The Third-Eye
A video-based helping aid for dementia patients

By

Sayed Mehedi Azim - 011162097
Sharfaraz Mahmood Jamee - 011161183
Shofiqul Islam - 011161079
Anika Tabassum - 011161150
Md. Mainul Islam - 011162101

Submitted in partial fulfilment of the requirements
of the degree of Bachelor of Science in Computer Science and Engineering

November 25, 2021

Department of Computer Science and Engineering
United International University

Abstract

Dementia is a medical condition which is responsible for the decline in a person's memory, problem-solving and thinking skills. In short, dementia is responsible for the reduction of a person's cognitive skills. This loss of cognitive skills results in the inability of a person to perform daily routine works because he or she can not completely remember all the tasks needed to be performed in a day. Therefore, these people need help from others such as caregivers or family members. Different researches and developments has been done in assistive technologies to help these patients but these existing helping aids are not generalized to provide assistance by observing a user's routine activities. In this paper, we propose a novel approach of real time reminder system from observing a persons daily activity from first person camera view. This proposed system is aimed to automatically identify a person's daily activity from first person camera view and by analyzing the activities provide real time notification in situations where he or she forgets to complete a routine task.

Acknowledgements

This work would have not been possible without the input and support of many people over the last two trimesters. We would like to express our gratitude to everyone who contributed to it in some way or other.

First, we would like to thank our academic advisor, Dr. Chowdhury Mofizur Rahman for supervising us on this project. Our sincere gratitude goes to our beloved course teacher Dr. Swakkhar Shatabda for his relentless support and guidelines.

Last but not the least, We owe to our family including our parents for their unconditional love and immense emotional support.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Project Overview

Dementia is a medical condition which is responsible for the decline in a person's memory, problem-solving and thinking skills. In short, this is the decrease of a person's cognitive skills. This loss of cognitive skills results in the inability of a person to perform normal daily work because he or she can not completely remember all the tasks needed to be performed in a day. There are different types of dementia such as Alzheimer's disease and Vascular dementia. When someone is diagnosed with any type of dementia, they slowly start to loss the ability to perform daily task by themselves. Therefore, caregivers are assigned to help these patients for completing day to day tasks. The involvement of caregivers increases the dependency of dementia patient on others. Current development in assisted living technologies has improved the quality of day to day lives of people of all aspects including elderly people and those that are suffering from diseases. In this project, we work towards developing an intelligent assisted living technology for dementia patients which will aid them to complete daily task without the involvement of other caregivers.

1.2 Motivation

People suffering from dementia face numerous problems in their daily life. They have to depend on others to finish their works. Their family members and caregivers need to keep an eye on them all the time which is quite difficult. In some situations, it is needed to note down the completion of tasks at the time of occurrence which can be labouring and time consuming for the family members or caregivers of the patients. The constant dependency on others can be very depressing and stressful for the patients. A system that helps these patients to complete tasks can reduce this dependency which in result can give some relief to the patients or the elderly people. At present, there is no such generalized system to monitor activities of daily living (ADL) in real time.

1.3 Objectives

Dementia patient needs help from others in completing their daily task. This helping process is very time consuming and needs a huge amount of labour. Keeping that in mind we are trying to build a helping aid that can help the dementia patient in a broader way. The objective of our research is to-

- Design a system that can identify activities of daily living of a user from first person camera view.
- Provide assistance by acting as an automated reminder when the user forgets to complete a task.
- Build a model that can notify in real time by analyzing activities.
- Reduce dependencies on caregivers.

Keeping these objectives in mind we are proposing a system that will be able to reduce some of the dementia patients or elderly people's dependencies on caregivers, reduce the time and effort that is need to help the patients. As most of the existing helping aid is not providing such proactive assistance, we are hoping that our proposed system will have a vary positive effects in people's lives.

1.4 Methodology

Our project contains the following steps of research methodology:

- Defining the problem
- Background study and benchmark analysis.
- Designing the software as well as hardware and develop both of them
- Data collection and pre-processing
- Testing on real user.
- Experimental result analysis

1.5 Project Outcome

Through our work we make an effort to create a system that can help dementia patient in daily life by acting as a smart reminder and task assistance tool. To that end, we have done two related works, recognising daily activity from first person video, recognising different steps within an activity to provide prompts for those step to help a user in completing an activity.

1.6 Organization of the Report

The rest of the report is organized as follows: Chapter 2 briefly presents a preliminary knowledge and literature review of the related work; Chapter 3 describes the methodology and system architectural design of the system proposed in this paper; Chapter 4 presents the evaluation metrics used for the project with environment setup, and results and discussion; In chapter 5 standards and constraints of our work is discussed, additionally the challenges we have faced during this project is discussed here; the paper is concluded in Chapter 6.

Chapter 2

Background

In this chapter we present an overview of the related works on activity recognition, prompting system and some other topics related to our work.

2.1 Preliminaries

In this section, we present a brief overview of some algorithms, working methods related to activity recognition and some terminologies that has been used in the following sections.

2.1.1 Dementia

Dementia is a medical condition which is responsible for the decline in a persons memory, problem-solving and thinking skills. In short, this is the decrease of a person's cognitive skills. This loss of cognitive skills results in the inability of a person to perform normal daily work because he or she can not completely remember all the tasks needed to be performed in a day. There are different types of dementia such as Alzheimer's disease and Vascular dementia. When a person is diagnosed with any type of dementia, they slowly start to loss the ability to perform daily task by themselves.

2.1.2 POMDP

A partially observable Markov decision process (POMDP) is a modified version of a Markov decision process (MDP). In a POMDP model an agent take decision in which it is presumed that the system dynamics are decided by an MDP, but the agent can not directly spot the underlying state. Instead, over the set of possible states it maintains a probability distribution, based on a set of findings and probabilities, and the underlying MDP.

2.1.3 Recurrent neural network

Recurrent neural networks are simply the network with a loop in them or we can think of it as multiple copies of the same network, where each copy passes a message to a successor which mainly allows the architecture to persist the information. RNN provides incredible results in a variety of fields such as speech recognition [1, 2, 3], handwriting recognition [4], translation [5], and image captioning [6], etc. A single block of RNN takes previous state output with the current input as input and passes its output to the next RNN block. A multiplication operation between the output of the previous state $\mathrm{h_{t\text{-}1}}$ and the input $\mathrm{X_t}$ is done and then it passes through an activation function such as tanh, relu, etc. The current state of an RNN can be formulated with tanh activation function as follows

$$h_t = tanh(W_h.h_{t-1} + W_X.X_t) \tag{2.1}$$

Where $\mathrm{W_h}$ denotes the weight at the previous hidden state, $\mathrm{h_{t\text{-}1}}$ denotes the output of the previous state, $\mathrm{W_X}$ is the weight at the current input state, and $\mathrm{X_t}$ is the current input.

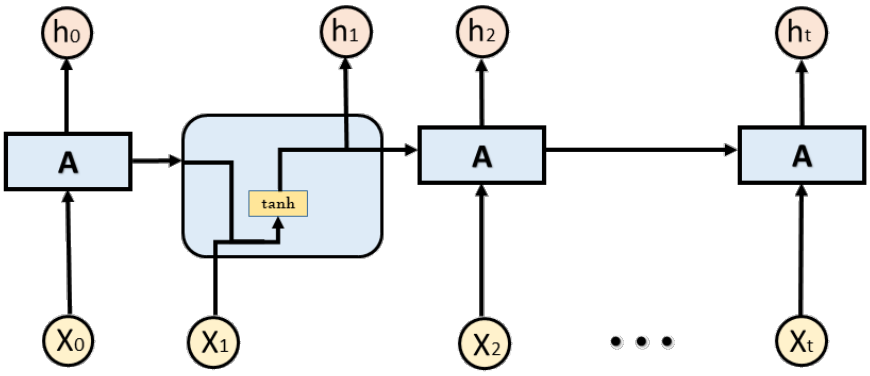


Figure 2.1: RNN model architecture

2.1.4 Long short-term memory

Long short-term memory (LSTM) [7] is a modified version of the recurrent neural network (RNN) model which widely used for sequential modeling. It is very popular for processing sequences of data such as speech and videos. LSTM was actually invented as the solution to the vanishing and exploding gradient problem. LSTM is used in recognizing handwriting [8], speech recognition [9], machine translation [10], etc. It trains using back-propagation through time. A block of LSTM consists of multiple gates instead of a simple activation function. The input gate decides which value from input should be used to modify the memory of the current block. Forget gate decides what details to be discarded from the block, and using the input and the memory of the block the Output gate decides the output. The sigmoid function is used in these gates for making these decisions.

