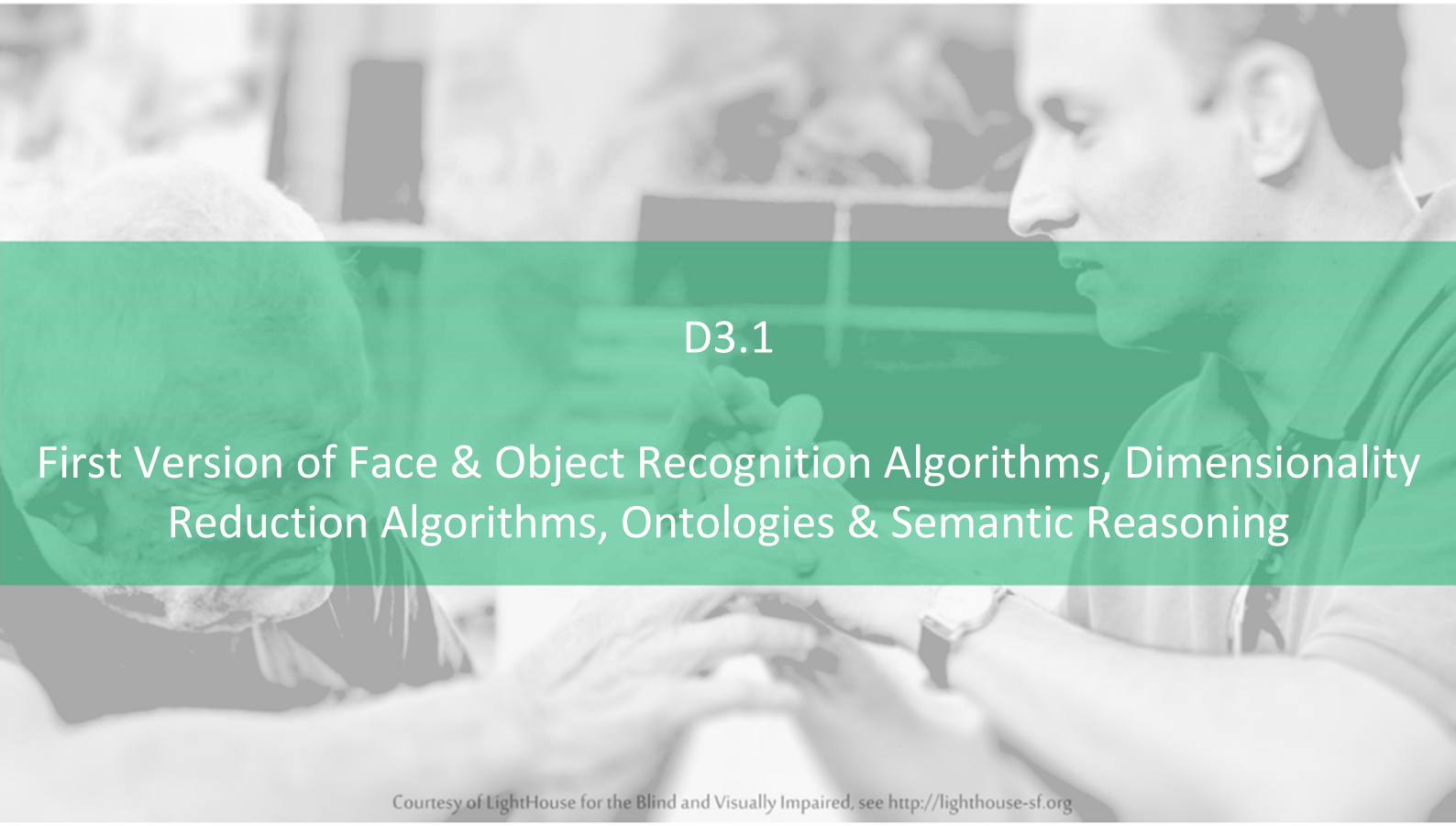
SUITCEYES

1 Jan 2018 - 31 Dec 2020

Smart, User-friendly, Interactive, Tactual, Cognition-Enhancer, that Yields Extended Sensosphere
Appropriating sensor technologies, machine learning, gamification and smart haptic interfaces



Courtesy of LightHouse for the Blind and Visually Impaired, see http://lighthouse-sf.org

D3.1

First Version of Face & Object Recognition Algorithms, Dimensionality Reduction Algorithms, Ontologies & Semantic Reasoning

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 780814.

Dissemination level		
PU	PUBLIC, fully open, e.g. web	X
CO	CONFIDENTIAL, restricted under conditions set out in Model Grant Agreement	
CI	CLASSIFIED, information as referred to in Commission Decision 2001/844/EC.	

Deliverable Type		
R	Document, report (excluding the periodic and final reports)	X
DEM	Demonstrator, pilot, prototype, plan designs	
DEC	Websites, patents filing, press & media actions, videos, etc.	
OTHER	Software, technical diagram, etc.	

Deliverable Details	
Deliverable number	D3.1
Part of WP	WP3
Lead organisation	CERTH
Lead member	E. Kontopoulos

Revision History			
V#	**Date**	**Description / Reason of change**	**Author / Org.**
v0.1	31-Aug-2018	Structure proposal	E. Kontopoulos / CERTH
v0.2	23-Nov-2018	1st draft for internal review	E. Kontopoulos / CERTH P. Petrantonakis / CERTH S. Darányi / HB
v0.3	10-Dec-2018	2nd draft addressing review comments	E. Kontopoulos / CERTH P. Petrantonakis / CERTH
v0.4	18-Dec-2018	Pre-final draft after PC's comments	E. Kontopoulos / CERTH
v1.0	18-Dec-2018	Final draft submitted to the EU	N. Olson/ HB

Authors	
Partner	**Name(s)**
CERTH	E. Kontopoulos, P. Petrantonakis
HB	S. Darányi

Contributors		
Partner	**Contribution type**	**Name**
CERTH	Internal revisions to Chapter 2	K. Avgerinakis, P. Giannakeris
TU/e	Internal review	A. Kappers
ULEEDS	Internal review	R. Holt, Zhengyang Ling

Glossary	
Abbr./ Acronym	**Meaning**
BoW	Bag-of-Words
CNN	Convolutional Neural Networks
CQ	Competency Question
DL	Description Logics
DR	Dimensionality Reduction
FOAF	Friend-Of-A-Friend
FPN	Feature Pyramid Network
FPV	First-Person View
GMM	Gaussian Mixture Modelling
HIPI	Haptic Intelligent Personal Interface
HOF	Histograms of optical flow
HOG	Histograms of Oriented Gradients
HSV	Hue, saturation, value
IoT	Internet of Things
IoU	Intersection-over-Union
KCF	Kernelized correlation filtering
mAP	mean Average Precision
MBH	Motion boundary histogram
MDS	Multidimensional Scaling
NMS	Non-maximum suppression
PCA	Principal Component Analysis
OWL	Web Ontology Language
R-FCN	Region – Fully Convolutional Network
ROI	Region-of-Interest
SIFT	Scale Invariant Feature Transform

SNE	Stochastic Neighbours Embedding
SOSA	Sensor, Observation, Sampler, and Actuator
SSH	Single-Shot Headless
SSN	Semantic Sensor Network
TP	True Positive
SURF	Speeded-Up Robust Features
TPV	Third-Person View
W3C	World Wide Web Consortium
XSD	XML Schema Definition Language

Executive Summary

This document constitutes SUITCEYES deliverable D3.1 presenting the work conducted during the period M1-12 within WP3 and reports on the following:

(a) The basic version of visual analysis algorithms evaluated in benchmark datasets for performing object detection and tracking, face detection and tracking, scene recognition, first person activity recognition and third person gesture recognition;

(b) The first version of the algorithms that achieve discrete low dimensional representations for different high dimensional concepts evaluated in the CIFAR-100 benchmark dataset;

(c) The preliminary version of the semantic knowledge graphs for semantically integrating the multimodal outputs from the various heterogeneous SUITCEYES sensors and components, along with some first insights into the associated interpretation and reasoning techniques.

Table of Contents

1 Introduction

Deafblindness which can be congenital or acquired, and complete or partial, presents research with an enormously broad spectrum of possible user needs, whose common denominator must be selected carefully. After initial discussions with the Project's Advisory Board, and in accord with the strategy taken in WP2, and, more specifically, in deliverables D2.1 "*Requirements for HIPI*" and D2.2 "*Personas, environments and use scenarios*", early on it was agreed upon in our consortium that the project must address three specific but related research tracks: **situational awareness**, **navigation**, and **communication**. This agreement anticipated that, from the perspective of the user, the three tasks are related, and in a sense, both navigation and eventual communication serve the purpose of situational awareness. This focus may be further refined once the process of determining the end-users' requirements within WP2 is concluded; nevertheless, potential refinements will not substantially affect work within WP3, which is primarily aimed at capturing, translating and semantically representing environmental cues.

As navigation by selecting and ontologically interpreting specific visual cues significant enough to pass a threshold for the user with deafblindness is not different from picking significant concepts from a vocabulary to transmit them between any two such users for communication, the single major difference between navigation and communication is that for the latter, sequences of such cues must be encoded for and decoded after transmission. Therefore, we envisage both research areas as addressable by the same entwined strategy, identifying visual vs. verbal signs, and relating their content – regardless if single signs or their sequences are transmitted – to a knowledge graph to secure a standard semantics for both tracks.

1.1 WP3 Overview

WP3 consists of three tasks, the interplay of which is illustrated in Figure 1.

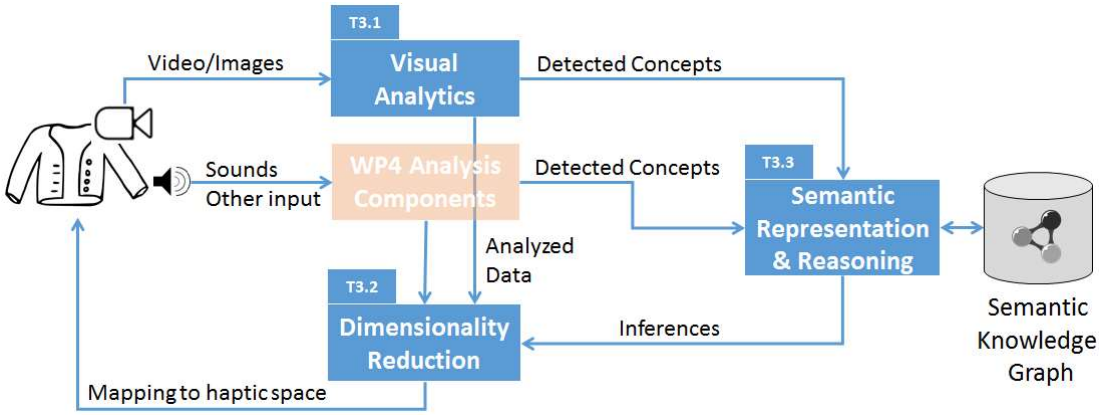


Figure 1: Schematic overview of the interrelationships between WP3 tasks.

Visual input from the camera mounted on the smart garment, called the HIPI (Haptic Intelligent Personal Interface), is fed to the visual analysis component (T3.1) which extracts the detected concepts (objects, faces, activities etc.). The latter are fed to the semantic representation and reasoning component (T3.3), which is coupled with a semantic knowledge graph (also called an "ontology"), semantically aggregating the multimodal information from the analyses and inferring higher-level derivations. The outputs from T3.1 and T3.3, along with other signals and outputs from non-WP3 com-

ponents, are submitted to the dimensionality reduction component (T3.2), which maps the respective information to the haptic space.

1.2 Document Outline

The rest of the document presents our work so far within WP3 and is structured as follows:

- **Chapter 2** presents the basic version of visual analysis algorithms evaluated in benchmark datasets for performing object detection and tracking, face detection and tracking, scene recognition, first person activity recognition, and third person gesture recognition;
- **Chapter 3** presents the first version of the algorithms that achieve discrete low dimensional representations for different high dimensional concepts evaluated in a benchmark dataset;
- **Chapter 4** presents the preliminary version of the semantic knowledge graphs for semantically integrating the multimodal outputs from the various heterogeneous SUITCEYES sensors and components, along with some first insights into the associated interpretation and reasoning techniques;
- **Chapter 5** concludes this deliverable with some final points for discussion.

