Manchester Metropolitan University

e-space

Manchester Metropolitan University's Research Repository

Please cite the Published Version

Sultan, Saba, Javed, Ali, Irtaza, Aun, Dawood, Hassan, Dawood, Hussain and Bashir, Ali Kashif (2019) A hybrid egocentric video summarization method to improve the healthcare for Alzheimer patients. Journal of Ambient Intelligence and Humanized Computing, 10 (10). pp. 4197-4206. ISSN 1868-5137

DOI: https://doi.org/10.1007/s12652-019-01444-6

Publisher: Springer Science and Business Media LLC

Version: Accepted Version

Downloaded from: https://e-space.mmu.ac.uk/623848/

Additional Information: This is an Accepted Manuscript of an article which appeared in Journal of Ambient Intelligence and Humanized Computing, published by Springer. The final publication is available at Springer via http://dx.doi.org/10.1007/s12652-019-01444-6.

Enquiries:
If you have questions about this document, contact openresearch@mmu.ac.uk. Please include the URL of the record in e-space. If you believe that your, or a third party's rights have been compromised through this document please see our Take Down policy (available from https://www.mmu.ac.uk/library/using-the-library/policies-and-guidelines)

ORIGINAL RESEARCH

A hybrid egocentric video summarization method to improve the healthcare for Alzheimer patients

Saba Sultan[1] · Ali Javed[1] · Aun Irtaza[2] · Hassan Dawood[1] · Hussain Dawood[3] · Ali Kashif Bashir[4]

Received: 1 May 2018 / Accepted: 26 August 2019
© Springer-Verlag GmbH Germany, part of Springer Nature 2019

Abstract
Alzheimer patients face difficulty to remember the identity of persons and performing daily life activities. This paper presents a hybrid method to generate the egocentric video summary of important people, objects and medicines to facilitate the Alzheimer patients to recall their deserted memories. Lifelogging video data analysis is used to recall the human memory; however, the massive amount of lifelogging data makes it a challenging task to select the most relevant content to educate the Alzheimer's patient. To address the challenges associated with massive lifelogging content, static video summarization approach is applied to select the key-frames that are more relevant in the context of recalling the deserted memories of the Alzheimer patients. This paper consists of three main modules that are face, object, and medicine recognition. Histogram of oriented gradient features are used to train the multi-class SVM for face recognition. SURF descriptors are employed to extract the features from the input video frames that are then used to find the corresponding points between the objects in the input video and the reference objects stored in the database. Morphological operators are applied followed by the optical character recognition to recognize and tag the medicines for Alzheimer patients. The performance of the proposed system is evaluated on 18 real-world homemade videos. Experimental results signify the effectiveness of the proposed system in terms of providing the most relevant content to enhance the memory of Alzheimer patients.

Keywords Alzheimer · Education · Egocentric data · Healthcare · Video summarization

✉ Ali Javed
ali.javed@uettaxila.edu.pk

Saba Sultan
sababukhari71@gmail.com

Aun Irtaza
aun.irtaza@uettaxila.edu.pk

Hassan Dawood
hassan.dawood@uettaxila.edu.pk

Hussain Dawood
hdaoud@uj.edu.sa

Ali Kashif Bashir
dr.alikashif.b@ieee.org

[1] Software Engineering Department, University of Engineering and Technology, Taxila, Pakistan

[2] Computer Science Department, University of Engineering and Technology, Taxila, Pakistan

[3] Department of Network and Computer Engineering, College of Computer Science and Engineering, University of Jeddah, Jeddah 23890, Saudi Arabia

[4] Department of Computing and Mathematics, Manchester Metropolitan University, Manchester, UK

1 Introduction

Technological advancement in digital video recording devices provide a motivation for the users to capture and record videos on daily basis. The evolution in digital capturing devices specifically wearable cameras, i.e., Looxcie, SenseCam, etc. nowadays make the users eager to capture their daily life activities frequently. Most of the data captured from the wearable cameras consist of unstructured, lengthy and redundant content. There exists a dire need to summarize this redundant video content into a concise representation that can provide only the relevant information to the users. Video summarization techniques are being applied to produce a succinct representation of a full-length video that must contain its most informative segments.

Existing video summarization approaches (Grauman and Lu 2013; Lee and Grauman 2015) generate the output either in the form of static images as key-frames (Lidon et al. 2017) or a dynamic representation as video skims (Lee and Grauman 2015; Javed et al. 2016). Egocentric video summarization provides a compact representation of the input

video captured from a wearable camera that offers a first-person view of the world. Egocentric video summarization has many practical applications in various domains such as healthcare (Grauman and Lu 2013), surveillance (Song et al. 2016), sports (Aghdam et al. 2015), media (Varini et al. 2015a, b), etc. Egocentric video summarization approaches have been proposed to solve various problems in the healthcare industry (Zhang et al. 2016a, b). Alzheimer is a common disease associated with memory loss, physiological and inventive abilities. Patients suffering from Alzheimer's face difficulties in remembering and performing daily life activities. It is difficult for Alzheimer patients to remember the identity of persons (even their family members), objects, things and so on. In the proposed research work, egocentric video summarization approach is presented to educate the Alzheimer's patients to learn and memorize their deserted memories. Existing techniques (Doherty et al. 2008; Zhang et al. 2016a, b) have used egocentric videos to propose the health care systems for various diseases, i.e., autism, obesity, and depression etc. Karaman et al. (2014) proposed a technique based on two-level Hidden Markov Model to recognize the daily-life activities of dementia patients captured through the egocentric camera. Doherty et al. (2008) proposed a technique for video summarization, where key-frames are selected based on the face and objects detection. The limitation of this approach (Doherty et al. 2008) is that the relationships between key-frames are completely ignored while generating the video summary. Meditskos et al. (2018) used the technique of multi-sensory data analysis along-with egocentric video recording from a bracelet to aid the dementia patients by recognizing the daily living activities. This technique was used to better understand the human behaviours captured through wearable devices. Nguyen et al. (2016) proposed an object recognition technique using wearable cameras to facilitate the daily life activities of older people. This method has a limitation of recognizing only limited objects and unable to accurately classify between similar objects. Zhang et al. (2018b) proposed an object-oriented video summarization technique using the unsupervised learning approach. This method detects the key motion of objects available in the online videos to extract the key-frames followed by summarizing the content in unsupervised manner. Method proposed by Zhang et al. (2018a) used the query-conditioned three pair generative adversarial network to learn the user queries and video contents. Three-pair loss was used to generate more compact summaries of the daily life videos.

