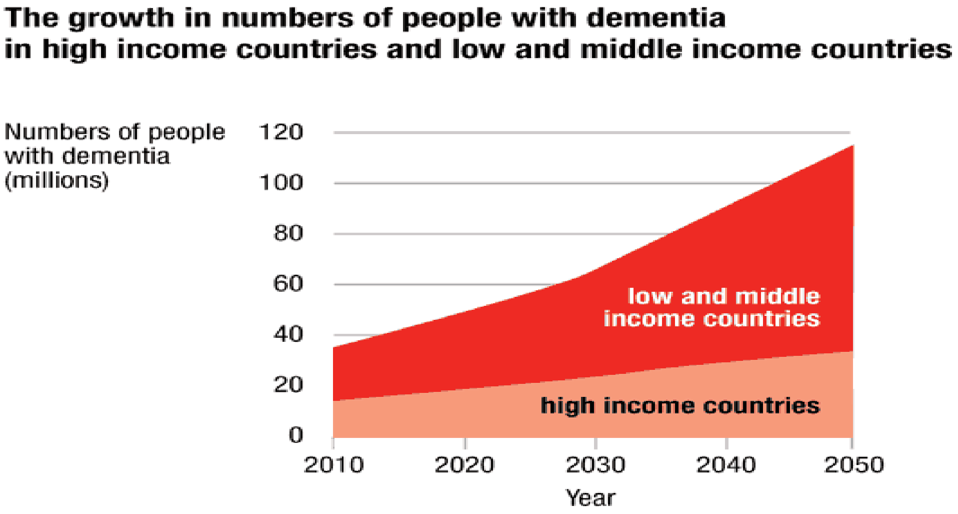


Fig 3. Increase in cases of dementia in high, middle and low income countries.

We propose a solution for remote monitoring that provides alarms when unusual events occur or the person cannot carry out usual daily activities. Recognition of activities of daily living has taken place in [1],[2],[3]. Our innovation lies in the fact that we successfully combine machine learning techniques (SVMs) with computer vision algorithms for extracting appearance/motion features and optimized trajectories via statistical signal change detection, to recognize daily activities. More specifically, the innovations that we introduce are:

1. A new statistical model, based on kurtosis metric, for separating static from moving pixels, so that we can reduce the computational cost of the action representation procedure.
2. A novel statistical model, based on CUSUM, for selecting optimal and meaningful lengths for the construction of each action trajectory vector.
3. The use of a hierarchical k-means tree, based on which a vocabulary tree could be built and will be used for defining the set of actions that performed in videos. Additionally, we propose the use of an inverted index file which will be used to describe fast and discriminatively the actions that occur in each video.

Early systems for real-time video monitoring are already in the market, increasing the independence of people with dementia and reducing related healthcare costs. People can continue to live in their homes unassisted, nursing homes can be decongested, friends and relatives can resume work. The cost of such a system is lower than providing help in the person’s home and will further decrease as technologies improve. The independence it offers will also increase as the accuracy of the monitoring and resulting feedback improve.

Methods

The concept is based on unobtrusive monitoring of a person with dementia, the only requirement being to place cameras at home. The caretaker can continuously monitor them in real time and is automatically notified of abnormalities. Furthermore, monthly reports from the automatic video monitoring are provided to doctors and an assessment-diagnosis can be compiled depicting the progression of each patient's condition.

Motion statistics and appearance features in conjunction with fast machine learning techniques are combined to deal with realistic scenarios. We extract dense trajectories for foreground subjects and appearance and motion descriptors represent actions after statistical change detection-based temporal segmentation. A vocabulary based on these features is used to cluster actions and train an SVM. This database of actions can be stored and used for testing new sequences in real-time.

- Trajectory estimation from dense optical flow

For separating static from moving pixels, motion estimations (optical flow) are analysed statistically using Kurtosis metric. When there is no real motion, non-zero motion values are induced by noise, corresponding to hypothesis H0, and are therefore assumed to follow a Gaussian distribution, while real motion introduce deviations from Gaussianity (hypothesis H1)

$$\mathrm{H}_0 \ : \ \mathrm{u}_\mathrm{k}^0(\mathrm{r}) = \mathrm{z}_\mathrm{k}(\mathrm{r})$$

$$\mathrm{H}_1 \ : \ \mathrm{u}_\mathrm{k}^1(\mathrm{r}) = \mathrm{u}_\mathrm{k}(\mathrm{r}) + \mathrm{z}_\mathrm{k}(\mathrm{r})$$

Where $u_k(r)$, $z_k(r)$ denotes true and noisy motion values respectively. The kurtosis G2 of Gaussian data is equal to zero and is used to detect whether motion is caused by noise or by changes in optical flow. The unbiased estimator of for kurtosis G2 is given by:

$$\mathrm{G}_2[\mathrm{y}] = \frac{3}{\mathrm{W(W-1)}} \sum_{i=1}^{W} (\mathrm{u}_\mathrm{i}(\mathrm{r})^4) - \frac{\mathrm{W}+2}{\mathrm{W(W-1)}} (\sum_{i=1}^{W} (\mathrm{u}_\mathrm{i}(\mathrm{r})^2)^2$$

Where W is a manual chosen temporal window from which motion values are obtained. Kurtosis values are significantly higher in regions of pixels whose motion changes.