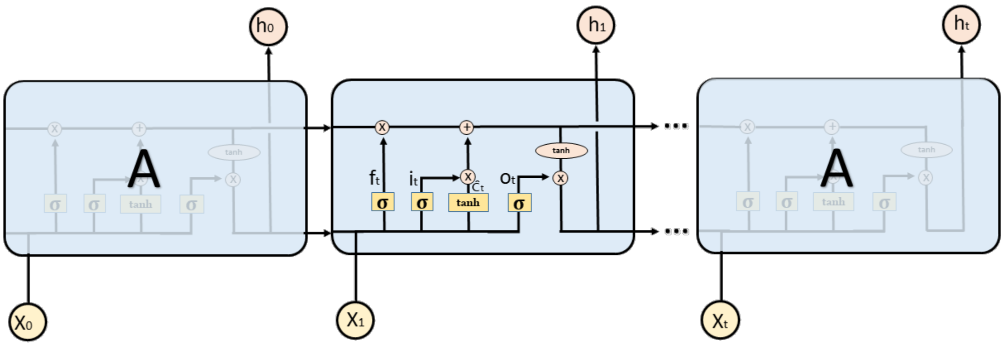


Figure 2.2: LSTM model architecture

2.1.5 Gated Recurrent Unit

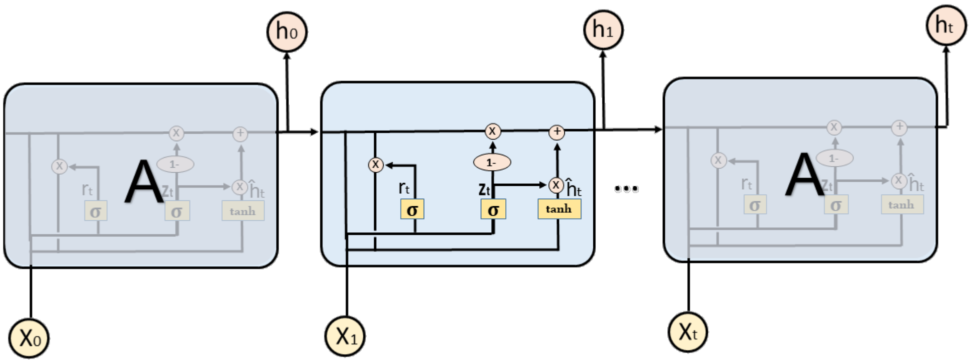


Figure 2.3: GRU model architecture

Gated Recurrent Unit (GRU) [11] is a variation of LSTM. Both GRU and LSTM are designed similarly and sometimes provide equally good results. To solve the vanishing gradient problem GRU uses two additional gates which are the update gate and reset gate. How much information from previous time steps needs to be passed is determined by the update gate. The reset gate is decisive in how much of the past information to forget.

2.1.6 MobileNetV2

MobileNetV2 [12] is a modified neural network specifically designed for mobile devices and environments with limited resources. MobileNetV2 [12] provides the state of the art accuracy of mobile tailored computer vision models. It requires a significantly lower number of operations and far less memory than the existing method needs to gain that result. They have introduced a novel layer module: the inverted residual with a linear bottleneck. A low dimensional compressed representation of data is feed to the model as input then upsampling is done and filtered with a lightweight depthwise convolution. After that with linear convolution, features are subsequently projected back to a low-dimensional representation.

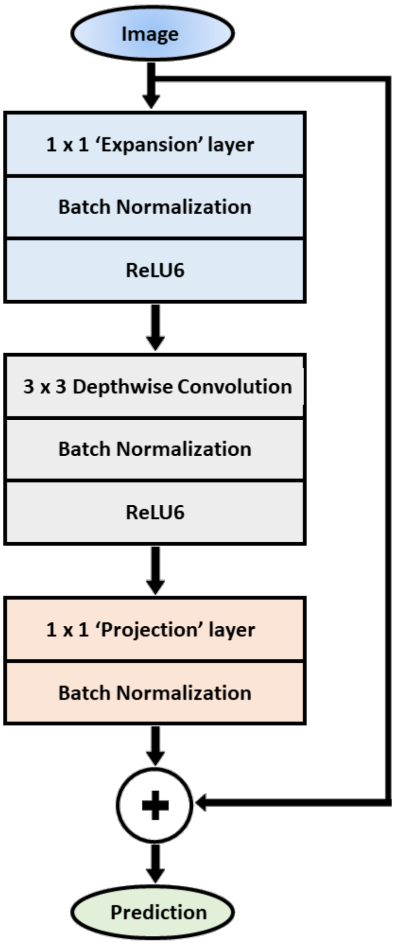


Figure 2.4: MobileNetV2 model architecture

2.2 Literature Review

In this subsection we present a review of a number of research papers. Among these papers some present different work about activity recognition from video inputs, some works are done with assisting during activity and prompting. We divided the papers into different types according to different works.

2.2.1 Activity Recognition

Pirsiavash et al. in [13], introduce a dataset of 1 million frames. The proposed paper [13] collects first person video stream from wearable camera. Detecting day to day activities from a vast data with a huge class variety is quite challenging. Data contains diversity as it was a collection of 10 hours of continuous video from 20 different homes of 20 people. Object bounding box for deduction of hand and objects from video stream is used here as the viewing angle is quite large (170 degrees). An object centric description using temporal pyramid representation, contains images and active and passive methods from top to bottom. This [13] model separate active and passive objects and temporal ordering the active and passive objects. Binary approach shortens the object presence duration and filter the unnecessary data. Linear SVM classifier on this [13] unique dataset gives the best result. Model is trained using SVM on a bag of detected object which contains pre-segmented video clips that initially gave a poor accuracy of 16.5%. After comparing results between pre-segmented and temporally continuous video clips it is shown that a perfect and idealized object detector (IO) interacted with active object increased the accuracy to 77%.

In another work [14], Kenji Matsuo et al. proposed an approach using attention to recognise activity from first person view videos of the user. The approach is discussed in several steps. From the egocentric video detecting objects is the first step. An object is classified as active if it is being interacted with by the user, otherwise it is classified as passive. In the next step, a visual attention map is created from the user's motion and the importance given to different objects within the video frames. The next step is involved in finding the objects with most focus from the user. The next step generates descriptors for the objects which is used for feature vector generation. User's activity is learned and recognised from these feature vectors using support vector machine (SVM). The experiments are done on a dataset [13] containing 20 people's videos. Videos of 6 people are used for training and the videos of other 14 people are used for testing. For evaluation leave-one-out cross validation method is applied. The experiments show that attention-based activity recognition performs better than hand gestures based recognition models. The authors also mentioned that activities with less motions of the user is harder to recognise using the proposed attention-based activity recognition approach.

Zhan et al. in [15] introduce a small and cost efficient camera embedded in glass that can automatically recognize a person's activity. It is useful for the betterment in healthcare system by assisting nurses and for taking care of elders who live alone. Caregivers and patients family members have access to monitor from a distant place. In this [15] system, a front facing CMOS camera with 60-degree field of view is used. Basic features are extracted by 'Lucas-Kanade based optical flow' which high cost computation. Zhan et al. [15] propose a new way of average pooling that improves the overall performance. 'Forward-Backward' algorithm is used to train the model. Authors in [15] set 4 major activities: walk, drink, go upstairs, go downstairs and run cross validation over a number of videos containing one activity on each video on different location to choose parameters with highest accuracy. To compare, authors in [15] classified the features with 3 different classifiers: k-Nearest Neighbor (k-NN), LogitBoost (with decision stumps) and Support Vector Machines (SVMs). Hidden Markov Model (HMM) was applied on every classifier for more polishing. After comparing parameterization and feature analyzation on 'Single-Activity' and 'Single and Multi-Activity' on training data where it showed that multi-activity videos improved the performance in general. SVM and HMM with average pooling acquired the highest accuracy with 82.1%. LogitBoost with HMM acquired 82% (with average pooling). They were also able to control natural periodic shivering during walking and other activities.

Egocentric activity recognition is hard because it requires fine-grained details about small objects and their manipulations. There can be a lot of noise in the data because of the articulated nature of our body. To handle all problems and to recognize activities, [16] presents a new recurrent network unit called long short-term attention (LSTA), which is a modified LSTM with a built-in attention mechanism and for that they had to change some gate function of the default LSTM. They then used two-Stream architecture to recognize the action. The first Steam has a simple object recognizer model that classifies the objects of a frame and where to attend and encode that, which is passed to the second steam which is used to recognize actions.

In [17], Bertasius et al. design and propose a convolutionl neural network EgoNet for action-object detection from first person data. Action-object is an object that triggers conscious visual as well as motor signals. The proposed model uses two types of information extracted from the data. One is the visual appearances of an object which through red-green-blue (RGB) coloured images of the object. Another information is the 3-dimensional (3D) information such as depth and height of the object with which the person is having some interactions. The proposed model combines these two information to detect action-object within video frames. The authors also present a dataset of RGB first person action-object which contains objects interaction that occurs during 7 activities. The datas are frames from videos captured by two participants. For the experiments, the EgoNet method is used on First Person Action-Object dataset, GTEA Gaze+ dataset [18] and SocialChildren Interaction [19] dataset. This work also includes a human-study where five human performs action-object detection on these datasets. The evaluation metrices that are used for performance are average precision (AP) and maximum F-score (MF). The results show that, while EgoNet does not outperform human detection accuracy, it performs better than other currently available action-abject detection methods.

In [20], a novel deep learning approach is discussed to trail a person's day to day activities from a new dataset. The dataset is collected with neck fitted smartphone camera. A number of 40,103 photos of two people's daily activities of 6 months is used to train the new model addressed as 'A late fusion ensemble'. To maintain the privacy and reduce the storage, the wearable camera is designed to snap image over a specific pause of around 30 seconds to 60 seconds. User had access to label data and remove private data of a day with an annotation tool that automatically received data and displayed in sequential order. Data is labeled into 19 activity classes such as working, family, reading, hygiene etc. The ImageNet [21] dataset is used here to fine-tune the dataset due to the shortage of data and the Caffe [22] convolutional framework is used to build model. K-nearest neighbors (KNN) doesn't give the best result where convolutional neural network (CNN) giving a good accuracy, over-fitted the model due to the lack of contextual information. Random decision forest (RDF) on contextual metadata solves the CNN over-fitting but doesn't increase accuracy. CNN probabilities merged with RDF trained with contextual metadata and color histogram shows the best result to decide the selected tasks.