Existing techniques for egocentric video summarization can be classified into two classes (1) learning-based approaches (Grauman and Lu 2013; Varini et al. 2015b; Zhang et al. 2016a, b; Bolanos et al. 2017), and (2) non-learning-based approaches (Lidon et al. 2017; Blighe et al. 2008). Grauman and Lu (2013) proposed a learning-based approach for story-based summarization. The input video was partitioned into small shots using a static-transit grouping approach. Each video shot was analysed to detect the selected objects in the input video frames. The worth of each shot and its influence on the other shot was determined and energy function of each shot was optimized to examine the significance of preserving the important events. Zhang et al. (2016a, b) proposed a learning-based subset selection criteria approach to generate the summarized video. Detrimental point process (DPP) was used for structured problems due to its accuracy as compared to other graphical models. Su and Grauman (2016) used ego-motion cues to develop a learning-based technique to detect the engagement of the camera wearer with other objects using the ego-motion cues.

Non-learning-based techniques (Lidon et al. 2017; Blighe et al. 2008) have been proposed to address the limitations of the learning-based methods (Varini et al. 2015b; Zhang et al. 2016a, b). For example, Lidon et al. (2017) used a semantic relevance criterion to generate the summary of egocentric photo streams captured through a wearable camera. CNN was applied to remove the non-informative images from the input stream. Semantic diversity was obtained by ranking the images according to the relevance. Finally, images were re-ranked to filter those images that represent the most diverse characteristics thereby reducing the redundancy while preserving the semantic information. Blighe et al. (2008) proposed an approach to compute the similarity index between the input video frames to identify the key-frames for video summarization. This approach selects the key-images within an event using a combination of MPEG-7 and Scale Invariant Feature Transform (SIFT) features. Similarly, Lu (1995) used textual features to develop an optical character recognition technique that can isolate and detect the characters from the words and robust to broken and overlapping characters.

An effective egocentric video summarization technique is proposed to detect and recognize the objects, persons, and medicines to aid the Alzheimer's patients. It has been observed that existing summarization techniques for Alzheimer's or dementia patients generate the summary in the form of video skims (concise videos). The main limitation of video skims to aid the Alzheimer patients is the difficulty of memorizing the provided content, i.e., person, object, event, etc. Based on this observation, a static video summarization approach is proposed to provide the most relevant key-frame that facilitates the Alzheimer patient to easily learn and memorize the identity of persons, objects or events. Moreover, the proposed method captures and processes the egocentric videos at real-time, therefore a light-weight and effective system is developed to generate the static video summaries through selecting the most relevant video frames.

The proposed research work comprises of face recognition, object recognition, and medicine recognition that educate the Alzheimer patients to memorize the daily life

activities. Face recognition is used to recognize the identity of persons related to Alzheimer patients. Face recognition is implemented in two stages. Firstly, haar classifiers are extracted to train a classifier to detect and localize the face segments from the input video frames. Secondly, HoG features are extracted from these localized face segments to train the multi-class SVM classifier for face recognition. Finally, the recognized faces are tagged to determine the relationship with the patient. For object recognition, SURF descriptor is applied to detect the specific objects used by the Alzheimer patients. The recognized objects are then tagged with a complete description regarding the use of each object. This feature is provided to facilitate the Alzheimer patients in terms of memorizing the identity and usage of any object. For medicine recognition, morphological image analysis and OCR are used to recognize the medicine names of Alzheimer patients. Moreover, the complete prescription of the recognized medicine is also provided along-with the reminder to take the correct dose of medicines at regular intervals. Performance of the proposed system is evaluated on the homemade egocentric video dataset of three different categories.

The paper is organized as follows. Section 2 describes the proposed method. Performance evaluation of the proposed method is provided in Sect. 3. Section 4 concludes the proposed research work.

2 Proposed method

The proposed technique is divided into three main modules that are face recognition, object recognition, and medicine recognition. The block diagram of the proposed framework is presented in Fig. 1.

2.1 Face recognition

It is very common for Alzheimer patients to forget the identity of family members, friends and relatives. To overcome this issue, face recognition is performed in the proposed work to educate the Alzheimer patients to memorize the identity of their closed ones (i.e., family, friends, etc.).