2 Visual Analysis

Visual analysis constitutes a core component in SUITCEYES. In order to extract knowledge using the visual information that a wearable camera can capture, several computer vision tasks are carried out within the component. In the current version these are: object detection and tracking, face detection and tracking, scene recognition, first person activity recognition, and third person gesture recognition. All tasks have been chosen and designed to work in line with SUITCEYES principals and intentions to support the users with deafblindness, and are generally accomplished using state-of-the-art computer vision algorithms. In this chapter we provide the background for each task in the form of brief literature reviews, as well as a detailed presentation of the algorithms used within SUITCEYES along with some qualitative and quantitative results.

2.1 Background

This subsection presents the current state-of-the-art for each individual computer vision task in SUITCEYES. Wherever needed, we also give clarifications for the motivation behind the selection of some tasks being part of the visual analysis component. This is done in order to form a basis of understanding with respect to what each task is supposed to achieve towards enriching the end users' experience.

2.1.1 Object Detection

Object detection will have a central role in SUITCEYES. Besides the obvious task of recognizing the objects that exist in front of the users and are of interest to them, object detection will also serve as a platform to build upon other visual analysis components, such as the activity recognition component (see later subsections). This dependency relates closely to the fact that we developed an object-centric activity detection algorithm that also uses users' motion patterns in order to understand what the user does. Moreover, face detection is inherently related to object detection, when considering faces as being just another type of "interesting object" that current algorithms can be trained to detect. More details are given in later sections about those topics.

Object detection has been consistently keeping computer vision researchers busy for decades, and, thus, there is a plethora of algorithms available in the literature. Naturally, powerful deep Convolutional Neural Networks (CNNs) were thoroughly examined for this task and were eventually established as the state of the art. The seminal work of (Girshick, 2015) includes a multi-scale bounding box proposal generation method like Selective Search (Uijlings et al., 2013), as a feeding mechanism of candidate object boxes to a deep CNN and then a Region-of-Interest (ROI) feature pooling layer leading to a Fully-Connected (FC) section with two branches that act as a classifier and a bounding box coordinate regressor. Later, the bounding box proposal network was incorporated into an end-to-end deep architecture in Faster R-CNN (Ren et al., 2015), achieving better performance and faster prediction during testing.

Another class of deep CNN object detectors that was called "region-free" – as opposed to the previously mentioned region-based methods – was proposed shortly thereafter, which skip the region proposal step and predicts classes and boxes coordinated directly using the latest convolutional feature maps, building a form of single shot detectors (Liu, et al., 2016; Redmon et al., 2016). Those models achieved a better trade-off between accuracy and speed, and their main advantage was their high time-efficiency.

Since then, a variety of techniques were proposed to further explore and improve upon the classic aforementioned architectures. More specifically, Dai et al. (2016) proposed R-FCN (Region – Fully Convolutional Network) as a region-based method that, contrary to the costly approach of the classic Faster R-CNN (Ren et al., 2015), was a shared, fully convolutional network with position-sensitive score maps that aid the purpose of making the ROI pooling operations sharable and, thus, faster than before. In addition, an effort was made by Kong et al. (2017) to combine the best of both the region-based and region-free worlds. A reverse connection between convolutional layers was designed, which enabled the network to detect objects on multi-levels of CNNs, and the above concern of being an object or not was introduced to significantly reduce the searching space of objects.

Later, Wang et al. (2017) proposed to learn an adversarial network that generates examples with occlusions and deformations which is essentially a hard-positive sample generation process. The authors followed this approach using the Fast R-CNN architecture and managed to achieve a performance boost equal to 2.3% mean Average Precision (mAP) compared to the original approach. In (Ren et al., 2017), the authors experimented with region-wise classifier networks that use shared region-independent convolutional features (NoCs). They emphasized that, aside from advances in deep feature extraction, a deep convolutional per-region classifier is of particular importance for object detection. Another architecture was developed in (Lin et al., 2017) that focused on the deep feature extraction step, called Feature Pyramid Network (FPN), where a top-down architecture with lateral connections was deployed in order to build high-level semantic feature maps at various scales in a pyramidal hierarchy.

More specifically, within the scope of works that focused on the analysis of wearable camera footage, a good number of datasets was made publicly available, including annotations of objects that can be found on videos, or images that capture everyday life through a first-person perspective. The ADL dataset (Pirsiavash & Ramanan, 2012) contains videos of indoor activities performed by a group of users, and provides annotations for object bounding boxes that the users interact with during those activities. Similarly, the Bristol Egocentric Object Interactions Dataset (Damen et al., 2016) is compiled of egocentric videos in indoor environments and provides object annotations for the instances found in the sequences. The EDUB-Obj dataset (Bolanos & Radeva, 2015) contains images of activities of daily life in unconstrained environments (outdoors) and object mask annotations. The NYU Depth Dataset V2 (Silberman et al., 2012) provides segmentation masks for objects obtained from RGBD images and is oriented towards general indoor segmentation tasks. Finally, the EPIC-KITCHENS large-scale dataset was published in (Damen et al., 2018) and an object detection competition was organized as well. The dataset contains 55 hours of video recordings related to kitchen activities as well as bounding box annotations of objects that are usually found in a kitchen.

2.1.2 Face Detection

Early face detection and recognition techniques were based on shallow representation frameworks, like the Haar cascades (Viola & Jones, 2001), or robust features like SURF (Speeded-Up Robust Features) (Li et al., 2011) and Histograms of Oriented Gradients (HOG) (Shu et al., 2011) or Local Binary Patterns (LBP) (Ahonen et al., 2006) in order to detect faces. However, their low recognition accuracy rate and high computational cost (i.e. they used exhaustive sliding window search techniques) led the computer vision community to search for faster and more accurate algorithms. Thus, experimentation started with facial points, such as the mixtures-of-trees (Zhu & Ramanan, 2012) and consen-

sus of exemplars (Belhumeur et al., 2013), that required much less computational time due to the more efficient point localization methods that they deployed.

With the rise of deep neural networks, especially in the tasks of image classification and generic object detection, a breakthrough in terms of higher performance models came to be soon thereafter. Deep architectures, such as the deep convolutional network cascade (Sun et al., 2013), almost solved the face detection problem, by achieving very low failure and average detection rates on some very challenging face detection datasets, such as LFPW (Belhumeur et al., 2013) and YouTube faces (Wolf et al., 2011). In addition, methods that simply approached the face detection problem as a generic object detection problem drove researchers to publish works, such as (Jiang & Learned-Miller, 2017; Markatopoulou et al., 2017) and (Sun et al., 2018), that were more straightforward applications to already classic object detection schemes, using only face boxes for training and at the same time techniques like hard negative mining, feature concatenation and careful fine-tuning to improve their results. More sophisticated works considered the spatial structure and arrangement of facial parts (Yang et al., 2015). Inspired by classic object proposal generation techniques, different CNNs were trained to detect different facial parts, while early convolution feature maps were shared for computational efficiency. Then, an object proposal ranking step followed that used “faceness” (i.e. the concept of what constitutes a face) scores calculated by measuring how well each proposal met the structural constrains that were posed by the detected facial parts. Later, the framework by Zhang et al. (2016) leveraged a cascaded architecture with three stages, each one deploying a deep CNN. Object proposals were extracted during the first stage using an FCN, false positive candidates were filtered using another CNN in the second stage, and further refinements were performed based on facial landmarks during the final stage.

Analogously to the object detection literature, following in the footsteps of the single-shot region-free philosophy for generic object detection, the SSH (Single-Shot Headless) face detector (Najibi et al., 2017) discarded the need for a face bounding box proposal generation step. It achieved that while using different layers of an FCN to predict various scales of faces simultaneously in one forward pass. Delving deeper into the relevant tasks adjacent to face detection, Ranjan et al. (2017) extended the functionality of their network constructing a multi-tasking framework for automatic facial landmark localization, pose estimation and gender recognition along with face detection. This method was categorized as region-based, in essence working on patches of the image (candidates). Deep convolutional features taken from different layers were first fused and then fed to a five-headed output, where each "head" was a Fully Connected (FC) network dedicated to a task. Recently, in the wider context of the general problem of detecting small objects, Hu & Ramanan (2017) proposed that, in order to accomplish tiny face detection, a separate training of dedicated detectors had to be done for a number of selected scales. They did this more efficiently by allowing sharing of features between multiple layers of a hierarchy. Finally, it was also concluded that context is very crucial in detecting tiny objects by making use of very large receptive fields.

2.1.3 Scene Recognition

Scene recognition for indoor environments is also an integral part of the SUITCEYES visual analysis component. Not only will it be able to provide the user with information about the environment they are currently in, but the knowledge of the scene can offer useful information for the workflows of other visual analysis tasks, such as the activity recognition task. These are important contributors to the task of situational awareness.

Earlier works proposed shallow colour descriptors to alleviate illumination variances of the scenes, and an overall boost in classification performance was reported compared to using Scale Invariant Feature Transform (SIFT) descriptors in (Van de Sande et al., 2010). At the same time, Quattoni & Torralba (2009) realized that models trained in outdoor scenes were performing poorly in indoor scenes; the exploitation of exclusively global spatial features in some cases or local visual cues in other cases, such as the existence of particular objects, was identified as the main reason for this performance gap. Their model achieved better results by incorporating both sources of information in the final descriptor using scene prototypes, which was a technique previously examined in (Quattoni et al., 2008). Another work that leveraged objects (Espinace et al., 2010) proposed a Bag-of-Words (BoW) dictionary-based model using features from ROIs that were found by object detectors. Part-based models were dominating the scene recognition literature at that time, like the work by Pandey & Lazebnik (2011). Later, following the same trend, a collection of region models in a reconfigurable pattern was the main idea behind the work proposed by Parizi et al. (2012). The authors suggested that each scene could generate specific patterns of visual parts arranged spatially in a discriminative way.

With the extensive work by the research community to design powerful deep CNN feature extractors, the task of classifying images with high accuracy rates has become the standard recently, as shown by the performance of state-of-the-art methods in well-known competitions such as the ImageNet visual recognition challenge (Russakovsky et al., 2015). The problem of defining the scene from visual content suddenly became much simpler than object or face detection. The state-of-the-art since then lies in the classic combination of a generic deep CNN feature extractor like VGG (Simonyan & Zisserman, 2014), ResNet (He et al., 2016), or Inception (Szegedy et al., 2015) and a classifier on top that can be simply trained with images depicting various scenes (Avgerinakis et al., 2018), like the ones found in the Places2 dataset (Zhou et al., 2017).

2.1.4 First-Person Activity Recognition

Wearable cameras can capture the activities that the user or other people in the same environment are doing. The camera view however is not the same for those two kinds of activities, since on one case the user plays the role of both the target and the observer (through the mounted camera). In other words, the actions of the user are captured from a First-Person View (FPV), while the actions of other people are captured from a Third-Person View (TPV). This distinction is naturally evident in the literature as well, and the development of two types of algorithmic approaches emerged that are designed to operate either for a First-Person or a Third-Person perspective. More specifically, the task of detecting activities from egocentric vision is well suited to SUITCEYES for a single user situation awareness scenario. This function will be developed in order to gather and analyse visual data with the aim of learning probable routines that the users may frequently follow throughout their daily lives. This knowledge can be further exploited later by experts with the prospect of learning people from their activities and improving their lives by improving their behavioural patterns.

