Wright State University

CORE Scholar

Browse all Theses and Dissertations | Theses and Dissertations

2022

Building an Understanding of Human Activities in First Person Video using Fuzzy Inference

Bradley A. Schneider
Wright State University

Follow this and additional works at: https://corescholar.libraries.wright.edu/etd_all

Part of the Computer Engineering Commons, and the Computer Sciences Commons

Repository Citation

Schneider, Bradley A., "Building an Understanding of Human Activities in First Person Video using Fuzzy Inference" (2022). *Browse all Theses and Dissertations*. 2566.
https://corescholar.libraries.wright.edu/etd_all/2566

This Dissertation is brought to you for free and open access by the Theses and Dissertations at CORE Scholar. It has been accepted for inclusion in Browse all Theses and Dissertations by an authorized administrator of CORE Scholar. For more information, please contact library-corescholar@wright.edu.

BUILDING AN UNDERSTANDING OF HUMAN ACTIVITIES IN FIRST PERSON VIDEO USING FUZZY INFERENCE

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

by

BRADLEY A. SCHNEIDER
B.S., Morehead State University, 2012
M.S., Wright State University, 2017

2022
Wright State University

Wright State University
GRADUATE SCHOOL

March 22, 2022

I HEREBY RECOMMEND THAT THE DISSERTATION PREPARED UNDER MY SUPERVISION BY Bradley A. Schneider ENTITLED Building an Understanding of Human Activities in First Person Video using Fuzzy Inference BE ACCEPTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF Doctor of Philosophy.

Tanvi Banerjee, Ph.D.
Dissertation Director

Yong Pei, Ph.D.
Director, Computer Science and Engineering Ph.D. Program

Barry Milligan, Ph.D.
Vice Provost for Academic Affairs
Dean of the Graduate School

Committee on
Final Examination

Tanvi Banerjee, Ph.D.

Yong Pei, Ph.D.

Michael Riley, Ph.D.

Mateen Rizki, Ph.D.

Thomas Wischgoll, Ph.D.

ABSTRACT

Schneider, Bradley A. Ph.D., Department of Computer Science and Engineering, Wright State University, 2022. *Building an Understanding of Human Activities in First Person Video using Fuzzy Inference.*

Activities of Daily Living (ADL's) are the activities that people perform every day in their home as part of their typical routine. The in-home, automated monitoring of ADL's has broad utility for intelligent systems that enable independent living for the elderly and mentally or physically disabled individuals. With rising interest in electronic health (e-Health) and mobile health (m-Health) technology, opportunities abound for the integration of activity monitoring systems into these newer forms of healthcare.

In this dissertation we propose a novel system for describing 's based on video collected from a wearable camera. Most in-home activities are naturally defined by interaction with objects. We leverage these object-centric activity definitions to develop a set of rules for a Fuzzy Inference System (FIS) that uses video features and the identification of objects to identify and classify activities. Further, we demonstrate that the use of FIS enhances the reliability of the system and provides enhanced explainability and interpretability of results over popular machine-learning classifiers due to the linguistic nature of fuzzy systems.

Contents

1 Introduction **1**
- 1.1 Need for In-Home Monitoring . . . 1
- 1.2 Activities of Daily Living . . . 2
- 1.3 Research Goals . . . 3
- 1.4 Thesis Statement . . . 6

2 Related Work **7**
- 2.1 Activities Being Detected by Vision Sensors . . . 7
  - 2.1.1 Anomaly Detection . . . 8
  - 2.1.2 Full-body Activities . . . 9
  - 2.1.3 Partial-body Activities . . . 10
  - 2.1.4 Discussion . . . 11
- 2.2 Types of Vision Sensors . . . 12
  - 2.2.1 Fixed RGB Vision Sensors . . . 12
  - 2.2.2 Fixed RGB-Depth Vision Sensors . . . 14
  - 2.2.3 Wearable RGB Vision Sensors . . . 16
  - 2.2.4 Wearable RGB-D Vision Sensors . . . 17
  - 2.2.5 Secondary Sensors . . . 18
  - 2.2.6 Discussion . . . 18
- 2.3 Computational Intelligence Methods . . . 19
  - 2.3.1 Stochastic Modeling . . . 19
  - 2.3.2 Fuzzy Logic and Clustering . . . 20
  - 2.3.3 Bayesian Methods . . . 21
  - 2.3.4 Gaussian Mixture Models (GMM) . . . 21
  - 2.3.5 Neural Networks . . . 22
  - 2.3.6 Other Methods . . . 22
  - 2.3.7 Discussion . . . 23
- 2.4 Key Takeaways . . . 24

3 Methods **25**
- 3.1 Fuzzy Inference Systems . . . 25
  - 3.1.1 Fuzzification and Defuzzification . . . 26

3.1.2 Fuzzy Rules . . . 29
3.2 Genetic Algorithms . . . 29
3.2.1 Chromosome Encoding . . . 30
3.2.2 Reproduction of Individuals . . . 31
3.2.3 Evaluating Fitness . . . 32
3.3 Optical Flow . . . 32
3.4 Convolutional Neural Networks . . . 33
3.4.1 MobileNet . . . 35

4 Experiments and Results **36**
4.1 Gait Speed Comparison from Wearable Camera and Accelerometer Vest in Structured and Semi-Structured Environments . . . 36
4.1.1 Overview . . . 37
4.1.2 Data Collection . . . 39
4.1.3 Signal Processing . . . 42
4.1.4 Statistical Analysis . . . 44
4.1.5 Results . . . 45
4.1.6 Conclusion . . . 46
4.2 Describing Object-centric Activities using Fuzzy Methods . . . 47
4.2.1 Overview . . . 47
4.2.2 Dataset . . . 51
4.2.3 Features . . . 53
4.2.4 Fuzzy Inference System . . . 61
4.2.5 Results . . . 67

5 Conclusions **78**
5.1 Contributions . . . 78
5.2 Future Work . . . 79

Bibliography **79**

List of Figures

1.1 Example of activities with varying spatio-temporal complexity which may occur during the larger activity of preparing breakfast. 4

3.1 Illustration of membership functions, fuzzification, and defuzzification process. 27
3.2 The pixel specified in the first frame at time t, $I(x, y, t)$, translated by (dx, dy) units and is located at $I(x + dx, y + dy, t + dt)$ in the next frame at time $t + dt$. 33

4.1 The motion capture laboratory in which data was collected from participants 40
4.2 Placement of Sensors on Subject and Alignment of Axes Between Sensors . 40
4.3 Mean Plots with Standard Error Bars for Recorded Features By Gait Speed 42
4.4 Illustration of our system. Features such as optical flow motion vectors and color histograms are extracted from video frames and passed to the fuzzy inference system, which calculates fuzzy membership in three outputs based on a series of rules within the system. 51
4.5 A sequence of still frames at 15 frame intervals from a sample activity annotated in the first three frames as ”take plate” and in the latter three frames as ”wash plate”. The change in activity label is an indication of the uncertainty inherent in activities, their definitions, and their spatiotemporal boundaries as they are performed. 52
4.6 Illustration of the distribution of action locations within the training data. . 56
4.7 Loss curves for the final training of the hand detection model on images from the EPIC KITCHENS dataset. 60
4.8 Example first person video frames of the hand detection model. Hands were detected in these images with a confidence ranging between 0.96 to 1.0, highlighting the robustness of the model to different orientations and poses of hands in the view. 68
4.9 Example hand detection result from the EgoHands dataset prior to training on EPIC KITCHENS data. The hands do not move from the vicinity of the table, leading to poor detection in datasets with more variation in background. 69
4.10 Progression of the GA in optimizing the fitness metric (f1-score) over generations of the GA. 70

4.11 Final membership functions output from evolution of the FIS using a GA. . 72
4.12 Video frame from the EPIC Kitchens dataset overlaid with results from a Mask R-CNN pretrained on the MS COCO dataset. Multiple objects have been segmented with more precision than a rectangular bounding box. . . . 75
4.13 Example first person video frame of subject washing spatula with overlaid bounding boxes, action, and object labels 76

List of Tables

4.1 Video Features 55
4.2 Input Variables to the Fuzzy Inference System 62
4.3 Output Variables for the Fuzzy Inference System 62
4.4 Rules for the Output "Interaction" (*HINT*=hand intersection, *OC*=object centrality, *INT*=interaction) 63
4.5 Rules for the Output "Modification" (*INT*=interaction, *CH*=color histogram, *CKP*=CenSurE keypoint similarity, *ORB*=ORB keypoint similarity, *MOD*=modification) 64
4.6 Rules for the Output "Relocation" (*MAG*=difference in motion magnitude, *DIR*=difference in motion direction, *INT*=interaction, *REL*=relocation) . . . 65
4.7 Results of activity classification on first person video 76

Introduction

1.1 Need for In-Home Monitoring

The number of people aged 65 and older in the United States is expected to reach 89 million by 2050, double the number of United States citizens in the same age group in 2011 [49]. As the elderly population continues to grow, so will the need to provide in-home care for the elderly. Elder care has traditionally been lacking qualified, certified practitioners and clinicians, and the demand for in-home health aides is very likely to exceed recruitment rates over this period of growth [85]. This will result in a growing deficit of qualified healthcare workers for in-home care, which will require the development of new models of healthcare delivery.

One response to this deficit of personnel is to leverage e-Health technology to provide in-home healthcare. In a 2005 report, the World Health Organization (WHO) defined e-Health as the "cost-effective and secure use of information and communications technologies in support of health and health-related fields, including health-care services, health surveillance, health literature, and health education, knowledge and research" and resolved to encourage long-term strategic plans to develop e-Health resources, noting that advances in technology have raised expectations for healthcare [1]. Since then, WHO has also formalized the idea of mobile-health, or m-Health, which is similar to e-Health but uses mobile technology. This includes in-home installation of sensors or imaging devices used to monitor patients with chronic illnesses [112]. A recent survey of m-Health tech-

nology found that m-Health has been found as an effective way to monitor elderly patients suffering from dementia and cognitive disorders including screening for cognitive decline and promoting healthy habits in terms of physical activity [105].

The need for in-home monitoring extends beyond caring for the elderly to monitoring individuals of any age who require assistance living independently due to a mental or physical disability or injury. In addition to the use of in-home monitoring for clinical purposes, it may be an important aspect of ubiquitous smart home systems. As more and more electronic devices are becoming internet-of-things (IoT)-enabled, including televisions, kitchen appliances, and security systems, the demand is growing for systems that automatically understand and react to users and their actions.

1.2 Activities of Daily Living

Activities of Daily Living (ADL's) are defined as activities routinely performed in daily life, often for basic hygiene and personal care as well as food preparation, housekeeping, and other aspects of independent living. These activities have been studied since at least the 1960s [53] as an assessment for the ability to perform physical self-maintenance, especially among the elderly. Several scales of assessment have been proposed, each with the common purpose of evaluating the cognitive and physical capabilities of the subject [53, 20, 22]. In the clinical setting, it has been shown that the ability to independently perform ADLs is an indicator of the onset and progression of conditions such as Dementia and Alzheimer's disease [33].

The manual assessment of human activities is dependent on the ability for the subject to be observed, necessitating the use of either in-home visits by a clinician or visits by the subject to clinical laboratories. This can be challenging when considering the expense of home visits and the limited ability of elderly to travel and coordinate appointments. The use of technology through eHealth and mHealth applications to monitor and observe ADLs

in a home environment can greatly reduce the impact to both the subject and clinician by providing an unobtrusive, automated method of gathering activity data.

A challenge to implementing automated activity detection and recognition systems is the need for a formalized model defining each activity under observation. The ability to recognize activities as they occur comes very naturally for humans, but it is much more difficult to define for a machine. Human activities are defined in overlapping temporal and semantic spaces. That is, one higher-level activity may include the performance of several lower-level activities at the same time. Thus the definition of these activities may be subjective to the observer.

For example, a subject may perform the high-level activity of preparing breakfast, which involves many actions and movements in a sequence over a significant period of time. Within the action of preparing breakfast, the subject may have accomplished the lower-level activity of making toast, and within that activity the subject may have accomplished the lower-level activity of inserting bread into the toaster. The chain of activities may continue decomposing all the way down to the activity of picking up a bag of bread. Figure 1.1 illustrates several possible activities involved with this scenario and places them on a scale of increasing complexity. Understanding the same or overlapping activities at all of these levels is crucial to a fully functioning activity monitoring system.

1.3 Research Goals

Research Question 1: Can wearable sensors be used to unobtrusively monitor activities of subjects in a semi-structured environment? A desired outcome of this work is to provide evidence that a system based on wearable sensors may be used in the home (i.e. not in a clinical setting or laboratory) to provide continuous monitoring of activities of daily living. The outputs of the system provide data on activities and patterns of activities to inform clinicians on possible changes in the health and wellness of the user. Data collected at

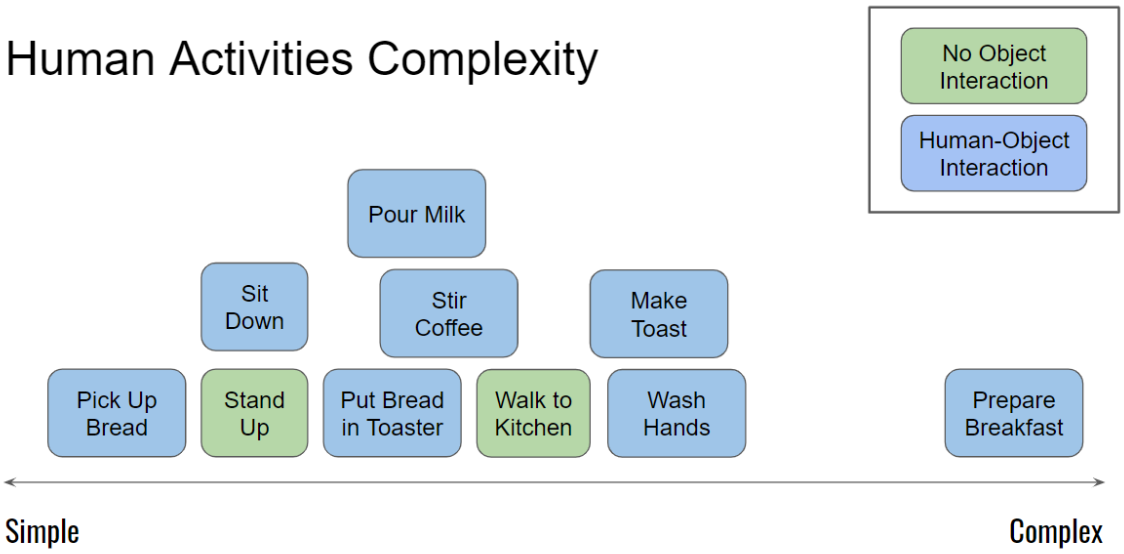


Figure 1.1: Example of activities with varying spatio-temporal complexity which may occur during the larger activity of preparing breakfast.

clinical laboratories is temporally sparse and involves scripted movements, bringing into question the validity of the data for assessing the subject's ability to perform the actions in a natural environment; by contrast, the data collected by in-home monitoring will show a more complete look into the subjects' abilities to perform activities on a day-to-day basis. Early work was focused on detecting and measuring gait in controlled and semi-controlled environments [92, 94]. While results were achieved with uninterrupted gait sequences of several meters, it was clear that long uninterrupted gait sequences are uncommon in many daily activities, which include starts and stops and navigation around household objects. For this reason, the ultimate focus of this work is on object interactions rather than gait, as object interactions provide a much richer context for activities that are being performed.

Research Question 2: Can features extracted from first person videos recorded in subjects' homes be used to categorize different interactions with objects? The use of object interactions to determine activities has been popular in recent automated activity recognition and classification systems [80, 98, 73, 110, 42]. Many in-home ADL's are at least partially defined by interactions with certain objects. For example, knowing that a subject

is interacting with a remote control should reveal that the subject is interacting with a television or other electronic device. The search space of candidate activities can be greatly culled by knowing which objects are involved. A desired outcome of this work is to exploit a combination of video and object features to characterize different interactions with objects. Building on an initial set of features from previous work with gait, an initial baseline set of features was identified for categorizing object interactions. In addition to the location of objects, this set of features was heavily dependent on optical flow features to identify changes in object location and color features to identify changes in the appearance of objects. Additional experiments are needed to extend and refine this set of features and the resulting Fuzzy Inference System (FIS) that was developed, but the initial work proved the feasibility of using first person video features to categorize object interactions.

Research Question 3: Can the activity identification system a) accurately describe activities and b) handle uncertainty in the results better than current state-of-the-art methods? While much recent work has been done on using deep learning and neural networks to identify human activities in video sequences (such as run, walk, sit down, throw, turn on faucet, open refrigerator, etc.) [2, 79, 56], there are two major drawbacks to the current deep learning trend:

1. The networks must be trained to recognize a very specific set of actions. To date, most work involving neural networks has been against small sets of precisely defined and scripted activities, limiting practical applications.

2. Since deep learning requires no features to be identified/extracted, the use of these technologies in isolation provides no ability to explain the outcome – only to identify it. That is, when the network fails to provide the correct answer, there is no explainability of the result or the misidentified activity.

We propose the use of fuzzy logic to address these limitations. Many activities are very similar in their basic movements and absolute identification may become difficult as the set

of activities grows. If an unknown activity occurs, fuzzy outputs may still provide some insight into the activity or similar activities. This is an advantage over machine learning approaches which will simply fail to identify the behavior. The simplicity of a fuzzy system can be seen in different lights; our goal is to show that the simpler fuzzy model is capable of providing a rich description of activities. In the course of attempting to identify an activity, the fuzzy system will produce interpretable fuzzy variables (as opposed to unexplainable deep learning features) which describe actions in the scene - for example, a specific object in the scene is being interacted with and the subject is manipulating the shape of the object. While the overall activity might be unidentified, a good amount of information may still be taken away from the fuzzy system.

1.4 Thesis Statement

This dissertation proposes to bridge the gap between in-home activity monitoring leveraging an object recognition framework and using a fuzzy linguistic framework to handle uncertainty in complex movements between humans and objects in an in-home unstructured setting.

Related Work

This chapter contains a summary of the existing body of work related to the objective of describing activities occurring in first person video. The related work varies based on several factors - the types of activities detected, the sensors used to detect the activities, and the computational intelligence methods used to convert sensor data into activity information. The chapter is organized into sections based on these factors. Each section provides an overview of the variation seen within the related work with respect to the given factor.

2.1 Activities Being Detected by Vision Sensors

A number of activities have been found in related work on detecting and classifying activities using vision sensors. These related works were grouped into three categories, and discussion of the types of activities follows below:

1. Anomaly Detection,
2. Full-body Activities, and
3. Partial-body Activities.

2.1.1 Anomaly Detection

A subset of related work focuses on the detection of anomalous activities. Each of these systems used full-body views of subjects from third-person vision sensors combined with computational intelligence to observe an environment and determine when either an unexpected activity occurred, or when an activity occurred in an unexpected way. Systems detecting such anomalies focus on actions that do not fit expectations based on previous observations, but do not often focus on describing precisely what these anomalous actions are - just that they are unexpected.

In [16], a vision sensor monitored a pedestrian area and used motion history images to identify activities that were determined to be anomalies, such as vehicles entering the space, by means of clustering. This method approximated the spatio-temporal distribution of activities that occurred frequently and detected activities that did not fit that distribution. Another outdoor environment - a loading dock - was monitored for anomalous activities in [43]. Activities were represented as n-grams of events and a distance metric based on these n-grams determined the dissimilarity between activities. Anomalous activities that were discovered included a truck leaving the dock with its door open and an unusual number of people unloading a truck.

Anomalous activity detection was also found in indoor environments. Fuzzy membership functions were used to captured several parameters of indoor human activities (location, time, perceived area of subject) from an omni-directional vision sensor and output a fuzzy determination of how ”normal” each activity was [95]. [116] presented a similar experimental setup with an indoor omni-directional vision sensor seeking to detect anomalous activities, but used Gaussian models of time and space to determine whether an activity was normal.

2.1.2 Full-body Activities

While detecting anomalies is useful for some applications, detecting the occurrence of a specific pre-defined activity is important for many applications. In the following reviewed studies, various vision sensors recorded subjects' full bodies to detect specific activities using a variety of computational intelligence. These works have the goal of detecting a specific set of activities and differentiating the occurrence from other activities. These related works are able in the best case to exploit full third-person views of the subjects which provide a maximal amount of input information about both the subject and surrounding environment.