Face detection is commonly employed initially in the input video frame to localize the faces that are further processed for various applications, i.e., face tracking, face recognition, etc. In the proposed work, face detection is performed to localize the faces that are then used for face recognition in the input video frame. It is a common observation that the eyes region in the face is darker than the cheeks region. Therefore, for face detection, we employed haar-based features (Viola and Jones 2004) to train a classifier. Haar features consist of a set of two adjacent rectangles that lie above the eye and cheek region. The location of these adjacent rectangles is defined relative to a detection window that acts as a bounding box to the target object. Integral

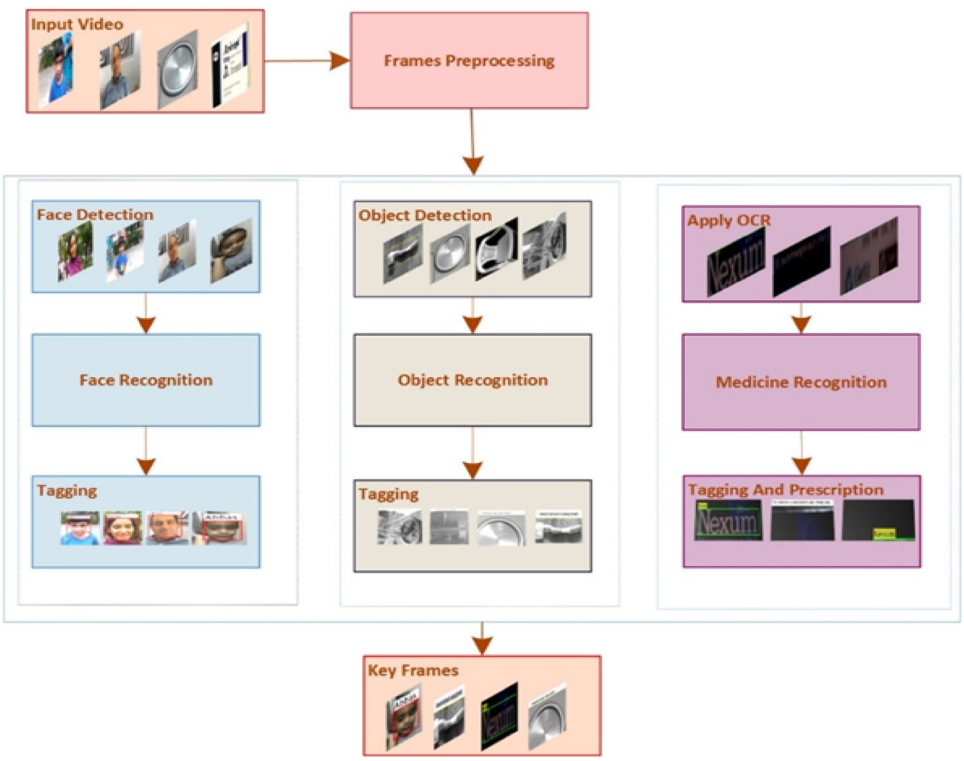

Fig. 1 Block diagram of proposed system

Fig. 2 Face detection

image representation is used to compute the features efficiently. Adaboost algorithm is applied for face detection and the detected faces are localized as shown in Fig. 2.

The localized face segments in the input video frames are then analysed to recognize the relatives of Alzheimer patients. For this purpose, we have used histograms of oriented gradient (HoG) descriptor to extract the features on the localized face segment to represent a 70×4680 feature vector that is used to train the SVM classifier. Seventy images (seven images per class) are used to train the classifier. In experiments, 70% of the images are used for training, whereas, remaining 30% images are used for testing. HoG features encode the edge information along with its direction that makes it very effective in terms of extracting the texture of an object.

Face recognition is a multiclass classification problem where one label is selected among more than two class labels. We employed the SVM classifier face recognition. SVM is commonly used for binary classification problems, however SVM can also be used to address the multi-class classification problem. We employed the Error correcting output codes (Dietterich and Bakiri 1995) that is an ensemble method designed to solve a multi-class problem by dividing it into a collection of two-class problems. A separate SVM base classifier is trained to solve each two-class problem. For error correcting output codes (ECOC) we trained 15 binary SVM classifiers to recognize the 10 test subjects. We adopted the one versus one approach to design the multi-class SVM in ECOC framework. We tested different SVM kernels, i.e., linear, quadratic, cubic, Gaussian RBF. It has been observed during experiments that the Gaussian RBF kernel achieves superior classification accuracy as compared to others. Therefore, Gaussian RBF kernel is used in the proposed SVM-based ECOC framework for face recognition that is represented as:

$$k\left(y_i, y_j\right) = e^{-\frac{1}{2\sigma^2} y_i - y_j^2} \tag{1}$$

where $y_i - y_j^2$ represents the Euclidean distance between the two feature vectors.

Ten class classification problem is illustrated in Table 1. From the table, it can be clearly observed that a 15-bit error correcting output code is created for a ten-class classification problem. A unique binary string/codeword of length 15 is allocated to each face class. One binary classifier is trained for each column so 15 binary classifiers are trained in this manner. We evaluated 15 binary classifiers against each given input face image to acquire a 15-bit codeword. Finally, the input face image is assigned to a class whose codeword is closest to the 15-bit codeword of the 10 face classes. Once the faces are recognized then the corresponding name and relation of the recognized person is tagged as shown in Fig. 3. In this way we recognize the identities of the relatives of Alzheimer patients.