Activity recognition from egocentric videos is a very hot topic in computer vision and a lot of works have been proposed in the last decade to solve this challenge. Many of them propose to describe activities by an object-centric manner following the information that can be derived from the existence of specific objects in the scene (Fathi et al., 2011; Pirsiavash & Ramanan, 2012; McCandless & Grauman, 2013; Zhou et al., 2016). Moreover, scene understanding is also used in (Vaca-Castano et al., 2017), in order to provide *a priori* knowledge to the system about activities that usually take

place in certain environments, e.g. being in the bathroom limits the possibilities of activities such as taking a bath, or washing teeth. Other works leverage the motion that appears in the scene and extract features so as to represent the activities that take place (Kumar & Bhavani, 2017; Kumar et al., 2017). Also, in the work of (Avgerinakis et al., 2015), camera motion that is present in videos from ego-motion was compensated using superpixels and multiple homographies, before sampling dense trajectories (Avgerinakis et al., 2016) to describe activity motion from salient parts of the video frames. In (Yan et al., 2015) a multi-task clustering framework tailored to FPV activity recognition is presented. Another more recent approach is to use deep CNN architectures (Wang et al., 2018) to learn deep appearance and motion cues. Deep CNNs are also used to learn hand segmentations in order to understand the activities that a user performs and his interaction with other users that might also appear in the video frame (Zhou et al., 2016; Bambach et al., 2015a; Bamback et al., 2015b). More recent works focus on multi-modal analysis of egocentric cameras and information from other wearable sensor equipment with the deployment of early or late fusion schemes (Meditskos et al., 2017; Crispim-Junior et al., 2016; Crispim-Junior et al., 2017).

2.1.5 Third-Person Gesture Recognition

The literature refers to gesture recognition as the task of recognizing meaningful expressions of motion by a human, involving a range of human body parts, like the hands, arms, face, head, or the body (Mitra & Acharya, 2007). Many vision-based gesture recognition surveys and works have stated that the field is destined to be the critical element in human-computer interaction. Although in the SUITCEYES project the purpose of gesture recognition is not the same, however, the algorithms remain the same. Most of these techniques focus on recognition of hand or finger gestures (Rautaray & Agrawal, 2015). The next most popular category is gestures using objects vs body gestures, including movements with body parts other than hands, which attract the interest of a small portion of researchers. Many works include the use of sensors or extra equipment to aid the purpose. However, here, we are going to review only some of the appearance-based models.

Appearance-based representation methods can be classified in two major subcategories: static model-based methods, and motion-based methods. Some early works on static-based methods focus on colour-based body markers to track the motions of a particular body part using particle filters for tracking (Bretzner et al., 2002). Though simple, this is not very practical in terms of flexibility, and cannot deal with recognizing gestures "in the wild". Another static-based method focused on the silhouette geometric parameters of hands like orientation, perimeter and centroid position (Birdal & Hassanpour, 2008). The core of the motion-based models is when the approach of the task falls into the category of general action recognition. In early such works, a gesture recognition pipeline involved hand segmentation and tracking and extraction of motion descriptors. Hand detection could be done using skin colour segmentation in many early works operating in alternative colour domains, such as HSV (hue, saturation, value), to compensate for illumination changes (Saxe & Foulds, 1996; Chai & Ngan, 1998; Yang, Lu, & Waibel, 1998). Other works, like (Belongie et al., 2002) focused on shape descriptors to detect the hands.

For the tracking task, various approaches have been proposed including correlation-based feature tracking (Crowley et al., 1995; Darrell et al., 1996), or contour-based tracking (Cootes & Taylor, 1992). Last, in order to model sequential states of the detected and tracked hands that make gesture phrases, the literature proposes Hidden Markov Models (Starner, 1995), Dynamic Time Warping (Corradini, 2001), and time delay neural networks (Sigal et al., 2004). More recent works make use

of 3D convolutional neural networks to model hand gestures, as in (Molchanov et al., 2015). Moreover, end-to-end trainable deep architectures incorporating temporal convolutions and bidirectional recurrence have been proposed in (Pigou et al., 2018). Finally, a method using a 2D map of motion energy from consecutive frames, and further training of PCA models optimizing a reconstruction-error, have also been proposed for one-shot gesture recognition (Escalante et al., 2017).

2.2 Algorithms and Results

This section discusses the current visual analysis algorithms that have been developed and provides qualitative and quantitative results for each task as a means of evaluation.

2.2.1 Object Detection and tracking

For the purpose of detecting objects of interest, we chose to extract deep image representations from a CNN and predict pixel coordinates of bounding boxes using a deep CNN object detector. To this end, we adopt a modification of the accurate Faster-RCNN. A thorough evaluation of this model and comparisons with other state-of-the-art deep object detectors presented in (Huang, et al., 2017) reveal that the Faster-RCNN-resnet101 architecture achieves a good trade-off between speed/accuracy. This model incorporates the resnet101 (He et al., 2016) deep feature extractor and a region proposal network along with a bounding box classifier and coordinate regressors. We chose this architecture because it achieves very fast object detection by using a single feed-forward convolutional network to directly predict classes and bounding boxes of objects. In order to speed up our object detection procedure during inference time, we tracked the detected objects found in a frame into the next T frames of the video. By assigning a detection rate of $T > 15$, our combined detector and tracker algorithm achieves real-time performance. Following an empirical evaluation after trials, we manually set the detection rate parameter to 15, with other values ranging from 15 to 30. Intuitively, the detection rate defines the temporal resolution of the continuous object detection function. Lower detection rate means higher temporal resolution of the detector and vice versa. Note that, by setting the detection rate to 15, the detector only runs once between half-second intervals and the tracker works the rest of the time, yielding an adequate temporal resolution considering that it is very unlikely that an object will appear and disappear in less than that.

The core functionality of our object tracker is based on the kernelized correlation filtering (KCF) tracking algorithm that was proposed in (Henriques et al., 2015). The detector is used initially in order to detect objects in every T video frames and initialize the new object candidate database with new entries. Bounding box coordinates are stored over time so that full trajectories can be built. For every new target ID, its corresponding class label and a detection score are saved as well. Afterwards, the algorithm checks the new detections from the candidate pool for overlaps with already existing recent trajectories. Then, based on an Intersection-over-Union (IoU) score check, it rejects found boxes that exceed an overlap threshold to avoid creating multiple identities for the same object. Next, we feed the KCF tracker with the remaining boxes in order to localize their position throughout sequential video frames. Future detections of already tracked objects are also utilized in order to rectify the bounding boxes of the monitored objects. When a detection is missed, we re-localize the bounding box relying only on KCF update coordinates, while, when the algorithm does not localize any tracked object for l sequential video frames, the object is presumed to have travelled off the frame. In the current version the parameters are set to T=15 and l=3 frames. To tackle

overlaps between True Positive (TP) cases, we chose to merge the trajectories at the current frame and assign the oldest ID to the resulting trajectory.

From the 48 different classes of objects that are available in the current version of the ADL dataset (Pirsiavash & Ramanan, 2012) we select the 30 most frequently annotated to train our object detector. Table 1 shows the list of objects that the object detector is trained for. Figure 2 shows a qualitative evaluation of the detector. The visualization uses coloured boxes to mark the location of the targets that have been found by the detector and their respective class.

Table 1. Object classes that the detector can recognize.

cell phone	tv remote	towel	door	pan	knife/spoon/fork
oven/stove	washer/dryer	vacuum	detergent/soap	tv	pills
water tap	fridge	blanket	microwave	container	food/snack
book	mug/cup	toothbrush	tooth paste	dish	comb
laptop	pitcher	trash can	kettle	bottle	person

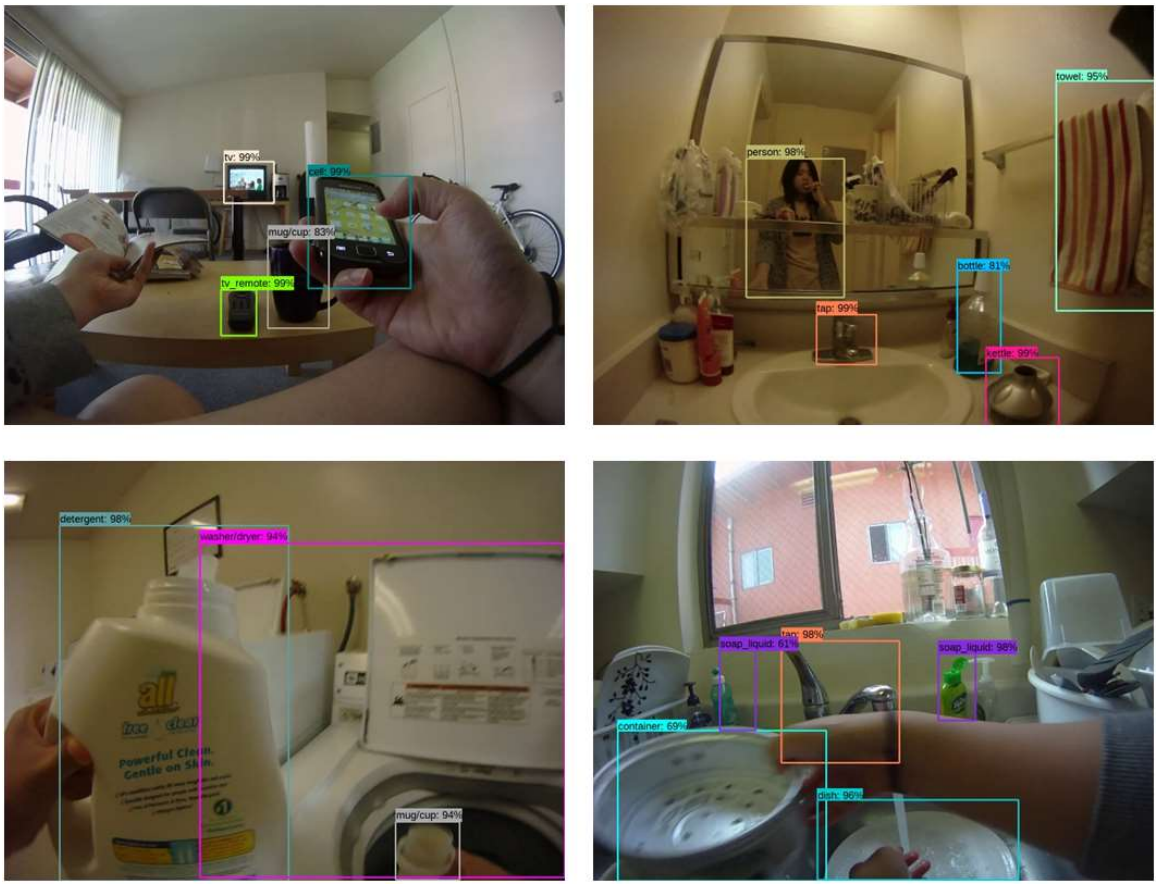


Figure 2: Object detection qualitative results.

2.2.2 Face Detection

As explained in detail in the background section, the face detection problem can be tackled by a solution for generic object detection like the one described previously. Besides its simplicity, this approach would also lead to a shared implementation and perhaps some shared calculations between those two subtasks, which is obviously a characteristic of great importance for real-time systems. This is a matter that will be explored further in future versions of the visual analysis component. In

the current version we chose to focus on performance rather than execution speeds and adopt the excellent performing TinyFaces deep CNN framework (Hu & Ramanan, 2017) for face detection. Its performance boost is due to the fact that it is designed to detect even the most distant faces. Moreover, the method can leverage from high quality video that the equipment is capable of capturing. Another important aspect is that we can always speed up the face detection process by combining it with our object tracker in the same manner as described in the previous section and still take advantage of the state-of-the-art performing model. Boxes of faces can be fed to the tracker in order to alleviate from the system the cost of performing the entire feed forward pass of the TinyFaces model for a short period of time.

In its core, TinyFaces is a binary multi-channel heatmap prediction problem, utilizing an FCN, where the predicted heatmap at a single pixel location resembles the confidence of a fixed size detection window centred at that pixel. Separate scale-specific detectors are trained that were designed carefully in order to maximize performance for each face scale with respect to the characteristics of a detector's template size and resolution.

Figure 3: Face Detection qualitative results.

In more detail, the design is based on two very important conclusions that were verified in (Hu & Ramanan, 2017) by performing an extensive cycle of experiments involving context and resolution. It was proven that using deep CNN features of multiple layers, hence information from different receptive fields, is of extreme importance especially for finding small faces. This can be related to context. This may seem counterintuitive at first, but practically it is reasonable that even humans need surrounding context to recognize if a real size 25x20 pixel portion of an image resembles a face and that they probably will fail to recognize when looking only at this tiny portion. In contrast, in order to de-

tect bigger faces (e.g. 300x300), it is enough to get no context at all but only what is provided inside the face portion. As a result, it was successfully demonstrated that deep CNN features extracted from multiple layers, are effective "foveal" descriptors that capture both high-resolution detail and coarse low-resolution information from large receptive fields. Foveal descriptors were proven to improve performance of small face detection by up to 33% compared to regular single layer-based descriptors.

