DemCare action dataset for evaluating dementia patients in a home-based environment

Avgerinakis Konstantinos[1,2], Kompatsiaris Ioannis[1]

Information Technologies Institute, CERTH, Thessaloniki, Greece[1],

Center for Vision, Speech and Signal Processing, University of Surrey, UK[2]

Abstract

Computer vision technologies and more specifically activity recognition can be considered one of the most helpful tools that computer science can provide to the society's disposal. Activity recognition deals with the visual analysis of video sequences and provides semantic information about the activities that may occur within them. In state-of-the-art literature, activity recognition deals with problems that vary from (a) video retrieval topics, which concentrate to the extraction of visual information concerning activities that exist within movies or youtube video samples, to (b) activity of daily living (ADL) topics which focus to the recognition of activities that may occur within a home or kitchen based environment. Although the great range of activities that current action datasets include, we have not yet encountered the implementation of any realistic scenario which deal with the real life problems, such as dementia and related diseases. Considering the above reasons and trying to encourage future studies on dementia disease, we propose DemCare action datasets which record a spate of human patients to perform a large set of daily activities in a home based environment.

1. Introduction

Dementia diseases tend to become one of the most usual health problems that are encountered in modern societies. Large amount of financial resources are spent every year in healthcare, so that dementia sufferers can be attended within nursing homes and hospitals. Recent technologies though provided the opportunity to these patients to remain on their own homes and be attended remotely without congesting healthcare institutes. These technologies usually use a spate of sensors for recording and accumulating patients' progress. However, their obstructive nature raises a great hindrance concerning their applicability in a realistic scenario. Most of them need to be worn to the patient in order to extract

useful information, such as accelerometers, gyroscopes, wearable camera or microphones, rendering them annoying and they are usually neglected and abandoned by the patient. On the other hand, activity recognition technology is based on visual analysis and can be easily passed unnoticed, creating a great benefit and renders it a very useful tool for remotely patients' attendance.

Our main interest on this work is to provide an activity recognition schema that will be able to provide accurate recognition results on a home based environment. Our algorithm should be able to analyze video samples that contain dementia patients' ADL recordings and provide useful results to the appropriate attendance physician. The paper is organized as follows: Section 2 elaborates on the existing related work on activity recognition and public available datasets. In Section 3 the two new activities of daily living (ADL) datasets, called henceforth DemCare1 and DemCare2, are presented. Section 4 presents the activity representation and recognition schema that we adopt in this work. Section 5 present experiment results in DemCare1 and DemCare2 datasets and Section 6 concludes the work with useful deductions.

2. Related work

Activity recognition has been evolved to one of the most active topics in computer vision within the last decade. Many action datasets can be spotted in the literature focusing on a large range of human activity aspects. **KTH** and **Weizman** action dataset [1, 2] are considered of the most early ones and deal with very simplistic actions, such as run, jog, walk, handclap and hand wave, with little variation within them, under very constrained indoor and outdoor environmental parameters. **IXMAS** and **HumanEva** [3, 4] followed and introduced a small set of simplistic activities with multiple cameras in a simplified room environment. Although their limited applicability in real life scenarios, the aforementioned datasets are considered of the most important indicators regarding the robustness of an activity recognition system and they can still be encountered in state-of-the-art works. More challenging action datasets are encountered in [5, 6] were actions are sampled from **Hollywood** movies and **youtube** video sequences in [7] where intra-class variations, moving camera and random viewpoint camera angle render them a very difficult problem to deal with. Nevertheless, they still deal with single actions and are mostly aimed at video retrieval purposes. Recently, more realistic and complex action datasets have been presented in the literature concerning activities that take place in kitchen based and home environments. **URADL** and **KIT** datasets, proposed in [8, 9] correspondingly, are considered of the most popular one ADL datasets and are used in many state-of-the-art works. In **CMU** action dataset in [10] the authors used several cameras in order to record a spate of human subjects preparing 5 different recipes. The most serious disadvantage of this recordings is that they only depict one action (i.e. cooking), limiting by this way its applicability to a very specific occasion. **TUM** action dataset, proposed in [11], also contains only one activity in a kitchen environment (i.e. set table) and is more focused on motion tracking purposes.