2.2 Object recognition

Object recognition module is designed to detect and recognize specific objects related to the Alzheimer's patient. This approach is used to educate the Alzheimer patients to

Table 1 Error correcting output code for ten class classification

Class	Code word														
	C_0	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}	C_{12}	C_{13}	C_{14}
Face 1	1	1	0	0	0	0	1	0	1	0	0	1	1	0	1
Face 2	0	0	1	1	1	1	0	1	0	1	1	0	0	1	0
Face 3	1	0	0	1	0	0	0	1	1	1	1	0	1	0	1
Face 4	0	0	1	1	0	1	1	1	0	0	0	0	1	0	1
Face 5	1	1	1	1	1	0	1	1	0	0	1	0	0	0	1
Face 6	0	1	0	0	1	1	0	1	1	1	0	0	0	0	1
Face 7	1	0	1	1	1	0	0	0	0	1	0	1	0	0	1
Face 8	0	0	0	1	1	1	1	0	1	0	1	1	0	0	1
Face 9	1	1	0	1	0	1	1	0	0	1	0	0	0	1	1
Face 10	0	1	1	1	0	0	0	0	1	0	1	0	0	1	1

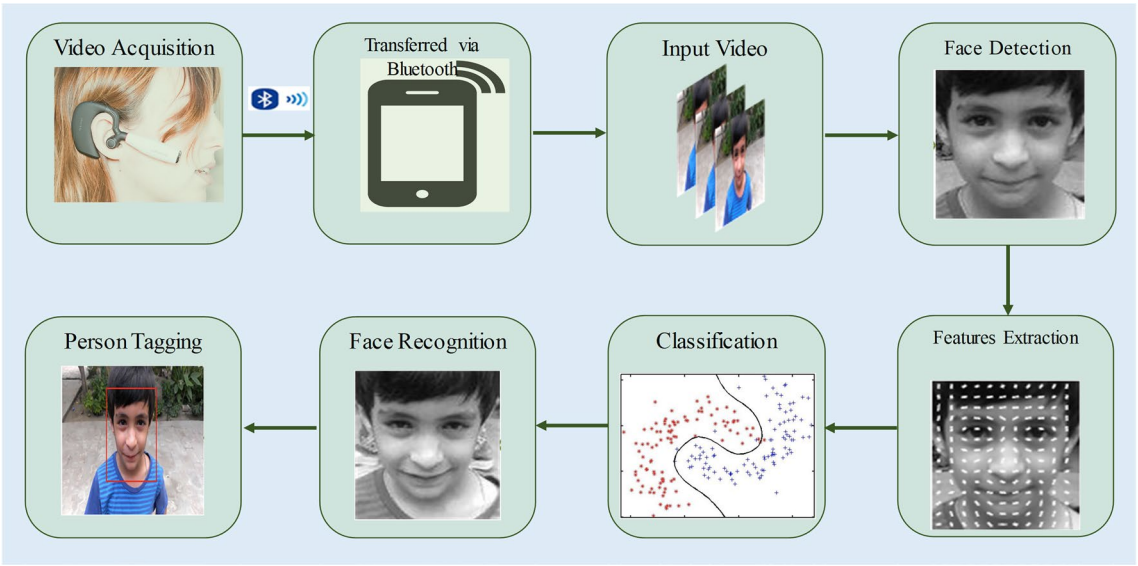

Fig. 3 Process flow of face recognition

memorize the use of objects (Bay et al. 2006). We proposed a light-weight non-learning-based method for object recognition due to the constraints of real-time video processing. The proposed method detects the specific objects based on finding the point correspondence between the reference and the target image. Reference objects correspond to the dataset images containing few objects relevant to the Alzheimer's patient. The proposed method is robust to variations in scale and illumination, in-plane and out-of-plane rotation, and occlusions.

The input egocentric video is pre-sampled by processing every 10th frame followed by the transformation into grayscale. Speeded up robust features (SURF) are used to extract the feature points from the input video frame where multiple objects exist in a cluttered scene. To this end, feature points of the stored images and the input frames are computed to detect an object of the class. The strongest feature point value determines the object from the referenced image. In the given input frame 300 strong feature points are extracted. Feature points are also extracted against each object image of the dataset and 100 strong feature points are selected. Feature descriptors are extracted at these interest points for both the input video frame and object images of the dataset to determine the corner locations in the images. Each object image in the dataset is matched with the given input video frame to get the matching points that are used to recognize the objects from the dataset. These matching points include the outliers as well, beside the object points. Geometric transformation is applied to remove the outliers and localize the object(s) in the input video frame. Each recognized object is tagged by assigning the name and use of that object. Process flow of the object recognition method is shown in Fig. 4.

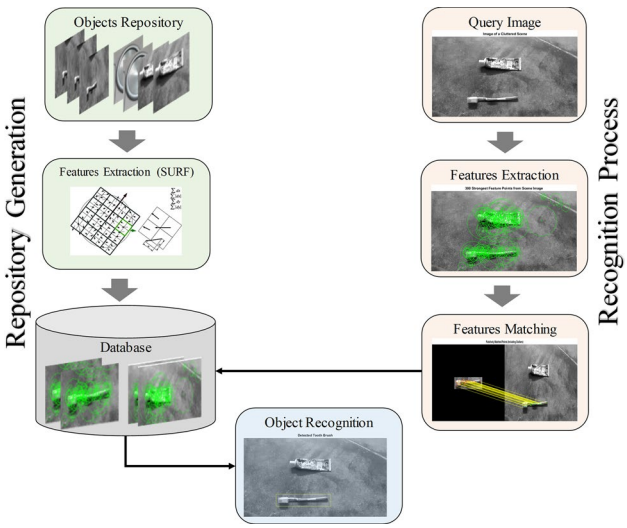

Fig. 4 Process flow of object recognition

2.3 Medicine recognition

Medicine prescription, recognition, and reminder setting for medicine intake is the best way to help Alzheimer patients to memorize their medicines and intake time. This module is specifically designed to educate the Alzheimer patients to remember the names, quantity, and intake time of medicines. The proposed egocentric video summarization method processes each frame to detect and localize medicine names that are then recognized using the optical character recognition (OCR) method. The corresponding frames of the input video where medicines are recognized are marked as key-frames.