Moreover, resolution can have different impacts on performance for various template scales. It was proven that building templates at the original resolution is not optimal. When finding small faces, the scale-specific template size is doubled in combination with 2x up-sampling in resolution, improving accuracy by 6.3% compared to standard template size with standard resolution. In contrast, for finding bigger faces, templates and resolution have to be downscaled by a factor of 2, resulting in an improvement in accuracy by 5.6%. This is done in order to accommodate the natural overfitting towards medium-sized objects a pre-trained CNN architecture has been exposed to, as a result of the fact that medium-sized objects are dominant in large scale datasets.

The TinyFaces model uses ResNet101 (He et al., 2016) as a deep feature extractor that is fed with the re-scaled input composed from a coarse image pyramid. Template responses at every resolution are made and non-maximum suppression (NMS) is applied at the original resolution to get the final detection boxes. The pre-trained ImageNet model is fine-tuned on the WIDER FACE dataset (Yang et al., 2016). Qualitative results are presented in Figure 3 using video frames from the Dem@Care action dataset for evaluating dementia patients in a home-based environment (Avgerinakis & Kompatsiaris, 2016). As shown by the results, even very small, partially occluded faces with challenging illumination settings, as well as the larger close-up faces can be detected.

2.2.3 Scene Recognition for Situational Awareness

Scene recognition is currently tackled in SUITCEYES by the classic image classification paradigm that was mentioned in the background section. Following the success of deep CNN convolutional image classification techniques in recent years, SUITCEYES adopts the deep VGG16 CNN architecture trained on the Places2 dataset (Zhou et al., 2017).

As the name suggests, VGG16 encapsulates 16 layers of analysis. The first 13 are convolutional, using 3x3 kernels and an increasing number of filters (from 64 to 512). In-between convolutional blocks there are 2x2 max pooling operations as well. The final 3 layers are two fully connected layers of width 4096, and an output layer that consists of a softmax function suitable for multi-class classification. The Places2 dataset contains over 10 million instances of pictures annotated from a set of 400+ scene categories. Among the possible categories, there are many indoor scenes that the SUITCEYES visual analysis component is mainly focused on providing predictions for.

Figure 4 provides some qualitative results when using the trained VGG16 network to detect scene categories from the ADL dataset. As shown, the prediction confidence can be low for some predictions, but this is due to the high number of classes that the model is trained for. In a later version, and based on the (now ongoing) user requirements analysis (WP2), the set of possible classes will be refined to include only indoor scenes relevant to SUITCEYES purposes and better performance will be expected. For visualization purposes, global average pooling, first proposed by (Zhou et al., 2016), is used to produce class activation maps and its results are shown in the right column of Figure 4. In addition to the original VGG16 architecture, a class-specific saliency map generator can be attached to the implementation that extracts salient parts in the scene for interpreting the decision of the

scene recognition network. In other words, a heatmap like the ones in the right column of Figure 4 for a particular category represents the regions that were the most discriminative in the image and were selected by the CNN in order to uniquely identify that particular category. Besides the current use of this technique as a utility to better visualize the outcome and mechanics of the scene recognition network, generation of salient features is a technique that will be explored in a future version as a way of providing additional information for other tasks of the visual analysis component.

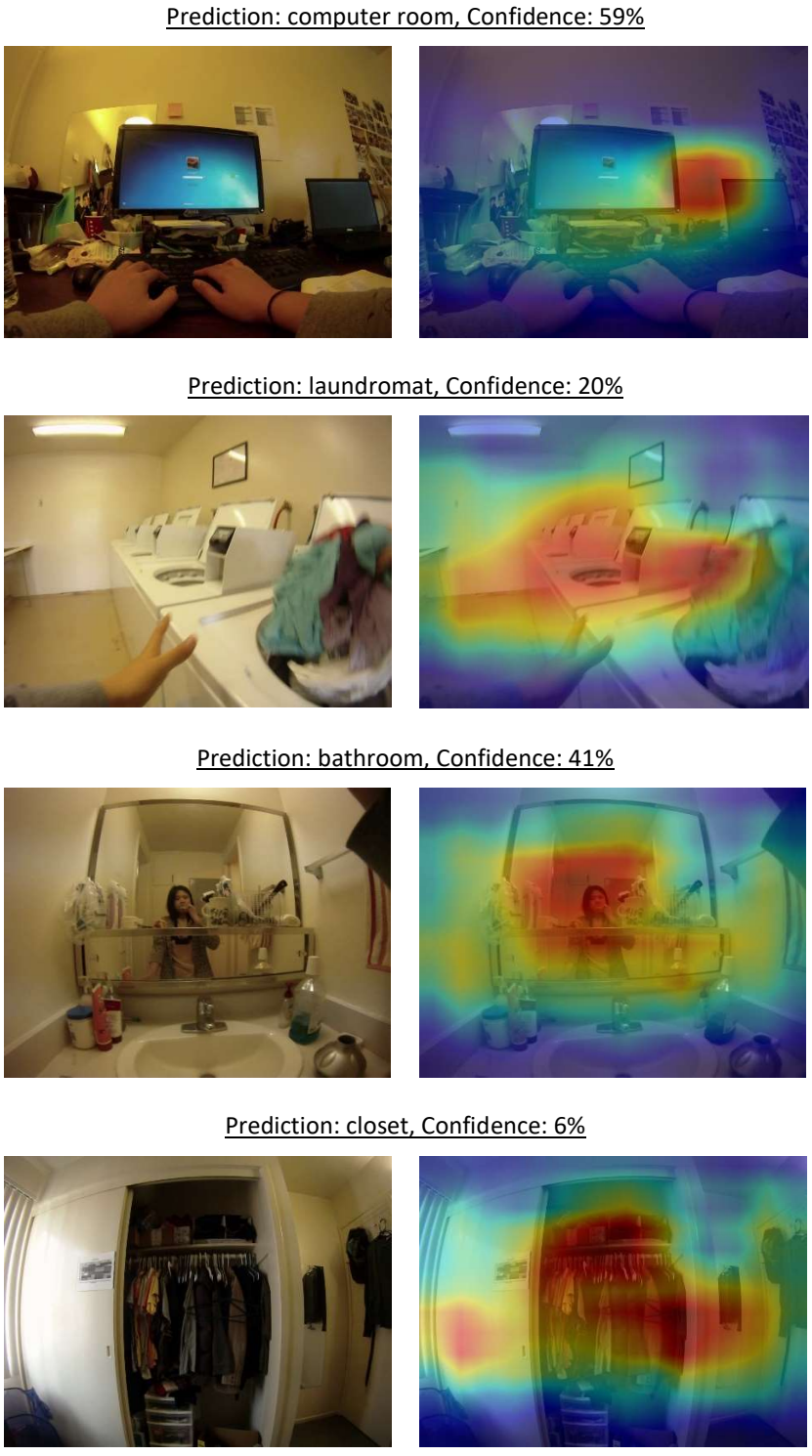


Figure 4: Scene Recognition qualitative results. Left column is the input images, right column is the activation maps and on top of the images are the predicted class and the prediction confidence.

2.2.4 First-Person Activity Recognition

In order to successfully recognize activities of daily living, such as "*book reading*", "*hand washing*" or "*preparing breakfast*" that take place inside an egocentric video, it is important at first to get a deep understanding of the short time lower level actions a person is performing sequentially in order to accomplish the bigger scale ones. For example, the "*preparing breakfast*" activity involves the short time actions "*opening the fridge*", "*grabbing butter*", "*closing the fridge*", "*taking a knife*", "*spreading the butter*" etc. This group of so-called "micro-actions" does not always need to form a complicated sequence for every activity. For example, the activity "*reading a book*" besides the actual reading usually involves only one micro-action performed repeatedly: "*turning the page*". For those reasons we seek a way of extracting a representation of the full duration of an activity video that will be informative towards the set of micro-actions that are included and have a strong ability to uniquely describe the activity. It is also very well established, as described previously in the background section, that objects are good indicators of certain activities such as the TV in the "*watching television*" activity or the book in the "*reading a book*" activity. We further elaborate this notion by hypothesizing that not only the presence but also the characteristic motion of the objects that is taking place in the scene is powerful enough to discriminate between active and passive ones and at the same time inform about the activities that are happening. For example, motion information from dishes that are being washed combined with the presence of a tap in the scene can uniquely describe the "*washing dishes*" activity.

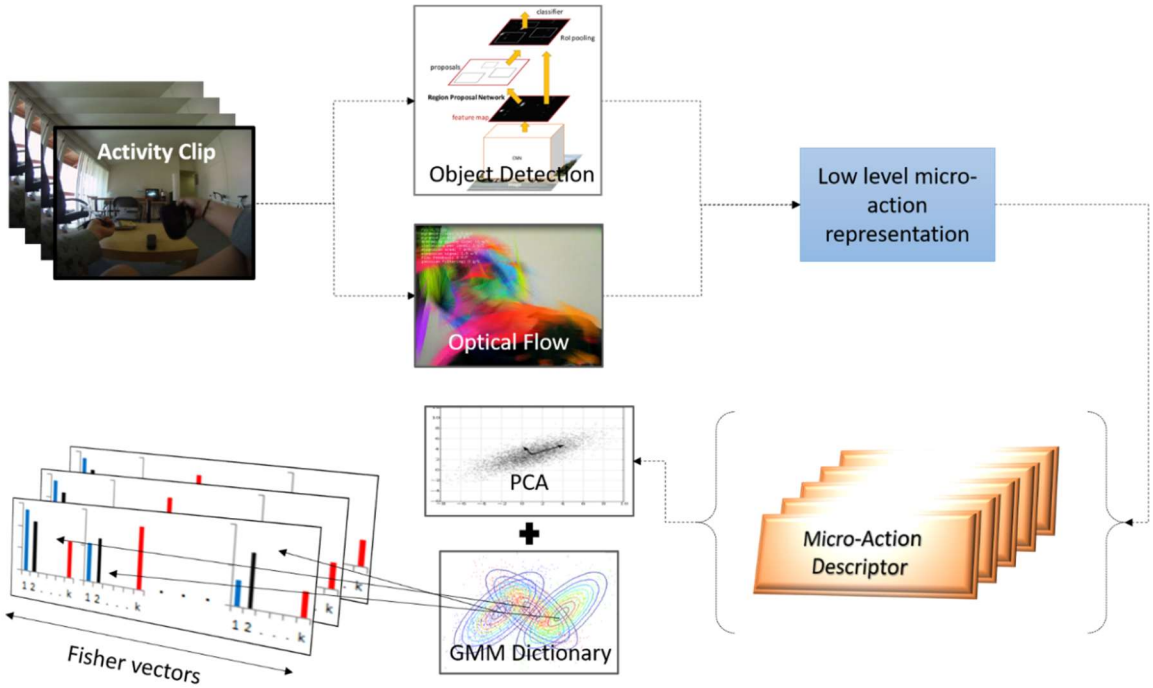


Figure 5: Block diagram of First-Person Activity Recognition pipeline.

The above assumptions are taken into account in our activity recognition method. The overall framework is shown in Figure 5. First, we detect objects and track them using the object detection pipeline. Then, every detected object's motion is analysed using HOF (histograms of optical flow) or MBH (motion boundary histogram) (Dalal et al., 2006) features so as to form the lower level micro-action representations that appear in short time windows over the full activity sequence. Finally,

GMM (Gaussian Mixture Modelling) clustering of the micro-action descriptors is performed in order to find the most discriminative of the full set. Given a set of micro-action descriptors extracted for a single activity sequence and the GMM clustering centres, a Fisher encoding scheme is used in order to yield the final descriptor of the full activity sequence in a Bag-of-Micro-Actions type of representation.