In another work [23] different neural networks are used for recognising activity and detection of abnormal behavior for elderly people with dementia. Three types of Recurrent Neural Networks (RNNs) are used in this work. These networks are Vanilla RNNs (VRNN), Gated Recurrent Unit (GRU) and Long Short term RNNs (LSTM). For the experiments the dataset of Van Kasteren [24] is used. This dataset has sensor data for daily life routine such as cooking, leaving home, sleeping, etc. The authors use publicly available Theano's and Keras Deep Learning libraries for the simulation. Several other (eg: LSTM) machine learning algorithm is also used to find out their accuracies. Finally the results are compared to see which algorithm works better in which condition. The results shows that LSTM works better some of the cases but the authors come to conclusion that RNNs are the option for recognition activity and detection of abnormal activity.

In [25], K. P. Sanal Kumar et al. tries to find the best way of task identifying using classifiers. Human activity recognition from self-centered videos can be used for monitoring patients. The authors of [25], extract local binary pattern (LBP) and gray level co-occurrence matrix (GLCM) which are textual features. The proposed work also extract a feature named speeded up robust features (SURF). Ambulation, daily tasks, office work and exercise are selected as the main four level categories. Each category holds 20 second level categories. After extracting the features from the video dataset, SVM, probabilistic neural network (PNN), kNN and combination of SVM and KNN are applied individually. From the trained classifiers task is recognised. In the presented work, 10 fold cross validation is applied. SVM with KNN classifier for GLCM feature shows the best result comparing with the accuracy of different applied classifier on different features.

In [26], Yan et al. present a noble multitasking learning framework to recognize daily activities from "first person vision" (FPV) using a wearable camera. The authors tries to solve the problem of huge data generated from continual recording using multitask clustering. It is more effective to work with the collected data from different individuals rather than each person independently. Home environment dataset and office environment both datasets are used in the model for generalization. These datasets are pre segmented into 15 minutes video clips. Data collection form wearable cameras are more challenging than motionless camera. Two effective algorithms for optimizing the problem are Earth Mover's Distance Multi-task Clustering (EMD-MTC) and Convex Multi-task Clustering (CMTC). EMD-MTC guarantees to cluster every single task data and shows the best performance in kernel version. CMTC makes sure the consistency among tasks and shows the related tasks in the corresponding situation.

In [27], a study is presented on the problem of personal context recognition from images captured by wearable devices. A personal context consists of one or more activities that may or may not be related to the context. Therefore, this is also related to the daily activity recognition problem from images captured by wearable devices. For the problem, a dataset is created using videos of five contexts which are car, TV, home office, coffee and machine, office. These videos are taken by wearable devices and the context are relevant to the task of routine analysis. Since different devices are used, the effects of device-specific properties on the recognition of personal context are assessed.Two device-specific properties are field of views (FOVs) and wearing modality. The dataset is built using egocentric videos captured by a single user. Four different types of devices are used, one smart glasses Recon Jet (RJ), two ear-mounted Looxcie LX2 and a wide-angular chest-mounted Looxcie LX3 video cameras. The RJ and one of the LX2 had narrow FOVs (70 and 65.5 degrees respectively), the other LX2 is equipped with a wide-angular converter and LX3 had larger FOVs (100 degrees). The training set is made using videos which are around 10 seconds long and of the five contexts. There is one video per context. For the test set, three to five videos of medium length (8 to 10 minutes) of the given contexts is used. Some short videos are also collected as negative samples which did not represent any of the five contexts. For training, all the frames of training videos and for testing 1,000 frames for each class evenly sampled from the testing videos are used. Four Independent device-specific dataset are made for comparing the four devices. Three type of representation was used for the videos :

a) Holistic representation: Mainly used for scene classification. The GIST descriptor which have 512 dimensions, was used for this.

b) Shallow representations: Using SIFT descriptor and Improved Fisher Vector (IFV) to encode the SIFT features, a Gaussian Mixture Model (GMM) was trained to create a representations ranging from 40960 to 83968 components.

c) Deep Representations: Features extracted from using three different publicly available convolutional neural networks (CNN) is used to create deep representation of 4096-dimensions.

For the experiments, a one class SVM (OCSVM) is used to classify the negative samples in to one class and all the other samples were used as input for a regular multi-class SVM (MCSVM). The MCSVM classified the input samples into 5 context classes. The experiments show that deep features outperform any other representation methods. The results also show that devices with larger FOVs has advantages over devices with narrow FOVs for modeling personal contexts recognition. It is also noted that, most of the errors were because of the negative samples which indicates that a good rejection mechanism is important for building effective systems that can not only separate known contexts, but also can recognize negative samples.

Moments in time [28] is a dataset with one million videos of dynamic events where each video is labeled with one class from 339 classes. This dataset is created for event and action recognition relate to not only humane but also objects, animals, natural phenomenon. The videos are 3 second in duration containing visual and in some cases auditory events. Most commonly used verbs such as 'opening', walking, 'driving' in English language is used as vocabulary for the classes. Each verb has more than 1,000 corresponding videos. From a list of 4,500 verbs, using K-means clustering and based on features containing information about the meaning and usage of the verbs, verbs with similar meaning is clustered. From each cluster, verbs with the highest frequency of use is taken to make the classes of the videos. The videos are collected from sources such as Youtube, Flicker, Bing, The Weather Channel and some other sources. A 3-second section is cut from each video and grouped with corresponding verbs. The groups are annotated by amazon mechanical turk (AMT). For the experiments 802,264 videos from the dataset are used as a training set, a validation set of 33,900 videos and 67,800 videos are used as a test set. Spatial, auditory, temporal and Spatial-temporal modality are used to train some action recognition models for action or event classification. Top-1 and Top-5 accuracy are used for scoring. Then all three modalities is combined using an ensemble model which is created using the best performing models of each modality. The ensemble model shows Top-1 accuracy of 31.16% and Top-5 accuracy of 57.67% which is better than the other models which are used separately.Additionally, cross dataset transfer experiments is done by using Kinetics dataset and moments in time dataset to pretrain two action recognition models. The models are tested on UCF101, HMDB51, and Something-Something datasets. The model pre trained using Moments in TIme dataset shows similar performance to the model pre trained using Kinetics when transferring to UCF101 dataset and a better performance of Top-1: 65.9 score, Top-5: 89.3 score when transferring to HMDB51 dataset. The model shows Top-1: 50.0 score, Top-5: 78.8 score when transferring to 'Something-Something' dataset.

In [29], Albukhary et al. propose a model that can detect basic activities in real time effortlessly. The model is effective for security camera like Closed-circuit Television (CCTV). The proposed method in [29] detects moving objects from collected video and analyze activity. Moving objects are detected from pixel by pixel differentiation in instance of background of current frame. Authors of [29] use geometric attribute with boundary box and color information to identify moving objects. Color histogram is used to differentiate the objects with one another. This paper identifies five activities: walking, running, sitting, standing and landing. Height and weight ratio along with centroid points are used to detect activity. In [29], the resolution is maximized to 384x586 for highest 3 objects in a frame with up to 61ms per frame processing speed.

2.2.2 Recognising Steps of Activity and Prompting

Boger et al. [30] develop a partially observable Markov decision process (POMDP) model for a specific activity of daily living or ADL (eg: hand washing). In [30] the COACH [31] have been modified in a decision theoretic fashion. Although the model is designed for one specific ADL, it can used for other ADLs.In [30], authors describe both fully and partially observable Markov decision process (POMDP) models of a specific ADL which is "hand washing task". Boger et al. in [30] propose a system that helps a dementia patient in completing a task by monitoring their approach to the task and offer assistance by giving prompts or reminders. This [30] model is the elongated version of COACH [31] which uses computer vision to learn to associate hand position (2D coordinate) with specified steps of handwashing and uses audio prompt (e.g. turning on/off water) for each possible step of the task. POMDP deals with partial observability and initiate the competency to plan a suitable course of action. The POMDP model calculates the moving probability from one state to another state for a taken action. In a stochastic observation model, there is an observation probability in a state. There is a reward for transition of one state to another by action. For a particular POMDP, Boger et el. in [30] try to identify the policy for maximizing the expected sum of rewards attained by the model. The authors calculate the reward function based on various state transitions. A large reward (+300) is given if a user can complete the full task and tiny reward (+3) are related to the completion of each step for encouraging even partial success.For the evaluation procedure, the solved MDP has 12 variables, 20 actions and a state space size of 25,090,560 states. State space size become 50,181,120 in the same action space because of the additional variable in POMDP. Authors in [30] simulate both model and in simulation POMDP model outperform the MDP model. For evaluating the result, six scenarios of guiding Dementia patients containing three human and three MDP are selected by 30 professionals. In different five criteria identification, detail, time,repetitions and overall effectiveness based on a five-point Likert scale each caregiver rate the used strategy of every scenarios. Though professional caregivers outperform the MDP model in all aspects but the performance suggest that the MDP model can work as a additive assistive device.