One such study described a system designed to detect a single activity (or a single *type* of activity) - falls [12]. This work focused on applications in elderly living facilities , where timely detection of falls is critical. It compared the use of three different vision sensors in the same facility and was shown to be capable of detecting three fall-related activities - being upright, sitting, and being on the floor.

Another two related works were focused on using full body video sequences to detect activities in a medical setting. These environments, such as a trauma unit or patient room, may be very busy with doctors and nurses, and activities may be happening frequently or simultaneously. Because of this, both of the reviewed studies required systems capable of monitoring several subjects at the same time to identify activities [23, 17]. The systems were focused on outputting activity information for a set of multiple actors in the scene, providing a summary or log of activities that may assist in the coordination of medical actions such as performing chest compressions, intubation, or checking pulse.

Many related studies which used full-body views from vision sensors were focused on detecting indoor activities of daily living. While the specific sets of activities varied somewhat from study to study, each of these were deployed in home-like environments and detected daily activities such as sitting on a sofa [37], eating dinner [115], or reading a book [34]. The sets of activities are pre-defined in each case (i.e. the system is trained to

recognize a specific limited set of activities), and the works all make use of scripted training and test datasets.

2.1.3 Partial-body Activities

A third class of related work focused on detecting activities using vision sensors with visibility of only part of the subjects' bodies. These limited views often complicate activity detection since the entire posture of the subject is not able to be known by the system. However, the trade-off usually comes with improved usability, portability, or some other convenience to the user. For example, applications using wearable hardware are better able to integrate into all aspects of the user's daily routine.

A fall detection system based on a wearable camera was described in [63]. This work differed from [12] in that it used only a partial view of the subject's body, since the camera was worn on the torso. This system was capable of detecting a set of three activities similar to the full-body fall detection work, including sitting down, lying down, and falling.

Several studies applied partial-body views from a vision sensor to detect speaking activities. These studies were concerned with identifying speakers in multi-user or noisy environments, citing use-cases such as human-robot interaction [117, 57] and video calling [90]. Though the recognition of the audio is important in each of these cases, the vision sensor provides valuable information for detecting the beginning and continuation of a speech activity, based on the movement of the subjects' mouths. In these studies, the detection of activities would not necessarily benefit from knowing the full-body posture of subjects, and the closely focused view of the face was preferred for the enhanced details.

Similarly, studies were discovered which used partial-body views to detect activities while driving. These studies had a similar goal of detecting drivers' actions while the car was in motion. These activities only require knowledge of the upper body posture. One study sought to build an understanding of actions that occur as the driver completes maneuvers such as turning or approaching an intersection [67]. The other study was concerned

with detecting both driving activities (steering, operating the shift level) and non-driving activities (eating, using cell phone) [119].

As expected, many reviewed papers described systems which used partial-body views from a vision sensor to detect activities of daily living. In these cases, the vision sensor was attached to the subject, either a human or a robot, and computational intelligence was applied to make a determination of the current activity being performed [56, 113]. Due to the positioning of the vision sensors, these studies have only a partial-body view of the subject. Unlike the studies detecting daily activities from full-body views, the partial-body view does not provide the visual information to determine the subject's complete posture. These studies choose instead to exploit the information of specific objects [80, 113, 69, 74] or other people within view [56] to detect the occurrence of an activity.

2.1.4 Discussion

Our results reveal a considerable degree of diversity in the types of activities to which activity detection via vision sensors and computational intelligence has been applied. Overall, the most common type of detected activities were indoor activities of daily living [70, 113, 115, 71, 34, 84, 37, 63, 77, 56, 69, 74]. However, we also found results focused on detecting activities in medical settings [23, 17], in vehicles [119, 67], in the workplace [43], and interactions with robots [117]. A conclusion may be drawn that the vision and computational tools discussed have broad utility to intelligent systems which must understand the actions of users, regardless of the particular domain, setting, or application of that knowledge.

A drawback to the diversity of the activity domains in our results is that it remains difficult to draw direct comparisons between methods which have seen success. A vision-based system which performs well inside of a home may not perform well when deployed to detect activities in an outdoor work environment. Even within the same activity domain, we did not see a standard dataset emerging to provide a common baseline. In most of

the studies, a new dataset was created, and little information was provided on how these datasets were captured. Further, datasets are typically gathered specifically for the evaluation of the system rather than from a truly operational deployment of the system.

2.2 Types of Vision Sensors

Our analysis revealed that several types of vision sensors have been used for activity detection. These sensors generally fell into the following four categories:

1. traditional fixed RGB vision sensors,
2. fixed RGB-depth vision sensors,
3. wearable RGB vision sensors, and
4. wearable RGB-D vision sensors

In some cases, several different devices were used either in concert or separately to provide the best results. The following subsections discuss the use of each of these types of hardware.

2.2.1 Fixed RGB Vision Sensors

Traditional RGB cameras come in a variety of styles, configurations, and sizes, and have been a frequent choice for vision-based activity recognition applications. The vast majority of related work incoporated one [117, 90, 16, 113, 95, 116, 71, 34, 84, 57, 23, 119] or more [70, 17, 67, 43] fixed (i.e. non-mobile) RGB vision sensors. In each of these studies, the vision sensors were statically deployed in the environment within which the recognized activity is taking place, providing a third-person view of the subject(s). Within the category of stationary RGB cameras we see variation amongst the types and number of devices used, depending on the specific challenge being addressed.

One challenge that researchers face when using fixed (stationary) RGB vision sensors is a limited view of the environment, either due to a limited view angle or due to occlusion by objects. However, some researchers have been able to alleviate this issue. For example, some studies used multiple cameras simultaneously to provide a more complete view of the monitored environment. In [70], a camera network was used to provide a comprehensive view of several seating stations within one room. The system mimics a home environment where activities such as dining, studying, and reading may be expected to occur in distinct locations. Individual camera nodes may provide sub-decisions, but communication between all nodes leads to a final result. Other studies that used multiple cameras examined differing use cases. For example, in [43] multiple cameras were deployed with partially overlapping viewpoints in order to provide complete coverage of activities that transpired in a loading dock. This allowed detection of multiple activities occurring in parallel at a given time. Multiple cameras were also used in [67] to provide detailed images of the face and hands of a subject driving a vehicle to analyze the coordination of the head and hands while maneuvering the vehicle. The cameras were positioned to optimally capture the relevant areas of the subject while excluding irrelevant details such as the dashboard or the passenger section of the car.

Our investigation also showed that researchers have found it possible to achieve an enhanced perspective without the use of multiple cameras. Two related studies incorporated omni-directional vision systems into the design of their activity recognition systems [95, 116]. These cameras map a complete 360 degree view onto a typical two-dimensional intensity image, providing full monitoring of an area with a single hardware device, typically mounted overhead in the center of the space. Though this results in a distorted image due to the 360-degree perspective, the image may still be processed and human postures are still reliably discernable [95]. A similar single-camera experimental setup is shown in [116] to effectively monitor an entire panoramic view of a room. The omni-directional vision device is capable of monitoring activities occurring in a seating area and four areas

of entry or exit to other rooms.

A challenge in detecting activities with fixed RGB vision sensors is the lack of depth information from the scene. That is, only two dimensions of data are recorded, leaving out potentially valuable information about the movement and positioning of objects. However, traditional RGB cameras (which do not directly record depth information) can recover approximate depth information when used in pairs as a stereo camera. A depth dimension may be constructed by comparing the perspective of each camera given the predetermined distance between them. Two of the studies included in our review take this approach [57, 17]. In [17], depth information from a stereo camera is exploited to detect the background in each frame and subsequently remove it, leaving only foreground objects for consideration in activity recognition. Depth information from stereo cameras has also been used to determine distances between multiple subjects in the same area [57]. The precision afforded by the depth information allows for improved facial tracking and differentiation between subjects in closer proximity to one another, improving the overall ability to differentiate the active speaker.

2.2.2 Fixed RGB-Depth Vision Sensors

As previously discussed, traditional RGB vision sensors record only a two-dimensional mapping of the three-dimensional environment. This is an important limitation because it is not possible to reconstruct the third dimension with high accuracy without the use of a standard vision system that integrates multiple vision sensors to provide different perspectives of the same area (i.e. a stereoscopic camera).

An alternative approach to providing depth information from a stereoscopic RGB vision sensor system is to use an RGB-Depth vision sensor. These sensors incorporate a traditional RGB camera, an infrared (IR) camera, and an IR projector. The IR projector projects a pattern of IR light onto the scene which is recorded by the IR camera to determine depth. The depth measures are combined as an extra channel of information with the

image captured by the RGB camera. Several studies within our sample used RGB-D vision sensors [12, 115, 5, 120, 37, 77].

The systems described in [77, 115] detect activities based on sets of joint features describing the flexion, extension, rotation, or position of a variety of human joints. The additional depth information from the vision sensor is required in this case to provide the necessary inputs to the recognition algorithms. [37] describes a similar approach where digitized three-dimensional skeletal and joint models are extracted from the depth images captured by the sensor and then used to detect activities based on a statistical Markov model. A multi-modal sensing system built a human motion model in [5]. An RGB-D vision sensor is paired with wearable measurement devices (IMU, gyroscope, accelerometer) to observe motions, and determine situations in which each device is better-suited to capture the data.

We also found that depth information may benefit systems that need to distinguish between multiple subjects. An RGB-D vision sensor was used in [120] to provide biometric identification of users based on bone lengths. The authors acknowledged that while previous work indicates that fingerprints or iris scans are more reliable means of identification, the model built by an RGB-D sensor is much more user-friendly and allows passive user identification from a distance, which may be important in a dynamic multi-user environment such as a workplace.

In addition to the extra dimension of positional input data, another convenience provided by the use of an RGB-D vision sensor is improved performance in poor lighting conditions. While RGB vision sensors may have difficulty properly adjusting exposure and maintaining a clear image in bright or low light, the IR projectors and cameras in RGB-D vision sensors provide a consistent result in most lighting conditions since they are not susceptible to changes in visible light and have an integrated source of IR light to illuminate the scene. [12] demonstrates this and gives a comparison of human silhouette extractions using a webcam with an IR filter and an RGB-D vision sensor in an eldercare facility. The

RGB-D sensor is shown to be effective, though the depth information begins to degrade as the subject moves farther from the vision sensor (due to natural scattering of IR light from the projector over this distance).

2.2.3 Wearable RGB Vision Sensors

A drawback to all of the RGB and RGB-D vision sensors discussed to this point is that they are not portable. They only record data within a fixed environment, which may not be sufficient for all use cases. For example in an indoor environment, if the person moved away from the room in which the sensor was placed, the system would fail to raise an alarm if the person fell down. Since many rooms may pose a high risk of falls (bedrooms, bathrooms, staircases, kitchens), providing full coverage is difficult. For this reason, many related works use wearable sensors to enable systems to detect and recognize activities wherever the user travels.

Many examples can be found which incorporate wearable vision sensors [83, 63, 56, 14, 69, 110, 30, 98, 42]. In all of these, the vision sensors were placed on the subject to provide a first person view of the activity being performed. In many cases, this meant that features describing the posture and joints of the subject were no longer available because they were out of view. However, the first person view provides an opportunity to identify objects that the wearer interacted with [80, 83, 69, 30], people with which the wearer interacted [56], areas of visual focus of the subject [98, 68], and estimated motion of the subject based on camera motion [63].

In [83], a system was built to identify the action of picking an object off of a shelf in a store. The action event was detected via non-vision (inertial) sensors, but the identification of the specific object was accomplished by a wearable camera worn on the user's wrist. This illustrates how wearable vision sensors may be used to narrow the focus of the input data by intentionally choosing the placement of the camera on the subject.

In other cases, the wearable vision sensor can provide information about specific hand

gestures instead of objects with which the subject is interacting. This is demonstrated in [14], where a vision sensor embedded in a pair of glasses provided a view of the subject's hands. Hand gestures were recognized and the gesture was applied to the object in the center of the subject's gaze. The glasses and hand gestures presented a natural means of interacting with the system which would not be achievable with a stationary vision sensor.

A final use case we found for wearable vision sensors was to describe the motion of the subject wearing the device. When mounted in a fixed position on the body, movement recorded by the sensor across temporal image sequences may be attributed to motion in that part of the body, supplying information about both the environment in front of the subject as well as the posture of the subject without the subject being directly in the field of view. In one study, an RGB sensor was worn on the subject's trunk and the movement of the camera was used to estimate the movement of the torso, enabling the detection of sitting and lying down without any part of the torso in view [63]. This is a good example of how wearable sensors may be used to capture derivative information from the scene, in addition to information of objects or subjects that are directly within the view of the sensor.

2.2.4 Wearable RGB-D Vision Sensors

Very few studies are found which employed wearable RGB-D vision sensors for the purpose of activity recognition. The approach described in on study was based on knowledge of object interactions, and a robot (rather than a human) was equipped with the wearable RGB-D vision sensor [56]. The purpose of the study was for the robot subject to become aware of the activity it was performing by using knowledge of nearby objects. The wearable sensor was used to detect the objects interacting with the robot's hands, and that data was fused with position and joint features provided from the robot's operating system, taking advantage of the availability of both object and posture data.

2.2.5 Secondary Sensors

In our survey of related work, secondary non-vision sensing devices were soemtimes used to enhance the data captured by the vision sensor [117, 83, 90, 113, 5, 71, 120, 57, 56]. Multiple studies used a microphone to assist in voice activity detection along with the vision sensors [117, 90, 57]. The vision sensor was important to these studies for identifying the speaker in the observed space, and the microphones were used to provide auditory data which can provide extra confidence in the number of voices occurring at a given time. Other reviewed studies made use of smart watch devices to provide wrist motion features as input to computational intelligence methods [83, 120]. In [113], radio frequency identification (RFID) tags were applied to objects and the RFID device was used to correlate vision sequences with proximity to objects in the environment. [71] combined an ear-worn accelerometer with a vision sensor and extracted non-overlapping discriminatory features from each device to detect activities. Similarly, [5] paired a vision sensor with an IMU device. The IMU device provided finer details of the motion the subject performed, allowing the activity detection to be fine-tuned based on the speed of the movement. These studies demonstrate situations in which multiple sensing devices and/or different sensing modalities may combine to provide a richer dataset for activity detection.

2.2.6 Discussion

In our survey of related work, we found a variety of vision sensors, including both fixed and wearable RGB and RGB-D sensors. We found that there were significant variations on the use and implementation of each of these, including special cases such as omni-directional or stereoscopic sensors. Our results indicate that each of these pieces of hardware are capable of providing positive results in activity detection when used with computational intelligence, though the choice of hardware requires careful consideration of the objectives of the system.

We note that the choice of vision sensor is largely dependent on two factors - 1) the scene in which activities will take place and 2) the types of activities that need to be detected. Wearable sensors are frequently chosen when the scene is not a fixed location (i.e. the activity may occur in any location in which the user is present) [14] or when objects within the user's perspective are more important than the overall posture or positioning of limbs and joints [14, 56, 83]. When choosing between RGB and RGB-D sensors, RGB-D sensors are a common choice when differentiation of activities seemingly necessitates precise three-dimensional data [77, 37] or when lighting conditions may affect the performance of visible-light-based RGB sensors [12].

2.3 Computational Intelligence Methods

Our analysis of related work indicated that a wide variety of computational intelligence methods have been used with vision sensors for the purpose of activity detection. We categorized studies by algorithm and found the use of both supervised and unsupervised methods. In general, supervised methods are a frequent choice when the desired output from the system is an explicit activity label from a set of predefined activities. Applications which need only to detect the abnormality of an activity often turn to unsupervised intelligence methods since these methods do not require the manual annotation of truth data, which can be time consuming. We review both methods in relation to the use of vision sensors and activity detection below.

2.3.1 Stochastic Modeling

Many of the reviewed studies employed some variant of a stochastic process model [84, 37, 57, 23]. The activity detection capability of a multiple vision sensor network described in [84] was demonstrated by a Markov model showing regions of activity across multiple

rooms with transition probabilities determined by a collected dataset. In another case, a large composite Hidden Markov Model (HMM) comprising smaller HMMs was built in [37]. The smaller HMMs individually specified the states within just a single activity and contained transitions back to a common initial and final state. The larger HMM was composed by allowing transitions between the common initial state, which represented transition between the smaller individual activities. The HMMs were trained on features representing the skeletal structure of the observed subject. In a different application of the Markov property, a Markov logic network (MLN) was employed in [23]. The MLN was driven by an activity grammar that described steps within an activity, and formed relationships between predicates and objects (e.g. "stethoscope approaches patient"). A Langevin process model, rather than a Markov model, was used in [57]. This work fused visual and audio data to determine speaking activities and identify speakers in a multi-user environment.

2.3.2 Fuzzy Logic and Clustering

Activity detection methods based on fuzzy logic were also used in several of the reviewed studies [12, 95, 115, 5]. Fuzzy methods introduce degrees of truth, rather than binary decisions. Fuzzy C-Means clustering was used in [115] to turn training posture feature vectors from an RGB-D vision sensor into fuzzy rules to support classification via type 2 fuzzy logic. An extension of Fuzzy C-Means clustering, Gustafson-Kessel clustering, was used to cluster silhouette and image moment features to recognize activities using RGB-D sensors [12].

In addition to fuzzy clustering, we also found other applications of fuzzy logic. Inputs from an omni-directional vision sensor were recorded in terms of fuzzy variables in [95]. The use of fuzzy variables improved the performance of the activity detection around the discretizing borders when compared with previous work using Bayesian methods [46].

2.3.3 Bayesian Methods

Our survey also included studies that used Bayesian methods to detect activities [70, 117, 113]. [70], which employed a network of several vision sensor nodes, proposed a greedy structure learning algorithm based on the Bayesian Information Criteria (BIC). Each sensor node built a Bayesian network to learn activities for which the node was a candidate for detecting. Candidate nodes contributed data to the network to make a final determination of an activity. The distributed approach allowed sub-decisions to be made at the node level, which positively impacted the fault tolerance of the network.

A Dynamic Bayesian Network (DBN) was used in [113] to automatically acquire models of activities and objects that are involved with them. The DBN was formed from internet-based knowledge repositories, object information, visual frames, and RFID inputs. Learning was done in an unsupervised manner so that data did not need to be labeled (except for the purpose of testing).

A third study used a Bayesian network to perform voice activity detection from a combination of RGB video and audio data [117]. The inputs included the log-likelihood of silence in audio data calculated by a speech decoder, a feature based on the height and width of the lips as captured by the vision sensor, and the confidence that a face was detected in the visual data. Probabilities for the Bayesian network were obtained from a Gaussian Mixture Model (GMM) trained in advance on a separate set of training data.

2.3.4 Gaussian Mixture Models (GMM)

Gaussian modeling, another probabilistic method, was also used in the examined studies [71, 34, 116]. In [71, 34], GMMs were used to perform fusion of data from multiple sensors for the detection of activities. The models were based on data provided by a vision sensor and an ear-worn accelerometer. Vision sensors provided features such as bounding box aspect ratio and eigenvectors of silhouettes, and average optical flow within the blob, while

the wearable device provided information on the tilt and movement of the head. [116] used Gaussian distributions to build spatial and temporal models of activities. Using the combination of these models, activities were determined to be either normal or abnormal. Recorded activity information was then used to update the models.

2.3.5 Neural Networks

Deep learning and neural networks were used for activity detection in multiple of the reviewed studies [56, 90, 41, 61, 54, 110]. Multiple types of Recurrent Neural Network (RNN) structures were used in [56]. These structures were shown to be well-suited to performing classification based on sequential video frames. Convolutional layers of the networks operated on pre-processed visual frames which were later combined with joint features as input to a Long Short-Term Memory (LSTM) layer. The LSTM layer determined when a tracked hidden state should be updated, contributing to the strong performance of the system against the temporal video data. Similar positive results were shown against temporal data in [90], where audio data was input to deep learning detectors after visual sensor data was used to detect speaking activities. Speaking activity was visually detected by analyzing motion near the mouth.

2.3.6 Other Methods

The previously discussed methods account for a majority of the computational intelligence methods found in the examined studies. However, a variety of other methods were also used to detect activities using vision sensors. A simple decision tree was used in [83] to detect the single activity of picking an item from a shelf. In another study, the K-means clustering algorithm was used to cluster unlabeled skeletal features into differentiable activities [77].