In Figure 6 we can see how our proposed object motion descriptors are computed. Our method builds representations of short-term low-level actions of fixed temporal window W from the motion patterns of the objects that are found in this window. More specifically, we compute dense optical flow to extract the full scene's motion between two consecutive frames. We use the OpenCV implementation of the dense inverse search algorithm proposed in (Kroeger et al., 2016). In addition, doing the calculation every other frame inside the window instead of every frame leads to W/2 calculations, yielding faster computation times. Having already detected the objects in a particular frame, we take each bounding box as our region of interest and crop the dense optical flow map accordingly, taking only the portion that belongs to the object. Consequently, we can calculate HOF descriptors that represent an object's motion.

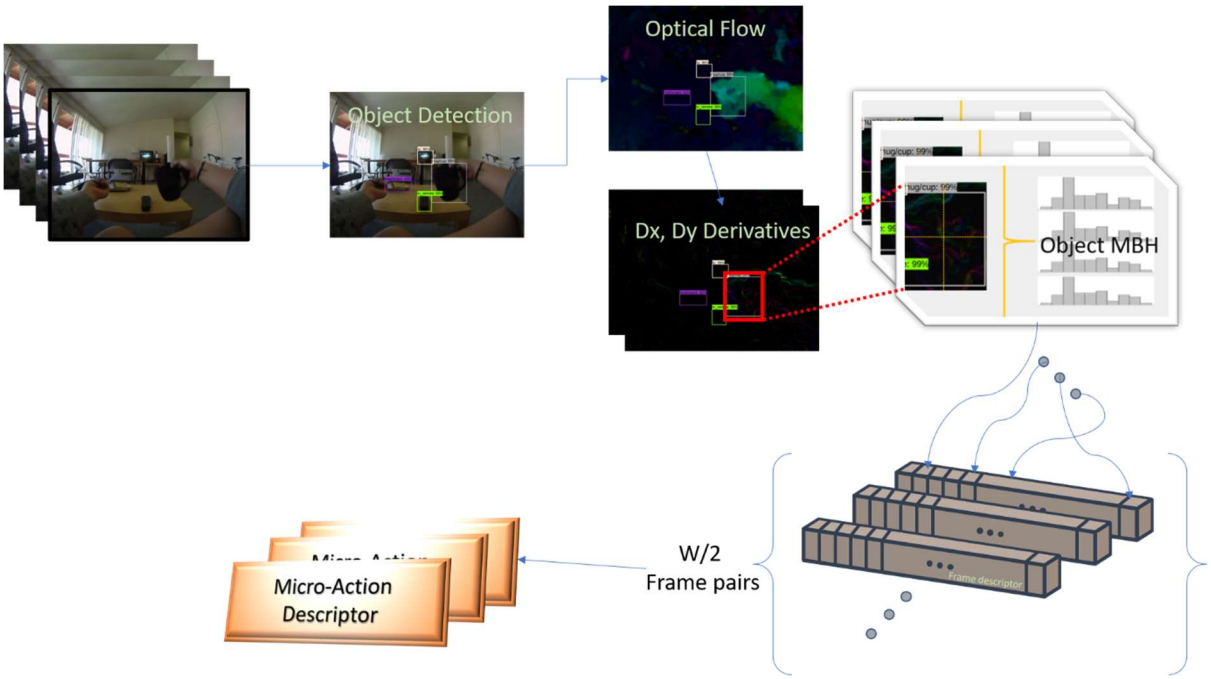


Figure 6: Computation of micro-action descriptors.

One problem with the accurate extraction of object motion from egocentric videos is that very frequently the wearable camera moves along with the person wearing it. As a result, global camera motion may overpower the delicate dynamics of the objects' motion that we are trying to capture. Therefore, we consider an alternative to the HOF descriptor that is the MBH descriptor, where the optical flow field is first separated into its x and y component and spatial derivatives are computed for each one of them. This time we obtain a 32-dimensional for each component (64-dimensional overall) and we follow the same procedure to obtain the final descriptor as in the HOF descriptor case. Because MBH is the gradient of the optical flow, any motion that is happening constantly (global motion) is suppressed and only information about changes in the flow field (i.e. motion boundaries) is kept (Wang et al., 2011). Compared to video stabilization and motion compensation this is a faster method of discarding global motion information.

For a given activity sequence, the extraction of micro-action descriptors that represents a small sequence of W frames takes place with a stride of S frames. We chose that value to be exactly 1 second in all our experiments. This simply means that for every micro-action descriptor M we skip 1 second into the video before we begin extracting the next micro-action descriptor. Contrary to using overlapping windows, the stride parameter was inserted to give our method a speed boost. Given that the micro-action window W is chosen sufficiently small, it is guaranteed that the number of micro-actions for an activity sequence will be enough for the activity to be adequately represented. Subsequently, all micro-action descriptors extracted from all the training activity sequences are fed into a Fisher encoding scheme. This way, a micro-action vocabulary based on the most discriminating ones is constructed. The computation of the most discriminating samples is performed by applying unsupervised clustering (Gaussian Mixture Modelling - GMM) in the micro-action representation hyperspace. As a means of dimensionality reduction, we perform Principal Component Analysis (PCA) on our low-level descriptors. PCA guarantees maximum variance of the samples in the lower dimensionality space. We chose two possible reductions in our experiments: 80 and 256 components. This way, our early micro-action descriptor's dimensionality reduces from some thousand components to only a couple of hundreds. Additionally, we experiment with two different vocabulary sizes using 32 or 64 words. For the final step, we deploy as our classifier a fully connected neural network (NN1) with a depth of two layers of width 512 and 256 accordingly, using RELU activations, 50% chance of dropout between layers and softmax activation in the output layer. Another similar architecture (NN2) was also deployed with half the number of neurons for each layer (256 in the first layer and 128 in the second) and a linear SVM classifier as well.

Table 2: Performance comparison of SUITCEYES FP Activity Recognition method with SoA in the ADL dataset.

Method	Performance (mAP%)
Boost-RSTP (McCandless & Grauman, 2013)	33.7%
Boost-RSTP + OCC (McCandless & Grauman, 2013)	38.7%
Bag-of-objects (Pirsiavash & Ramanan, 2012)	32.7%
Bag-of-objects + Active model (Pirsiavash & Ramanan, 2012)	36.9%
Cascaded Interactional Network (Zhou, Ni, Hong, Yang, & Tian, 2016)	55.2%
Bag-of-Micro-Actions with HOF	52.86%
Bag-of-Micro-Actions with MBH	**57.14%**

We performed our experiments in the ADL dataset (Pirsiavash & Ramanan, 2012), which is composed of videos recorded with a wearable camera from 20 different persons. The videos contain very realistic scenes of daily living and is challenging due to the existence of global camera motion as a result of the camera movement. The objects are also in many cases occluded. To evaluate the action recognition performance as in (Zhou et al., 2016), we performed the leave-one-person-out cross-validation method for every parameter combination we discussed and we calculate the per-class average precision (mAP) score. Overall, the best models came from the combination of 256 PCA components coupled with a GMM vocabulary of size 32 and the neural network architecture with the most learnable parameters (NN1). Finally, the MBH descriptor almost entirely outperformed the HOF descriptor for every experiment with a temporal window of 60 frames and the performance of the two was comparable for a window of 90 frames. This is an indication the MBH is more promising when micro-action extraction is more refined in time. In Table 2, we compare the accuracy rates of our best models to the ones that are mentioned in the literature with experiments made in the same dataset. We followed the evaluation procedure in (Zhou et al., 2016) in order to

present comparable results. As we can see, the MBH version of our method outperformed every other state-of-the-art method. The HOF descriptor is also highly ranked.

2.2.5 Third-Person Gesture Recognition

Regarding gesture recognition, SUITCEYES currently adopts the OpenPose framework for human pose estimation (Cao et al., 2017). That is, the estimation of the location of human parts, (e.g. right/left wrist, right/left elbow, etc.) and how they are linked. Note that gesture recognition has not been directly implemented here, but the current version's output is essential for future versions of this task. We plan to utilize these human body part detections later to understand what other people in the room are signalling using a set of predefined body stances. In this way, the design of the set of gestures will be more flexible to meet the users' requirements and also the signals could be more evident and distinct for the model to capture in comparison to a set constructed entirely out of look-alike hand gestures.

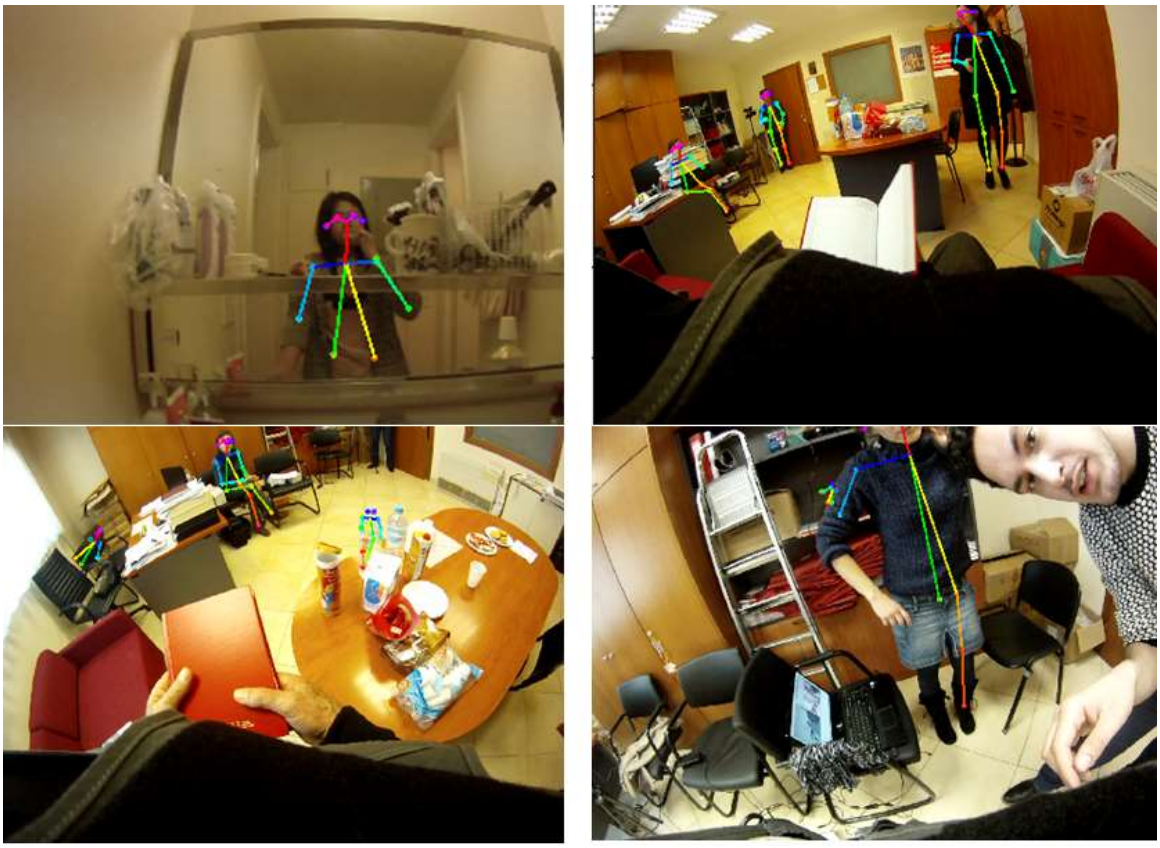


Figure 7: Qualitative evaluation of the pose estimation algorithm.

The architecture of this model is based on a two branch multi-stage CNN. There are two kinds of predictions that are calculated during each stage by the two branches. The first one predicts a confidence heatmap of possible locations of body parts (one for each part), whereas the second branch predicts a part affinity 2D vector field that represents at each point the most probable direction (if any) of a link between parts of the same human. Two L2 loss functions are applied at the end of each stage, one for each branch respectively. At first a deep CNN feature map F is extracted using the first 10 layers of the VGG19 architecture, and is fed to the first two-branch stage. Each stage after that gets as input the concatenation of the original VGG feature map F with the two output feature maps

S1 and S2 of the previous stage's two-branch network. This is done in order to provide more refined predictions after each stage.