Yi Chu et al in [32] introduces a model for persons with cognitive disabilities by using of interactive activity recognition and prompting system. The main objective of the model is to help the user from prompts or questions to resume a daily task that has been interrupted in their daily schedule. This model [32] is built upon a hierarchical partially-observed Markov decision process (POMDP) which runs with the help of MDP. This system is also able to identify complex prompting behavior. It uses some device such as RFID object touch sensors, appliance operating sensors and motion sensors as input device. One of the main features of this system is able to reduce uncertainty for the user by asking the user what he/she is doing. In this system there are several algorithms are used

used like as Bellman-Ford algorithm, Hidden Markov Model (HMM), Selective-inquiry based Dual Control algorithm (DC-SC), Viterbi algorithm. In the evaluation part, this system has gone through two sets of experiments one is Adaptive option and the other one is Unified control model. Both cases they have founded a very good result. The user responds around 85 % of the time Within 5 steps of a prompt. Conducting the experiment of Unified control model the system has considered a HMM with 3 states one is breakfast (BF) 2nd one is take medicine (TM) and the last one is no activity (NO). In the system demonstrate two students has been participating as volunteer. The system also pass the test to guide the agent. This system can instruct the user when the task should start, finish or resume the task.

Jesse Hoey et al, in [33] demonstrates a method for the dementia patients. This method has been generating a working POMDP prompting system. This method also recognition activity and for complex tasks context-sensitive prompting systems. In some of the recent works, COACH system is used to try and solve this problem. A POMDP is used for COACH system. This model uses some effectors and sensors for analyzing ongoing activities and prompting the users. This model is tested on a task of making a cup of tea in a particular kitchen. This model generates automatically a POMDP-based prompting system. Prompting system is using a description which captures the circumstances of a selective user, task and environment. The description describes the subjective and environmental forerunner for every step of the task. This model uses some video as the input which is third person view. Then analyses of the problems with dementia have with kitchen tasks. For the evaluation part two volunteers were asked to prepare a cup of tea. Most of the time the model can recognize the behavior, move alter etc.

In [34], Karaman et al. propose a model from wearable cameras to monitor people struggling with dementia. The model is established using video data based on HMM. Data is collected through a mobile to acquire the best quality. The mobile device is placed in shoulder. It is found that from shoulder view can solve the problem of unnecessary motion without disturbing the user. Karaman et al. [34] use Fish-Eye lens camera that covers a beneficial angle of 150 degrees. Camera motion detector (CMD) method extracts the details of motion related data. MPEG-7 color layout descriptor (CLD) takes color information to recognize details in case of blur images during movement. A cut histogram cuts the frames after specific time duration and cuts numbers are evaluated by temporal segmentation based on motion. Bag of feature approach is used for image based localization. Localization detects whether the works are happening in their respective place or not. A HMM taxonomy of states divide the stages for step by step understanding of activity. Hierarchical HMM holds states related to semantic activities like “making coffee” in the upper level. Elementary states are in lower level. The model gives highest accuracy with 5 states of HMM, cut histogram and localization. It is quite challenging to train the model as the huge diversity of home environments.

In [35], Barnan Das et al. proposed an automated prompting system called “Prompting Users and Control Kiosk” (PUCK). The authors also discuss about different challenges of building such a system. The objective of this system is to learn the timing for when to give prompts. The prompts are given during the occurrence of an activity. The proposed system PUCK is a framework that has a lot of sensors to collect data. It also includes components for data preparation, machine learning algorithms and also prompting device. This system does not rely on user feedback which allows it to better imitate human interventions such as the caregivers instruction to a dementia patient. An open source Python Visualizer software is made for annotating the data collected from the sensors. The visualizer shows where the sensors are and whether a sensor is active or not. The experiments is done on a set of eight Activity of daily living (ADL), each of which is subdivided into different number of steps performed by some participants. All the steps of an activity is considered as individual training instances. An instance for which a prompt is given is of class 1 and if no prompt is given then the instance is considered of class 0. As performance metrics true positive rate (TPR), true negative rate (TNP), area under curve (AUC) is considered. As learning models J48 decision trees, SVM using sequential minimal optimization (SMO), LogitBoostboosting are used. The authors discussed that one of the challenges of building an automated prompting system is the imbalanced data. Oversampling and undersampling can be used to handle imbalanced data but they have some drawbacks. SMOTE method is a combination undersampling and oversampling. The over sampling is done by generating new instance instead of simply replicating them. In this work a new approach is discussed about an algorithm which is a variation of SMOTE. The experiments show that the classification models using proposed SMOTEvarient approach performs better than the same models using cost-sensitive learning (CSL) on the dataset.

2.2.3 Other Related Works

Older people and patients of dementia face difficulties remembering certain information and may have to take help from others.To reduce taking help from other people,in [36] the authors presented a design of Fiducial Marker Tracker (FMT)—a real-time capture and access application that opportunistically captures video clips of objects the user interacts with. Recalling “Objects States” such as ”If the door is closed or not” and “Routines or Interactions” such as watering Plants, taking Medicine, etc are the two most common types of information for which older adults used memory aids. In [36], the authors made ”Fiducial Marker” and attached those markers around important objects they want to know the state of, or the objects that are related to certain activities they want to remember related information of, in a later time. In the system, when the markers (objects of interest) come across to the body-worn camera’s field of view it starts capturing video until the markers are no longer visible(for 3 seconds) to the camera. When the patients need memory aid about something, they open the FMT-APP and open the videos of the last few interactions with the object labeled by the markers. Thus they try to remember things and it can help older adults determine if they have completed certain actions with these objects and what their states are.

Another work [37] is done where the authors proposed the design of a system that aids the memory of dementia patients and older adults using audio-visuals in order to connect with their loved ones. Additionally, a daily task reminder system such as when the next medication is due, doctor's appointment, appointment with friends and other loved ones amongst others. The proposed design [37] of an assistive system is based on Internet of Things' concept. The authors also discuss current technological solutions for older adults and dementia patients. One of them is Basis B1 smartwatch Which is used to monitor the health of dementia patients by tracking their sleep, heart rate and stress levels. Another one is Computerized help and information project (CHIIP) that consists of a near field communication (NFC) tag and an android app, and various reminders can be set by the patient's caregiver or the patient, which makes it particularly beneficial for dementia sufferers. The third one is Tech@home which works by using magnetic sensors to monitor the environment in which a patient is domiciled. In this work [37] a system is made where a doctor or a nurse from a patient's hospital could upload detailed prescribed medication which is synced to the cloud while the patient gets reminders when they need to take their medication. This system will send a reminder (emails or text messages) to the mobile device using a cloud-based e-mail server. This system can be implemented using an audio-visual display which is connected to a Raspberry Pi system. An easy-to-upload interface is also provided, which could be used to upload inputs like appointments, medications, visuals, audios. A voice module was also provided to allow healthcare workers to create voice recordings of medication intake (type and amount of medication to take and when), or loved one's messages.

In [38] a system named DeepMon is proposed that will make it feasible to run many vision-based deep neural network(DNN) models on mobile devices. The authors [38] state that it is seen that in many models, much processing time is used in the CNN part which makes the prediction slow. To reduce the CNN part's processing time DeepMon is proposed. This is a system that uses CNN layer decomposition, which is a technique to reducing layers with redundancy weights, smart caching convolutional layers by using intermediate convolutional layers's output of the previous frame, and other optimization technique including implementation and kernel-based optimization that ultimately reduce the model's latency by fastening the CNN part. In using the DeepMon system, the model losses some accuracy but it reduces CNN bottleneck. The authors run VGG-VeryDeep-16 on Galaxy S7 which took 644ms to classify image which normally takes 100 seconds to run on S7. Their system can also automate model conversation from many normal desktops based model to mobile GPU(OpenGL) based model.

2.3 Summary

A summary of the reviewed literature is given in the following table 2.1:

Table 2.1: Summary of the reviewed works

Author[Ref]	**Recognising Activity**	**Recognising Steps of Activity**	**Prompting**
Pirsiavash et al. [13]	Linear SVM	–	–
Kenji Matsuo et al. [14]	SVM, Leave-one-out Cross validation	–	–
Zhan et al. [15]	SVM, Knn, LogitBoost	–	–
Sudhakaran et al. [16]	LSTA (modified LSTM)	–	–
Bertasius et al [17]	EgoNet (CNN)	–	–
Damla Arifoglu et al. [23]	Three RNNs (VRNN, GRU, LSTM)	–	–
D Castro etal. [20]	Ensemble model (CNN + RDF)	–	–
K. P. Sanal Kumar et al. [25]	SVM with KNN, 10 fold cross validation	–	–
Yan et al. [26]	Earth Mover's Distance Multi-task Clustering (EMD-MTC), Convex Multi-task Clustering (CMTC)	–	–
Boger et al. [30]	–	POMDP	Audio prompt for different steps of an activity
Yi Chu et al in [32]	–	Bellman-Ford algorithm, HMM, DC-SC, Viterbi algorithm	–
Jesse Hoey et al. [33]	–	POMDP	Text prompt for different steps of an activity
Karaman et al. [34]	–	Hierarchical HMM	–
Barnan Das et al. [35]	–	–	J48 DT (Weka), SVM(using SMO), LogitBoost

2.4 Overall Observation

In the current chapter we review the some of the existing methods on activity recognition, step by step recognition for a single activity and prompting systems. Our overall observation on these works are:

- SVM is used in many of the works for activity recognition but recent works try to use different neural network models such as CNN, RNN for recognising activity.
- It is also understandable that for recognising activity from first person view, the camera in use for data input has a big effect on the performance of the models.
- POMDP is widely used for recognising steps within an activity. In some works HMM is also used.
- There are some existing work for prompting system using machine learning model on sensor based data but prompting using video based data is very few.