A more complex density-based clustering algorithm was used in [16] to detect rare activities using binary space-time descriptors. Two of our included studies used a Random

Forest classifier to detect driving activities [119, 67]. These were the only studies in our results to use this method.

2.3.7 Discussion

We found that numerous computational intelligence methods have been used with vision sensors to detect activities. As with the choice of vision sensor, this may be driven by the set of activities that need to be detected. In applications requiring only a measure of normality for an activity (e.g. normal vs. abnormal) rather than a specific classification of the activity in a given set, clustering methods are shown to be robust [16]. In applications where activities of daily living are detected, we see several distinct methods including Bayesian methods [113, 70], Stochastic modeling [84, 37], and Gaussian Mixture Models [34, 71]. The differences in the experimental setup of these similar studies present a difficult comparison of methods against each other.

A commonly cited factor in the success of computational intelligence methods in classifying an activity into a predefined set of possible activities is the ability for the method to take into account the various states within an activity. Activities typically have both spatial and temporal components, and our results indicate strong performance with methods that can either consider time-aggregated data, have a memory of previous outcomes, or in some way indicate change of pose or environment over time. Examples of this are seen in neural network and deep learning methods [90, 56], stochastic modeling methods [84, 37, 23, 57], and Bayesian methods [113].

Comparing computational intelligence methods and the results associated with each was complicated by the lack of consistency in reporting results. Several of the reviewed studies failed to report an accuracy metric or the details of the datasets that were used.

2.4 Key Takeaways

The survey of related work reveals trends in the hardware, computational methods, and experimental approaches used to detect, classify, and describe human activities. There are a wide variety of methods applied in different combinations to accomplish very distinct tasks, and the field appears to suffer from inconsistency in experimental methods and reporting of results. However, there are clear common limitations that are identified:

1. The use of black-box methods such as convolutional neural networks limits the generalizability of the related works. Many reviewed publications describe systems built for very specific purposes and environments, but do not classify a set of activities that is encompassing of instrumental daily activities.

2. Most of the surveyed methods offer very little in the way of interpretability of the learned input features. As a result, classification of activities is either completely correct or completely incorrect; an unrecognized activity is given no additional context or description by these systems. If input features are able to be interpreted, then some context may be discerned about the activity even in the event of an incorrect classification.

3. Related works acknowledge that complex activities are often defined as the composition of smaller activities in sequence, but do not address the detection of activities at these various levels.

Methods

This chapter describes the theory and concepts behind methods that are used throughout the experiments that follow in later sections. Our final proposed system of classifying actions is implemented using a Fuzzy Inference System that takes input from a variety of feature extraction techniques described below.

3.1 Fuzzy Inference Systems

Fuzzy Inference Systems (FIS) are a type of decision making system that accepts a set of inputs, processes them according to a set of rules, and produces an output decision. Unlike other types of classifiers or decision making systems, FIS uniquely model the 'fuzziness' of truth values that are encountered in real-world applications. A huge advantage this provides over current state-of-the-art techniques, such as machine learning and deep learning, is the ability to clearly apply intuition and interpretation of the problem domain to the performance of the system. While other machine learning techniques adapt inputs to a feature space that is not clearly interpretable or meaningful outside the system, the FIS is built around a set of rules and linguistic labels that retain meaning. For this reason, our solution is implemented as a FIS.

The two common types of FIS are Mamdani [64] and Sugeno [99]. In this work, we discuss Mamdani-style inference.

The variables in a fuzzy system are called fuzzy variables (as opposed to *crisp* values

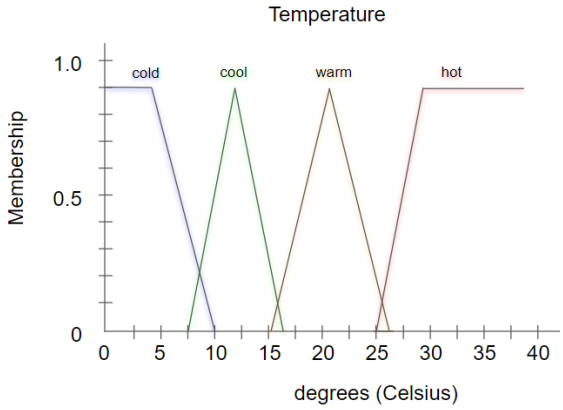
found in systems employing traditional binary logic), and take on levels of truth rather than precise numeric values. The levels of truth are defined linguistically. For example, in a fuzzy system, a temperature variable may be defined to take values in the set [‘cold’, ‘cool’, ‘warm’, ‘hot’]. For any given crisp value, the corresponding fuzzy value is a set of membership levels in each of the terms in the set. Following the example of temperature, a crisp value of 8 degrees Celsius may correspond to the fuzzy temperature value that has membership levels $\{cold : 0.4, cool : 0.75, warm : 0, hot : 0\}$. Membership values are not required to sum to any given value (though they each typically exist in the range $[0, 1]$), and the power of fuzzy systems comes from ability to have membership in multiple terms, which is not possible in Boolean systems.

3.1.1 Fuzzification and Defuzzification

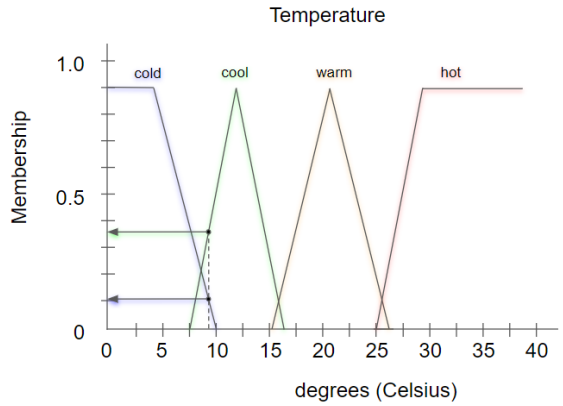
In order to determine membership levels of fuzzy variables, membership functions exist for each linguistic term in the set to define a mapping of crisp values to amounts of membership in the fuzzy terms. Figure 3.1a illustrates membership functions for the example temperature variable. It is often the case that membership functions overlap - if they did not, then values will only have membership in one term, which is essentially the same case as Boolean logic. The overlapping regions provide membership in multiple fuzzy terms. This is a critical distinction for fuzzy systems and allows a variable to express some uncertainty.

Membership functions can take any form or shape, but are commonly triangular or trapezoidal. More complex functions, such as Gaussian curves are also common candidates, especially when the problem domain is based on probability or statistics [6]. Much like the choice of activation functions in a neural network depends on the specific problem domain, the choice of membership functions may be driven by the domain to which the FIS is being applied.

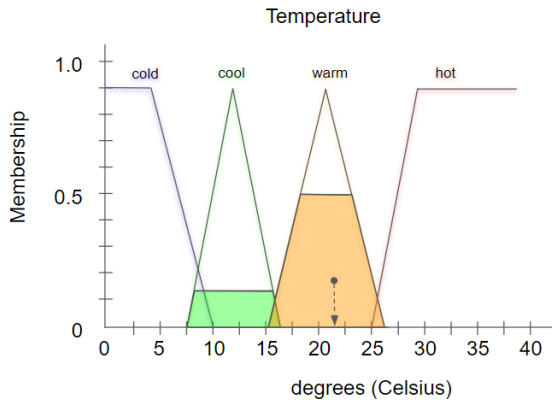
Before a FIS can operate on input data, it must transform crisp input values into the



(a) Example membership functions for the fuzzy variable temperature



(b) Fuzzification of the crisp value 10 degrees to a fuzzy temperature value



(c) Defuzzification of a fuzzy temperature value to 22 degrees.

Figure 3.1: Illustration of membership functions, fuzzification, and defuzzification process.

fuzzy value space, i.e. map the values into the membership functions. The process of transforming crisp numeric values to the fuzzy values is referred to as *fuzzification*. Similarly, the output of the FIS must be transformed back to crisp values for most applications. This reverse process, transforming a fuzzy value back to a crisp value, is called *defuzzification*. Each of these processes are illustrated in 3.1b and 3.1c.

Fuzzification is performed by evaluating each membership function for a variable on the crisp input. The resulting output of the membership function indicates the crisp value's level of membership in that particular variable. The crisp value is represented by the set of membership levels from all membership functions in the set, as shown in the earlier example.

Defuzzification requires the combination of multiple membership levels to be reduced to a single crisp value. There are many methods for accomplishing this, and the selection of defuzzification method greatly impacts the results produced by the system. Common defuzzification methods include maximum membership, centroid, and bisector of area. The maximum membership method simply discards all but the highest membership term, and returns a scalar falling in that membership function (most commonly, mean-max membership is used to return the mean value of the range of the membership function). This means that for a given fuzzy variable, the same scalar will be returned in every case where that membership term is the maximum, regardless of what the membership level was. More complicated methods are based on the area under each membership function. For example, the centroid method returns the geometric centroid of the shape formed by the area under all of the membership functions, as illustrated in 3.1c. Similarly, the bisector of area method considers this shape and returns the point that bisects the area. These techniques are capable of returning a continuous range of crisp values, unlike the maximum membership method.

3.1.2 Fuzzy Rules

In this work, we use only Mamdani-type inference in which all variables are expressed as fuzzy sets. Rules in a Mamdani-style inference system are expressed in the form "*IF* X *is* x_i *and* Y *is* y_j *THEN* Z *is* z_k", where X, Y, and Z are each fuzzy variables and x_i, y_j, and z_k are terms belonging to the fuzzy sets. An example in a fuzzy system controlling a thermostat might be "*IF outsideTemp is cold and indoorTemp is low THEN heatSetting is high*." Rules may be conjunctive or disjunctive and may contain any number of operands in the rule.

During the evaluation of a rule, the level of membership of the input terms are considered to determine the strength of membership of the output. The Fuzzy 'and' operator results in the minimum level of membership strength between the operand terms being passed to the output. By contrast, the fuzzy 'or' operator results in the maximum level of membership strength between the two operands being passed to the output.

To illustrate the evaluation of rules, consider the previous example "*IF outsideTemp is cold and indoorTemp is low THEN heatSetting is high*." Let $outsideTemp['cold'] = .35$ and let $indoorTemp['low'] = .1$. The rule is evaluated as $heatSetting['high'] = min(.35, .1) = .1$. The final output of the fuzzy system depends on the evaluation of the rest of the rules in the system and the defuzzification method used.

3.2 Genetic Algorithms

Genetic algorithms (GA) define a process for evolving high-quality solutions to a problem by mimicking the biological processes of natural selection and mutation [47]. This process is often applied to large optimization problems as a means to approximate an optimal solution. Genetic algorithms begin with a set of sub-optimal solutions (the *population*) and iterate through generations of offspring, evaluating generated child solutions against a fitness metric. By allowing the most fit solutions to combine and replicate at each generation,

the final population will represent higher-quality solutions than the initial population.

The structure of fuzzy systems, whose performance is dictated by a number of membership function parameters and rule sets, makes them natural candidates for optimization by a GA. There is much support in related literature for applying genetic algorithms to fuzzy systems [96, 50, 102].

3.2.1 Chromosome Encoding

Candidate solutions in a genetic algorithm must be expressed as a set of properties (chromosomes) to facilitate the reproductive process. The method of encoding properties of a solution into chromosomes is highly dependent on the particular problem to which the GA is being applied. Popular choices include value encoding (e.g. binary, hexadecimal, real-valued) which directly encodes values in an equation or system into different genes on the chromosome, permutation encoding which encodes the ordering of properties into the chromosome, or tree encoding which encodes tree or graph properties of a solution with a complex or graph-like structure (e.g. when using a GA to write an algorithm based on a graph of operations) [52].

Since membership functions within a fuzzy system are easily expressed as sets of real values (e.g. for a triangular membership function, a triplet of real values (x_1, x_2, x_3) define the function), the genes within the chromosomes representing fuzzy systems typically use simple value encoding. This allows the GA to evolve the solutions by altering the structure of membership functions within the system. Fuzzy systems are commonly initialized with rules based on intuition applied to the problem domain, but the mapping of crisp input values to fuzzy terms (i.e. the boundary of membership functions) is not always immediately clear. The GA approach provides support for finding appropriate boundaries for the terms. Additionally, GA optimization may find unnecessary terms by evolving a higher-quality solution in which one membership function completely envelopes another. In this case, the GA is capable of optimizing the structure of the system in addition to fine-tuning the

membership parameters.

3.2.2 Reproduction of Individuals

The primary operations by which subsequent generations of a population are created by a GA are crossover and mutation. These genetic operations mimic biological reproduction and accomplish two goals: 1) the preservation of good features in the population (i.e. features which contribute high fitness), and 2) the introduction of random mutations (i.e. new features) into the population.

Crossover is the process of selecting genes from each parent to build a chromosome for a child. Since only one parent may contribute a given gene, the method of mixing these genes can greatly impact the performance of the GA. Single-point crossover is performed by choosing a random point along the chromosome and including all genes from the first parent before the crossover point, and all genes from the second parent after the crossover point [47]. This process generalizes to k-point crossover, where any number of points may be selected to alternate between genes from each parent. For real-valued gene encoding, different forms of blend crossover, where child genes are determined by some operation incorporating both parent values (such as averaging), may be used instead [35]. However, the success of blend crossover depends on the level of correlation between output variables in the fitness function. Additional techniques may be required in non-separable cases [103, 19].

The crossover operation alone investigates solutions in new regions of the solution space. However, to fully mimic evolutionary processes and introduce some amount variability into the population, GA optimization also implements the mutation operation. After crossover is performed, mutation introduces new genetic information with some predetermined probability. Mutations help to ensure that the population of individuals does not become entirely homogeneous and converge in a sub-optimal solution space [47]. The implementation of mutation is dependent on the chromosomal structure. Real-valued genes

may mutate by increasing or decreasing in value within some allowable range of the original value. Binary-encoded genes may mutate by simply flipping a bit in the target gene.

3.2.3 Evaluating Fitness

In each generation of the GA, the population is evaluated against a fitness metric to determine the most fit solutions. The fitness function represents the problem the GA is attempting to optimize. Fitness functions typically include accuracy, precision, recall, and/or cross-entropy, as would be expected in the training of any classifier. After evaluating each individual, the fitness may be used to inform which individuals are dropped from the population and how likely each individual is to reproduce. Techniques like roulette-wheel selection or tournament selection are popular choices that reward fit individuals with a higher probability of mating [121]

3.3 Optical Flow

Several of the input features to our proposed fuzzy system are derived through optical flow. Optical flow is a category of algorithms that take two frames of video as input and provide an estimate of the motion vector for pixels in the frames, allowing for the approximation of the movement of objects in the video. For our activity recognition system, which is built on the theory that objects are a key indication of activities being performed, the movement of objects in the view of the actor is a key piece of information.

The motion estimation can be computed for every pixel in the input (*dense*) or for a specific subset of pixels which have been identified as being important (*sparse*). The motion estimates for pixels are computed by considering the linear displacement of corresponding pixels in the two frames as shown in Figure 3.2. The optical flow algorithm provides an estimation of the vector $< dx, dy >$.

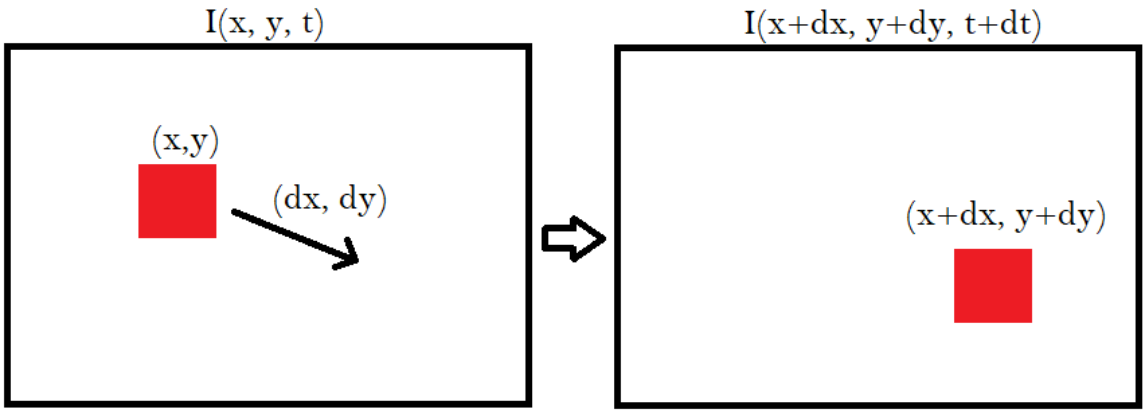


Figure 3.2: The pixel specified in the first frame at time t, $I(x, y, t)$, translated by (dx, dy) units and is located at $I(x + dx, y + dy, t + dt)$ in the next frame at time $t + dt$.

To compute optical flow, it is necessary to assume that pixels in the image $I(t)$ have equal intensity in the image $I(t + dt)$ for small values of dt. This is valid in practice because a small dt represents a time difference where the appearance of the object and lighting conditions are unlikely to change drastically. Based on this, we can set the equation for each pixel in the two images as equal (known as the constant intensity assumption): $I(x, y, t) = I(x + dx, y + dy, t + dt)$. This equation can be solved for the resulting motion vector using polynomial expansion, Taylor series approximation, and a variety of other methods [36, 48, 62].

3.4 Convolutional Neural Networks

In order for our proposed system to build an understanding of which objects are being interacted with, we rely partially on information about the actor's hands in relation to objects around them. We assert that most activities occur as interactions with objects in an environment, and hand position is a strong indicator (from a first person perspective) of which objects are being interacted with. To this end, our system uses a convolutional neural network to determine hand location as input data to our FIS.

Neural networks are a set of machine learning techniques inspired by the structure of

the human brain, where input data results in the stimulation of neurons, which in turn leads to information being passed on in a network of other neurons, which eventually lead to an intelligent output. Neural networks are composed of a variable number and type of layers of neurons (nodes) which are activated when the input they receive exceeds a threshold value. The activated neurons pass data through the layers of the network, performing various operations at each layer, until the output layer is reached. The weight given to each layer or neuron in determining the output is learned by the network through a training process. Neural networks can be used to classify and cluster data very efficiently, as they are very good at approximating complex functions.

Convolutional neural networks (CNN) are a special class of neural networks that contain convolutional and pooling layers. They are designed to learn spatial relationships of features in an input, which has lead to their popularity in applications such as computer vision and natural language processing [65].

Unlike more traditional neural networks, where features are extracted from data prior to feeding inputs to the network, CNN's expect raw data and produce the features within the network. Features are extracted through convolution operations, which involve computing a kernel function as it slides over a provided input. The kernels functions are learned parameters of the CNN, and each kernel theoretically serves to extract a different type of feature. While the user of a CNN does not specify the features and kernels, visualizing the trained kernels on images often reveals that features learn to extract similar information to traditional features, such as color, edge, and texture descriptors.

In addition to convolution layers, pooling layers are used within the network to reduce the dimension of features extracted in later layers. Pooling is an operation that condenses information within a neighborhood of each data point, through operations like averaging or taking the maximum value. The result is that convolutional layers deeper in the network have lower dimensionality, and therefore represent more abstract features about the input than earlier layers [118]. This allows CNN architectures to perform robust detections on

data that may require knowledge of features that have varying spatial or temporal correlations by extracting features at multiple resolutions.

3.4.1 MobileNet

In the field of CNN's, there are several commonly used and notable architectures, including AlexNet [51], ResNet [45], and VGGNets [97] among others [55]. These models have gained popularity for solving different challenges within CNN's, such as optimizing network size or speed, efficiency of training, or overall performance in specific types of tasks. The hand detection network included in our proposed solution uses the MobileNetV2 [89] architecture. This network is specifically designed for low-resource applications, such as mobile or edge computing while retaining accuracy competitive with larger networks. While our stated goals for this work do not include performance optimizations, the MobileNetV2 model provides a convenient and accessible solution for an activity recognition system that may someday have such constraints.

Our MobileNetV2 model uses a version of the Single Shot Detector (SSD) [60], specifically the Feature Pyramid Network (FPN) [58]. While traditional object detection models approach detection by proposing bounding boxes, then resampling features for each box, and finally applying a classifier to each box, the SSD method eliminates bounding box proposal and the feature resampling steps by producing predictions of different scales using features maps of different scales, computing the information in a single pass through the network. The accuracy of SSD is shown to be more accurate than previous state-of-the-art while being much faster.

Experiments and Results

This chapter describes the experimental methods and results achieved in the process of resolving the research questions posed earlier.