In order to create ground truth to train the two branches of the network, a dataset pre-processing step has to be performed. For the first branch, the ground truth map is created by centring a Gaussian distribution to the pixels where a body part location exists. For the second branch, the ground truth map is constructed using a part affinity field that is a 2D vector field for each body part link. Each pixel inside the support region of a body part link votes a unit vector in the direction of the link starting from the first part's centre and pointing towards the second part's centre. During testing the alignment of the predicted part affinity field with the candidate link that would be formed by connecting the detected body parts is measured and the strongest links are kept. The network is trained on the MPII human pose dataset (Andriluka et al., 2014). Figure 7 displays some qualitative results obtained by evaluating the OpenPose model on still frames from the Dem@Care and the ADL datasets. As shown, the algorithm performs mostly well, and even estimates the human pose of an almost completely occluded human body in the top-right picture. There are some false positives in the bottom-left picture, where a human body is found in a plastic chair. Also, the behaviour of the algorithm can be unpredictable in occluded bodies, as shown in the bottom-right picture, where it has failed to recognize the left side of the human in the right, and the right leg of the human in the left.

2.3 Chapter Summary and Future Work

In this chapter the basic computer vision algorithms that will carry the visual analysis tasks within the SUITCEYES project were presented, along with some qualitative and quantitative performance evaluation results. In the upcoming releases each specific task will be carried out by more mature and evolved solutions. A form of nearest-neighbour search inside a database of known faces will be developed, in order to evolve face detection into a face recognition task. A set of gestures will be defined, in order to develop a new pipeline for gesture recognition using human poses and the possibility to incorporate movement information will be explored. Scene recognition will also be leveraged by the first person activity recognition pipeline in order to narrow down the possible activity outcomes based on the environment the user is in. Furthermore, all the algorithms will be properly modified in order to process automatically fixed temporal segments (or temporal windows) of the video feed rather than still frames, and will produce results that describe the events of the most recent window, with a task-depended frequency.

3 Dimensionality Reduction

In order to deal with the translation of environmental cues to meaningful haptic signals (e.g. vibrations), this task aims at utilizing already widely used embedding algorithms to reduce the dimensionality of images (i.e. high dimensional signals). The dimensionality reduction (DR) methodologies mainly aim at dramatically reducing the dimensionality of high dimensional signals, paving the way for the visualization of the inner structure properties of high dimensional data in low dimensional manifolds, mainly in two or three dimensions. In other words, the common idea underlying the DR methods is that of providing a mapping which preserves the similarity, e.g. pairwise distances, between instances of the original high dimensional space to a new, low dimensional one. For the SUITCEYES project, such algorithms will be useful in order to map high dimensional signals (i.e. environmental cues, such as images captured from a wearable camera) to low dimensional (e.g. 2D) spaces, and subsequently to a haptic space where the two dimensions will refer to, e.g. the tempo and the pitch of the vibration. That way, semantically same environmental concepts, e.g. images of beds or of apples, would be grouped in the haptic space and lead to potentially recognizable patterns from the users as far as they have become familiar with certain activation patterns in the haptic space.

The following subsections present the aforementioned DR methodologies used for the reduction of the dimension of images from the CIFAR-100 benchmark database[1]. More particularly, we present the Principal Components Analysis method (Van Der Maaten et al., 2009), a methodology widely used for decades, along with Isomap (Tenenbaum et al., 2000) and t-SNE (Van Der Maaten & Hinton, 2008), two of the most famous current state-of-the-art methodologies for DR. Subsequently, the CIFAR-100 benchmark dataset is presented along with the concepts that will be utilized for mapping to low dimensional spaces, i.e. the objects or notions to be mapped in low dimensional spaces, e.g. images of an apple. Next, the DR results are presented for five object categories, i.e. 'Food Containers', 'Fruits and Vegetables', 'Household electrical devices', 'Household furniture', and 'People'. Finally, we discuss issues on the results and future steps.

3.1 Dimensionality Reduction Algorithms

In this subsection we will briefly present the three most established methods for DR in literature. The description of the methods covers the fundamental principles of the techniques. For all techniques we consider $X \in \mathbb{R}^{N \times D}$ being the high dimensional instances (N instances) with dimensionality D and $Y \in \mathbb{R}^{N \times d}$ the low dimensional ones with dimensionality d.

3.1.1 Principal Component Analysis

The Principal Component Analysis (PCA) method is a seminal technique which dates back to the early 20th century. In essence it is an orthogonalization procedure and not a dimensionality reduction method per se. The reduction occurs when projecting only on a subset of the orthogonal components. More formally, the PCA technique works as follows: first the covariance matrix $C_X = X^T X$ is computed; second, using the Single Value Decomposition method, the covariance matrix is decomposed, i.e. $C_X = U \Lambda U^T$. U is the eigenvector matrix whereas Λ is a diagonal matrix of the eigenvalues of C_X. Finally, the orthogonal representation is accomplished by a linear mapping, i.e. $Z = XU$

[1] Cifar-100, available online at: https://www.cs.toronto.edu/~kriz/cifar.html

whereas the dimensionality reduction is done when using only the first d components, i.e. $Y = XU_d$, where U_d is a matrix that contains only the d columns of U (sorted in decreasing order of the corresponding eigenvalues). By definition, such a reduction minimizes the $\|X - YU^T\|_2^2$ error.

3.1.2 ISOMAP

The ISOMAP method works as follows: First, a k-Nearest Neighbour graph is constructed for data X. A complete matrix of pairwise *geodesic* distances is computed using traditional shortest–path algorithms such as Djikstra or Floyd-Warshall. As soon as the pairwise matrix is estimated, the Multidimensional Scaling (MDS) procedure is followed. Thus, the distance matrix is doubly centred and its eigenvectors are extracted by spectral decomposition (Van Der Maaten et al., 2009).

3.1.3 t-SNE

A non-convex approach for dimensionality reduction, namely Stochastic Neighbours Embedding (SNE) (Tenenbaum et al., 2000), has been proposed using probabilistic modelling. In particular, SNE's principles are as follows: the high dimensional points are modelled by a pairwise probability matrix P where nearby instances are assigned high transition probabilities and distant instances are assigned low probability values. Similarly, the embedded instances are modelled using another probability matrix Q which is constructed with the same procedure. The objective of the method is to minimize the Kullback-Leibler divergence between the two distributions. Despite the fact that the original SNE work used Gaussian distributions for both matrices, a more recent work used a heavy tailed, t-Student distribution for the Q matrix, which led to the one of the most famous DR techniques, i.e., the t-SNE method which is utilized here (Van Der Maaten & Hinton, 2008).

3.2 Benchmark Dataset and Concepts to be Recognized

The CIFAR-100 dataset consists of 60000, 32x32 pixel images of 100 classes, i.e., 600, images per class. The classes are grouped into 20 super-classes (categories). A sample of the CIFAR-100 images is shown in Figure 8.

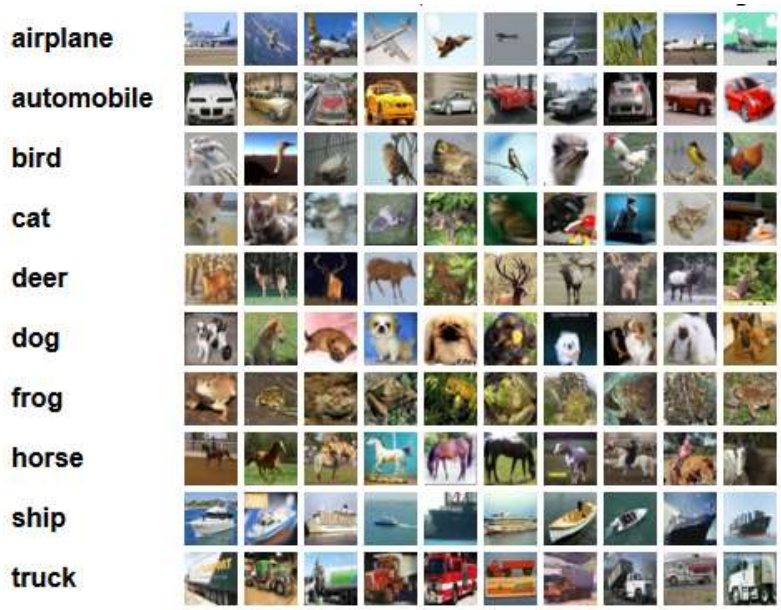


Figure 8: CIFAR-100 image sample.

In order to test the DR algorithms we used 5 categories of the CIFAR-100 database with 5 classes each. The selection of the categories and the respective classes were based on the reference book of 103 haptic signals from the Danish Association of the Deafblind (2012) and the intersection of the available objects and categories in the CIFAR-100 database. Thus, in the results section we present the DR mapping of images from five categories with 5 classes each. Table 3 presents the aforementioned categories and the corresponding classes/concepts used.

Table 3: Categories and classes used within SUITCEYES.

	Food containers	**Fruits and vegetables**	**Household electrical devices**	**Household furniture**	**People**
Concept 1	*Bottle*	*Apple*	*Clock*	*Bed*	*Baby*
Concept 2	*Bowl*	*Mushroom*	*Keyboard*	*Chair*	*Boy*
Concept 3	*Can*	*Orange*	*Lamp*	*Couch*	*Girl*
Concept 4	*Cup*	*Pear*	*Telephone*	*Table*	*Man*
Concept 5	*Plate*	*Sweet pepper*	*Television*	*Wardrobe*	*Woman*

3.3 Results

The three DR techniques described in section 3.1 were applied on images of all categories (see Table 3) and the mapping results are presented for each category separately. Moreover, the embeddings of different concepts are presented in the following subsection in all possible combinations of two, three, four and five concepts-classes, respectively. The dimensionality was reduced to two and three dimensions for all cases. In addition, for the 2D embeddings where two concepts are presented, a black line of separation between the two concepts is depicted along with the classification outcome (top of the figure, e.g. 0.78 means 78% accuracy of discrimination between the two concepts). For instance, the separation line could be used as a rule for the controller of the HIPI for the activation of different vibrators for different objects. The line is computed using a simple perceptron network.

3.3.1 Category I

3.3.1.1 Combinations of 2 concepts

In this subsection we present the results for the 2D and 3D embedding of 2 concepts from category I in all possible combinations and for all dimensionality reduction approaches. Figure 9 depicts the DR results of the PCA algorithm for all combinations of 2 objects for Category I.

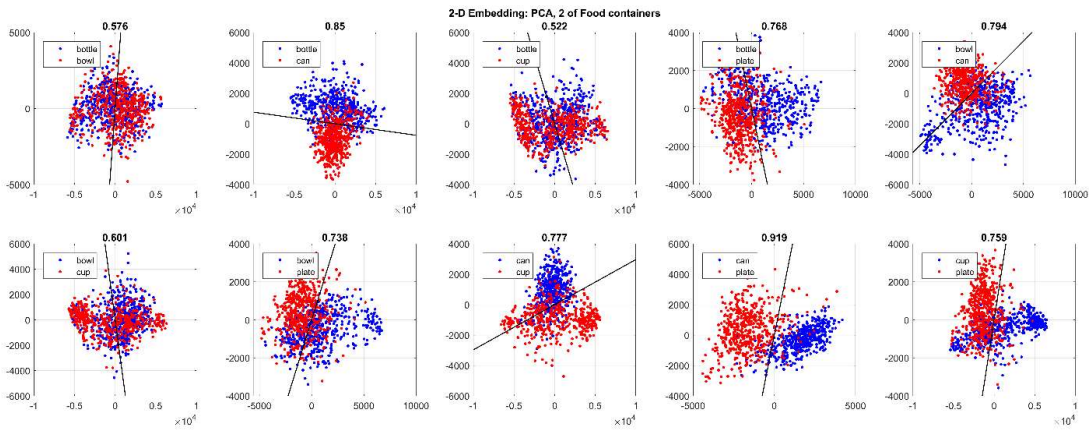


Figure 9: PCA 2D embedding.