Chapter 3

Project Design

In this chapter, we present the methodology and system architectural design of our project.

3.1 Design

We present the architectural design of our system which is shown in 3.1. We pre-train an action recognizer model from the real time data. The video data is collected through a camera. A task list with completion time is stored in server. The model is trained after learning a user's daily activity for a few days. After these the model reminds forgotten tasks to a user in real time.

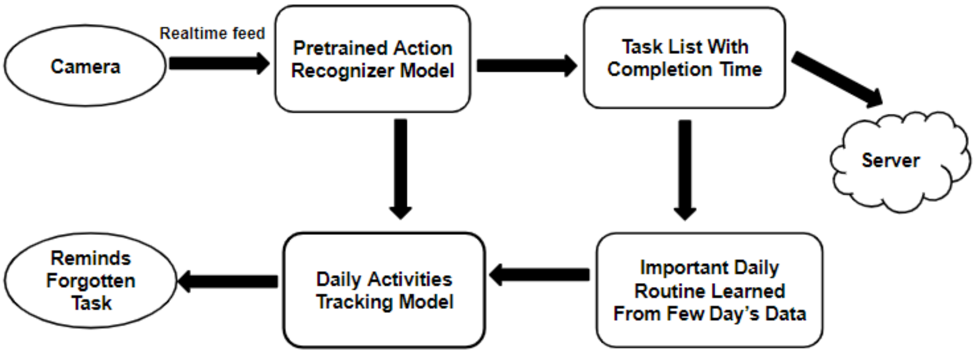

Camera
Realtime feed
Pretrained Action Recognizer Model
Task List With Completion Time
Server
Reminds Forgotten Task
Daily Activities Tracking Model
Important Daily Routine Learned From Few Day's Data

Figure 3.1: System Architectural Diagram

3.2 Methodology

In this section, we discuss the data collection, data pre-processing, and the models of the system. The collected data is used to train RNN, LSTM model. For achieving state-of-the-art accuracy on activity recognition, we have used MobileNetV2 [12] for feature extraction from the video frames. This trained LSTM works as the pre-trained model that is used in the system for recognizing activities from real-time video data. The complete system can be divided into two major parts:

1. Activity recognition model

2. Activity reminder model

3.2.1 Data collection and pre-processing

We have collected some video data of different activities and constructed a data set for this work. In order to collect these data, a Jetson Nano and a night vision camera is used. The data is collected by taking videos consisting of a set of activities in it. These videos are taken at 8 frames per second (fps) rate and at a resolution of 512*512 pixels.

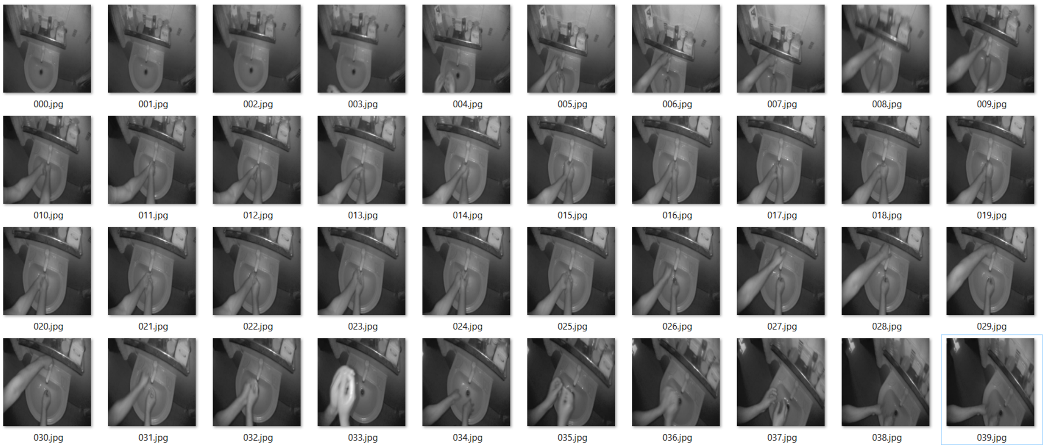


Figure 3.2: Data instances of ”Hand washing” activity

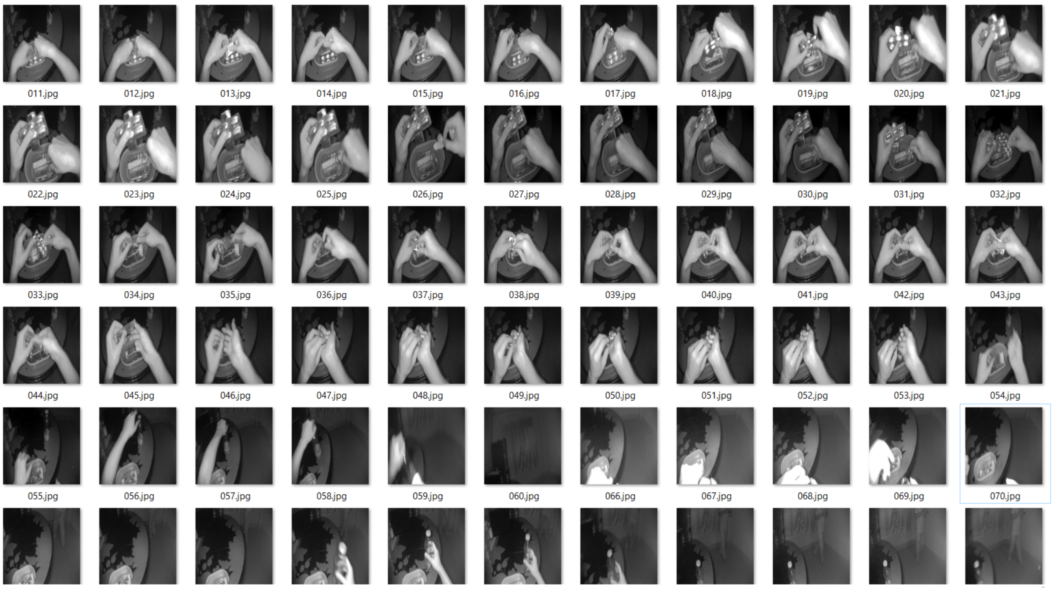


Figure 3.3: Data instances of ”Taking medicine” activity

We reduce the videos to 4 fps. We remove alternative frames from original 8 fps data to make 8 fps to 4 fps video.

The next step is to clean the data by splitting videos into small segments. Each segment consists of only one activity. Based on the length of activities these videos produce different number of frames for each activity.

The following step is to extract features from the videos. MobileNetV2 is used to extract features from each image frame of each video. After extracting the features of the frames, the features of the frames are grouped in sequences of 5 seconds each which produces 20 frames for 4 fps video in each sequence. If an activity has n number of frames, we can get N number of 5 seconds sequence from it where,

$$N = \frac{(n - 16)}{4}$$

For instance, if an activity has 100 frames in it, we create 21, 5-second sequences out of it. The following figure 3.4 shows our segmentation process for a video with 28 frames.

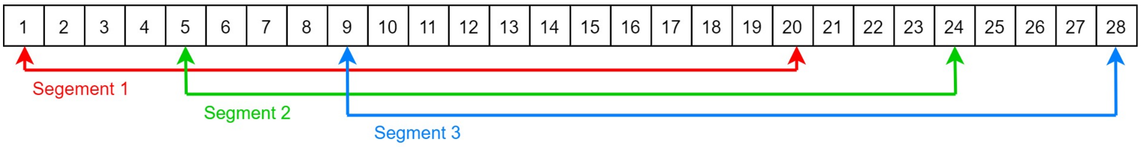


Figure 3.4: Video frame segmentation

Some recorded activities had an odd number of frames. In these cases we added the last frame 1 to 3 more times to get the maximum number of data instances with 20 frames. Following the aforementioned process resulted in making a dataset with 1908 instances. Thus we create our final data set. We collect data on the following list of activities:

Table 3.1: Activity list of the dataset

No.	Activity Name	No.	Activity Name
1	Checking Fridge	8	Putting into Fridge
2	Closing Door	9	Reading
3	Drinking Water	10	Taking from Fridge
4	Hand Wash	11	Taking Medicine
5	Hand Wipe	12	Tooth Brush
6	Opening Door	13	Using Oven
7	Phone Charging		

3.2.2 Activity recognition model

This first part of the system is the most decisive for the model's performance. It takes real-time video as input. The video frames are then split into segments of 5 seconds. MobileNetV2 [12] extracts features from these frames for the next step where an LSTM model predicts the activities. We train a simple LSTM model on the dataset and achieve a great success rate on activity recognition. We focus on keeping the model as small as possible as it is supposed to be run on Jetson nano and yet achieve a decent accuracy. These predicted activities are saved in a file ”Activity list”, with their corresponding name and occurrence time. These tasks are listed in the ”Activity list” file when the LSTM model recognizes those activities.