4.1 Gait Speed Comparison from Wearable Camera and Accelerometer Vest in Structured and Semi-Structured Environments

To address our first research question *(Can wearable sensors be used to unobtrusively monitor activities of subjects in a semi-structured environment?)*, we limited the scope of activities to a single activity - gait. By considering only a single activity, much of the complexity of the problem was eliminated, instead focusing on determining the ability of various sensors to describe the parameters of the activity in semi-structured environments, i.e. home-like, controlled, indoor setting without the visual noise of environmental motion found in public settings. While the focus was on a single activity and the primary concern was validating the use of wearable sensors - in particular, vision-based sensors - for activity recognition, this work still presented potential for meaningful outcomes due to the strong correlations of gait quality to a number of medical concerns. *This work was described in [94]. A modified version is provided below in this section.*

4.1.1 Overview

Gait analysis is an area of research that has seen an increasing focus due to its applicability to a wide range of age-related health issues which may impact the growing elderly population [78]. For example, Beauchet et. al found evidence that dementia can be predicted by poor gait performance [15]. Similarly, Valkanova and Ebmeier show that evidence strongly supports a relationship between gait and impairment of cognitive functions in patients with mild cognitive impairment and Alzheimer's disease [104]. These findings illustrate the potential of using gait analysis in detecting symptoms of age-related illnesses.

While many objectively quantifiable gait parameters could be used for effective decision support and automated monitoring, the simple measure of *gait speed* has shown to be an accurate predictor of mobility, health, and even mortality [3, 106, 38]. Gait speed is thus a critical parameter for evaluating the utility of candidate gait analysis systems. Therefore, the goal of this study was to determine the feasibility of using our wearable system, comprising an affordable and non-invasive wearable camera and computer-vision based processing methods, to classify the gait speed of healthy individuals. Samples of gait were collected at three self-determined over-ground walking speeds (slow, medium, fast). Since accelerometer-based methods have been successfully used to quantify gait, we deployed an accelerometer near the subject's right hip for the purpose of providing a direct point of comparison for our system to a more widely-used device. We also compared the capabilities of both devices to a research-grade optical motion capture system which represents the "gold standard" for gait analysis. The comparison to both another previously studied wearable system and a high-precision standard provides validation for using our single-camera system for gait analysis tasks.

Expensive laboratory-based gait analysis systems can provide extremely robust quantification of human gait and locomotion. For example, highly precise three-dimensional motion capture systems have recently been used to study detailed gait features across age groups in healthy individuals [27] as well as individuals with Parkinson's disease [29] and

Alzheimer's disease [87]. While these systems can often provide an in-depth description of gait, they are not feasible for the use-case of continuous monitoring due to their size, complexity, and cost. Continuous monitoring of patients in their natural environments during everyday activity provides a more constant and natural sampling of gait activity, enabling the detection of changes in performance over time. Additionally, user-friendly and lower cost pervasive health monitoring systems reduce the burden on patients to make trips to a physician's office, which may be especially cumbersome for the elderly.

Recent work focused on providing convenient, in-home solutions for activity recognition, eliminating the need for a laboratory setup. Video-based methods may offer inexpensive solutions with performance similar to more sophisticated motion capture systems or floor sensors [107], but suffer from obstructed line-of-sight within the home environment. Audio-based systems have also been used to analyze gait in indoor environments [40, 7]. While both audio- and video-based systems have shown promise, their use is limited to a specific preconfigured location. To overcome these issues, gait analysis is also being performed with wearable devices, such as smart watches [100], shoe-based wearable sensors [66], and wearable accelerometers [39, 44, 32, 28]. By instrumenting the subject instead of the environment, the systems become portable and problems such as line-of-sight obstruction can be avoided. While these solutions provide promising results for gait analysis, they tend to be either still too complicated for in-home use (e.g., due to the number of components), or incapable of matching the level of precision and capturing gait performance as comprehensively as laboratory-grade motion analysis systems. Thus, there are trade-offs between accuracy and cost for laboratory-grade motion analysis systems and in-home wearable systems for gait analysis.

Our system used a single head-worn camera to collect first person video. Using computer vision techniques, we were able to extract optical flow output from the video that mimics the ability of a low-resolution accelerometer to register movement parameters [92]. A benefit of a vision-based sensor such as this is that the video data can provide addi-

tional context to the in-home monitoring scenario. For instance, if an unexpected event occurs with respect to gait speed, the monitoring system could notify additional automated or manual review processes to analyze the specific related video segment and determine whether the event was a clinically significant event, such as a fall, or simply an abrupt stop. It may also be possible to analyze the coarse direction of the subject's visual attention while walking and to identify objects that are being interacted with - a key capability that is exploited in the FIS described later in this document - or other factors that may impact gait performance. Methods based entirely on accelerometer or pressure sensors are unable to explain variations or interruptions that are seen in daily gait activities and would not be able to inform further analysis processes. While vision-based in-home monitoring systems may raise privacy concerns by recording subjects and others in their home, automated methods of processing the video upon recording would eliminate the need to store the raw video, mitigating the privacy risk. The storage or transmission of computed motion-based features removes all identifying information as these features are essentially equivalent to those recorded by inertial sensors, which do not raise the same concerns.

Our previous work based on the use of a wearable camera (and accelerometer for comparison as a well-established device) involved collecting data from subjects on a treadmill [91]. However, limiting the data collection to occur on a treadmill was an artificial constraint, especially for the in-home use-case being described. In this experiment, we remove the limitation and also investigate the impact of changing this aspect of the experimental design to incorporate semi-structured, natural, over-the-ground gait sequences from nineteen participants.

4.1.2 Data Collection

Accelerometer, motion capture, and first person video data were collected from nineteen participants as they walked over ground six times, covering a distance of four meters, in a

Figure 4.1: The motion capture laboratory in which data was collected from participants

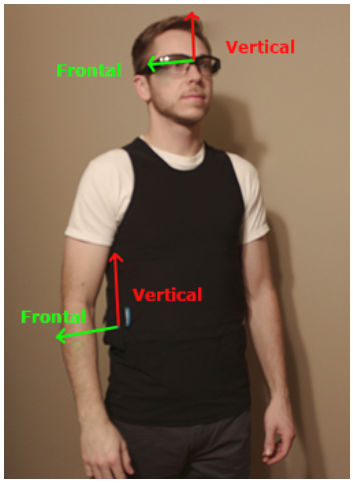


Figure 4.2: Placement of Sensors on Subject and Alignment of Axes Between Sensors

large motion capture laboratory. The space was free from any physical obstructions such as furniture or walls (Figure 4.1). The participants were healthy college students ranging from 18 - 21 years old. Ten of the participants were male and the remaining nine were female. Participants were instructed to walk at three self-determined speeds: slow, medium, and fast. Categorical speeds were used for multiple reasons. First, as participants were walking over the ground, and not on a treadmill, it would have been difficult to adequately control gait speed. Second, the time series data from inertial sensors and the processed video features indicated the frequency of each subject's gait, which directly correlates to gait speed. Estimating the continuous gait speed requires knowledge of the exact stride length

of a subject. Categorical gait speed was thus more appropriate as slow, medium, and fast gaits naturally correspond to an increasing step frequency regardless of stride length or distance covered.

Each subject wore two commercial devices during data collection (see Figure 4.2). The Pivothead SMART Architect Edition glasses [81] were used to record video of the activity in high-defintion resolution (1920 x 1080) at 30 frames per second. The device is a pair of eyeglasses with a camera located in the center of the glasses, above the nasal bridge, aimed directly forward. The glasses are nearly indistinguishable in shape and weight from a normal pair of glasses, providing a comfortable and natural sensor that is easily integrated into daily routine with no encumbrance or health risks to the wearer. The Hexoskin smart shirt [21, 13] was also worn, providing tri-axial accelerometer readings at 64 Hz (data from the remaining sensors in the Hexoskin shirt were not analyzed for this study). The accelerometer was located near the right hip on the torso of the subject inside a pocket of the shirt. Gait data were also recorded with a 20-camera Motion Analysis Corporation Kestrel motion-capture system at a sampling rate of 120 Hz, and motion data were processed using Cortex v. 6.2 software (Motion Analysis Corp., Santa Cruz, CA). Each participant was instrumented with motion capture markers according to the Cleveland Clinic marker set. This model includes markers tracking the position of the feet, legs, trunk, arms, and head. While the position of each marker was recorded, our analysis focused on the head marker (for comparison to the head-worn glasses results) and estimates of whole-body center of mass derived from the global marker set using a whole-body mass model calculated from Zatsiorsky-Seluyanov's body segment inertia parameters [31]. Each of the three systems independently and simultaneously recorded the gait sequences that were performed.

The Pivothead glasses were purchased for $300 USD and the Hexoskin vest and device may be purchased together for $499 USD, making them easily available to consumers. While the exact price of the motion capture system is not immediately available and will vary based on configuration, the cost of the 20-camera system is roughly $100,000 USD. A

Figure 4.3: Mean Plots with Standard Error Bars for Recorded Features By Gait Speed

minimal set of four lower-precision cameras could be obtained for less than $10,000 USD. Even considering the lower-cost motion capture option, the consumer-grade devices have the advantage of being easy to use, while the motion capture system requires an expert user, calibration of the system, and extensive instrumentation of the subject. While the motion capture system presents an excellent means for collecting our high-precision truth data, the cost and complexity dictate that the motion capture system will only be feasible in a controlled clinical laboratory, and not for continuous, in-home monitoring. While we did not investigate real-time processing, the computational requirements for data from the single camera and inertial sensor would also be much lower than the 20-camera system.

4.1.3 Signal Processing

The intent of incorporating the glasses-style camera into the experiment was to use the collected video to describe the subject's head motion in two dimensions (the frontal and vertical planes) throughout the gait sequence. Since the camera faced directly forward from the participant, the frontal and vertical axes in physical space (relative to the participant) correspond to the x- and y-axes of the recorded video. While the camera device collected less data and operated at a lower sampling rate and spatial resolution than a highly accurate 3-D motion capture system, the device was extremely portable, affordable, and simple to operate. However, the lower-fidelity visual data produced by the camera required careful

processing in order to extrapolate information about the movement of the subject who is not in the view of the camera. The Lucas-Kanade optical flow technique was applied to the collected video samples in order to estimate participant motion from the videos [62]. This method produced a displacement vector for a series of significant keypoints within a given video frame. An average vector was computed from all keypoint vectors per frame, resulting in a single two-dimensional vector which represented the overall displacement of the participant in each frame of video. This approach was previously validated against other possible computer vision techniques and was found to provide the most accurate representation of the actual displacement between frames [92].

The two-dimensional (frontal, vertical) components of the optical flow displacement vectors were considered over time to generate two separate sets of time series data. Three-axis time series data were also collected from the body-worn accelerometer, and head and center of mass data (each in three dimensions) from the motion-capture system were also analyzed. The time series data were manually separated into segments collected during each of the six trial walks performed by each participant. The collected data were manually segmented into six trial walks by examining the video and audio, and then recording the start and stop times of each gait segment within the video. None of the systems directly provided a determination of when a gait sequence occurred, though it would be feasible to automate this detection based on the collected features and the video data. For this initial work, such a system was not developed as the focus remained on considering only the data recorded during known gait activities. Motion-capture data were filtered in Cortex using a 4th-order low-pass Butterworth filter with a cut-off frequency of 6 Hz, which is the default setting for the low-pass filter in Cortex. Cut-off frequencies in motor control research generally range from 6 to 10 Hz, depending on the behavior being observed. Given that the observed behavior was walking, 6 Hz was much higher than the frequencies of interest and only filtered out sensor noise.

Because the wearable devices were commercial devices which operated independently

of each other, it was not possible to guarantee perfect synchronization of the data collection across devices. This precluded any direct comparison of the raw time series data from each of the sensing devices, since errors in synchronization of the collected data would negatively impact any calculations. However, it is not necessary to directly compare the time series data - gait speed alone has been shown in clinical applications to be a predictor of cognitive disorders such as dementia [18]. To overcome the limitation on direct comparison between the time series data from each device, we moved data out of the time domain and instead derived frequency-based features in the following manner.

A periodogram calculation was applied over the entirety of each walk segment for each channel of data being considered. As shown in [92], the periodogram transformation can be used to identify main frequencies that occur in each time series. We identified for each time series the frequency with the highest amplitude from the computed periodogram to serve as our gait metric. While the periodogram provides an analysis of constant gait speed in [92] and a successful indication of gait speed in temporally short gait sequences collected for this study, longer gait sequences containing changes in gait speed may be better analyzed with methods that consider time locality, such as a spectrogram. For this study we have assumed that gait speed was constant in each gait segment.

4.1.4 Statistical Analysis

The goal of this work was to determine the feasibility of classifying gait samples categorically by their speed with a specific interest in the performance of the video-based wearable system. We determined whether the collected video-based features were impacted by gait speed in a manner similar to the features from more traditional gait analysis devices. We used analysis of variance (ANOVA) to determine whether our gait metric was significantly impacted by gait speed. Separate ANOVAs were conducted for each plane of motion for each gait measurement system. Data were screened for outliers ($\pm$ 2.5 standard deviations from the median) prior to analysis; 6 trials were identified as outliers, all for the

motion capture system in the sagittal plane and likely resulted from obstruction of one or more markers from the motion capture cameras' view. Violations of the sphericity assumption were resolved by correcting the degrees of freedom of the statistical test using the Greenhouse-Geisser method.

4.1.5 Results

ANOVA on the frequency-based gait metric derived from the eyeglass camera data revealed no significant differences across gait speed conditions in the vertical plane ($p = .55$, $\eta_p^2 = .04$). However, in the frontal plane, there was a significant effect of gait speed, $F(1.49, 22.39) = 36.0$, $p < .001$, $\eta_p^2 = .71$.Pairwise post-hoc comparisons revealed significant differences among all three speed conditions (fast vs. medium: Cohen's $d = 1.96$; fast vs. slow: $d = 1.76$; medium vs. slow: $d = 0.75$; all $p < .05$). This indicated that using the frontal plane data (u for the glasses), the sensor was able to distinguish between gait speeds across the three categories.

As was the case with the camera data, the accelerometer data did not discriminate gait speed conditions in the vertical plane ($p = .26$, $\eta_p^2 = .09$). There was a significant effect of gait speed in the frontal plane, $F(1.6, 23.98) = 11.43$, $p < .001$, $\eta_p^2 = .43$. Post-hoc tests again identified significant differences among all three speed conditions (fast vs. medium: $d = 0.69$; fast vs. slow: $d = 0.99$; medium vs. slow: $d = 0.68$; all $p < .05$).

For the motion capture system, we first considered data from the head marker. A significant effect of gait speed condition was observed in the vertical plane, $F(1.41, 21.1) = 160.2$, $p < .001$, $\eta_p^2 = .91$. All three speed conditions were found to differ significantly according to post-hoc tests (fast vs. medium: $d = 2.78$; fast vs. slow: $d = 3.51$; medium vs. slow: $d = 2.65$; all $p < .001$). A significant effect was also observed in the frontal plane, $F(1.25, 18.79) = 66.97$, $p < .001$, $\eta_p^2 = .82$. All three speed conditions were found to differ significantly according to post-hoc tests(fast vs. medium: $d = 0.78$; fast vs. slow: $d = 3.11$; medium vs. slow: $d = 2.81$; all $p < .01$).

For the center of mass displacements calculated from the motion capture data, ANOVA revealed significant effects of gait speed in each plane of motion. In the vertical plane [$F(1.37, 20.04) = 175$, $p < .001$, $\eta_p^2 = .92$], post-hoc tests revealed significant differences among all conditions (fast vs. medium: $d = 3.17$; fast vs. slow: $d = 3.64$; medium vs. slow: $d = 2.69$; all $p < .001$). Likewise, for the frontal plane [$F(1.19, 17.88) = 68.52$, $p < .001$, $\eta_p^2 = .82$], all pair-wise post-hoc comparisons were significant (fast vs. medium: $d = 1.06$; fast vs. slow: $d = 2.3$; medium vs. slow: $d = 3.81$; all $p < .05$).

4.1.6 Conclusion

In this particular experiment, we evaluated the performance of a gait analysis system which used only a wearable camera to collect two-dimensional, first person video. Optical flow and frequency domain analysis were used to generate a dataset from video, and this dataset was then compared to data collected with a wearable tri-axial accelerometer and a 20-camera motion capture system. This was a crucial step to validate the use of a single-camera wearable system against the vest device and the gold standard motion capture system. While both of the latter devices (the accelerometer and motion capture systems) have seen more use in the domain of wearable gait analysis than the wearable camera, the camera-based system has several advantages including cost, simplicity, and the ability to analyze the visual scene to provide context for the gait analysis data. Although the motion-capture system provided superior discrimination of gait speed in all planes of motion, as expected, the eyeglass-based camera system nonetheless discriminated gait speed significantly and outperformed the vest-based accelerometer system. This suggests considerable promise for its use in unobtrusive activity monitoring in a semi-structured environment.

The results of this experiment demonstrated the ability to detect a clinically significant factor of gait performance using only a wearable camera. Additionally, the experimental setup validated that such a device did not interfere or alter the method of performing the activity. This was an essential outcome for positively concluding that the wearable vision-

based device is both convenient and useful for detecting activities in an unstructured, in-home environment.

4.2 Describing Object-centric Activities using Fuzzy Methods

To address our remaining research questions, (*RQ 2: Can features extracted from first person videos recorded in subjects' homes be used to categorize different interactions with objects?*), and *RQ 3: Can the activity identification system a) accurately describe activities and b) handle uncertainty in the results better than current state-of-the-art methods?*, we extended the limited scope of a single activity from the previous experiment and brought forward the use of a wearable camera and some of the input features from the first person video. After validating the wearable device (Section 4.1), the remaining experimentation focused on identifying the occurrence of activities based on the users' interactions with objects.

4.2.1 Overview

Recognizing and understanding human activities occurring in video is an important enabling technology to a variety of computationally intelligent systems. The spectrum of all possible human activities is extremely broad, with great variance in both the temporal and spatial aspects of activities. There are also visual challenges to consider - a limited first person perspective means objects or the person performing activities may not be visible, activities may involve a number of objects that are challenging to detect based on their size or variation in appearance, and a moving camera may contribute noise such as blurring or over/under-exposure. Visual challenges aside, these activities can be placed on a scale of increasing complexity, with simple (or *atomic*) actions occurring on the low end

of the scale, i.e. activities occurring for a short period of time or with a limited number of movements, and highly complex or composite activities occurring on the high end of the scale, i.e. activities which contain other simpler activities as sub-components occurring in some order over a longer period of time. Identifying the exact point in time when an atomic activity (e.g. picking up a spoon) turns into a complex activity (e.g. making a cup of coffee) with any amount of objective precision is nearly impossible. This dilemma illustrates an inherent amount of fuzziness in the domain of activity recognition. That is, an activity may be described in multiple ways, at multiple levels of specificity, with each overlapping description retaining full accuracy.

In this experiment, we sought to address the challenge of handling the natural fuzziness in activity recognition that arises from the overlapping boundaries between activities and their definitions. We used fuzzy logic to discover object-based activities in first person video of daily activities in kitchen environments. Using a variety of extracted features and object annotations from the video, the system was able to describe a diverse set of activities and provide information on the nature of activities without requiring extensive training. We sought to describe activities more completely and reliably than current uninterpretable or black box methods. Such a system can be used to detect activities more effectively in an in-home setting with applications such as aging in place for older adults, or other rehabilitation applications. A description of a preliminary version of this system was provided in [93].

Traditional work in the field of first person activity recognition is often focused on exploiting information about objects in the scene in order to classify activities being performed [74, 69, 80, 98, 110]. The use of object-based activity models has shown promise, since many of the activities that people perform in their homes on a daily basis are largely based on the use of some object. However, visual noise, such as occlusion of the objects, negatively affects the object recognition accuracy, and thus greatly dampens the accuracy of the activity recognition framework as it relies heavily on the identification of objects the

human is interacting with. Related works have shown that this noise can be reasonably handled with current object recognition techniques. Deep learning and convolutional networks can be very robust in classifying objects which are partially occluded or may have changing appearance based on the state of the object or the angle of the view. [80] takes advantage of visual differences in the appearance of objects over time to determine whether an interaction is occurring.