Each subfigure illustrates also a black line of separation between the two concepts which is computed using a simple perceptron network. Moreover, the classification outcome is provided (top of the figure, e.g. 0.78 means 78% accuracy of discrimination between the two concepts). Different pairs of objects lead to different discrimination results. For instance, in the *Bottle vs. Bowl* example the two objects are discriminated with approximately 57% confidence. In essence, the discrimination is slightly better than a random classifier. Nevertheless, in the *Can vs. Plate* example the discrimination rate is over 90%. The separation line could be used as a rule for the controller of the HIPI for the activation of different vibrators for different objects.

Figure 10 depicts the 3D DR results for the same approach, i.e., PCA. The discrimination between the two concepts is not enhanced when using 3D PCA embeddings.

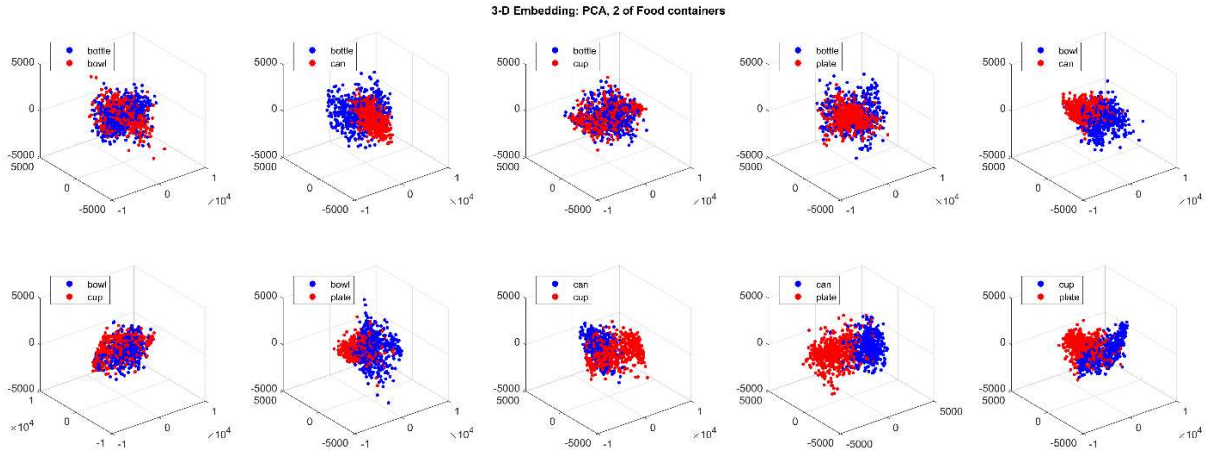


Figure 10: PCA 3D embedding.

Figure 11 and Figure 12 illustrate the 2D and 3D results for the Isomap methodology for Category I. It is noteworthy that similar discrimination results for 2D case are provided by PCA and Isomap. For instance, *Bottle vs. Bowl* and *Can vs. Plate* cases exhibit similar discrimination rates.

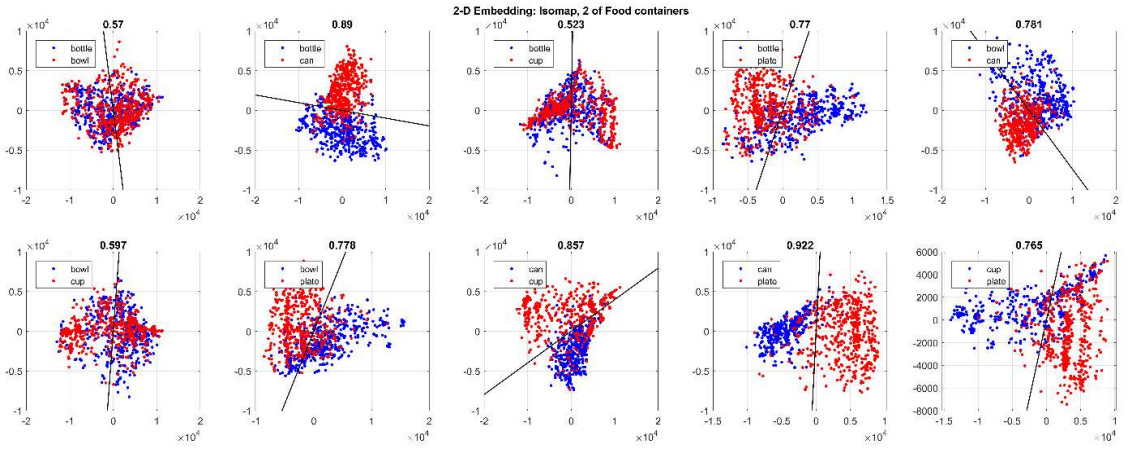


Figure 11: Isomap 2D embedding.

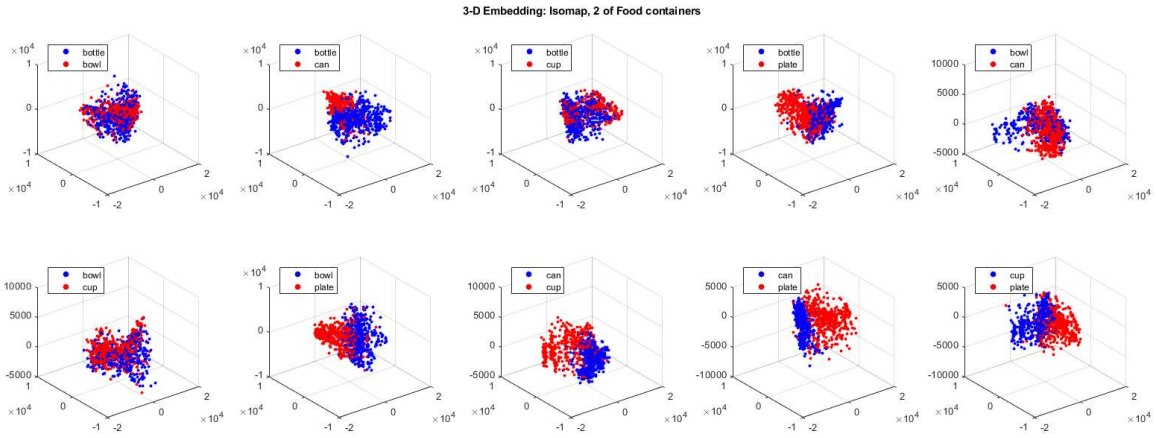


Figure 12: Isomap 3D embedding.

Finally Figure 13 and Figure 14 show the results using the t-SNE approach in 2D and 3D, respectively. The performance of this approach is overall similar to the two previous approaches in terms of discrimination approach in the 2D case.

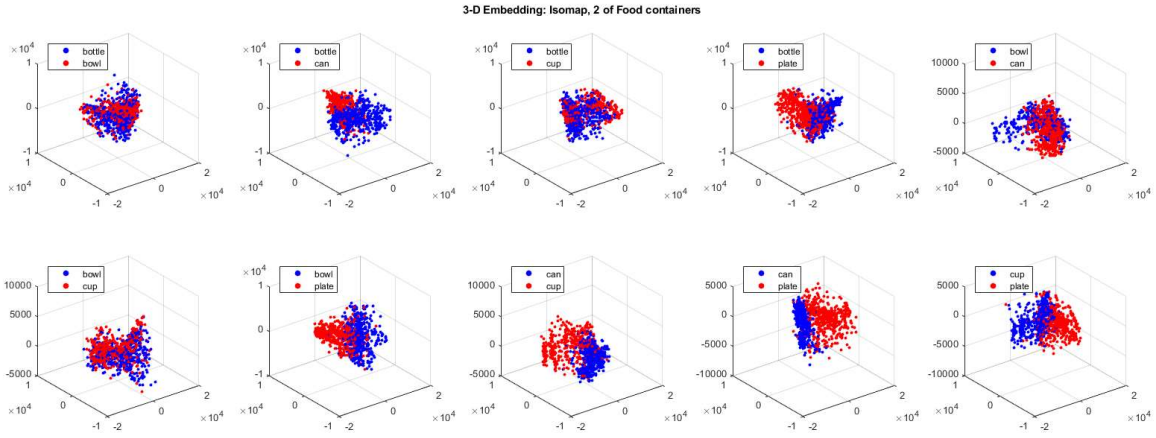


Figure 13: t-SNE 2D embedding.

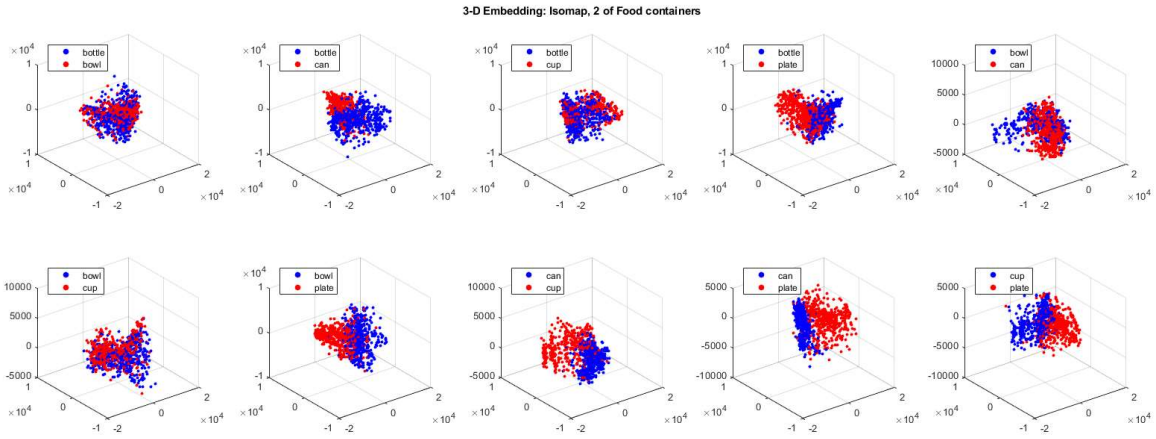


Figure 14: t-SNE 3D embedding.

3.3.1.2 Combinations of 3 concepts

In this subsection the 3-object/concept results of dimensionality reduction are presented using the three DR approaches, i.e., PCA, Isomap, and t-SNE. Figure 15 shows the results for the PCA approach in 2D space. Now the overlap of the concepts in the low dimensional embedding is more intense. Hence, discrimination of different objects in the haptic space, e.g., different vibration frequency for different objects, would be much more difficult for the 3-object case.

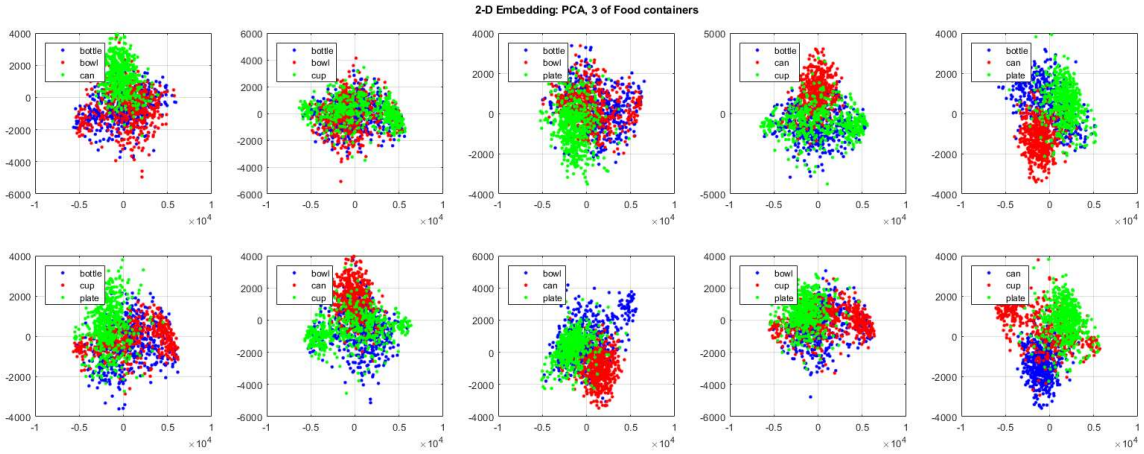


Figure 15: PCA 2D embedding.

Figure 16 shows the 3D case for the PCA approach. Overlap between different concepts is again extensive.