Regarding the technologies that exist in activity recognition literature, we can classify them **either** to holistic based approaches which represent activities as space-time shapes [2], motion history volumes [3] and trajectory vectors [8, 9] **or** to local based approaches, which represent actions as 3D appearance volumes and are extensions of image local patches to the temporal space, such as SIFT3D, SURF3D and HOG3D proposed in [13, 14, 15] Other local based approaches decode action based on motion information. Thus, they create motion histograms around spatio-temporal interest points such as in [5, 6, 12]. Local based approaches when combined with a Bag-of-Words (BoW) and SVM classification schema lead to best recognition rates in state-of-the-art literature.

3. DemCare ADL dataset

In order to turn activity recognition attention to patients with dementia problem, we set a room-kitchen environment for developing a complete set of activities that may occur in a home based environment. Recordings were held on Alzheimer's institute premises in Thessaloniki and were completed in two phases, leading into two activity datasets.

3.1 DemCare1 ADL dataset

In the first set of recordings, henceforth mentioned as DemCare1, we used two different types of video sensors, including a HD camera and a Kinect sensor. In Figure 1, we can see the room setup.

Figure 1. Left image depicts the HD and Kinect camera setup, while on the right one depicts the room that the activities took place.

In this first dataset 32 patients with Mild Cognitive Impairment (MCI) were called to perform a set of predefined activities. Activities were designed so that information concerning the patients' capabilities could be extracted. **Feeding** and **refreshing** capability was observed by an eating and drinking scenario, in a kitchen based environment. Eating scenario included the preparation of a meal and its consumption, while on the drinking scenario, a beverage was served in a glass and

later on was consumed. Both scenarios was followed by a cleaning up activity, so that it can be observed if the patient is capable to leave the table condition in a proper state. **Socializing** capability of the patient was checked by initializing two different scenarios. On the first one the human subject was called to use a phone to contact with another person. The scenario included the start phone-call action which detected when the patient picks up the telephone handset and the end phone-call action which detected when the patient hang ups the phone and terminates the conversation. A visiting activity was the second scenario that initialized for checking socializing capability. On this case, a visitor enters the room that the patient stays, has a handshake or hug with him and starts a conversation. Finally, patients' capability of allocating **recreational** time within their day was checked by a reading a paper activity. On this scenario a patient is called to sit in a sofa or chair, grab a book and read it. Activities description and further information are aggregated in Table 1.

3.2 DemCare2 ADL dataset

In the second set of recordings, henceforth mentioned as DemCare2, we used two static cameras for recording activities that occur in a room (i.e. HD camera and Kinect sensor). We also recorded data by using a wearable camera, by adjusting a GoPro camera in a proper designed vest. An accelerometer watch was finally worn to the patients for checking their stability. In Figure 2, we can see the room setup and the sensors that were used for recording purposes.

Figure 2. On the left picture a patient with dementia performing drinking action. Kinect sensor and wearable camera records her actions. On the top right image the room setup with the Kinect, HD and GoPro wearable camera. While on the bottom right are depicted the sensors from a closer view.

In this second dataset, 35 human subjects, including patients with dementia and healthy ones, were called to perform a set of activities similar to those in DemCare1, introducing a great anthropometric variance to the dataset. An extra activity was included in the experiment called: use closet and is classified to the recreation ability section. In this activity scenario, the human subject is called to open a closet, grab an object from inside and close it. The activities description and other information for each one separately are aggregated in Table 1.