Many of the more recent works in this field of study have abandoned traditional video features in favor of deep learning techniques to identify activities occurring in first person video. [54] features a recurrent convolutional neural network (R-CNN) trained on RGB images, optical flow, and gaze annotations. Similarly, [88] describes the use of CNN features in a pooled time series representation in combination with a non-linear SVM to identify activities. [82] applies the Hilbert-Huang transform to CNN features and uses an SVM classifier to classify activities. CNN-based techniques are favored for their ability to discover useful features in images and video, but the networks require extensive training. Due to the amount of training required, the set of activities classified by each of these CNN-based systems tends to be fixed and small relative to the complete set of daily human activities.

While the deep learning techniques may perform well for a small set of actions when trained against extensive annotated datasets, the requirement of massive amounts of labeled data is a huge limitation in the utility of such methods for activity detection and recognition. Given the multitude of activities that may occur on a daily basis, it is not simple to build a training dataset for a system that recognizes a complete set of these activities. Additionally, when these deep learning methods fail to recognize an activity, there is often no meaningful output - the deep learning techniques operate as a "black box" and the correlation between features and the output of the system is not usually well understood outside of that box. Consequently, the identification of an activity has a binary outcome, and when the classification is incorrect, the system provides little or no useful information.

We seek to remove these limitations through our use of fuzzy logic. In building a FIS to classify activities, extra work must be done to extract and understand video features and their correlation to activities recognized by the system, but the understanding of these features and parameters provides insight into the type of activity being performed even if the actual name of the activity is unknown. That is, a fuzzy system can provide a result which indicates aspects of the occurring activity without the need to positively identify the activity, incorporating uncertainty that is extremely useful for reliable performance in dynamic environments [8]. For many purposes of activity detection, including in-home monitoring of elderly [11, 9] and medical patients, human-machine interaction, or life-logging, the fuzzy information may be informative where artificial intelligence techniques produce no useful information.

We took inspiration from existing methods and extracted both motion- and object-based features from a set of first person videos. Following is a description of a fuzzy inference system that uses these features to classify activities occurring in these videos. To our knowledge, the application of fuzzy methods to egocentric activity classification within this context is a novel approach.

Our goal was to create a fuzzy inference system that classifies a set of activities being performed by the subject of first person videos. We acknowledge that the recognition of an activity is not always a binary decision given the relationships between various atomic and complex activities. For example, multiple atomic activities may occur simultaneously as part of a more complex activity. For this reason, we demonstrate a fuzzy system of activity recognition which provides a depiction of multiple aspects of the activities being performed. A simple diagram of the system is provided in Figure 4.4.

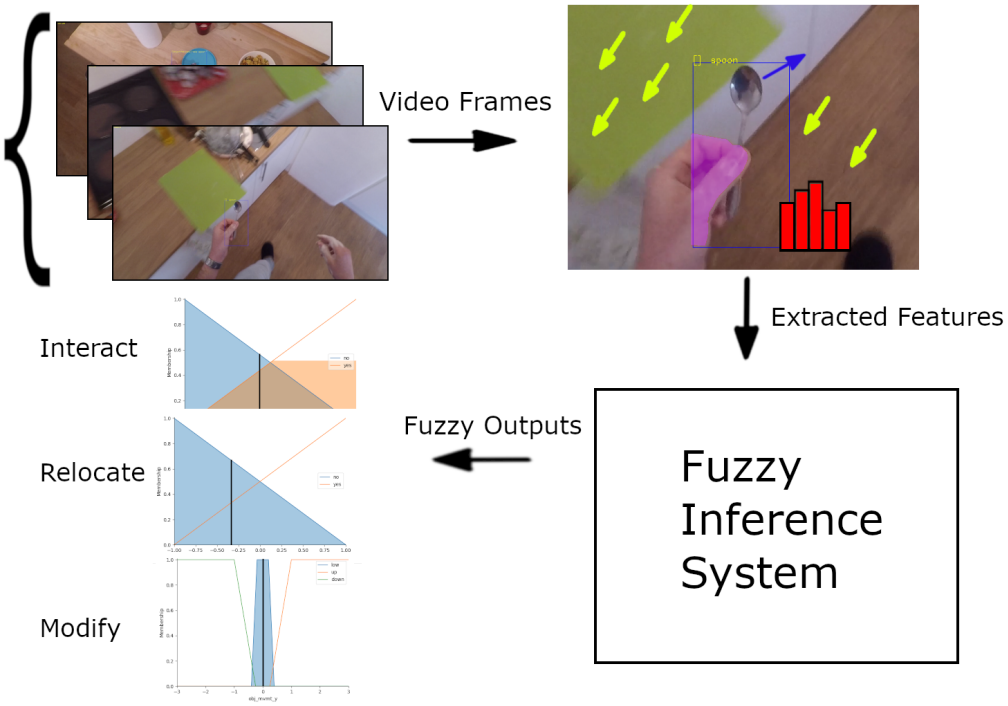


Figure 4.4: Illustration of our system. Features such as optical flow motion vectors and color histograms are extracted from video frames and passed to the fuzzy inference system, which calculates fuzzy membership in three outputs based on a series of rules within the system.

4.2.2 Dataset

In previous experiments, the target environment for activity detection was either structured or semi-structured. This control of environmental variation lessened the number of factors which impacted the results, but was not realistic for a system that is capable of daily, in-home use, which is a key piece of research question 2. For this experiment, the environmental control was completed lifted to include video that was recorded in a true home environment. To avoid the complication of collected a custom dataset for our work, we surveyed existing datasets to find one that provided appropriate videos and labeled truth.

For this work, we used the EPIC-KITCHENS 2018 dataset [30]. This dataset contains unscripted first person videos of subjects performing daily activities such as cooking, washing dishes, and cleaning in their kitchens. The dataset provides annotations of objects

Figure 4.5: A sequence of still frames at 15 frame intervals from a sample activity annotated in the first three frames as ”take plate” and in the latter three frames as ”wash plate”. The change in activity label is an indication of the uncertainty inherent in activities, their definitions, and their spatiotemporal boundaries as they are performed.

of interest, as well as annotation of the action involving the objects. The action annotations provided with the dataset comprise over 100 different verbs and highlight the amount of variability found in daily in-home actions in just a single room. The dataset is unscripted, which was an important feature to ensure that the data truly represents the intended use-case for our research questions. Figure 4.5 shows a series of still images from a sample activity to show the perspective of the dataset.

While the intent of the proposed FIS system is to describe a potentially unconstrained set of daily activities, it would not have been feasible to simply build the FIS to output a single classification out of hundreds of activities. The FIS was instead built to output fuzzy

descriptive features of the activities, classifying them into multiple categories of intent. While this meant that the FIS was not producing a specific activity label, this classification method provides insight into what is actually happening in the activity, which can provide description of even activities that the system was not trained on, alleviating the limitation that many classifiers have with only performing on a specific set of activities.

The output of the FIS focuses on three types of object-actions: interaction, modification, and relocation. Modification intent was defined as an action that intends to change the appearance or structure of an object. Relocation intent was defined as an action which intended to move an object in space. Other interactions which did not seek to perform modification or relocation were simply labeled 'interaction'. A categorization of main intent was determined for each annotated truth action in the dataset and appended to the annotations. While our expectation was that actions will have some membership in more than one output category, each action was given a single label for the most likely intent. For example, the action 'pick up' provided in the original annotations received an intent label of 'relocation', the 'cut' action was labeled as modification, and the 'wash' action was labeled as 'interaction'. The additional intent labels were determined per action class and not per individual instance of the action, i.e. a given verb was considered to always have the same intent.

4.2.3 Features

As previously mentioned, building an FIS to analyze actions in video required manual feature selection and extraction. We build our fuzzy inference system around a set of features that are extracted from the three-channel first person videos described in the prior section. The recognition of activities via the fuzzy rules in our system is predicated largely on the identification of objects in the video. The presence and motion of individual objects in the scene and their relation to the subject are critical inputs to the fuzzy inference system. The set of input features was built intuitively from this observation.

Table 4.1 lists each of the video features which are used as inputs to the fuzzy inference system.

Object Features

The fuzzy system was provided with a rectangular bounding box for each object in the scene. The bounding box for each candidate object was read from the dataset annotations, i.e. no object detection was implemented for this work. The annotations include bounding boxes at all times when objects involved in actions as well as some cases when they are not in use. The bounding box information was used to isolate the object for calculating several of the object-descriptive features discussed below.

Centrality The one feature which is computed directly from the bounding box is the centrality metric. An analysis of action locations in a subset of the training data revealed that the distribution of actions tends toward the center of the camera's perspective. This result is expected, since the center of the video represents the center of a subject's field of view, and thus a likely location for the initiation of an action. Figure 4.6 illustrates the distribution of actions in this training data. The starting positions of all objects involved in activities in our training data were plotted to verify the assumption that the center of the video represents the main area of the subjects' attention and that this is were activities tend to occur. As shown, few activities occur in the corners of the field of view. Centrality of each object is computed by the two-dimensional distance from the center of the video, (v_x, v_y) to the center of the bounding box (b_x, b_y), subtracted from 1 to achieve high centrality when the distance is low and low centrality when the distance is high:

$$centrality = 1 - \sqrt{(v_x - b_x)^2 + (v_y - b_y)^2} \tag{4.1}$$

Table 4.1: Video Features

Feature	***Description***
Object bounding box	the location of a rectangular bounding box around each identified object in the video
Object label	the name of each identified object
Frame optical flow	the mean, 2-dimensional optical flow vector in the video frame computed between frames t_0 and t_1
Object optical flow	the mean, 2-dimensional optical flow vector in the video frame computed between frames t_0 and t_1 within the bounding box of an object
Color histogram difference	the distance metric of the color histogram within the bounding box of an object in frames t_0 and t_1
Hand intersection	the ratio of pixels within the bounding box of an object that are identified as belonging to a hand
Keypoint similarity	a ratio of matching keypoints to all keypoints detected on an object over a time step
Object centrality	a measure of how near the object is to the center of the visual window.

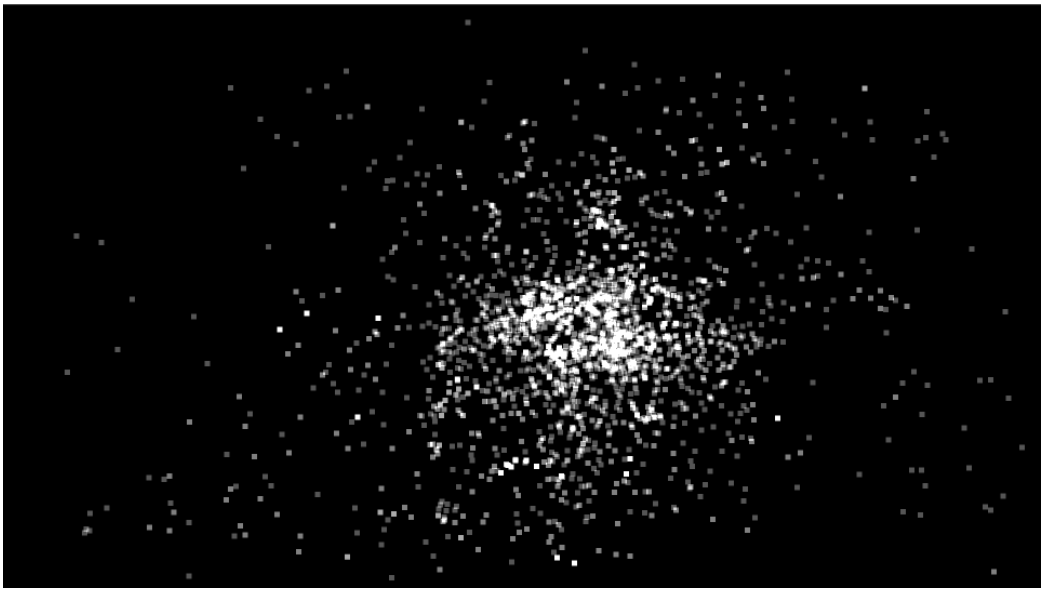


Figure 4.6: Illustration of the distribution of action locations within the training data.

Object Appearance The change in appearance of an object over time is another logical feature that may inform whether the object is being interacted with, and whether the object is being modified or manipulated in some way. Previous work has been done on exploiting the change in appearance of objects to indicate the occurrence of activities [80]. Two types of features were implemented to provide insight into different types of change to the object.

Change in color is one of the simplest ways to perceive changes in an image. Features derived from the distribution of colors within a region of an image have long been used in literature to provide a measure of object similarity for use in object comparison and tracking applications [101, 24, 108, 122]. While more complex methods are required for complex tasks such as tracking objects, our system implemented a color histogram difference feature. This feature was useful for approximating the amount of change in an object's appearance over a time period. The color values were split into eight bins in each channel of the image to build the histogram, and the histogram was computed for all pixels within the object bounding box for two different frames. The final difference feature provided to the FIS was given by applying the intersection distance metric between the two histograms.

While color information is a simple indicator of change in object appearance over

time, it is frequently the case that an object may be modified in a way that does not change the color of the object. In these situations, the color histogram feature alone was unable to inform the FIS on whether the object had been modified. To improve the robustness of the overall indication of object change, keypoint-based features were also introduced to the FIS. Keypoint detectors are responsible for choosing points within an image that are most meaningful to the image. Most often keypoint detectors identify points like corners, edges, or other complex shapes or textures that are integral to the appearance of the image. These keypoints can be compared across images to perform object recognition [76, 75].

In the case of our FIS, the keypoints were detected within an object bounding box and matched between two different frames, and the ratio of matching keypoints is used as a keypoint similarity metric. Two algorithms were implemented for keypoint detection - CenSurE [4] and ORB [86]. While the information provided by each of these keypoint detectors is similar and likely redundant in some cases, each detector performs differently in different scenarios so both were provided as inputs to the system.

Motion Features

In addition to features describing aspects of the objects like their location and appearance, our FIS also required knowledge knowledge of motion within the input videos.

Optical Flow The motion features referred to above were implemented using optical flow. There are two pieces of information being provided by optical flow - the motion within the object bounding box, and the motion within the entire video frame.

Since the first person video is recorded by a mobile, body-worn camera, there is expected to be overall motion in the video which is a consequence of the movement of the camera itself. We equate overall motion in the video with the movement of the camera, which equates to motion of the subject wearing the camera. While the movement of the subject is not directly used to inform the activity classification process (i.e. no mapping of

the environment or the subject within the environment was considered), it was necessary to consider it in order to discern the motion of the objects in the scene relative to this global motion factor.

Since the camera motion was additive to the motion of the objects, the fuzzy system considered an object to be moving with the scene only if the perceived motion of the object did not match the global motion of the scene. This implied that the object had some motion relative to its surroundings (presumably through interaction with the object by the subject). Both optical flow features were calculated using the dense Farneback method [36] since computation time was not a priority. The dense optical flow calculation is far less time efficient than sparse implementations, but performs more reliably in cases where high-quality keypoints are limited. Furthermore, due to the limitation of bounding boxes to rectangular shape, the boxes were not always tightly fit to the object in all cases. Using sparse optical flow risked the situation where the perimeter of the bounding box may have picked up keypoints from a separate object, and the motion of that other object may have skewed the resulting motion estimation. It was not possible to guarantee that a majority of the keypoints belonged to the object of interest. Using dense flow helped the true object motion to reduce the contribution of such noise since a majority of pixels in the box should belong to the object of interest.

Hand Detection

The object motion and appearance features were useful for indicating information about the result of interactions on objects, i.e. how they move or change, but it is intuitively possible for an object that is not involved in an activity to move or change appearance. The centrality measure mitigated the risk of misclassification, but is just one metric supporting such a decision. To further avoid the false detection of activities, a final feature was introduced to compute the overlap between objects in the scene and a subject's hands. While hands are not always visible in the given experimental setup, many object-centric activities involve

contact between the hand and the object.

The major hurdle in computing contact between hands and objects was detecting the hands in the frame. For this, we used machine learning to implement a hand detection network. Two initial attempts were made using a pre-trained, open-source hand-detection network trained on the EgoHands Dataset [10]. The first attempt used the model provided by the authors of [10]. This model was built to classify a provided region of an image as either hands or background. The selection of windows is critical with this method, since hands can only be detected in proposed windows. A window proposal method was also identified in [10]. Resulting detections were much too sparse for use in our FIS as a vast majority of interactions were being missed. Attempting to slide windows over the entirety of each frame, which was computationally inefficient but guaranteed all windows containing hands were proposed, revealed that the detector was not sensitive enough for identifying hands.

To eliminate the need to propose windows for the hand classifier, a second attempt was made which used a Region-based CNN (RCNN). RCNN's are capable of receiving the entire image as input and outputting a set of object bounding boxes and classification scores per bounding box. The RCNN was an open-source implementation which was also pre-trained on the EgoHands dataset. This solution proved equally limited and produced too few hand detections to be useful, leading to the conclusion that the data in the EgoHands dataset was not similar enough to the EPIC KITCHENS dataset.

The final, working hand detection solution was produced using an implementation of the MobileNetV2 single shot detection network [89]. Transfer learning was used to gain efficiency in the training process. The network was initialized with a version of the model included in the Tensorflow Model Zoo which was trained for object detection on the Microsoft Common Objects in Context (COCO) 2017 dataset [59]. This initial network was capable of classifying over 90 object types, and modified to classify only a single label - hands. To ease the data demands for the training process, the model was trained on the

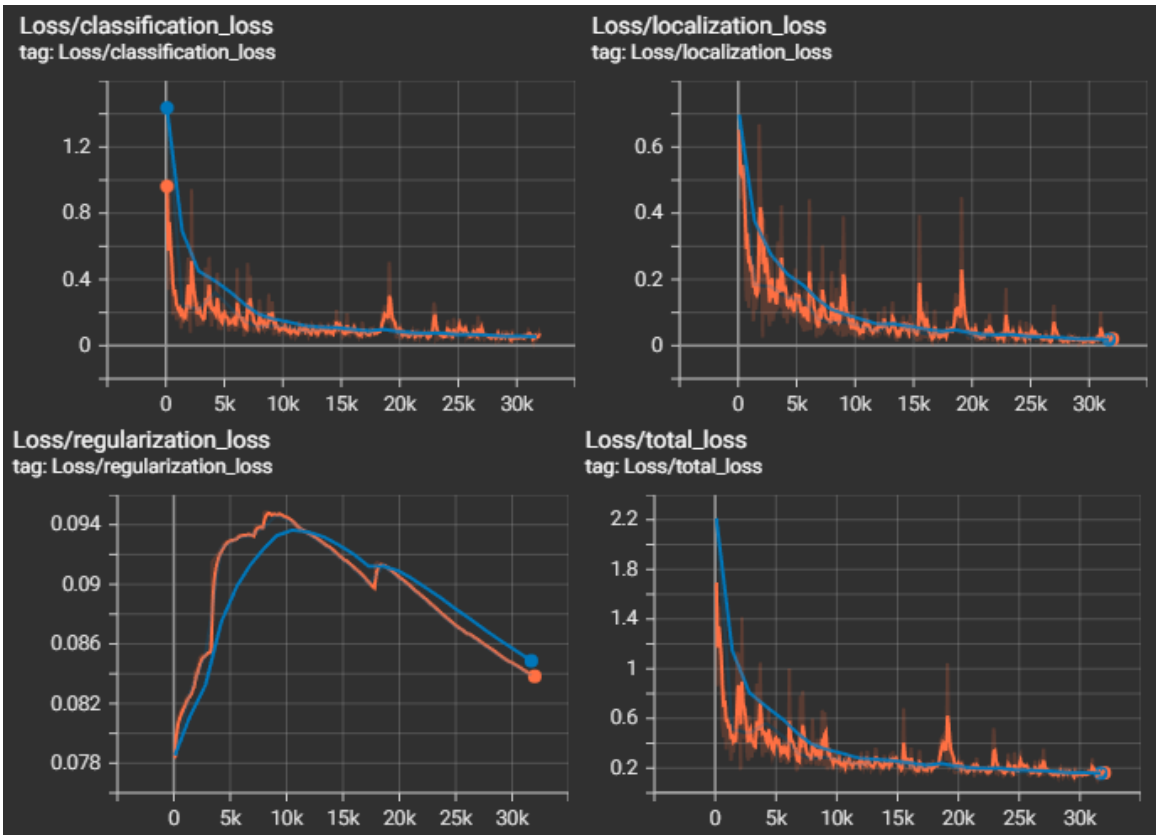


Figure 4.7: Loss curves for the final training of the hand detection model on images from the EPIC KITCHENS dataset.

same EgoHands dataset that the prior two models were trained on. Because the network had previously been trained for object detection, the network reached high accuracy within a few hours of training. Because the earlier models had difficulty detecting hands in the EPIC KITCHENS dataset (even after successfully training on EgoHands), our final model was then also trained on a set of 660 manually annotated images of hands from the EPIC KITCHENS videos. As shown in the loss plots for the final training run (Figure 4.7), the model trained very quickly on this final dataset despite being a very lightweight and computationally efficient model.