These marked key-frames are then used to generate the static summary of the input egocentric video.

The input video frames are transformed into grayscale images by applying the weighted average scheme on each colour channel as follows:

$$I_{gs}^{(i)}(x,y) = 0.298 \times I_r^{(i)}(x,y) + 0.587 \times I_g^{(i)}(x,y) + 0.114 \times I_b^{(i)}(x,y) \quad (2)$$

where $I_{gs}^{(i)}$ represents the grayscale frame,$I_r^{(i)}$,$I_g^{(i)}$, and $I_b^{(i)}$ represents the red, green, and blue components of the colour video frame respectively. The grayscale frames are transformed into binary images as follows:

$$I_{bin}^{(i)}(x,y) = \begin{cases} 0, & if\ \left(\mu^{(i)} - \rho \times \sigma^{(i)}\right) \le I_{gs}^{(i)}(x,y) \le \left(\mu^{(i)} + \rho \times \sigma^{(i)}\right) \\ 1, & otherwise \end{cases} \quad (3)$$

where $I_{bin}^{(i)}$ represents the binary frame,$\mu^{(i)}$ and $\sigma^{(i)}$ represents the mean and standard deviation of the gray-scale frame. ρ is a parameter.

We have applied few pre-processing steps on these gray-scale frames to transform input video frames in a way that are more suitable to be processed by the OCR. For this purpose, morphological operators are applied to enhance the image before feeding it into the OCR algorithm. Morphological erosion operation is applied to remove the noise and reduce the blurriness as follows:

$$I_e^{(i)}(x,y) = I_{bin}^{(i)}(x,y) \ominus S \quad (4)$$

where $I_e^{(i)}$ is the morphed image obtained after applying erosion, S is the structuring element, and Θ is the erosion operator. The shape of structuring element S is set to disk for faster processing as the morphological operations using disk approximations run much faster. The size of the structuring element is set to 3×3 for effective noise removal.

Morphological erosion creates small gaps in characters that reduce the accuracy rate of OCR. To resolve the issue of broken characters dilation operator is applied on the eroded image as follows:

$$I_d^{(i)}(x,y) = I_e^{(i)}(x,y) \oplus S \quad (5)$$

where $I_d^{(i)}$ is the dilated image, and $\oplus$ represents the dilation operator. S is used with the same settings (disk shape, 3×3 size) to effectively bridge the gaps between characters and preserving the fine details of each character.

Optical character recognition algorithm (Smith 2007) is applied on this dilated image to recognize the name of medicines. The recognized words obtained from the OCR is compared with the names of Alzheimer patients medicine already stored in the database. In case, the recognized word from the input video frame matches with any of the patient's medicine then the medicine name is tagged in the corresponding input video frame as shown in Fig. 5. The recognized medicines are then prescribed for the patient. A prescription consisting of the details of the medicines, i.e., quantity, time etc. is displayed for patient's reference. A reminder is also set to notify the Alzheimer patients regarding the time of medicine intake at regular intervals. The process flow of the medicine recognition approach is shown in Fig. 5.

3 Performance evaluation

We evaluated the performance of the proposed video summarization system on a video dataset consisting of 18 real-world homemade videos. Objective evaluation is used to measure the performance of the proposed method. *Precision,*

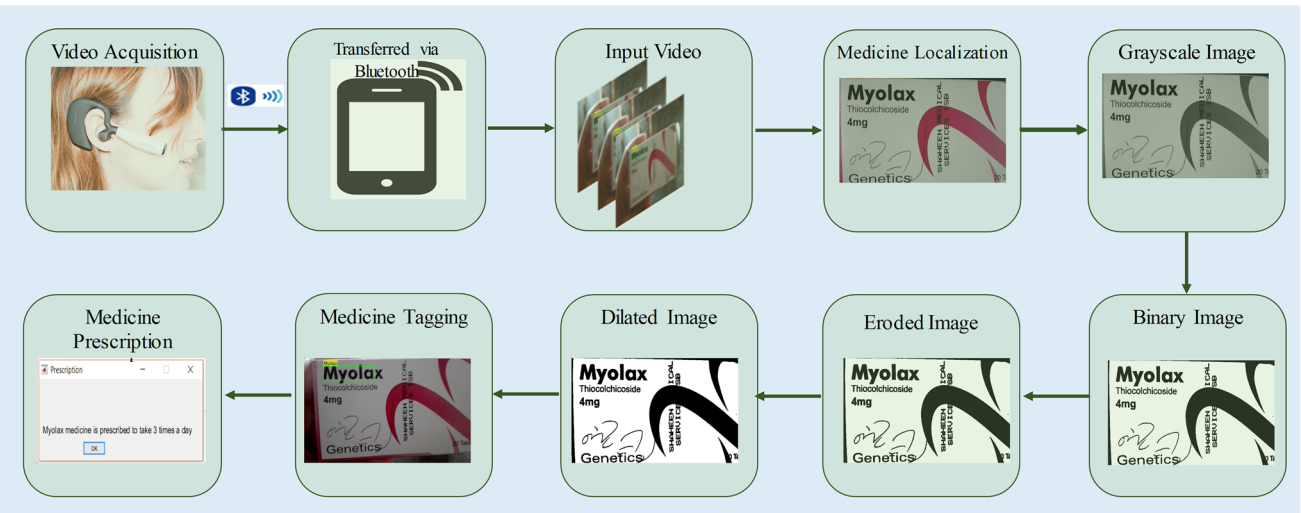

Fig. 5 Process flow of medicine recognition

recall, accuracy, error rate, and *F-1 measure* are used for performance evaluation. This section provides the details of different experiments performed along-with a comprehensive discussion on the obtained results. The proposed video summarization system is implemented in MATLAB.