Abilities	**Initials**		**Description**
Feed ability (Kitchen)	PS	Prepare Snack	The patient is called to grab a plate and a snack from the table and put it in front of him.
	ES	Eat Snack	The patient picks the snack that is placed in front of him and eats it. (bring it to his mouth)
Drink ability (Kitchen)	SB	Serve Beverage	The patient grabs a bottle of water or orange juice and pours it inside a glass. He brings the glass in front of him.
	DB	Drink Beverage	The patient drinks the liquid that his has served in his glass by bringing the glass to his mouth.
Cleanup ability (Kitchen)	CU	Clean Up	The patient cleanup the table in front of him, by discarding the glass and the plate to a bin.
Phone ability (Social)	SP	Start phone-call	The patient picks up the phone and dials a number, indicating that he initializes a phone-call.
	EP	End phone-call	The patient puts down the phone, indicating the termination of the phone-call.
Having visitor ability (Social)	ER	Enter room	An activity which indicates that a person opened the door and entered the room.
	HS	Handshake	The patient greets the visitor by having a handshake with him.
	TV	Talk to visitor	The patient talks to his visitor, by standing in front of him and making some gestures.
Recreation ability (Reading)	RP	Read paper	The patient sits to a sofa or chair and reads a book that is placed in a table next to him.
	UC	Use closet	The patient opens up a closet, pick a book and close its door. (exists only in DemCare2)

Table 1. The set of activities that are observed in DemCare1 and DemCare2 action datasets and their description.

4. Activity recognition

Activity recognition technologies consist of two basic parts. On the first part, action representation is performed, while on the second one, action recognition classifies video sequences to the appropriate action class.

For **action representation** we choose to follow a local based approach combined with a holistic one. For local approaches, it is prerequisite to sample interest points from the spatio-temporal space. Harris3D [16] is one of the most popular spatiotemporal interest point detectors and basically extends Harris corner detector to the temporal space for that purposes. The main disadvantage of this technique is that it provides very few and sparse interest points, leading to low discriminative representation power. Inspired from recent state-of-the-art work [12], we follow a dense sampling technique for collecting spatio-temporal interest points. A foreground/background subtraction algorithm, called Activity Area (AA), appropriate for separating static from moving pixels [17], is initially applied on consecutive image frames. We sample interest points within these regions (i.e. AA) by using a dense spatial grid and track them throughout time using a KLT tracker, producing a trajectory vector. Each region around sampled interest points is described by a HOG-HOF descriptor [5] providing appearance and motion information to our representation schema. A spatio-temporal volume is ultimately produced for each trajectory vector and concatenated with raw trajectory's coordinates in order to include global spatial information to our action descriptor.

For **action recognition** we use a K-means combined with Chi-Square schema. K-means produces K cluster centers from the training data and quantizes provided video sequences using a hard binning approach. Thus, a frequency histogram is produced for each video sequence. A Chi-Square kernel is later on produced by comparing the training histograms in a pairwise manner and fed to a SVM for producing a multi-class classifier. In our experiments, we use K=4000 cluster centers for partitioning our feature vector space. To limit complexity, cluster centers are clustered on a randomly selected subset of 100.000 feature vectors acquired from the training set. K-means is finally initialized 10 times in order to provide the most discriminative cluster centers.

5. Experiments

In the following two sub-chapters we elaborate on the activity recognition system evaluation based on the two ADL datasets that we propose in this work. HOGHOF descriptor, boosted with raw trajectory coordinates feature vector, is used for action representation. K-means is used for constructing a visual vocabulary of 4000 cluster centers, for quantizing action descriptors extracted from given video sequences. Chi-square distances among BoW visual histograms are used for creating an appropriate kernel and fed to an SVM classifier for recognition purposes. Several tests were conducted and confusion matrixes for each different experimental setup lead to meaningful conclusions.

5.1 DemCare1

In DemCare1 ADL dataset, 32 human subjects with mild cognitive impairment (MCI) were called to perform 11 different activities. For experimental evaluation of the activity recognition system, we split DemCare1 dataset in two different ways. On the first one split, we separated activity dataset in 20 train to 12 test video samples. While one the second split, we followed a leave-one-subject-out technique, leading to far better results than the former. Confusion matrixes for each splitting are depicted in Table 2 and Table 3. Activities bellow are encoded as: **CU**: clean up table, **DB**: drink beverage (i.e. water-orange juice), **EP**: end phone-call, **ER**: enter room, **ES**: eat snack, **HS**: handshake, **PS**: prepare snack, **RP**: read paper on the couch, **SB**: serve beverage, **SP**: start phone-call, **TV**: talk to visitor.