4.2.4 Fuzzy Inference System

In this section we detail the structure of the proposed FIS. Each of the previously discussed input features was created to support an intuitive set of fuzzy rules which were used to produce the final outcome of membership levels in the three action categories of interaction, modification, and relocation.

Fuzzy Membership Functions

The input layer of the proposed FIS comprised a set of fuzzy definitions for each of the input features. Tables 4.2 and 4.3 summarize the sets created for the fuzzy input and output features. The two types of the motion vectors were reduced into two features - one describing the difference in the amount of motion between the object and the global frame, and one describing the difference in the direction of motion between the object and the global frame. Each of the other features discussed above was directly represented by a single input variable.

We were able to capture the uncertainty in the feature space using the linguistic variables in the fuzzy model. For example, to compensate for the level of noise expected in the optical flow calculations, a "low" membership term was used to represent a lack of significant motion, or motion within an amount of error that a direction could not confidently be determined. The number of terms belonging to each input set was initially determined through manual experimentation and knowledge of the feature extraction methods. Analysis of the genetic evolution of the FIS was used to eliminate terms which were not useful or to expand terms which appeared to provide improved results.

Triangular membership functions were used for all membership functions within the system in order to keep the definition of the functions as simple as possible, which simplified the encoding of the system for the genetic algorithm discussed later. Trapezoidal functions were tested with no apparent improvement in the results.

Table 4.2: Input Variables to the Fuzzy Inference System

Input Linguistic Variable	***Terms***
Motion Magnitude Difference (MAG)	{low (LO), medium (MD), high (HI)}
Motion Direction Difference (DIR)	{low (LO), medium (MD), high (HI)}
Color Histogram (CH)	{very low (VLO), low (LO), medium (M), high (HI), very high (VHI)}
Hand Intersection (HINT)	{low (LO), high (HI)}
CenSurE Keypoint (CKP)	{ low (LO), medium (M), high (HI)}
ORB Keypoint (ORB)	{ low (LO), medium (M), high (HI)}
Object Centrality (OC)	{ low (LO), medium (M), high (HI)}

Table 4.3: Output Variables for the Fuzzy Inference System

Output Linguistic Variable	***Terms***
Object Interaction (INT)	{unlikely (UN), somewhat likely (SL), very likely (VL) }
Object Modification (MOD)	{unlikely (UN), somewhat likely (SL), very likely (VL) }
Object Relocation (REL)	{unlikely (UN), somewhat likely (SL), very likely (VL) }

Table 4.4: Rules for the Output "Interaction" (*HINT*=hand intersection, *OC*=object centrality, *INT*=interaction)

Inputs			***Outputs***
HINT	**OC**		**INT**
LO		$\longrightarrow$	UN
	LO	$\longrightarrow$	UN
	MD	$\longrightarrow$	SL
HI		$\longrightarrow$	VL
	HI	$\longrightarrow$	VL

Fuzzy Rules

The rules in the fuzzy system control the way in which the inputs are transformed to the outputs. The set of rules used in the system are built intuitively from knowledge of the inputs and how they are expected to change with different activity types. Following is a discussion of the rules in the system grouped by which output action class they relate to.

Interaction The level of interaction occurring in a candidate object-action pair was determined by two of the inputs to the system - object centrality, which represents an approximation of the amount of attention being paid to the object by the subject, and hand interaction, which represents an approximation of the amount of hand contact the subject has with the object. The rules affected by these input values are represented in Table 4.4.

Modification The membership level of the modification output class was determined by the color- and keypoint-based features. The larger the amount of detected change in either feature, the larger the confidence that the action was a modification action. Additionally, the membership of the action in the 'interaction' class is considered for the modification class, i.e. if the object was not being interacted with, there should be low confidence that the subject is modifying the object. Note in this case that for the keypoint-based features, the terms 'low', 'medium', and 'high' refer to the level of matching, where low matching

Table 4.5: Rules for the Output "Modification" (*INT*=interaction, *CH*=color histogram, *CKP*=CenSurE keypoint similarity, *ORB*=ORB keypoint similarity, *MOD*=modification)

Inputs					***Outputs***
INT	**CH**	**CKP**	**ORB**		**MOD**
LO				$\longrightarrow$	UN
	VLO			$\longrightarrow$	UN
	LO			$\longrightarrow$	UN
		HI		$\longrightarrow$	UN
			HI	$\longrightarrow$	UN
	M			$\longrightarrow$	SL
		M		$\longrightarrow$	SL
			M	$\longrightarrow$	SL
	HI			$\longrightarrow$	VL
	VHI			$\longrightarrow$	VL
		LO		$\longrightarrow$	VL
			LO	$\longrightarrow$	VL

would provide stronger support for a change having occurred to the object. The rules affecting the modification output are show in Table 4.5.

Relocation The motion features naturally formed the basis for the output membership level in the relocation class. Similarly to the modification rules, the level of interaction informs the level of relocation, since it is assumed that a relocation cannot occur without an interaction. The rules affecting the relocation output are show in Table 4.6. For rules with multiple inputs on the same line are created using the conjunctive operator AND.

The relocation rules are simply based on a difference between the magnitude and direction of both the object and global frame motion vectors. The magnitude difference is the absolute difference between the magnitudes of the vectors, and the direction difference is computed as the angle between the two vectors using the inverse cosine of the dot product.

Table 4.6: Rules for the Output "Relocation" (*MAG*=difference in motion magnitude, *DIR*=difference in motion direction, *INT*=interaction, *REL*=relocation)

Inputs				***Outputs***
INT	**MAG**	**DIR**		**REL**
LO			$\longrightarrow$	UN
	LO	LO	$\longrightarrow$	UN
	MD		$\longrightarrow$	SL
		MD	$\longrightarrow$	SL
	HI		$\longrightarrow$	VL
		HI	$\longrightarrow$	VL

Fine Tuning with Genetic Algorithm

While the inputs, outputs, membership functions, and rules described thus far were built on intuition and experimentation within the problem domain of activity classification, there still remained possibility for optimizing the performance of the system. The optimization was performed by using a genetic algorithm to evolve the membership functions for each variable in the system.

While the FIS derives power from interpretability and explainability of the results, tuning the membership functions with a GA does not remove these benefits. It is difficult to intuitively determine, for example, where the lower or upper bounds of a membership function should reside for a 'low' movement term. Clearly the relation of 'low' to 'high' is intuitive, but the number of pixels of motion per frame that should represent 'low' amounts of motion is unclear. The GA provides a method of extracting sensible values for each membership function while optimizing the performance of the system.

Genetic Representation In order to evolve the FIS using a GA, a genetic/chromosomal representation of the functions defined within the system was required. Since the triangular membership function for each term was defined by a set of three values representing the lower, middle, and upper extents of the triangle, a real-valued triplet encoding was used to encode each function. The entire gene, then, was represented as a collec-

tion (dictionary) of real-valued triplets. An excerpt of one such genetic representation is provided in Listing 4.1. This is represented in short form as a nested list of values: $\{[-3.915, -3.00, 2.424], [-2.353, -0.132, 2.024], [-2.799, 3.477, 3.794], ...\}$

Listing 4.1: Example JSON for Genetic Representation

```
{
  ‘‘frame_movement_x’’: {
    ‘‘ left ’’: [−3.915, −3.000, 2.424],
    ‘‘low’’: [−2.353, −0.132, 2.024],
    ‘‘ right ’’: [−2.799, 3.477, 3.794]
  },
  ‘‘obj_movement_x’’: {
    ‘‘ left ’’: [−3.820, −3.729, 3.868],
    ‘‘low’’: [−3.655, 0.786, 2.917],
    ‘‘ right ’’: [1.096, 3.00, 4.258]
  },
 ...
}
```

For the process of evolving a high-accuracy solution, the population was initialized with 20 randomized individuals. Each individual was tested against a set of actions from the EPIC KITCHENS dataset to evaluate fitness. In an attempt to retain negative prediction value, annotations of objects with no action were also included in this training set since many objects in the scene may be part of no action.

Roulette wheel selection was used to select parents for new generations, meaning that each parent had a chance of selection proportional to their fitness relative to the fitness of

the other individuals. Single-point crossover was used in the mating process, and mutation occurred with a probability of $p = 0.20$. When selected for mutation, a value in the system was allowed to randomly move within $\alpha = 1.2$ times the distance to it's neighbors. Since $\alpha > 1$, it was possible for a term to surpass it's neighbor. This was allowed in order to not constrain the evolutionary process. As an implementation detail, care had to be taken to keep the system well-formed, e.g. that the lower bound on a triangular function did not become greater than an upper bound. Following all mutations, the values in each triplet were re-sorted to ensure that the lower, middle, and upper values were appropriately ordered. In this way, a mutated lower value was able to become the middle value of a triangle if it happened to mutate past the original middle value, and so on.

Several fitness functions were experimented with during the training process. As is the case with any classification problem, the accuracy of the system must be weighed against the impact of different types of errors. Training for high accuracy often led to a bias in positive or negative results for a particular category, but the F1 score provided an appropriate balance between accuracy, precision, and recall and was ultimately used as the fitness metric during training.

4.2.5 Results

This section provides details on the results of the methods used to classify activities using the FIS, including the custom built hand detector, the tuning of the FIS with genetic algorithms, and finally the overall results of classifying activities with the FIS.

Hand Detection

A key factor in the overall output of the FIS is the hand detector accuracy. Using the model and training methods described earlier, hand detection reached a mean average precision (mAP) of 0.995 at 50% intersection over union (IOU), and a mAP of 0.990 at 75% IOU

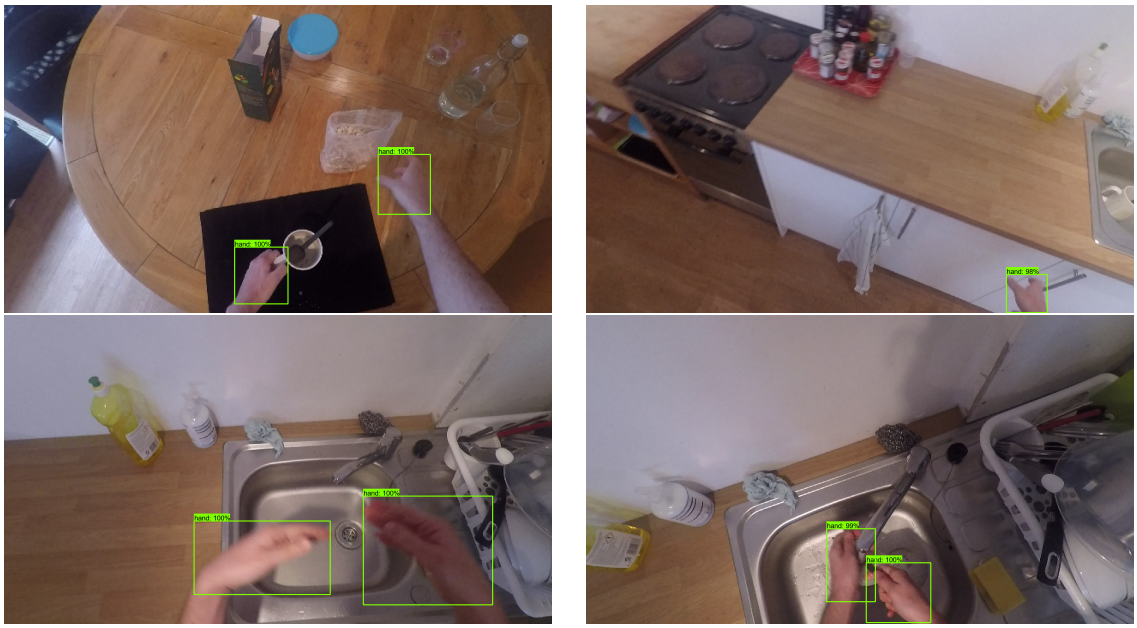


Figure 4.8: Example first person video frames of the hand detection model. Hands were detected in these images with a confidence ranging between 0.96 to 1.0, highlighting the robustness of the model to different orientations and poses of hands in the view.

when trained on the samples from the EPIC KITCHENS dataset. Figure 4.8 illustrates samples of hand detections obtained from the final dataset.

Prior to training on the EPIC KITCHENS data, the model was trained on the EgoHands dataset to a mAP of 0.970 at 50% IOU and mAP of 0.864 at 75% IOU. While further training would have likely increased the mAP metric, prior experimentation had revealed that the EgoHands dataset samples were not completely representative of the types of data in the final target dataset, so training was stopped at this level of accuracy. Figure 4.9 provides an illustration of the types of images in the EogHands dataset, which contains both first- and third-person hands playing tabletop games. The background of the hands tend to be more uniform than in our dataset, which was likely the cause of poor detections using models trained only on EgoHands.

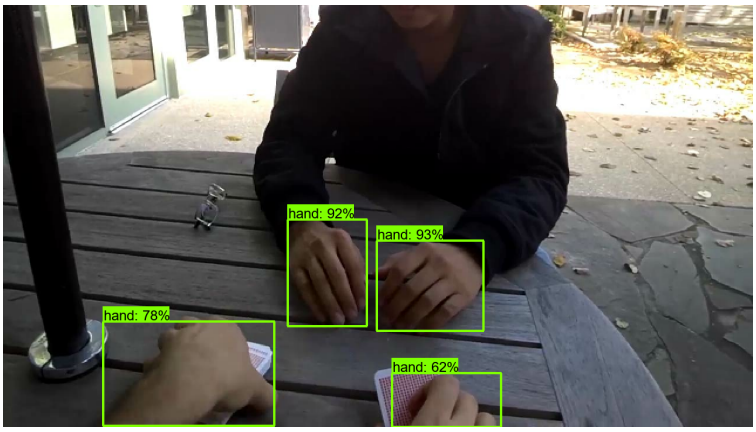


Figure 4.9: Example hand detection result from the EgoHands dataset prior to training on EPIC KITCHENS data. The hands do not move from the vicinity of the table, leading to poor detection in datasets with more variation in background.

Genetic Algorithm Tuning

As mentioned above, the FIS was fine-tuned using a genetic algorithm. Membership functions within the system were evolved to find a final solution that optimized the F1 score. While the rules of the system were intuitive by design, the shape of the membership functions was not always so, especially for features such as color histogram difference and keypoint similarity ratios. While the relative values and that came from these features are intuitive (i.e. higher score indicated higher change in appearance), it was not immediately clear which values might constitute a 'low', 'medium', or 'high' value for the feature. The GA method of tuning the system mitigated the risk of choosing poor values.

For the tuning, a random population of 20 individuals was initialized. While the initial population was random, care was taken to structure the membership functions in a sensible way, such that the entire universe of values for each feature was covered by at least one membership function. A training dataset of 3,560 actions was sampled from a seven videos in the EPIC KITCHENS dataset. The set of actions was used to evaluate the fitness (F1 score) of each individual in the population. Due to the randomization of the individuals, initial results varied slightly, but in the final training run, the best randomly initialized individual began at an overall F1 of 0.235 and overall accuracy of 0.531. By the end of training,

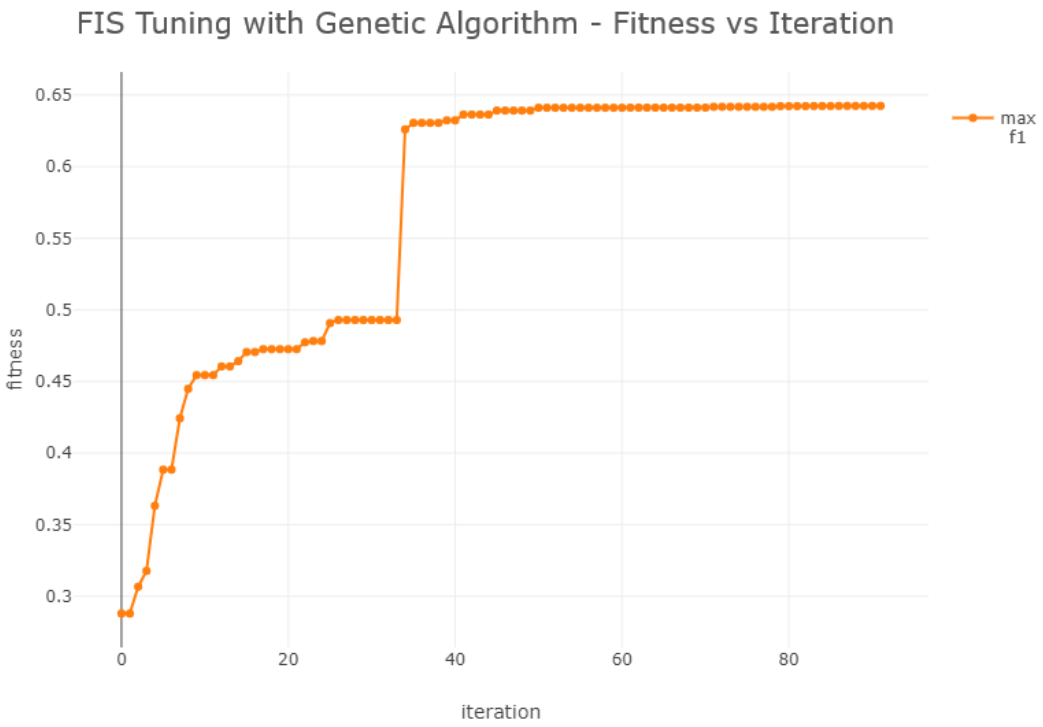


Figure 4.10: Progression of the GA in optimizing the fitness metric (f1-score) over generations of the GA.

the population achieved an overall F1 score of 0.642 and an overall accuracy of 0.782 on the training set. Running multiple rounds of GA showed that the results consistently converged near this solution.

Figure 4.10 illustrates the results over each iteration of training. The training results follow the expected curve of early increases in performance which diminish over time as the system reaches an optimized output. Training was halted after several rounds with no progression on the fitness measure.

Discussion The GA was an extremely useful tool for optimizing the performance of the FIS. While it was only used in this work to optimize the membership functions, the training results were also useful for performing feature selection and pruning. For example, the GA training was used to determine whether both keypoint features were necessary, or if one

consistently outperformed the other. Both were kept because the solutions evolved with both features achieved better results. Figure 4.11 shows that the GA evolved very similar functions for these inputs, indicating that they are providing similar information to the system. However, a closer look at the data revealed that despite producing similar information, each method of keypoint extraction tended to perform well in different situations, so both features were necessary to keep consistent results through this redundancy.

Initially, the FIS was structured to take four motion-based inputs - object movement in two dimensions and global frame movement in two dimensions. The fuzzy rules were structured to detect a difference in these inputs, i.e. if object movement in the vertical axis is 'down' and global frame movement in the vertical axis is 'up', then the object must be moving relative to its environment so relocation is likely. Earlier versions of the FIS included several terms for each direction of movement for both object and global frame motion, i.e. 'very left' and 'left', 'slightly left', but many of these were dropped from the system when their membership functions were absorbed into their neighboring functions. Similarly, Figure 4.11 illustrates how 'low' histogram difference was nearly absorbed by 'very low'. This result likely meant that the terms could be combined.

After further use of the GA to fine-tune the FIS, it was apparent that preserving the information in each dimension for each input motion feature was confusing the results of the system. In many cases, the system evolved very different ideas of motion with respect to the object and global frame. This is likely due to error in the measurements causing conflicts in the training data. To eliminate this issue, the difference in both magnitude and direction of the two motion vectors was used instead. This allowed the system to train a two membership functions (one for each feature) instead of four (two for each dimension of each feature, giving much better results. Pre-processing the motion estimates into the appropriate inputs was a trivial amount of overhead.