3.1 Dataset

A dataset consisting of 18 real-world homemade videos of a total duration of 10 h is created for performance evaluation. The dataset videos are captured via Looxcie wearable camera mounted on the ear. Each recorded video in the dataset has a frame resolution of 640×480 pixels and a frame rate of 30 fps. The dataset videos contain the family members, medicines, and various objects of interest for Alzheimer's patient. Some frames of the dataset are provided in Fig. 6.

For face recognition a dataset consisting of images of ten persons is created. Ten images of each subject are collected at different poses and emotions for this purpose. Some snapshots of face dataset are shown in Fig. 7.

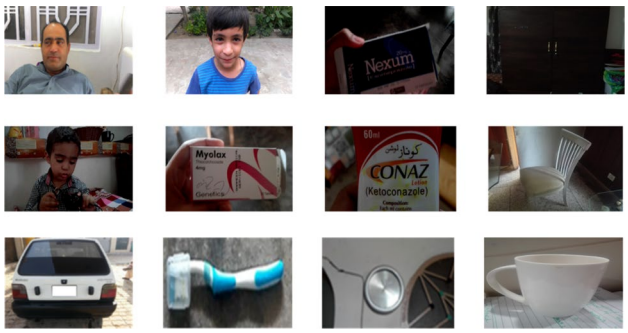

Fig. 6 Snapshots of dataset

Fig. 7 Snapshots of face dataset

3.2 Results and discussion

The proposed video summarization method is evaluated on each video of the dataset to measure its effectiveness in terms of recognizing persons, medicines and various objects related to the Alzheimer patients. In this section we provided the details of the evaluation metrics and results of different experiments along-with the discussion.

3.2.1 Evaluation metrics

Precision, recall, accuracy, error, and F-1 measure are computed in the first experiment to measure the detection performance of face, object, and medicine recognition.

Precision for face recognition represents the correctly tagged face, to the total detected faces; and similarly, in case of object recognition, it is computed as the ratio of correctly labelled objects to total number of detected objects and same is for medicines. Precision is computed as follows:

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

where *TP* represents the True Positive values, and *FP* represents the false positive values in terms of face, object or medicine frame recognition from the videos. Recall represents the ratio of true recognition and tagging of faces in case of face recognition against the total number of face frames in the videos. In case of object and medicine recognition, recall represents the ratio of correctly detected objects and medicines against the total number of objects and medicine frames in the videos. Recall is computed as follows:

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

where *TP* represents the True Positive values and *FN* represents the false negative values. False Negatives are the faces, objects and medicine frames in the video that are misclassified or wrongly labelled. *F-1* score or *F-measure* represents the weighted average of precision and recall. Some methods have higher precision and lower recall and vice versa. So, to overcome the situations where precision and recall amongst the relative methods are overlapping, we have computed the *F-1* score. *F-1* score is computed as follows:

$$F1 - Score = 2 \times \frac{(Precision \times Recall)}{Precision + Recall} \tag{8}$$

Accuracy is the ratio of correctly identified and tagged face/non-face frames, correctly detected and labelled object/non-object frames and rightly labelled medicine/non-medicine frames to the total number frames in each of the video. Accuracy rate is computed as follows:

$$Accuracy = \frac{TP + TN}{P + N} \tag{9}$$

where P represents the total count of positive frames/samples and N represents total count of negative samples. *Error* is computed as the ratio of False positive and False negative frames, i.e., mislabelled face, object and medicine frames to the total number of frames in a video. We computed the error rate as follows:

$$Accuracy = \frac{FP + FN}{P + N} \tag{10}$$

3.2.2 Detection results

The recognition performance of the proposed method against each input video of the dataset is shown in Table 2. The proposed method effectively recognizes the important persons, medicines, and objects relevant to Alzheimer patients. More specifically, the proposed method achieves an average precision, recall, accuracy, error rate and F-measure of 0.93%, 0.87%, 0.89%, 0.11% and 0.89%, respectively for all videos. We achieved the average accuracy of 90% and 91% for face and medicine recognition which shows the effectiveness of the proposed techniques to recognize the face and medicines. Whereas the average accuracy of object recognition is around 87%. The slight decrease in the performance of the object recognition module is due to the fact that majority of the selected egocentric videos contain visually similar objects. In addition, some objects are also similar in color of the background and often remains undetected. However still the proposed object recognition method achieves better recognition accuracy. From the results presented in Table 2, we can clearly observe that the overall performance of proposed system is remarkable in terms of recognizing the faces, medicines and other objects relevant to Alzheimer patients.

3.2.3 Performance comparison with existing methods

Performance of the proposed system is compared against the existing state-of-the-art recognition methods. Precision, recall, and F-1 measure are used for performance comparison. For this purpose, we compared the performance of the proposed method against existing techniques of face, object, and text recognition. The statistical comparison of the proposed and existing methods is provided in Table 3. From the results it can be observed that the proposed method provides superior recognition performance as compared to the existing state-of-the-art methods.