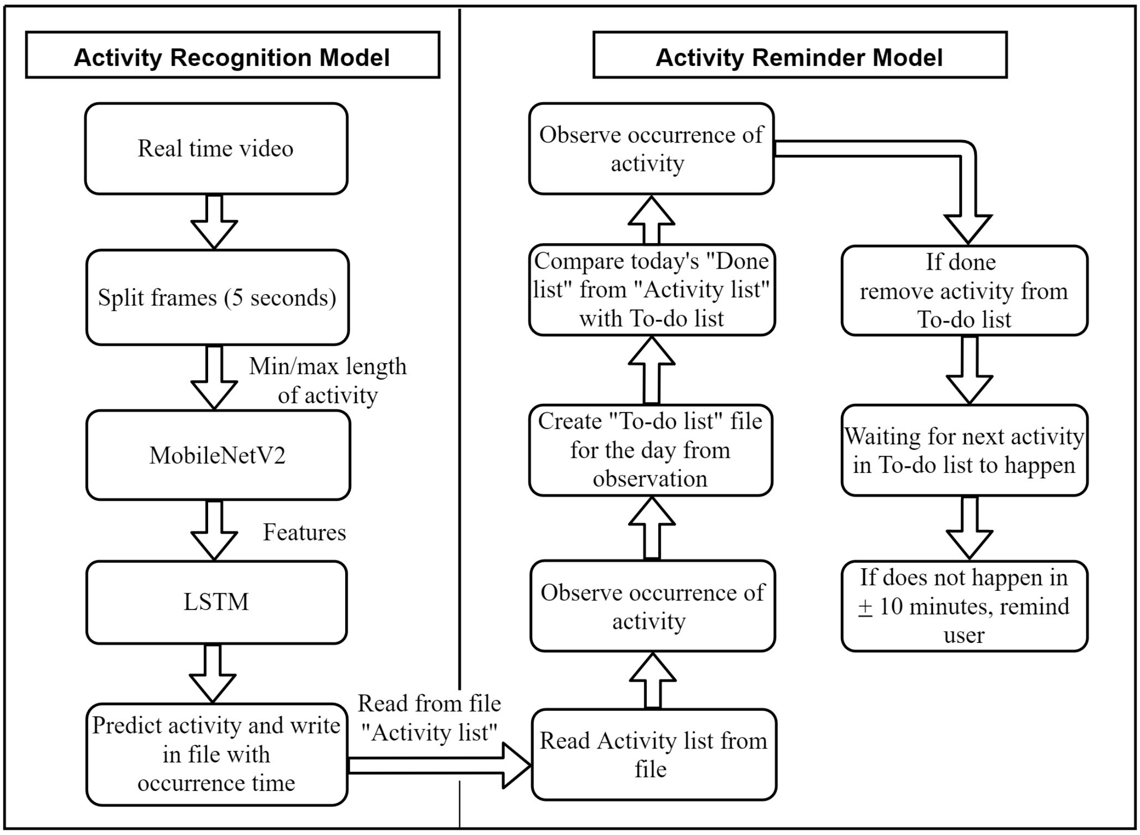

Activity Recognition Model
Real time video
Split frames (5 seconds)
Min/max length of activity
MobileNetV2
Features
LSTM
Predict activity and write in file with occurrence time
Read from file "Activity list"
Activity Reminder Model
Read Activity list from file
Observe occurrence of activity
Create "To-do list" file for the day from observation
Compare today's "Done list" from "Activity list" with To-do list
Observe occurrence of activity
If done remove activity from To-do list
Waiting for next activity in To-do list to happen
If does not happen in ± 10 minutes, remind user

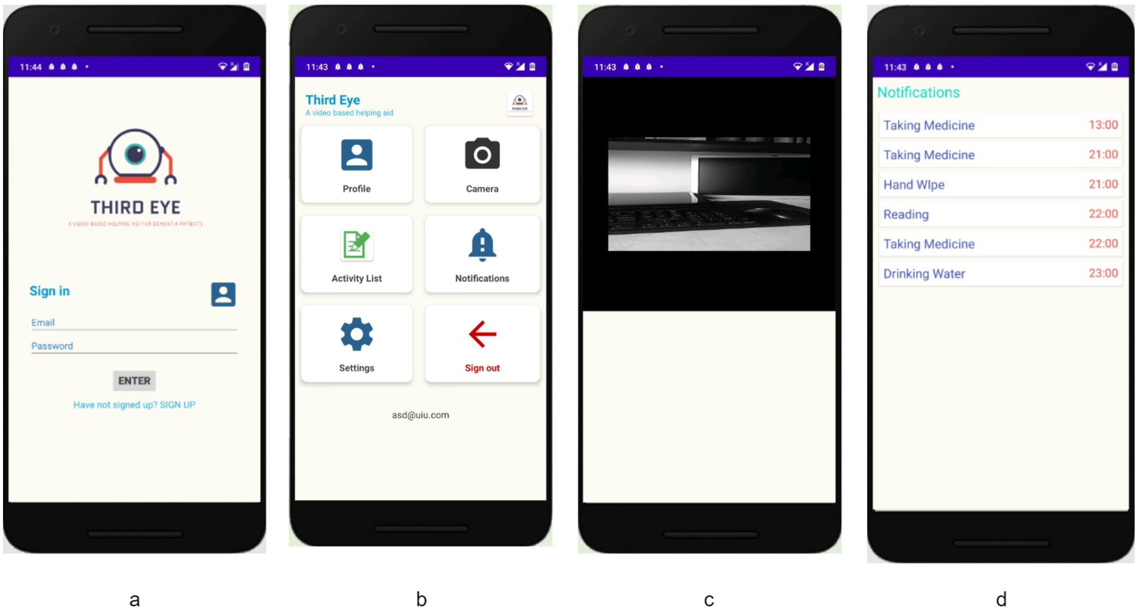
Figure 3.5: Model diagram

3.2.3 Activity reminder model

The activity reminder model reads the “Activity list” file created by the activity recognition model. Based on the occurrence of activities of the last N days, a ”To-do list” file is created with a list of activities to complete in a single day. This file contains the most frequent activities which have been done by the dementia patient in the last N days in the different timestamps. The model then compares the “Activity list” file with the ”To-do list” file. If an activity in the ”To-do list” occurs, then that activity is removed from the ”To-do list”. The model waits for the next activity to be completed. If the system does not detect the occurrence of an activity within a time frame, the model provides a reminder for the user.

3.2.4 Mobile Application

A mobile application is introduced that runs on android operating systems to implement the software model of our project. Figure 3.6 shows some pages from the application. A user can create a personal account where his/her daily activities are stored. This app ask permission for camera usage. As the videos taken by the camera are not stored but the predicted activities are written in a file for reducing memory utilization, this app allows users to see their activity routines and past notifications.



Taking Medicine	13:00
Taking Medicine	21:00
Hand WIpe	21:00
Reading	22:00
Taking Medicine	22:00
Drinking Water	23:00

Figure 3.6: Application user interface. (a) Sign in interface. (b) Home page. (c) Camera interface. (d) Upcoming notification page.

3.3 Summary

This chapter is focused on the elaborated discussion on the procedure of data collection, and data pre-processing of our research work. We also discussed the workflow of the two major parts of the system here.

Chapter 4

Implementation and Results

This chapter focuses on the environment setup, experimental analysis, evaluation and results on our work.

4.1 Environment Setup

Our project is focused on human's daily activity. But there was no dataset available that was suitable for our intended project to work on. As a result, we decided to build up our small dataset for completing the project. We collected a Jetson Nano from our University and set up the Linux OS provided by Nvidia which it operates on, and the necessary libraries to use the Jetson nano's camera to collect some image frames while doing some activities around the home.

Our own collected daily activity dataset which after feature extraction using MobileNetV2 becomes a larger dataset (using 16bit float) containing around 10 GB of data. We initially tried to use Google Colab for our data processing and model training but google colab's 12 GB RAM was not enough. We then used the UIU GPU server which had an Nvidia Titan XP with 11GB of VRAM and a 32GB of available system RAM which was enough for our work. We created a conda environment on the server with our necessary deep learning libraries such as TensorFlow, Keras, and used those libraries for creating our model.

For deploying our model on the android application we had to convert deep learning models into TensorFlow Lite since android applications don't support Tensorflow code in it.

We implemented the machine learning models using an application called ”The Third Eye”. For developing the application we used the Android Studio IDE. The application is written in the Java programming language. While developing the application, we wanted to make it compatible with the majority of android devices. The application can be installed in any mobile device that runs on any version of android from 6.0 (Marshmallow) to 10.0 (Q). The device must also have a functioning rear camera for video input.

4.2 Evaluation

For evaluating the performance of our model in recognizing activities we used accuracy, precision, recall, and F1 score as evaluation metrics. We can define these metrics as follows.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.1}$$

$$Precision = \frac{TP}{TP + FP} \tag{4.2}$$

$$Recall = \frac{TP}{TP + FN} \tag{4.3}$$

$$F1 - score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{4.4}$$

Where, TP denotes true positive, and TN, FP, and FN denotes true negative, false positive, and false negative respectively.