Further improvement may have been possible by encoding the structure of the rules and evolving the entire structure of the system rather than just the parameters within it

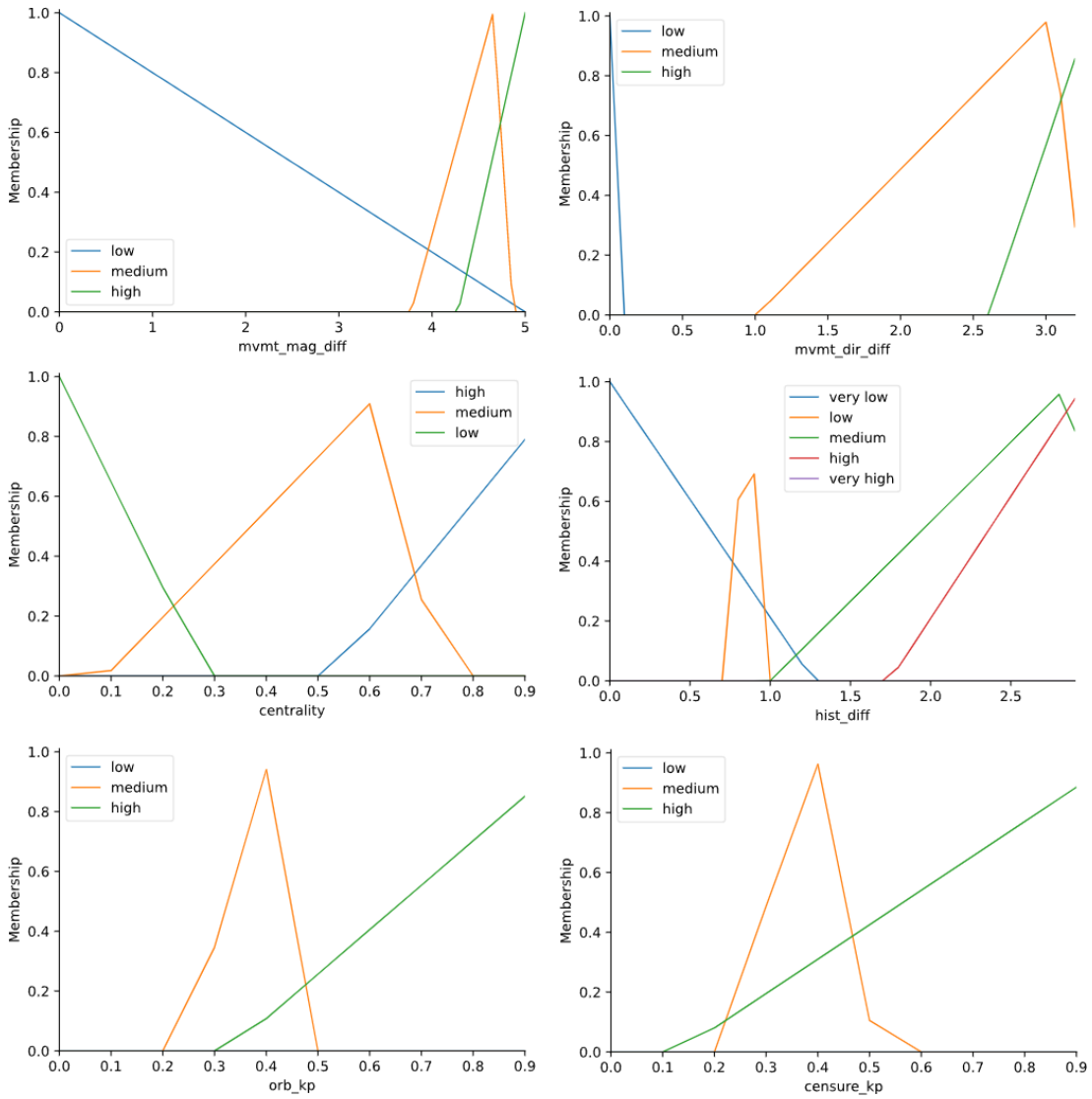


Figure 4.11: Final membership functions output from evolution of the FIS using a GA.

[111]. However, for the proposed system with a small number of features, it was decided that the intuition in the manually formed rules was more beneficial than rules evolved by a GA which may lack explainability.

Fuzzy Action Classification

After tuning both the feature extraction methods and the FIS parameters, we evaluated the final performance of the FIS in classifying activities into the three action intent classes:

'interact', 'modify', and 'relocate'.

We evaluated our work against a set of 17 videos from the EPIC KITCHENS dataset. The selected videos contained 7,186 object-action annotations. Of these actions, 2396 were annotated as 'interact', 1612 were annotated as 'modify', and 1385 were annotated as 'relocate'. The remaining 1793 were labeled as no action (objects that were annotated but not being interacted with). Table 4.7 summarizes the final metrics overall and for each class of activity. Overall, the system reached an accuracy of 74%, which appears to be competitive with state of the art approaches. It is difficult to draw direct comparisons on accuracy since we are taking a different approach in describing various aspects of actions rather than producing a single definitive activity label, but the best performing deep learning technique on the same dataset reached an accuracy of 35.8% [109], highlighting the need for a change in perspective to produce useful information across the entire dataset rather than simply mislabeling a large majority of actions.

From Table 4.7, we observe that the fuzzy system produces the highest precision, recall, and f1 metrics for actions in the 'interact' class. This can be attributed to the robust hand detector and the centrality metric being both accurate and positively predictive of interaction with objects. However, because objects do pass through the center of view and may appear to contact hands from a two-dimensional perspective, the FIS tended to err on the side of producing false positives for this class.

On the flip side, our fuzzy system produced the highest specificity metric for the 'relocate' class, indicating a strong negative predictive value. The system is identifying a much lower rate of false positives for this class, but this comes at the expense of identifying fewer true positives as well since the classifier errs toward negative results for this class; it also has the lowest precision, recall, and f1 scores. This indicates that there is a need for higher precision in the inputs pertaining to the rules for this class.

Error Analysis

A manual analysis of the samples annotated as relocation action reveals two related but distinct sources of error, both stemming from the use of the bounding boxes provided with the dataset.

Annotation Error The bounding boxes provided with the dataset are only updated every 30 frames of video. Figure 4.13a illustrates this issue with the bounding box annotation. The low frequency of updates to the bounding box lead to poor results when the object moves out of the area before the bounding box is updated. This is illustrated by the green highlighted portion of the object. The impact on the system is that the object optical flow estimation is no longer strongly correlated to the object it is intended to describe; taking the motion estimate within the box is estimating motion of either the background or other objects (in the case of this example, the subject's hand).

Bounding Box Error A second type of error stemming from the use of annotated object bounding boxes is that they are limited to rectangular shapes. This is comparable to the output of many deep learning object detection techniques, such as RCNN's. However, this also leads to noise in the motion features by guaranteeing that some background motion is included in the object motion estimate unless the object is perfectly rectangular. The loose fit of bounding boxes for particularly eccentric objects, i.e. objects with a large ratio of one dimension to another, is easily seen in Figure 4.13b. The bounding boxes for both the spatula and faucet include substantial amounts of other objects and background despite being well-aligned to the objects they describe (unlike the example in Figure 4.13a). Pixel-level segmentation of the objects, such as that which could be obtained from a mask R-CNN, would eliminate the issue of loose fitting bounding boxes by segmenting only the pixels belonging to the intended object from its surroundings.

A generalized object detection model was implemented to investigate the possibility

Figure 4.12: Video frame from the EPIC Kitchens dataset overlaid with results from a Mask R-CNN pretrained on the MS COCO dataset. Multiple objects have been segmented with more precision than a rectangular bounding box.

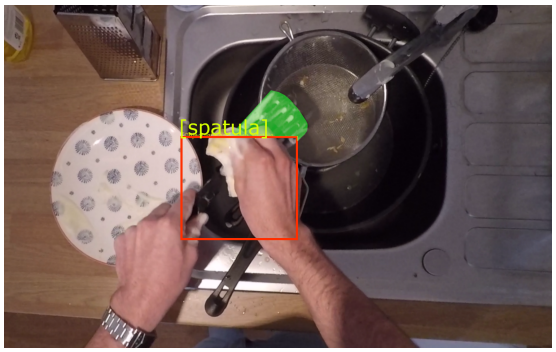
of providing pixel-level masks of objects to 1) avoid the errors in the annotated rectangular boxes and 2) implement a system that is capable of functioning in an environment. Mask RCNN's are a common choice for this task (instance segmentation) and have been shown to perform well in detecting and classifying diverse sets of objects[26, 25, 72]. The object detection method was implemented using the Detectron2 library and used a pre-trained FPN model trained on the MS COCO dataset [114].

Figure 4.12 shows a sample of the results from the object detection method. Several objects have been segmented with high accuracy including the spatula, fork, dish soap, hand soap, and the colander in the sink. The countertop was also detected, albeit with rather poor segmentation results. Compared to the rectangular annotations provided with the dataset, such segmentation results provide a more accurate set of input features for the FIS since very few surrounding pixels will be included in the feature extraction phase.

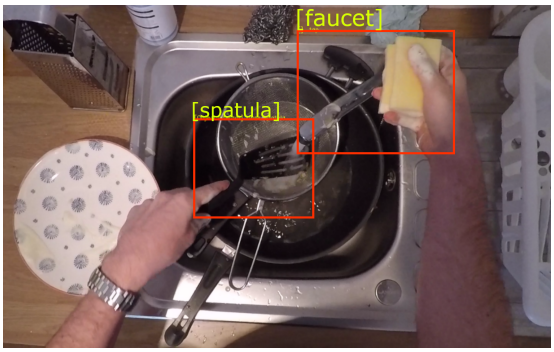
While the object detection seems to solve issues arising from the use of annotated bounding boxes, there are additional aspects of the technique that require solving, which

Table 4.7: Results of activity classification on first person video

Activity Class	*accuracy*	*precision*	*recall*	*specificity*	*f1*
interact	0.767	0.766	0.991	0.909	0.864
modify	0.651	0.268	0.319	0.748	0.291
relocate	0.775	0.242	0.078	0.942	0.118
overall	0.740	0.702	0.598	0.779	0.646



(a) Video frame with object exceeding the rectangular bounding box (the portion of the spatula outside of the bounding box is highlighted in green)



(b) Video frame with loose rectangular bounding boxes around a spatula and faucet

Figure 4.13: Example first person video frame of subject washing spatula with overlaid bounding boxes, action, and object labels

is the reason that further investigation was not prioritized. The runtime of Mask R-CNN models tends to be very slow. While performance optimization was not a concern in our work, the method requires a lot of computing (GPU) resources and would ideally be optimized to achieve greater than 2-3 frames per second. More consequential to our work was the need to implement object tracking in order to use this method. The human-provided annotations include an identifier for each object through multiple frames so our system is aware that an instance of an object in one frame is the same instance of the same object in a later frame. The object detection model does not correlate objects between frames, which is essential for computing any features that occur over time, e.g. optical flow.

During the analysis of our results, we also observed subjectivity in the human-generated action annotations provided with the dataset. The annotation of actions on a per-frame ba-

sis is very difficult since there is rarely a definitive start and end to an activity at such high temporal resolution. We observed inconsistencies in the application of action labels stemming from the imprecision of natural language. For example, opening a faucet and opening a refrigerator are very different types of actions and motions from the subject. While the verb for each of these actions is the same linguistically, they are clearly not the same action. The variance seen between these results via the membership in our fuzzy outputs captures the difference much more precisely than the verb label. Taking a sample of 9 'open tap' actions and 21 'open refrigerator' actions, we computed the average membership in interaction, relocation, and modification as [0.544, -0.471, -0.750] for 'open tap' and [0.607, -0.703, -0.965] for 'open refrigerator', respectively. Variance within these samples in each output term is below 0.01 for each class, so these represent separable clusters of outputs. While the results of both are a positive identification of an interaction, there are clear tendencies that differentiate the two actions, likely stemming from the difference in size of the objects and the type of motion involved from the perspective of the first-person camera. With this knowledge, we have identified a possible application of fuzzy methods in the refining of annotated truth. By detecting outliers or multi-modal distributions of output memberships within the same original verb class (e.g. 'open'), different applications of the same word can be realized.

Conclusions

Through the experiments and results presented in Chapter 4, we have reached positive conclusions to our research questions. To address our first research question, we presented a validation study of a wearable camera being used in a clinical field by measuring gait parameters consistent with much more complex, expensive sensors representing the current standard. Broadening the scope to general human activities for the remaining research questions, we then presented work which demonstrated the viability of our technique using first person video features and fuzzy inference (along with several other feature extraction methods using machine learning and traditional algorithms) to provide data describing the activities that were occurring in the video. Additionally, we identified potential uses of the FIS for refining human annotations and detecting possible linguistic sources of ambiguity.

5.1 Contributions

In addition to answering the research questions outlined at the outset of our work, we made several other intermediate contributions along the way. These include:

- Validation of the wearable vision sensor for use in fields related to activity detection and the measurement of activity parameters,
- A fuzzy inference system which described human activities in first person video with 74% accuracy,

- A lightweight yet accurate hand detection network implemented using MobileNetv2,
- An annotation review tool which allows for review of annotations, actions, and object bounding boxes for the EPIC Kitchens dataset (which would also be generalizable to arbitrary datasets containing similar information). The tool was implemented as a web application with the back-end deployed on a cloud-based, containerized web hosting solution and the user interface served via a traditional web server.

5.2 Future Work

While the conclusions drawn from our experiments provide satisfactory outcomes for our research questions, there is opportunity to further expand the work in several directions. First, the system may be able to describe more specific activities, moving toward the more complex end of the activity scale (Figure 1.1). Possible solutions to explore are adding more outputs to the FIS, or using the FIS outputs as features to another classifier that produces information on the more complex activities, e.g. by clustering activities via the FIS output membership levels and/or considering the outputs as time series data. Another extension to our work could be to fully implement object detection using a deep network, fully eliminating the use of annotated inputs.

Bibliography

[1] 58th World Health Assembly. WHA58.28 eHealth. Technical report, World Health Organization, Geneva, Switzerland, 2005.

[2] Girmaw Abebe, Andrea Cavallaro, and Xavier Parra. Robust multi-dimensional motion features for first-person vision activity recognition. *Computer Vision and Image Understanding*, 149:229–248, 8 2016.

[3] Jonathan Afilalo, Mark J Eisenberg, Jean-François Morin, Howard Bergman, Johanne Monette, Nicolas Noiseux, Louis P Perrault, Karen P Alexander, Yves Langlois, Nandini Dendukuri, Patrick Chamoun, Georges Kasparian, Sophie Robichaud, S Michael Gharacholou, and Jean-François Boivin. Gait Speed as an Incremental Predictor of Mortality and Major Morbidity in Elderly Patients Undergoing Cardiac Surgery. *Journal of the American College of Cardiology*, 56(20):1668 – 1676, 2010.

[4] Motilal Agrawal, Kurt Konolige, and Morten Rufus Blas. Censure: Center surround extremas for realtime feature detection and matching. In *European conference on computer vision*, pages 102–115, 2008.

[5] Ali Akbari, Xien Thomas, and Roozbeh Jafari. Automatic noise estimation and context-enhanced data fusion of IMU and Kinect for human motion measurement. In *2017 IEEE 14th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*, pages 178–182. IEEE, 5 2017.

[6] Omar Adil M Ali, Aous Y Ali, and Balasem Salem Sumait. Comparison between the effects of different types of membership functions on fuzzy logic controller performance. *International Journal of Emerging Engineering Research and Technology*, 3:76–83, 2015.

[7] M. Umair Bin Altaf, Taras Butko, and Biing Hwang Fred Juang. Acoustic gaits: Gait analysis with footstep sounds. *IEEE Transactions on Biomedical Engineering*, 62(8):2001–2011, 8 2015.

[8] Derek Anderson, Robert H. Luke, James M. Keller, Marjorie Skubic, Marilyn Rantz, and Myra Aud. Linguistic summarization of video for fall detection using voxel person and fuzzy logic. *Computer Vision and Image Understanding*, 113(1):80–89, 1 2009.

[9] Derek Anderson, Robert H. Luke, James M. Keller, Marjorie Skubic, Marilyn J. Rantz, and Myra A. Aud. Modeling human activity from voxel person using fuzzy logic. *IEEE Transactions on Fuzzy Systems*, 17(1):39–49, 2009.

[10] Sven Bambach, David J. Crandall, and Chen Yu. Viewpoint Integration for Hand-Based Recognition of Social Interactions from a First-Person View. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction - ICMI '15*, pages 351–354, New York, New York, USA, 2015. ACM Press.

[11] Tanvi Banerjee, James M. Keller, Mihail Popescu, and Marjorie Skubic. Recognizing complex instrumental activities of daily living using scene information and fuzzy logic. *Computer Vision and Image Understanding*, 140:68–82, 11 2015.

[12] Tanvi Banerjee, James M. Keller, Marjorie Skubic, and Erik Stone. Day or night activity recognition from video using fuzzy clustering techniques. *IEEE Transactions on Fuzzy Systems*, 22(3):483–493, 6 2014.

[13] Tanvi Banerjee, Matthew Peterson, Quintin Oliver, Andrew Froehle, and Larry Lawhorne. Validating a commercial device for continuous activity measurement in the older adult population for dementia management. *Smart Health*, 2017.

[14] Lorenzo Baraldi, Francesco Paci, Giuseppe Serra, Luca Benini, and Rita Cucchiara. Gesture Recognition Using Wearable Vision Sensors to Enhance Visitors' Museum Experiences. *IEEE Sensors Journal*, 15(5):2705–2714, 2015.

[15] Olivier Beauchet, Cédric Annweiler, Michele L Callisaya, Anne-Marie De Cock, Jorunn L Helbostad, Reto W Kressig, Velandai Srikanth, Jean-Paul Steinmetz, Helena M Blumen, Joe Verghese, and Gilles Allali. Poor Gait Performance and Prediction of Dementia: Results From a Meta-Analysis. *Journal of the American Medical Directors Association*, 17(6):482 – 490, 2016.

[16] C Beleznai, A N Belbachir, and P M Roth. Density-based rare event detection from streams of neuromorphic sensor data. In *Distributed Smart Cameras (ICDSC), 2012 Sixth International Conference on*, pages 1–6. IEEE, 2012.

[17] Domenico Bloisi, Luca Iocchi, Luca Marchetti, Dorothy N. Monekosso, and Paolo Remagnino. An adaptive tracker for assisted living. In *6th IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS 2009*, pages 164–169. IEEE, 9 2009.

[18] E Bramell-Risberg, G Jarnlo, L Minthon, and S Elmstahl. Lower Gait Speed in Older Women with Dementia Compared with Controls. *Dementia and Geriatric Cognitive Disorders*, 20:298–305, 2005.

[19] Dirk Buche, Nicol N Schraudolph, and Petros Koumoutsakos. Accelerating evolutionary algorithms with Gaussian process fitness function models. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 35(2):183–194, 2005.

[20] R S Bucks, D L Ashworth, G K Wilcock, and K Siegfried. Assessment of Activities of Daily Living in Dementia: Development of the Bristol Activities of Daily Living Scale. Technical report, 1996.

[21] Inc Carre Technologies. Hexoskin Smart Shirts, 2017.

[22] Anne Carswell and Robin Eastwood. Activities of Daily Living, Cognitive Impairment and Social Function in Community Residents with Alzheimer Disease. *Canadian Journal of Occupational Therapy*, 60(3):130–136, 8 1993.

[23] Ishani Chakraborty, Ahmed Elgammal, and Randall S. Burd. Video based activity recognition in trauma resuscitation. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, FG 2013*, pages 1–8. IEEE, 4 2013.

[24] Peng Chang and John Krumm. Object recognition with color cooccurrence histograms. In *Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149)*, volume 2, pages 498–504, 1999.

[25] Xinlei Chen, Ross Girshick, Kaiming He, and Piotr Dollár. Tensormask: A Foundation for Dense Object Segmentation. In *The International Conference on Computer Vision (ICCV)*, 2019.

[26] Bowen Cheng, Maxwell D Collins, Yukun Zhu, Ting Liu, Thomas S Huang, Hartwig Adam, and Liang-Chieh Chen. Panoptic-DeepLab: A Simple, Strong, and Fast Baseline for Bottom-Up Panoptic Segmentation. In *CVPR*, 2020.

[27] J H Chien, J Yentes, N Stergiou, and K C Siu. The Effect of Walking Speed on Gait Variability in Healthy Young, Middle-aged and Elderly Individuals. *Journal of Physical Activity, Nutrition, and Rehabilitation*, 2015.

[28] Pau Choo Chung, Yu Liang Hsu, Chun Yao Wang, Chien Wen Lin, Jeen Shing Wang, and Ming Chyi Pai. Gait analysis for patients with Alzheimer'S disease using a triaxial accelerometer. In *ISCAS 2012 - 2012 IEEE International Symposium on Circuits and Systems*, pages 1323–1326, 2012.