HOGHOF_Bruhn (+coords) Kmeans(4000) Chi-Square											
	CU	DB	EP	ER	ES	HS	PS	RP	SB	SP	TV
CU	70,0%		15,0%		5,0%				5,0%	5,0%	
DB		47,8%	13,0%		30,4%					8,7%	
EP			100,0%								
ER				100,0%							
ES		12,0%	28,0%	8,0%	48,0%			4,0%			
HS				9,1%		63,6%					27,3%
PS		17,6%		5,9%	5,9%		41,2%		29,4%		
RP								100,0%			
SB					8,3%		8,3%		83,3%		
SP			16,7%					8,3%		75,0%	
TV				8,3%							91,7%
AA	74,6%										

Table 2. Our activity recognition algorithm when using a 20 train to 12 test splitting on DemCare1 ADL dataset.

From Table 2, it is obvious that most classes can be distinguished easily except for those that occur in the kitchen environment, which are quite similar to each other. Thus, drink beverage (DB) is confused with eat snack (ES), eat snack (ES) is confused with drink beverage (DB) and end phonecall (EP), while prepare snack is confused with drink beverage (DB) and serve beverage actions. Another observation that has been observed is that in some occasions actions of the similar concept can be confused, but in a lower level then the former ones (i.e kitchen). For instance, handshake (HS) is confused with talk to visitor (TV) and start (SP) with end phone-call (EP).

			HOGHOF_Bruhn(+ coords) kmeans(4000) Chi-Square								
	CU	DB	EP	ER	ES	HS	PS	RP	SB	SP	TV
CU	**83,8%**		4,4%		7,4%		1,5%			2,9%	
DB		**78,4%**			17,6%				2,0%	2,0%	
EP			**82,8%**		6,3%					10,9%	
ER				**100,0%**							
ES		27,8%	3,9%		**68,3%**						
HS						**75,0%**					25,0%
PS	2,9%	5,7%			5,7%		**77,7%**		8,0%		
RP	3,1%							**93,8%**			3,1%
SB		1,5%					10,3%		**88,2%**		
SP		3,0%	3,0%		3,0%			3,0%		**87,9%**	
TV				3,2%	3,2%	3,2%		3,2%			**87,1%**
AA	0,8391										

Table 3. Our activity recognition algorithm when using a leave-one-subject-out splitting on DemCare1 ADL dataset.

When we use a leave-one-subject-out technique, it is obvious that our recognition rates are increased significantly (i.e. almost 9%). The activities are classified far better and no matter the large anthropometric variation, we can acquire realistic recognition results. On this split, we can observe a similar confusion to the previous split when we have kitchen located activities. Thus, drink beverage (DB) is confused with eat snack (ES) and vice-verca. Prepare snack (PS) is mixed up with serve beverage (SB) and vice verca. Handshake (HS) also is confused with talk to visitor (TV) activity.

5.2 DemCare2

In DemCare2 ADL datasets the human subject population included patients with dementia and healthy subjects. All the human subjects were called to perform a set of 12 activities of daily living. For experimental evaluation of the activity recognition system, we split DemCare2 dataset with a leave-one-subject-out technique. The only new included activity is **UC**: use closet. Table 4 depicts the corresponding confusion matrix.

		HOGHOF (+Coords) Kmeans(4000) Chi-Square										
	CU	DB	EP	ER	ES	HS	PS	RP	SB	SP	TV	UC
CU	**81,8%**				13,6%					4,5%		
DB	1,4%	36,1%			**52,1%**		2,1%		8,3%			
EP			**92,0%**		4,0%					4,0%		
ER				**96,0%**	4,0%							
ES	3,2%	14,6%			**80,2%**				2,0%			
HS					4,2%	**87,5%**					8,3%	
PS		6,3%			11,8%		66,7%		15,3%			
RP					4,0%			**80,0%**			16,0%	
SB					3,2%		12,0%	4,0%	**76,8%**		4,0%	
SP			12,0%							**88,0%**		
TV					12,5%	8,3%		8,3%			**70,8%**	
UC												**100,0%**
AvAcc:	79,7%											

Table 4. Our activity recognition algorithm when using a leave-one-subject-out splitting on DemCare2 dataset.