- Sequential change detection on trajectories for temporal segmentation

Sequential change detection, namely Cumulative Sum (CUSUM) method is applied to HOF descriptors, so that we can obtain optimal trajectories. HOFs of the first w0 frames H_0 = f{h1,h2, …,hw0} are considered to follow an initial distribution f0, approximated as a multi-variate Gaussian, with mean and covariance matrix respectively given by:

$$\mu_0 = \sum_{i=1}^{W_0} \mathrm{h}_\mathrm{i}, \mathrm{C}_0(\mathrm{i, j}) = \mathrm{E}[(\mathrm{h}_\mathrm{i} - \mu_0)^\mathrm{T}(\mathrm{h}_\mathrm{i} - \mu_0)] = \frac{1}{\mathrm{W}_0} \sum_{i=1}^{W_0} (\mathrm{h}_\mathrm{i} - \mu_0)^\mathrm{T}(\mathrm{h}_\mathrm{i} - \mu_0)$$

And is compared at each frame k, with the “current”, estimated from the w_0 frames before, and including, the current one, i.e. frames k-w0+1 to k. Corresponding pdfs are given by:

$$\mathrm{f}_{\mathrm{pdf}=0,1}(\mathrm{h}_\mathrm{i}) = \frac{1}{(2\pi)^{\mathrm{N}/2} \, | C_{pdf=0,1} |^{1/2}} \exp(-\frac{1}{2}(\mathrm{h}_\mathrm{i} - \mu_{\mathrm{pdf}=0,1})^\mathrm{T} \mathrm{C}^{-1}_{\mathrm{pdf}=0,1}(\mathrm{h}_\mathrm{i} - \mu_{\mathrm{pdf}=0,1}))$$

Where pdf=0,k=1 denote the initial and current pdf correspondingly. $\mu_1 = \frac{1}{W_0} \sum_{i=k-W_0+1}^{k} h_i$ and $C_1(i, j) = 0$ for i ≠ j, $C1(i, i) = E[(hi-\mu_1)^T(hi–\mu_1)] = (\sigma_{i,1})^2$. In order to determine whether a change in motion occurs, the log-likelihood ratio can be used as a test statistic Tk, to be incorporated into the CUSUM test, so we have: $T_k = \log(\frac{f_1(h_k)}{f_0(h_k)})$. The CUSUM test is given by the iterative form: Sk = max(0,Sk-1+Tk), S0 = 0. For Gaussian data under each hypothesis H0 and H1, the test statistic, i.e. the log-likelihood ratio, becomes:

$$\mathrm{Tk} = \frac{1}{2} \ln(\frac{|\mathrm{C}_0|}{|\mathrm{C}_1|}) + \frac{1}{2}((\mathrm{h}_\mathrm{k} - \mu_0)^\mathrm{T} \mathrm{C}_0^{-1}(\mathrm{h}_\mathrm{k} - \mu_0)) - (\mathrm{h}_\mathrm{k} - \mu_1)^\mathrm{T} \mathrm{C}_0^{-1}(\mathrm{h}_\mathrm{k} - \mu_1)$$

with diagonal covariance given by: Ci = diag$[(\sigma_{i,1})^2,(\sigma_{i,2})^2,\ldots,(\sigma_{i,N})^2]$,i = {0, 1}, where $(\sigma_{i,k})^2=E[(hi-\mu_1)^T(hi–\mu_1)]$, k=[1,2,…,N]. The inverse of each diagonal matrix is given by: C_i^{-1}= diag$[1/(\sigma_{i,1})^2 , 1/(\sigma_{i,2})^2,\ldots, 1/(\sigma_{i,N})^2]$, and the determinant of each diagonal matrix is given by $|C_i|=\prod_{j=1}^{N}\sigma_{i,j}^2$, i={0,1}. By plugging in Tk equation into S_k at each frame, we get a value for the test statistic which significantly increases when there is a change in our data. This leads to the temporal segmentation of the extracted trajectories based on actual changes in motion, rather than their segmentation using a manually selected constant threshold.

Results

Evaluation uses a standard benchmark dataset[1] (Figure. 4) and results are analysed on Table 1 and Figure 5. We can observe the boost that hierarchical k-means introduce to the recognition system as the size of the vocabulary words increases. Furthermore, HOGHOF action representation seems to perform quite well when the vocabulary is large and the trajectories length have been set manually while on the other hand HOGHOF with optimal trajectories (ChDetHOF) perform better when the vocabularies are smaller (Figure 5). Activities that were confused were either too similar to each other, such as answering phone with dialling phone, or were induced by small motions, like use silverware and peel banana (Table 1). Detected activities and patterns can be fed back to the caretaker and combined with other monitoring devices to draw conclusions on the person's condition.