Table 2 Recognition Performance

Video type	True positive	True negative	False positive	False negative	Precision	Recall	Accuracy	Error	F-measure
Face 1	88	13	0	7	1	0.9	0.92	0.08	0.94
Face 2	36	2	0	4	1	0.9	0.90	0.1	0.94
Face 3	25	4	3	5	0.89	0.83	0.85	0.15	0.84
Face 4	7	31	1	3	0.87	0.7	0.90	0.1	0.76
Face 5	10	46	2	3	0.83	0.76	0.91	0.09	0.79
Face 6	30	10	1	1	0.96	0.96	0.95	0.05	0.95
Average					0.92	0.84	0.90	0.1	0.87
Med 1	19	28	3	2	0.86	0.90	0.92	0.09	0.87
Med 2	32	15	1	5	0.96	0.86	0.92	0.11	0.90
Med 3	25	20	2	4	0.92	0.86	0.88	0.11	0.88
Med 4	36	1	1	2	0.97	0.94	0.92	0.07	0.95
Med 5	20	36	2	1	0.9	0.95	0.94	0.05	0.92
Med 6	24	49	2	4	0.92	0.85	0.92	0.07	0.88
Med 7	19	23	1	2	0.96	0.90	0.93	0.06	0.92
Med 8	36	37	3	4	0.92	0.9	0.91	0.08	0.90
Average					0.92	0.85	0.91	0.10	0.90
Object 1	33	1	1	6	0.97	0.84	0.82	0.18	0.89
Object 2	23	3	1	3	0.95	0.88	0.86	0.14	0.91
Object 3	20	10	2	1	0.90	0.95	0.90	0.1	0.92
Object 4	15	14	1	1	0.93	0.93	0.93	0.07	0.92
Average					0.93	0.90	0.87	0.13	0.91
Average recognition					0.93	0.87	0.89	0.11	0.89

Table 3 Performance comparison with existing state-of-the-art methods

Techniques	Custom dataset details					Precision	Recall	F-measure
	No. of videos	Length (h)	Format	Frame rate	Resolution			
Lidon et al. (2017)	25	10	MP4	2 fps	320×720	0.84	0.86	0.84
Zhang et al. (2016a, b)	10	3–5	AVI	Not specified	384×216	Not used	Not used	0.49
Jeong et al. (2016)	04	3–5	Not specified	Not specified	Not specified	0.72	0.83	0.76
Zhang et al. (2018a, b)	04	3–5	Not specified	Not specified	Not specified	0.47	0.48	0.46
Tang et al. (2018)	937			300 fps	Not specified	0.80	Not used	Not Used
Toshev et al. (2009)	42	–	Not specified	50 fps	Not specified	0.8	0.86	0.82
Crandall et al. (2002)	15	–	MPEG	30 fps	320×240	0.46	0.48	0.46
Shivakumara et al. (2012)	–	–	–	–	–	0.74	0.87	0.79
Proposed system	18	10	MP4	30 fps	640×480	0.93	0.87	0.89

Table 4 Confusion Matrix

	Face recognition	Object recognition	Medicine recognition
Face recognition	0.9	0.1	0
Object recognition	0.01	0.87	0.12
Medicine recognition	0	0.09	0.91

3.2.4 Confusion matrix analysis

In this experiment we evaluated the performance of the proposed method using the confusion matrix analysis as shown in Table 4. Confusion matrix analysis is used to depict the classification accuracy of the proposed method in terms of recognizing face, medicine, and objects. As we can observe from Table 4 that our face recognition method achieves true positives of 90% and falsely assigns the face to object class for only 10%. In addition, medicine recognition method is more superior and accurately recognizes the objects 91% whereas, only 9% falsely recognize as other objects. Finally, object recognition achieves true positives of 87%. The classification accuracy of the proposed method for face, objects, and medicine recognition is remarkably well. We can argue from these results that the proposed method can reliably be used to generate the static summaries of egocentric videos.

3.2.5 ROC curve analysis

Performance of the proposed method is also evaluated using receiver operating characteristic (ROC) curve analysis. ROC curves of the proposed method for face, object, and medicine recognition is presented in Fig. 8. ROC curves are plotted against the true positive rate and false positive rate. It can be observed from the area under the ROC curves that the proposed method is very effective in terms of recognizing faces, objects, and medicines to facilitate the Alzheimer patients.

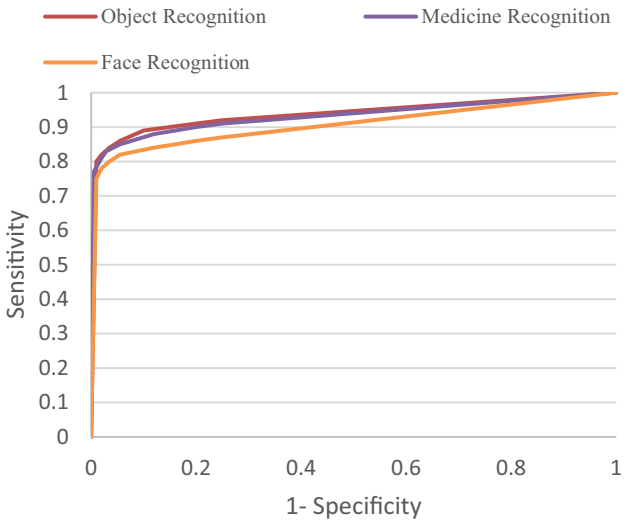

Fig. 8 ROC curve analysis for object, medicine, and person recognition

4 Conclusion

We have proposed an effective method for video summarization to aid the Alzheimer's patients to recall their blurred memories. The proposed method provides a static summary of the egocentric video data to educate the Alzheimer patients in terms of recognizing the identities of persons (i.e., family, friends, etc.), objects and their usage, and medicines along with the required information of dosage and intake time. Our method is robust to illumination conditions and camera jitters and successfully recognize the persons, objects, and medicines for videos containing severe illumination variations and shaky movements. A diverse dataset of real-world homemade wearable camera videos is used to measure the performance of the proposed method. The average recognition accuracy of 89% illustrates the effectiveness of the proposed method.