4.3 Results and Discussion

Our model's result mainly depends on user satisfaction. How well the system helped the dementia patient matters the most. Recognizing a person's activity accurately and then finding patterns of recurrent activities is the main target. Since the system is divided into two major parts, we describe the results for both parts separately which are as follows.

4.3.1 Result analysis of Activity recognition model

Our dataset has 13 activities collected from the first-person camera view. We collected data for each activity multiple times. The second column of table 4.1 shows the number of 4fps videos we collected for each type of activity. The third column shows the number of frames of each videos. Finally the last column shows the total number of 20 frames segments created from all the videos of a single activity. We split this dataset into three subsets of the training set, testing set, and validation set. This splitting has been on the ratio of 70:15:15 for training, validation and testing respectively. Table 4.2 shows the number of data instances of each set.

To recognize activities from video data, we have tried different deep learning algorithms implemented in TensorFlow. A two-layer architecture of SimpleRNN, LSTM, and GRU with different combinations of 8, 16, 32, and 64 units has been used. with the increment of the number of units in each layer, the accuracy of these three models increased significantly. A simpleRNN(64, 64) model achieves an accuracy of 86.81% in recognizing activities from video data and the LSTM (64, 64) achieves an accuracy of 98.96 % with the precision and recall of 97.31% and 98.65%. As the GRU works better than LSTM on a small amount of data, we have tried that also and have gotten a noteworthy improvement in result.

Table 4.1: Dataset summary

Activity	#Videos	Number of frames in each video	#Segments
Checking Fridge	2	46, 36	13
Closing Door	11	51, 50, 55, 51, 53, 63, 88, 39, 32, 30, 35	96
Drinking Water	10	56, 28, 42, 41, 45, 73, 63, 88, 43, 48	95
Hand Wash	20	69, 53, 84, 56, 54, 67, 66, 54, 76, 52, 128, 38, 64, 98, 66, 91, 55, 39, 36, 46	249
Hand Wipe	11	53, 37, 54, 53, 82, 56, 75, 65, 67, 57, 56	125
Opening Door	14	50, 50, 36, 31, 52, 51, 41, 41, 52, 30, 32, 29, 39, 40	92
Phone Charging	6	40, 29, 36, 49, 36, 79	45
Putting into Fridge	9	69, 52, 41, 28, 56, 61, 67, 68, 53	91
Reading	11	307, 71, 181, 199, 104, 165, 94, 153, 158, 205, 145	407
Taking from Fridge	8	47, 44, 34, 36, 46, 61, 73, 91	79
Taking Medicine	13	73, 103, 94, 102, 124, 165, 122, 71, 124, 118, 114, 111, 150	321
Tooth Brush	7	67, 61, 74, 73, 86, 93, 37	99
Using Oven	11	76, 116, 48, 98, 93, 106, 83, 79, 68, 113, 68	196

Table 4.2: Data distribution in training, testing, and validation set

Set	Number of instances
Training	1332
Validation	288
Test	288

Table 4.3: SimpleRNN results in activity recognition

Model	Accuracy	Precision	Recall	F-1 Score
SimpleRNN-64-64	0.868056	0.775827	0.772371	0.765800
SimpleRNN-32-32	0.854167	0.734825	0.734873	0.732879
SimpleRNN-64-16	0.829861	0.726689	0.707713	0.683513
SimpleRNN-32-16	0.826389	0.704967	0.663202	0.663928
SimpleRNN-64-32	0.815972	0.714387	0.677153	0.668333
SimpleRNN-64-8	0.791667	0.712781	0.632096	0.633063
SimpleRNN-32-8	0.746528	0.646299	0.611388	0.602213
SimpleRNN-16-8	0.704861	0.593476	0.536620	0.481168
SimpleRNN-16-16	0.513889	0.172657	0.243296	0.190031
SimpleRNN-8-8	0.392361	0.139428	0.156787	0.099403

Table 4.4: LSTM results in activity recognition

Model	Accuracy	Precision	Recall	F-1 Score
lstm-64-64	0.989583	0.9731098	0.986501	0.978464
lstm-32-16	0.979167	0.9760684	0.972247	0.971013
lstm-32-32	0.975694	0.9704925	0.938276	0.947412
lstm-64-16	0.972222	0.9650812	0.952564	0.954164
lstm-64-32	0.961806	0.9501006	0.911384	0.912281
lstm-64-8	0.947917	0.8550789	0.872294	0.855139
lstm-16-8	0.920139	0.8254995	0.824685	0.821376
lstm-32-8	0.881944	0.7859527	0.798621	0.755044
lstm-16-16	0.857639	0.7303337	0.736132	0.723299
lstm-8-8	0.739584	0.5037738	0.553492	0.518183

Table 4.5: GRU results in activity recognition

Model	Accuracy	Precision	Recall	F-1 Score
GRU-64-32	0.989584	0.990029	0.990045	0.989647
GRU-32-16	0.975694	0.961946	0.933974	0.940773
GRU-64-64	0.965278	0.961528	0.939238	0.946582
GRU-8-8	0.965278	0.886693	0.895947	0.890532
GRU-64-16	0.958334	0.944890	0.927966	0.927203
GRU-32-8	0.951389	0.858390	0.876584	0.865784
GRU-64-8	0.947917	0.858022	0.866503	0.857460
GRU-32-32	0.920139	0.901452	0.860325	0.851360
GRU-16-16	0.916667	0.832621	0.836696	0.823794
GRU-16-8	0.684028	0.373142	0.447905	0.388142

GRU(64, 32) model gives a significantly better result of the accuracy of 98.96% as well as precision and recall of 99.00% and 99.01% respectively. All three models have f1-score of 76.58%, 97.85%, and 98.96% respectively. Table 4.3, 4.4, 4.5 shows the results of these models with different numbers of units in each layer.

Table 4.6 shows the best result of these three architectures. Since GRU outperforms other models we used GRU while implementing our application.

Table 4.6: Summary of activity recognition model results

Model	Accuracy	Precision	Recall	F-1 Score
GRU-64-32	0.989584	0.990029	0.990045	0.989647
lstm-64-64	0.989583	0.9731098	0.986501	0.978464
SimpleRNN-64-64	0.868056	0.775827	0.772371	0.765800

4.3.2 Result analysis of Activity reminder model

After recognizing the activities, the next step is to check whether the user is performing the actions. We have created an action reminder(RemAct) model which checkes and reminds if an individual neglected to accomplish something in a specific time span. This model's outcome relies on the activity recognition model's outcome. RemAct model prepares a to-do list file for each day at the beginning of a day based on the activity list file data of the last N days. At the end of each activity completion, the to-do list is updated.

4.4 Summary

This research shows the path for developing an activity reminder which can be used as a replacement of caregivers for dementia patients. By monitoring a person's daily activity for a few days our model will be able to predict a person's next day's activities and if the person misses any predicted activity at a given time our model will notify the user. Although our model got a staggering accuracy of 98.95% in predicting activity recognition the result falls drastically when implementing in the application due to the data variations of the collected pre-trained model and the user devices. Since no work has been done before acknowledging this problem our research opens a new door for real-time activity recognition from video data for creating activity reminder models.

Chapter 5

Standards and Design Constraints

In this chapter, we mention few standards and constraints that are related to our project. We also discussed the challenges we faced doing this project in this chapter.

5.1 Compliance with the Standards

In this section, we mention few standards that are being followed during our thesis.

5.1.1 Software Standard

We are following the coding standards mentioned below-

1. Python - Python enhancement proposal (PEP 8) is used as coding standard. It provides a guideline on writing code in python.

2. Java - Code Conventions for the Java Programming Language is the guideline presented by Oracle for coding in Java programming language.

5.1.2 Hardware Standard

We will use Jetson Nano for our project. The Nvidia Jetson Nano Developer Kit was announced as a development system in mid-March 2019.

5.2 Design Constraints

In this section, we mention the constraints which we are working through in our thesis.

5.2.1 Economic Constraint

Currently, a lot of patients need full-time aid from some other people. Our system helps them to depend less on others which will save them a lot of money.

5.2.2 Environmental Constraint

To train our system we will use some computation power which will emit some extra CO2 to the environment. After deployment, the system will use some electricity to run.

5.2.3 Ethical Constraint

Most of the computation and work of our system will be done in the device itself which will ensure the privacy of the user. We will only save some of the important data of users upon their consent.

5.2.4 Health and Safety Constraint

Our system will not harm any user as we are expecting that our system will not emit any types of radiation.

5.2.5 Social Constraint

Our project will reduce human labor that has been given otherwise to dementia patients as nursing. It will also increase the self stream of the patients.

5.2.6 Political Constraint

As far as we are concern, there are no political issues.

5.2.7 Manufacturability and Cost Analysis

Cameras and hardware devices are becoming smaller every day. With the current technologies in hand, the system can be easily made and it can be cost-efficient.

5.2.8 Sustainability

We expect our product to be sustainable, it will be able to provide a long term service and it will not require frequent updates .

5.3 Challenges

Due to the Covid-19 pandemic situation, we had to change our plan of implementing the model using Jetson nano to implementing the model on a mobile application. Fitting the TensorFlow model in the mobile app due to the lack of experience was challenging for us. Communicating with the team members was also difficult in the pandemic. We had to work from different cities and the internet connectivity problem was a major setback. On top of that, current android devices at the time, do not have the adequate processing power to process TensorFlow model in real-time. Android devices can process TensorFlow lite models, so we had to convert our TensorFlow models to TensorFlow Lite versions.