The most important issue is interference in the user's personal life and privacy: Users are volunteers fully informed of the monitoring. The Charter of Fundamental Rights of the EU [4] will be respected and data will be stored and transmitted complying with the provisions of EU Directive 95/46/EC [5]. Private locations and faces are not recorded to ensure ethical regulations of individual countries are respected.

ACKNOWLEDGEMENT

This work was funded by the European Commission under the 7th Framework Program (FP7 2007-2013), grant agreement 288199 Dem@Care.

[1] http://www.cs.rochester.edu/~rmessing/uradl/

Fig 4. From top to bottom and left to right we have the activities: answer phone1, answer phone2, dial phone, chop banana, eat banana, drink water.

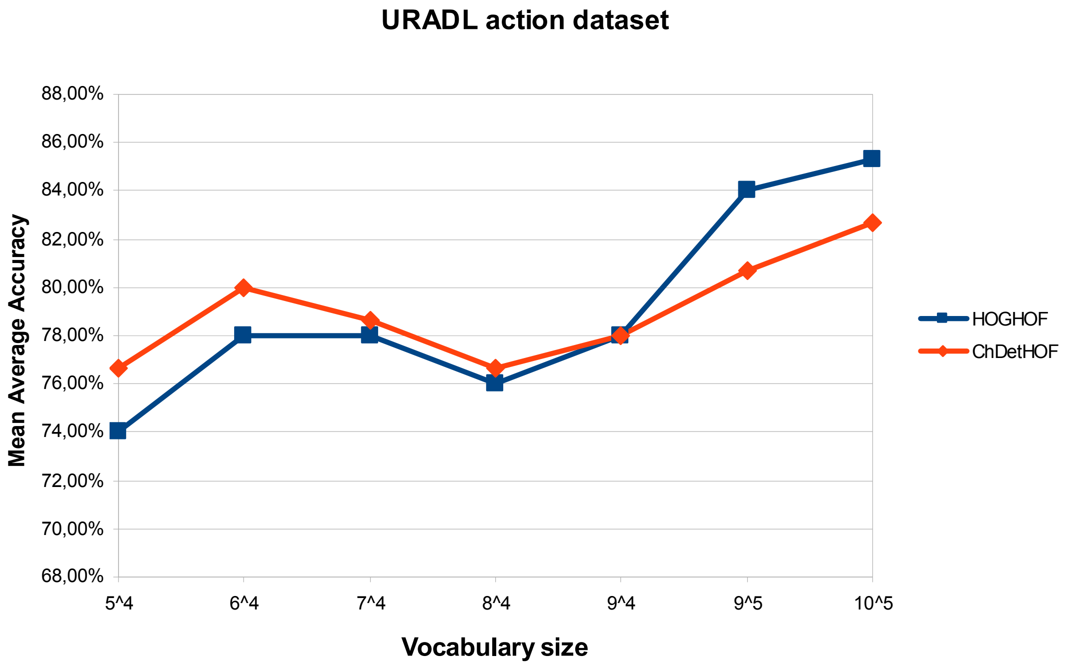


Figure 5. Action recognition results for each trajectory implementation when using different vocabulary sizes.

			HOGHOF ChDetHOF							
	AP	CB	DP	DW	EB	ES	LiP	PB	US	WoW
AP	**0,67**	0	0,27	0	0,07	0	0	0	0	0
CB	0	**0,93**	0	0	0	0	0	0,07	0	0
DP	0,27	0	**0,73**	0	0	0	0	0	0	0
DW	0	0	0	**1**	0	0	0	0	0	0
EB	0	0	0,07	0	**0,8**	0,13	0	0	0	0
ES	0	0	0	0	0	**1**	0	0	0	0
LiP	0	0	0	0	0	0	**1**	0	0	0
PB	0	0,07	0	0,07	0	0,13	0	**0,73**	0	0
US	0	0	0	0,13	0	0	0	0,2	**0,67**	0
WoW	0	0	0	0	0	0	0	0	0	**1**

overall accuracy :	**0,853**

Table1. Best recognition rates for 10 activities: AP(Answer Phone), CB(Chop Banana), DP(Dial Phone), DW(Drink Water), EB(Eat Banana), ES(Eat Snack), LiP(Lookup in Phonebook), PB(Peel Banana), US(Use Silverware), WoW(Write on Whiteboard).

[1] Messing, R., Pal, C. & Kautz, H., 2009 "Activity recognition using the velocity histories of tracked keypoints" *ICCV 2009.*

[2] N. Zouba, F. Bremond and M. Thonnat. Monitoring Activities of Daily Living (ADLs) of Elderly Based on 3D Key Human Postures. In the *4th International Cognitive Vision Workshop, ICVW 2008* - Santorini, Greece, May 12-15, 2008.

[3] Human Daily Activities Indexing in Videos from Wearable Cameras for Monitoring of Patients with Dementia Diseases, ICPR 2010, Turquie (2010), pp. 4113-4116

[4] http://www.europarl.europa.eu/charter/pdf/text_en.pdf

[5]http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:en:HTML