[29] Federica Corona, Massimiliano Pau, Marco Guicciardi, Mauro Murgia, Roberta Pili, and Carlo Casula. Quantitative assessment of gait in elderly people affected by Parkinson's Disease. In *2016 IEEE International Symposium on Medical Measurements and Applications, MeMeA 2016 - Proceedings*. Institute of Electrical and Electronics Engineers Inc., 8 2016.

[30] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling Egocentric Vision: The EPIC-KITCHENS Dataset. In *Proceedings of European Conference on Computer Vision (ECCV)*, pages 720–736, 2018.

[31] Paolo de Leva. Adjustments to Zatsiorsky-Seluyanov's segment inertia parameters. *Journal of Biomechanics*, 29(9):1223–1230, 1996.

[32] Silvia Del Din, Alan Godfrey, and Lynn Rochester. Validation of an Accelerometer to Quantify a Comprehensive Battery of Gait Characteristics in Healthy Older Adults and Parkinson's Disease: Toward Clinical and at Home Use. *IEEE Journal of Biomedical and Health Informatics*, 20(3):838–847, 5 2016.

[33] Abhilash K. Desai, George T. Grossberg, and Dharmesh N. Sheth. Activities of daily living in patients with dementia: Clinical relevance, methods of assessment and effects of treatment, 8 2004.

[34] M. ElHelw, J Pansiot, D. McIlwraith, R. Ali, B. Lo, and L. Atallah. An integrated multi-sensing framework for pervasive healthcare monitoring. In *Proceedings of the*

3d International ICST Conference on Pervasive Computing Technologies for Healthcare. ICST, 2009.

[35] Larry J Eshelman and J David Schaffer. Real-coded genetic algorithms and interval-schemata. In *Foundations of genetic algorithms*, volume 2, pages 187–202. Elsevier, 1993.

[36] Gunnar Farneback. Two-Frame Motion Estimation Based on. *Lecture Notes in Computer Science*, 2749(1):363–370, 2003.

[37] Jose I. Figueroa-Angulo, Jesus Savage-Carmona, Ernesto Bribiesca-Correa, Boris Escalante, Ronald S. Leder, and Luis E. Sucar. Recognition of arm activities based on Hidden Markov Models for natural interaction with service robots. In *2013 16th International Conference on Advanced Robotics, ICAR 2013*, pages 1–8. IEEE, 11 2013.

[38] Annette L Fitzpatrick, Catherine K Buchanan, Richard L Nahin, Steven T DeKosky, Hal H Atkinson, Michell C Carlson, and Jeff D Williamson. Associations of Gait Speed and Other Measures of Physial Function with Cognition in a Healthy Cohort of Elderly Persons. *Journals of Gerontology Series A: Biological Sciences & Medical Sciences*, 62(11):1244–1251, 2007.

[39] Emma Fortune, Vipul Lugade, Melissa Morrow, and Kenton Kaufman. Validity of using tri-axial accelerometers to measure human movement – Part II: Step counts at a wide range of gait velocities. *Medical Engineering & Physics*, 36(6):659 – 669, 2014.

[40] Jurgen T. Geiger, Martin Hofmann, Bjorn Schuller, and Gerhard Rigoll. Gait-based person identification by spectral, cepstral and energy-related audio features. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pages 458–462, 10 2013.

[41] Labiba Gillani Fahad, Arshad Ali, and Muttukrishnan Rajarajan. Long term analysis of daily activities in a smart home. In *European Symposium on Artificial Neural Networks*, pages 419–424, Bruges, Belgium, 4 2013.

[42] Zeynep Gokce and Selen Pehlivan. Human action recognition in first person videos using verb-object Pairs. In *27th Signal Processing and Communications Applications Conference, SIU 2019*. Institute of Electrical and Electronics Engineers Inc., 4 2019.

[43] Raffay Hamid, Amos Johnson, Samir Batta, Aaron Bobick, Charles Isbell, and Graham Coleman. Detection and explanation of anomalous activities: Representing activities as bags of event n-grams. In *Proceedings - 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005*, volume I, pages 1031–1038. IEEE, 2005.

[44] Antonia Hartmann, Susanna Luzi, Kurt Murer, Rob A de Bie, and Eling D de Bruin. Concurrent validity of a trunk tri-axial accelerometer system for gait analysis in older adults. *Gait & Posture*, 29(3):444 – 448, 2009.

[45] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[46] Hirokazu Seki and Susumu Tadakuma. Abnormality detection monitoring system for elderly people in sensing and robotic support room. In *2008 10th IEEE International Workshop on Advanced Motion Control*, pages 56–61. IEEE, 3 2008.

[47] John H Holland. Genetic algorithms. *Scientific american*, 267(1):66–73, 1992.

[48] Berthold K.P. Horn and Brian G. Schunck. Determining Optical Flow. *AI Memos*, 4 1980.

[49] Linda A Jacobsen, Mary Kent, Marlene Lee, and Mark Mather. America's Aging Population. *Population Bulletin*, 66(1), 2011.

[50] Chyck Karr. Applying genetics to fuzzy logic. *AI expert*, 6(3):38–43, 1991.

[51] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[52] Anit Kumar. Encoding schemes in genetic algorithm. *International Journal of Advanced Research in IT and Engineering*, 2(3):1–7, 2013.

[53] M Powell Lawton and Elaine M Brody. Assessment of Older People: Self-Maintaining and Instrumental Activities of Daily Living. *The Gerontologist*.

[54] Yin Li, Miao Liu, and James M Rehg. In the Eye of Beholder: Joint Learning of Gaze and Actions in First Person Video. Technical report.

[55] Zewen Li, Fan Liu, Wenjie Yang, Shouheng Peng, and Jun Zhou. A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[56] Zijia Li, Chi Wun Au, Yohei Kakiuchi, Kei Okada, and Masayuki Inaba. What am i doing? Robotic self-action recognition. In *IEEE-RAS International Conference on Humanoid Robots*, pages 165–170. IEEE, 11 2016.

[57] Yoonseob Lim and Jongsuk Choi. Speaker selection and tracking in a cluttered environment with audio and visual information. *IEEE Transactions on Consumer Electronics*, 55(3):1581–1589, 8 2009.

[58] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of*

the IEEE conference on computer vision and pattern recognition, pages 2117–2125, 2017.

[59] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C Lawrence Zitnick, and Piotr Dollár. Microsoft COCO: Common Objects in Context, 2015.

[60] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37, 2016.

[61] Minlong Lu, Ze Nian Li, Yueming Wang, and Gang Pan. Deep Attention Network for Egocentric Action Recognition. *IEEE Transactions on Image Processing*, 28(8):3703–3713, 8 2019.

[62] Bruce D Lucas and Takeo Kanade. An Iterative Image Registration Technique with an Application to Stereo Vision. In *Proceedings of Imaging Understanding Workshop*, pages 121–130, 1981.

[63] Anvith Mahabalagiri, Koray Ozcan, and Senem Velipasalar. A robust edge-based optical flow method for elderly activity classification with wearable smart cameras. In *2013 7th International Conference on Distributed Smart Cameras, ICDSC 2013*, pages 1–6. IEEE, 10 2013.

[64] Ebrahim H Mamdani and Sedrak Assilian. An experiment in linguistic synthesis with a fuzzy logic controller. *International journal of man-machine studies*, 7(1):1–13, 1975.

[65] D T Mane and Uday V Kulkarni. A survey on supervised convolutional neural network and its major applications. In *Deep Learning and Neural Networks: Concepts, Methodologies, Tools, and Applications*, pages 1058–1071. IGI Global, 2020.

[66] Benoit Mariani, Mayté Castro Jiménez, François J.G. Vingerhoets, and Kamiar Aminian. On-shoe wearable sensors for gait and turning assessment of patients with parkinson's disease. *IEEE Transactions on Biomedical Engineering*, 60(1):155–158, 1 2013.

[67] Sujitha Martin, Akshay Rangesh, Eshed Ohn-Bar, and Mohan M. Trivedi. Preparatory coordination of head, eyes and hands: Experimental study at intersections. In *Proceedings - International Conference on Pattern Recognition*, pages 2783–2788. IEEE, 12 2017.

[68] Kenji Matsuo, Kentaro Yamada, Satoshi Ueno, and Sei Naito. An attention-based activity recognition for egocentric video. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 565–570. IEEE Computer Society, 9 2014.

[69] Tomas McCandless and Kristen Grauman. Object-Centric Spatio-Temporal Pyramids for Egocentric Activity Recognition. In *Proceedings of the British Machine Vision Conference*. BMVA Press, 2013.

[70] Douglas G. McIlwraith, Julien Pansiot, James Ballantyne, Salman Valibeik, Ahmed Elsaify, and Guang Zhong Yang. Structure learning for activity recognition in robot assisted intelligent environments. In *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2009*, pages 4644–4649. IEEE, 10 2009.

[71] Douglas G. McIlwraith, Julien Pansiot, Surapa Thiemjarus, Benny P.L. Lo, and Guang Zhong Yang. Probabilistic decision level fusion for real-time correlation of ambient and wearable sensors. In *Proc. 5th Int. Workshop on Wearable and Implantable Body Sensor Networks, BSN2008, in conjunction with the 5th Int. Summer School and Symp. on Medical Devices and Biosensors, ISSS-MDBS 2008*, pages 117–120. IEEE, 2008.

[72] Claudio Michaelis, Ivan Ustyuzhaninov, Matthias Bethge, and Alexander S Ecker. One-shot instance segmentation. *arXiv preprint arXiv:1811.11507*, 2018.

[73] Tomoya Nakatani, Ryohei Kuga, and Takuya Maekawa. Preliminary Investigation of Object-based Activity Recognition Using Egocentric Video Based on Web Knowledge. In *Proceedings of the 17th International Conference on Mobile and Ubiquitous Multimedia - MUM 2018*, pages 375–381, New York, New York, USA, 2018. ACM Press.

[74] Tomoya Nakatani, Ryohei Kuga, and Takuya Maekawa. Object-based Activity Recognition Using Egocentric Video Based on Web Knowledge. In *2019 IEEE International Conference on Pervasive Computing and Communications Workshops, PerCom Workshops 2019*, pages 620–625. Institute of Electrical and Electronics Engineers Inc., 3 2019.

[75] Tomasz Nowak, Patryk Najgebauer, Jakub Romanowski, Marcin Gabryel, Marcin Korytkowski, Rafał Scherer, and Dimce Kostadinov. Spatial keypoint representation for visual object retrieval. In *International Conference on Artificial Intelligence and Soft Computing*, pages 639–650, 2014.

[76] Reza Oji. An automatic algorithm for object recognition and detection based on ASIFT keypoints. *arXiv preprint arXiv:1211.5829*, 2012.

[77] Wee Hong Ong, Takafumi Koseki, and Leon Palafox. Unsupervised human activity detection with skeleton data from RGB-D sensor. In *Proceedings - 5th International Conference on Computational Intelligence, Communication Systems, and Networks, CICSyN 2013*, pages 30–35. IEEE, 6 2013.

[78] Jennifer M Ortman and Victoria A Velkoff. An aging nation: The older population in the United States, 5 2014.

[79] Fatih Ozkan, Mehmet Ali Arabaci, Elif Surer, and Alptekin Temizel. Boosted Multiple Kernel Learning for First-Person Activity Recognition. 2 2017.

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

[81] Pivothead. Pivothead, 2017.

[82] Didik Purwanto, Yie Tarng Chen, and Wen Hsien Fang. Temporal aggregation for first-person action recognition using Hilbert-Huang transform. In *Proceedings - IEEE International Conference on Multimedia and Expo*, pages 895–900. IEEE Computer Society, 8 2017.

[83] Meera Radhakrishnan, Sougata Sen, S. Vigneshwaran, Archan Misra, and Rajesh Balan. IoT+Small Data: Transforming in-store shopping analytics & services. In *2016 8th International Conference on Communication Systems and Networks, COMSNETS 2016*, pages 1–6. IEEE, 1 2016.

[84] Anthony Rowe, Dhiraj Goel, and Raj Rajkumar. FireFly Mosaic: A vision-enabled wireless sensor networking system. In *Proceedings - Real-Time Systems Symposium*, pages 459–468. IEEE, 12 2007.

[85] John W. Rowe, Terry Fulmer, and Linda Fried. Preparing for better health and health care for an aging population, 10 2016.

[86] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. ORB: An efficient alternative to SIFT or SURF. In *2011 International conference on computer vision*, pages 2564–2571, 2011.

[87] Rosaria Rucco, Valeria Agosti, Francesca Jacini, Pierpaolo Sorrentino, Pasquale Varriale, Manuela De Stefano, Graziella Milan, Patrizia Montella, and Giuseppe

Sorrentino. Spatio-temporal and kinematic gait analysis in patients with Frontotemporal dementia and Alzheimer's disease through 3D motion capture. *Gait & Posture*, 52(Supplement C):312 – 317, 2017.

[88] Michael S. Ryoo, Brandon Rothrock, and Larry Matthies. Pooled Motion Features for First-Person Videos, 2015.

[89] 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.

[90] Arman Savran, Raffaele Tavarone, Bertrand Higy, Leonardo Badino, and Chiara Bartolozzi. Energy and computation efficient audio-visual voice activity detection driven by event-cameras. In *Proceedings - 13th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2018*, pages 333–340. IEEE, 5 2018.

[91] Bradley Schneider. *Gait Analysis from Wearable Devices using Image and Signal Processing*. PhD thesis, Wright State University, 2017.

[92] Bradley Schneider and Tanvi Banerjee. Preliminary investigation of walking motion using a combination of image and signal processing. In *Proceedings - 2016 International Conference on Computational Science and Computational Intelligence, CSCI 2016*, pages 641–646. Institute of Electrical and Electronics Engineers Inc., 3 2017.

[93] Bradley Schneider and Tanvi Banerjee. Bridging the Gap between Atomic and Complex Activities in First Person Video. In *2021 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–6, 2021.

[94] Bradley Schneider, Tanvi Banerjee, Michael Riley, and Francis Grover. Comparison of Gait Speeds from Wearable Camera and Accelerometer in Structured and Semi-Structured Environments. *Healthcare Technology Letters*, 11 2019.

[95] Hirokazu Seki. Fuzzy inference based non-daily behavior pattern detection for elderly people monitoring system. In *Proceedings of the 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society: Engineering the Future of Biomedicine, EMBC 2009*, pages 6187–6192. IEEE, 9 2009.

[96] Teo Lian Seng, M Bin Khalid, and Rubiyah Yusof. Tuning of a neuro-fuzzy controller by genetic algorithm. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 29(2):226–236, 1999.

[97] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[98] Swathikiran Sudhakaran and Oswald Lanz. Attention is all we need: Nailing down object-centric attention for egocentric activity recognition. In *British Machine Vision Conference 2018, BMVC 2018*. BMVA Press, 2018.

[99] Michio Sugeno. *Industrial applications of fuzzy control*. Elsevier Science Inc., 1985.

[100] Young Soo Suh, Ebrahim Nemati, and Majid Sarrafzadeh. Kalman-Filter-Based Walking Distance Estimation for a Smart-Watch. In *Proceedings - 2016 IEEE 1st International Conference on Connected Health: Applications, Systems and Engineering Technologies, CHASE 2016*, pages 150–156. Institute of Electrical and Electronics Engineers Inc., 8 2016.

[101] Michael J Swain and Dana H Ballard. Color indexing. *International journal of computer vision*, 7(1):11–32, 1991.

[102] Hideyuki Takagi and Michael Lee. Neural networks and genetic algorithm approaches to auto-design of fuzzy systems. In *Austrian Conference on Fuzzy Logic in Artificial Intelligence*, pages 68–79. Springer, 1993.

[103] Masato Takahashi and Hajime Kita. A crossover operator using independent component analysis for real-coded genetic algorithms. In *Proceedings of the 2001 Congress on Evolutionary Computation (IEEE Cat. No. 01TH8546)*, volume 1, pages 643–649, 2001.

[104] Vyara Valkanova and Klaus P Ebmeier. What can gait tell us about dementia? Review of epidemiological and neuropsychological evidence. *Gait & Posture*, 53(Supplement C):215 – 223, 2017.

[105] Andrea Vázquez, Cristina Jenaro, Noelia Flores, María José Bagnato, Ma Carmen Pérez, and Maribel Cruz. E-Health Interventions for Adult and Aging Population With Intellectual Disability: A Review. *Frontiers in Psychology*, 9(NOV):2323, 11 2018.

[106] Laura J Viccaro, Subashan Perera, and Stephanie A Studenski. Gait Speed at Usual Pace as a Predictor of Adverse Outcomes in Community-Dwelling Older People: an International Academy on Nutrition and Aging (IANA) Task Force. *Journal of the American Geriatrics Society*, 59(5):887–892, 2011.

[107] Fang Wang, Erik Stone, Marjorie Skubic, James M. Keller, Carmen Abbott, and Marilyn Rantz. Toward a passive low-cost in-home gait assessment system for older adults. *IEEE Journal of Biomedical and Health Informatics*, 17(2):346–355, 3 2013.

[108] Hanzi Wang, David Suter, Konrad Schindler, and Chunhua Shen. Adaptive object tracking based on an effective appearance filter. *IEEE transactions on pattern analysis and machine intelligence*, 29(9):1661–1667, 2007.

[109] Lei Wang and Piotr Koniusz. Self-supervising action recognition by statistical moment and subspace descriptors. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4324–4333, 2021.

[110] Meng Wang, Changzhi Luo, Bingbing Ni, Jun Yuan, Jianfeng Wang, and Shuicheng Yan. First-Person Daily Activity Recognition with Manipulated Object Proposals and Non-Linear Feature Fusion. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(10):2946–2955, 10 2018.

[111] S V Wong and A M S Hamouda. Optimization of fuzzy rules design using genetic algorithm. *Advances in Engineering Software*, 31(4):251–262, 2000.

[112] World Health Organization. mHealth: New horizons for health through mobile technologies: second global survey on eHealth. *Global Observatory for eHealth Series*, 3, 2011.

[113] Jianxin Wu, Adebola Osuntogun, Tanzeem Choudhury, Matthai Philipose, and James M. Rehg. A scalable approach to activity recognition based on object use. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1–8. IEEE, 2007.

[114] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019.

[115] Bo Yao, Hani Hagras, Daniyal Alghazzawi, and Mohammed J. Alhaddad. A big bang-big crunch type-2 fuzzy logic system for machine-vision-based event detection and summarization in real-world ambient-assisted living. *IEEE Transactions on Fuzzy Systems*, 24(6):1307–1319, 12 2016.

[116] Tang Yiping, Jin Shunjing, Yang Zhongyuan, and You Sisi. Detection elder abnormal activities by using Omni-drectional Vision Sensor: Activity data collection and modeling. In *2006 SICE-ICASE International Joint Conference*, pages 3850–3853. IEEE, 2006.

[117] Takami Yoshida, Kazuhiro Nakadai, and Hiroshi G. Okuno. Two-layered audio-visual speech recognition for robots in noisy environments. In *IEEE/RSJ 2010 In-*

ternational Conference on Intelligent Robots and Systems, IROS 2010 - Conference Proceedings, pages 988–993. IEEE, 10 2010.

[118] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833, 2014.

[119] C.H. Zhao, B.L. Zhang, J. He, and J. Lian. Recognition of driving postures by contourlet transform and random forests. *IET Intelligent Transport Systems*, 6(2):161, 2012.

[120] Wenbing Zhao, Roanna Lun, Connor Gordon, Abou Bakar M. Fofana, Deborah D. Espy, M. Ann Reinthal, Beth Ekelman, Glenn D. Goodman, Joan E. Niederriter, and Xiong Luo. A Human-Centered Activity Tracking System: Toward a Healthier Workplace. *IEEE Transactions on Human-Machine Systems*, 47(3):343–355, 6 2017.

[121] Jinghui Zhong, Xiaomin Hu, Jun Zhang, and Min Gu. Comparison of performance between different selection strategies on simple genetic algorithms. In *International conference on computational intelligence for modelling, control and automation and international conference on intelligent agents, web technologies and internet commerce (CIMCA-IAWTIC'06)*, volume 2, pages 1115–1121, 2005.

[122] Zoran Zivkovic and Ben Krose. An EM-like algorithm for color-histogram-based object tracking. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 1, page I–I, 2004.