From Table 4 we observe again confusion in Kitchen environment activities. On the one hand, drink beverage (DB) is mixed up with eat snack (ES) and vice verca, while prepare snack (PS) cannot be distinguished from eat snack (ES) and serve beverage (SB). Despite the great variance that was introduced from anthropometrics and illumination variance, the general conclusion that most activities can be discriminated from others, producing good and realistic recognition results.

6. Conclusions

In this work two novel ADL datasets were introduced to activity recognition community, dealing with people who suffer from dementia or related diseases. The goal of this dataset is to turn activity recognition to more realistic scenarios such as this of monitoring incompetent people in a home environment. It is obvious that using a video based technology, such as activity recognition, can provide a great deal of implementing of an unobstructive and realistic tool. Our results in the experiment section prove our allegations, as we reach accurate and realistic recognition rates.

Acknowledgement

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

References

[1] Christian Schuldt, Ivan Laptev and Barbara Caputo, “Recognizing Human Actions: A Local SVM Approach”, (ICPR), 2004.

[2] Moshe Blank and Lena Gorelick and Eli Shechtman and Michal Irani and Ronen Basri, “Actions as Space-Time Shapes”, (ICCV), 2005.

[3] D. Weinland, R. Ronfard, and E. Boyer, “Free viewpoint action recognition using motion history volumes,” Computer Vision and Image Understanding (CVIU), 2006.

[4] L. Sigal and M. J. Black, “HumanEva: Synchronized Video and Motion Capture Dataset for Evaluation of Articulated Human Motion,” Tech. Rep., 2006.

[5] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, “Learning realistic human actions from movies,” in CVPR, 2008.

[6] M. Marszalek, I. Laptev, and C. Schmid, “Actions in context,” in CVPR, 2009.

[7] J. Liu, L. Jiebu, and M. Shah, “Recognizing realistic actions from videos in the wild,” in CVPR, 2009.

[8] R. Messing, C. Pal, and H. Kautz, “Activity recognition using the velocity histories of tracked keypoints,” ICCV, 2009.

[9] L. Rybok, S. Friedberger, U. D. Hanebeck, and R. Stiefelhagen ,”The KIT Robo-Kitchen Data set for the Evaluation of View-based Activity Recognition Systems”, in IEEE-RAS International Conference on Humanoid Robots, 2011.

[10] F.De la Torre, J.Hodgins, J.Montano, S.Valcarcel, J.Macey, “Guide to the Carnegie Mellon University Multimodal Activity Database,” Tech. Rep., 2009.

[11] M.Tenorth,J.Bandouch, and M. Beetz, “The TUM Kitchen Data Set of everyday manipulation activities for motion tracking and action recognition,” in ICCV, 2009.

[12] H.Wang, A.Kläser, C.Schmid, Liu Cheng-Lin, “Action Recognition by Dense Trajectories”, Computer Vision & Pattern Recognition(CVPR), pp.3169-3176, 2011.

[13] P. Scovanner, S. Ali, and M. Shah, “A 3-Dimensional SIFT Descriptor and its Application to Action Recognition”, ACM MM 2007.

[14] A. Klaser; M. Marszaek; C. Schmid, M. Everingham and C. Needham and R. Fraile, “A Spatio-Temporal Descriptor Based on 3D-Gradients”, (BMVC), 2008.

[15] G. Willems, T. Tuytelaars, L. Van Gool, “An efficient dense and scale-invariant spatio-temporal interest point detector”, (ECCV), 2008.

[16] I. Laptev, "On Space-Time Interest Points", IJCV,(2005),

[17] A. Briassouli, I. Kompatsiaris, "Robust Temporal Activity Templates Using Higher Order Statistics", on Image Processing, Vol.18, pp. 2756-2768, 2009.