But this conversion affects the model's strength of recognizing the activities. That is why the performance of the model falls drastically. The TensorFlow models are trained with the data collected from jetson nano and the camera of the jetson nano which has a 77 degree feild of view. On the other hand, the mobile app is taking data from the respective phone's camera. Since different devices have different camera hardware, it hugely affects the model's correctness.

5.4 Summary

We followed the aforementioned constraints and standards during this research work. We tried to build a model that will be effective, reliable, and sustainable. Regardless of the challenges we faced, we were able to build an effective model that matches our goal.

Chapter 6

Conclusion

This chapter contains the summary and the future plan of our thesis.

6.1 Summary

Dementia is a medical condition, which is responsible for the decline in a person's memory and cognitive skill. Through the work presented in this paper, we proposed a system that can help a person with dementia or mild cognitive impairment by detecting his or her daily tasks and acting as a smart reminder as well as a prompting system to help with task completion. A mobile application was developed through which a user's everyday activity can be tracked, and from the past few days of data, our RemAct model can identify a set of important activities for that particular person. If the user misses any activity of a timestamp the RemAct model notifies the person within a time frame of 10 minutes. In recognizing activity from the video data GRU Outperformed other deep learning algorithms by a significant margin. Among other versions GRU(64, 32) model gives a staggering result of the accuracy of 98.96% with precision, recall, and f1-score of 99.01%, 99.00%, and 98.96% respectively. This research shows the path for developing an activity reminder which can be used as a replacement of caregivers for dementia patients. Although our model got a staggering accuracy of 98.96% in predicting activity recognition the result falls drastically when implementing in the application due to the data variations of the pre-trained model and the user devices. Since no work has been done before acknowledging this problem our research opens a new door for real-time activity recognition from video data for creating activity reminder models.

6.2 Future work

Through an android application, we tried to provide a user-friendly tool to tackle dementia. Although our primary aim was building an embedded system using jetson nano and raspberry pi camera due to the pandemic situation we were not able to complete the system using these hardware devices. Since the result of the model falls drastically while implementing in the mobile application due to the lack of processing power, there is a huge chance of improvement in the future with more processing power in hand. With the advancement of science and technology, we can see the ray of hope and can eloquently say that in the near feature mobile phones will be able to provide us the adequate processing power to reach the desired output. We aim to continue with this project and complete the hardware implementation since the jetson nano provides more processing power compared to the mobile phones available now and easy to work with. With the aim of making the dementia patient's life a bit better making a wearable device is the next goal that we hope to accomplish.

References

[1] Xiangang Li and Xihong Wu. Constructing long short-term memory based deep recurrent neural networks for large vocabulary speech recognition. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4520–4524. IEEE, 2015.

[2] Hasim Sak, Andrew W Senior, and Françoise Beaufays. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. 2014.

[3] A. Graves, A. Mohamed, and G. Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6645–6649, 2013.

[4] A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke, and J. Schmidhuber. A novel connectionist system for unconstrained handwriting recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(5):855–868, 2009.

[5] Michael Auli, Michel Galley, Chris Quirk, and Geoffrey Zweig. Joint language and translation modeling with recurrent neural networks. 2013.

[6] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). *arXiv preprint arXiv:1412.6632*, 2014.

[7] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[8] Victor Carbune, Pedro Gonnet, Thomas Deselaers, Henry A Rowley, Alexander Daryin, Marcos Calvo, Li-Lun Wang, Daniel Keysers, Sandro Feuz, and Philippe Gervais. Fast multi-language lstm-based online handwriting recognition. *International Journal on Document Analysis and Recognition (IJDAR)*, pages 1–14, 2020.

[9] Alex Graves, Navdeep Jaitly, and Abdel-rahman Mohamed. Hybrid speech recognition with deep bidirectional lstm. In *2013 IEEE workshop on automatic speech recognition and understanding*, pages 273–278. IEEE, 2013.

[10] Yiming Cui, Shijin Wang, and Jianfeng Li. Lstm neural reordering feature for statistical machine translation. *arXiv preprint arXiv:1512.00177*, 2015.

[11] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[12] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[13] Hamed Pirsiavash and Deva Ramanan. Detecting activities of daily living in first-person camera views. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2847–2854. IEEE, 2012.

[14] Kenji Matsuo, Kentaro Yamada, Satoshi Ueno, and Sei Naito. An attention-based activity recognition for egocentric video. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2014.

[15] Kai Zhan, Fabio Ramos, and Steven Faux. Activity recognition from a wearable camera. In *2012 12th International Conference on Control Automation Robotics & Vision (ICARCV)*, pages 365–370. IEEE, 2012.

[16] Swathikiran Sudhakaran, Sergio Escalera, and Oswald Lanz. Lsta: Long short-term attention for egocentric action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9954–9963, 2019.

[17] Gedas Bertasius, Hyun Soo Park, Stella X Yu, and Jianbo Shi. First person action-object detection with egonet. *arXiv preprint arXiv:1603.04908*, 2016.

[18] Yin Li, Zhefan Ye, and James M Rehg. Delving into egocentric actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 287–295, 2015.

[19] Hyun Soo Park and Jianbo Shi. Social saliency prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4777–4785, 2015.

[20] Daniel Castro, Steven Hickson, Vinay Bettadapura, Edison Thomaz, Gregory Abowd, Henrik Christensen, and Irfan Essa. Predicting daily activities from egocentric images using deep learning. In *proceedings of the 2015 ACM International symposium on Wearable Computers*, pages 75–82. ACM, 2015.

[21] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[22] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture

for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 675–678. ACM, 2014.

[23] Damla Arifoglu and Abdelhamid Bouchachia. Activity recognition and abnormal behaviour detection with recurrent neural networks. *Procedia Computer Science*, 110:86–93, 2017.

[24] Tim LM van Kasteren, Gwenn Englebienne, and Ben JA Kröse. Human activity recognition from wireless sensor network data: Benchmark and software. In *Activity recognition in pervasive intelligent environments*, pages 165–186. Springer, 2011.

[25] KP Sanal Kumar and R Bhavani. Human activity recognition in egocentric video using pnn, svm, knn and svm+ knn classifiers. *Cluster Computing*, pages 1–10, 2017.

[26] Yan Yan, Elisa Ricci, Gaowen Liu, and Nicu Sebe. Recognizing daily activities from first-person videos with multi-task clustering. In *Asian Conference on Computer Vision*, pages 522–537. Springer, 2014.

[27] Antonino Furnari, Giovanni M Farinella, and Sebastiano Battiato. Recognizing personal contexts from egocentric images. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1–9, 2015.

[28] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Yan Yan, Lisa Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, et al. Moments in time dataset: one million videos for event understanding. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[29] N Albukhary and YM Mustafah. Real-time human activity recognition. In *IOP Conference Series: Materials Science and Engineering*, volume 260, page 012017. IOP Publishing, 2017.

[30] Jennifer Boger, Pascal Poupart, Jesse Hoey, Craig Boutilier, Geoff Fernie, and Alex Mihailidis. A decision-theoretic approach to task assistance for persons with dementia. In *IJCAI*, pages 1293–1299. Citeseer, 2005.

[31] Alex Mihailidis, Geoffrey R Fernie, and Joseph C Barbenel. The use of artificial intelligence in the design of an intelligent cognitive orthosis for people with dementia. *Assistive Technology*, 13(1):23–39, 2001.

[32] Yi Chu, Young Chol Song, Richard Levinson, and Henry Kautz. Interactive activity recognition and prompting to assist people with cognitive disabilities. *Journal of Ambient Intelligence and Smart Environments*, 4(5):443–459, 2012.

[33] Jesse Hoey, Thomas Plötz, Dan Jackson, Andrew Monk, Cuong Pham, and Patrick Olivier. Rapid specification and automated generation of prompting systems to assist people with dementia. *Pervasive and Mobile Computing*, 7(3):299–318, 2011.

[34] Svebor Karaman, Jenny Benois-Pineau, Rémi Mégret, Vladislavs Dovgalecs, Jean-François Dartigues, and Yann Gaëstel. Human daily activities indexing in videos from wearable cameras for monitoring of patients with dementia diseases. In *2010 20th International Conference on Pattern Recognition*, pages 4113–4116. IEEE, 2010.

[35] Barnan Das, Diane J Cook, Maureen Schmitter-Edgecombe, and Adriana M Seelye. Puck: an automated prompting system for smart environments: toward achieving automated prompting–challenges involved. *Personal and ubiquitous computing*, 16(7):859–873, 2012.

[36] Franklin Mingzhe Li, Di Laura Chen, Mingming Fan, and Khai N Truong. Fmt: A wearable camera-based object tracking memory aid for older adults. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(3):95, 2019.

[37] David Airehrour, Samaneh Madanian, and Alwin Mathew Abraham. Designing a memory-aid and reminder system for dementia patients and older adults. 2018.

[38] Loc N Huynh, Youngki Lee, and Rajesh Krishna Balan. Deepmon: Mobile gpu-based deep learning framework for continuous vision applications. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*, pages 82–95. ACM, 2017.