Currently, we are examining the performance of the proposed system on a more diverse dataset. In the light of the results achieved after performance evaluation, we are also investigating other efficient feature descriptors and lightweight machine learning algorithms for recognition purposes. We are directing our efforts to further enhance our system by proposing more efficient yet effective approach to facilitate the Alzheimer patients.

References

Aghdam HH, Heravi EJ, Puig D (2015) An unsupervised method for summarizing egocentric sport videos. In: Eighth international conference on machine vision (ICMV 2015)

Bay H, Tuytelaars T, Van Gool L (2006) SURF: speeded up robust features. In: Computer Vision—ECCV 2006. Austria

Blighe M, Doherty A, Smeaton AF, Connor NEO (2008) Keyframe detection in visual lifelogs. In: Conference on pervasive technologies

Bolanos M, Dimiccoli M, Radeva P (2017) Towards storytelling from visual lifelogging: an overview. IEEE Trans Hum Mach Syst 47:77–90

Crandall D, Antani S, Kasturi R (2002) Extraction of special effects caption text events from digital video. Int J Doc Anal Recognit 5:148–150

Dietterich TG, Bakiri G (1995) Solving multiclass learning problems via error-correcting output codes. J Artif Intell Res 263–286

Doherty AR, Byrne D, Smeaton AF, Jones GJF, Hughes M (2008) Investigating keyframe selection methods in the novel domain of passively captured visual lifelogs. In: Proceedings of the 2008 international conference on content-based image and video retrieval, pp 259–268. ACM

Grauman K, Lu Z (2013) Story-driven summarization for egocentric video. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). Texas

Javed A, Bajwa KB, Malik H, Irtaza A (2016) An efficient framework for automatic highlights generation from sports videos. IEEE Signal Process Lett 23(7):954–958

Jeong D, Yoo HJ, Cho NI (2016) A static video summarization method based on the sparse coding of features and representativeness of frames. EURASIP J Image Video Process 2017(1):1

Karaman S, Benois-Pineau J, Dovgalecs V, Mégret R, Pinquier J, André-Obrecht R, Gaëstel Y, Dartigues J-F (2014) Hierarchical Hidden Markov Model in detecting activities of daily living in wearable videos for studies of dementia. Multimedia Tools Appl 69(3):743–771

Lee YJ, Grauman K (2015) Predicting important objects for egocentric video summarization. Int J Comput Vis 114(1):38–55

Lidon A, Bolanos M, Dimiccoli M, Radeva P, Garolera M (2017) Semantic summarization of egocentric photo stream events. In: LTA'17 Proceedings of the 2nd workshop on lifelogging tools and applications, Mountain View, California, USA, 23–24 October 2017. ACM, New York

Lu Y (1995) Machine printed character segmentation—an overview. Pattern Recognit 28(1):67–80

Meditskos G, Plans P-M, Stavropoulos TG, Benois-Pineau J, Buso V, Kompatsiaris I (2018) Multi-modal activity recognition from egocentric vision, semantic enrichment and lifelogging applications for the care of dementia. J Vis Commun Image Represent 51:169–190

Nguyen T-H-C, Nebel J-C, Florez-Revuelta F (2016) Recognition of activities of daily living with egocentric vision: a review. Sensors (Basel) 16:72

Shivakumara P, Sreedhar RP, Phan TQ, Lu S, Tan CL (2012) Multioriented video scene text detection through bayesian classification and boundary growing. IEEE Trans Circuits Syst Video Technol 22(8):1231–1233

Smith R (2007) An overview of the tesseract OCR engine. In: Proceedings of 9th international conference on document analysis and recognition (ICDAR)

Song X, Sun L, Lei J, Tao D, Yuan G, Song M (2016) Event-based large scale surveillance video summarization. J Neurocomput 187(C):66–74

Su Y-C, Grauman K (2016) Detecting engagement in egocentric video. In: Proceedings of the European conference on computer vision (ECCV). Amsterdam

Tang P, Wang C, Wang X, Liu W, Zeng W, Wang J (2018) Object detection in videos by short and long range object linking. arXiv :1801.09823

Toshev A, Makadia A, Daniilidis K (2009) Shape-based object recognition in videos using 3D synthetic object models. In: 2009 IEEE conference on computer vision and pattern recognition

Varini P, Serra G, Cucchiara R (2015) Egocentric video summarization of cultural tour based on user preferences. In: MM '15 Proceedings of the 23rd ACM international conference on Multimedia. Brisbane

Varini P, Serra G, Cucchiara R (2015) Personalized egocentric video summarization for cultural experience. In: Proceedings of the 5th ACM on international conference on multimedia retrieval. New York

Viola P, Jones MJ (2004) Robust real-time face detection. Int J Comput Vis 57(2):137–154

Zhang K, Sha F, Chao W-L, Grauman K (2016) Summary transfer: exemplar-based subset selection for video summarization. In: IEEE conference on computer vision and pattern recognition (CVPR). Las Vegas

Zhang K, Chao W-L, Sha F, Grauman K (2016) Video summarization with long short-term memory. In: Proceedings of European conference on computer vision (ECCV), California, 2016

Zhang Y, Kampffmeyer M, Liang X, Tan M, Xing EP (2018a) Query-conditioned three-player adversarial network for video summarization. Computer Vision and Pattern Recognition. BMVC 2018, pp 1–9

Zhang Y, Liang X, Zhang D, Tan M, Xing EP (2018b) Unsupervised object-level video summarization with online motion auto-encoder. Pattern Recogn Lett. https://doi.org/10.1016/j.patrec.2018.07.030