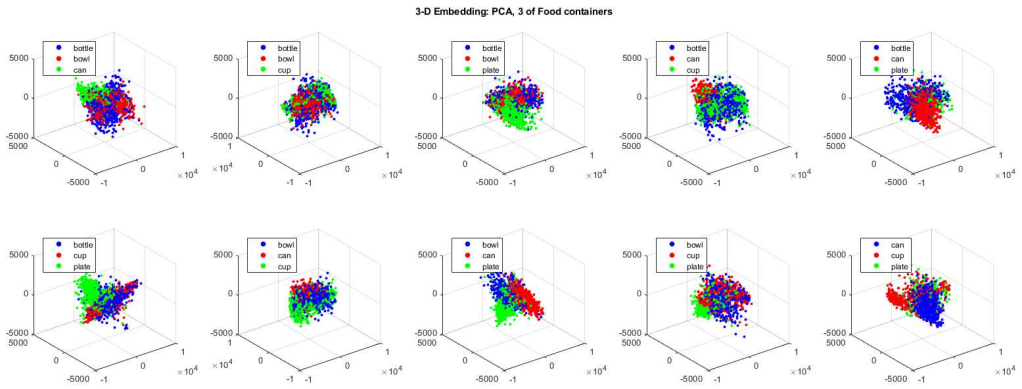


Figure 16: PCA 3D embedding.

The following figures show the results for Isomap (Figure 17 2D case, Figure 18 3D case) and t-SNE (Figure 19 2D case, Figure 20 3D case) approaches. Overall the overlap between concepts is extensive and it can be assumed that direct mapping to the haptic space (vibrations) would not lead to affective discrimination by people with deafblindness.

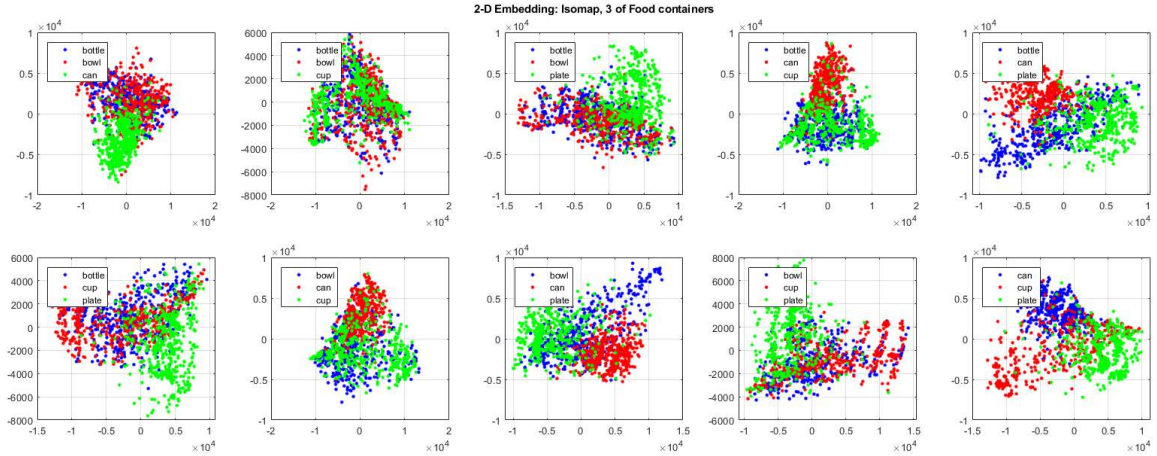


Figure 17: Isomap 2D embedding.

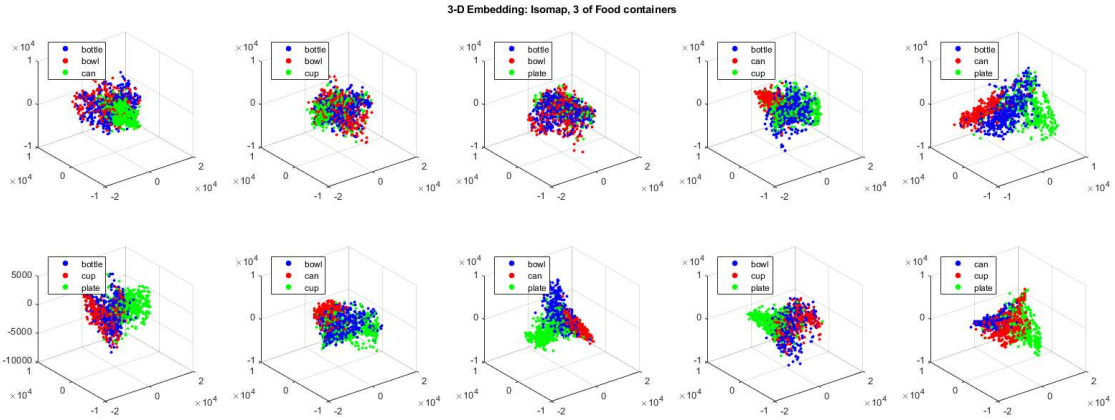


Figure 18: Isomap 3D embedding.

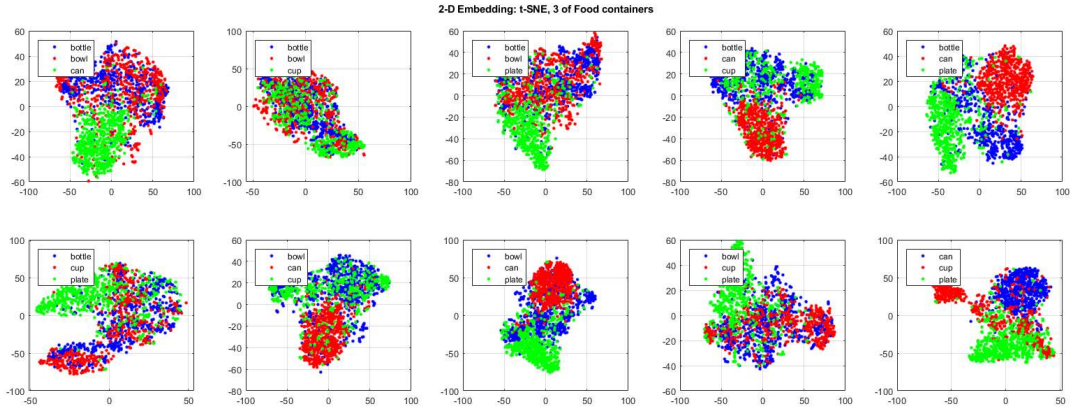


Figure 19: t-SNE 2D embedding.

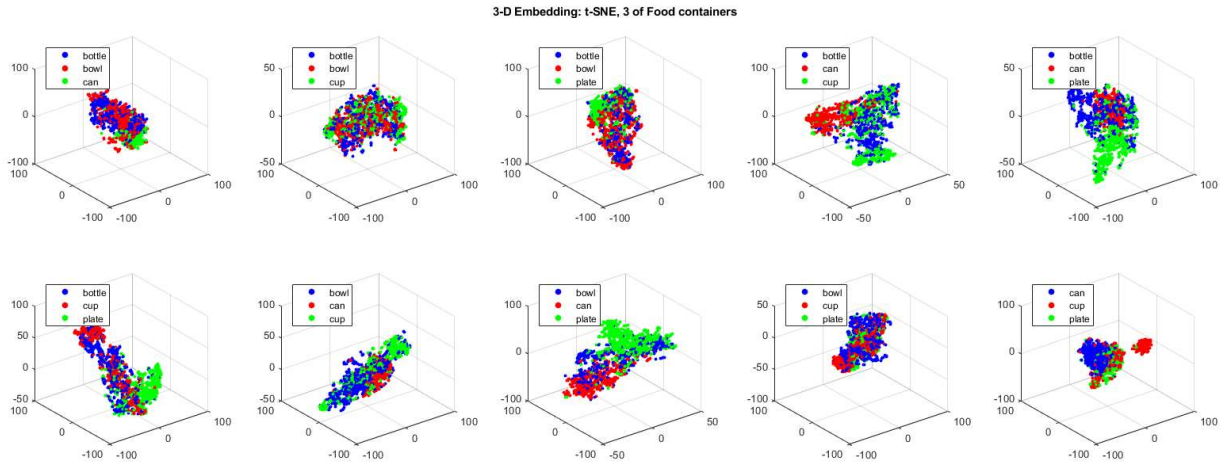


Figure 20: t-SNE 3D embedding.

3.3.1.3 Combinations of 4 concepts

In this subsection the results concerning the 4-object case are presented. It is obvious from the following figures (Figure 21 - 2D PCA, Figure 22 - 3D PCA, Figure 23 - 2D Isomap, Figure 24 - 3D Isomap, Figure 25 - 2D t-SNE, Figure 26 - 3D t-SNE) that the 4-object/concept overlap in the low-dimensional embedding is even more prominent than the 3-object/concept case. Thus, discrimination in the haptic space (vibrations etc.) will be even more difficult for people with deafblindness.

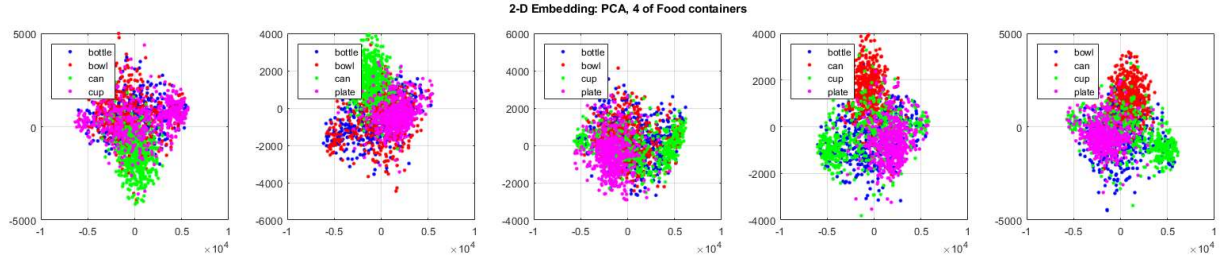


Figure 21: PCA 2D embedding.

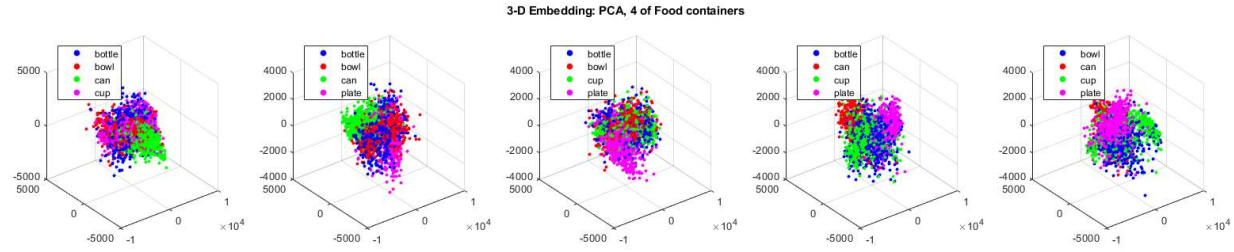


Figure 22: PCA 3D embedding.

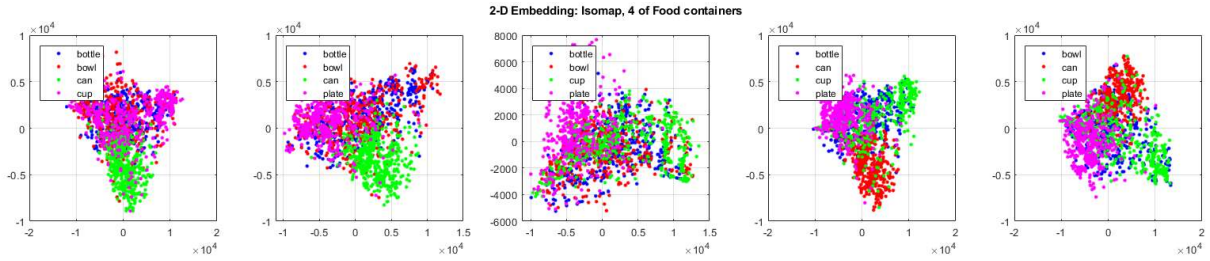


Figure 23: Isomap 2D embedding.

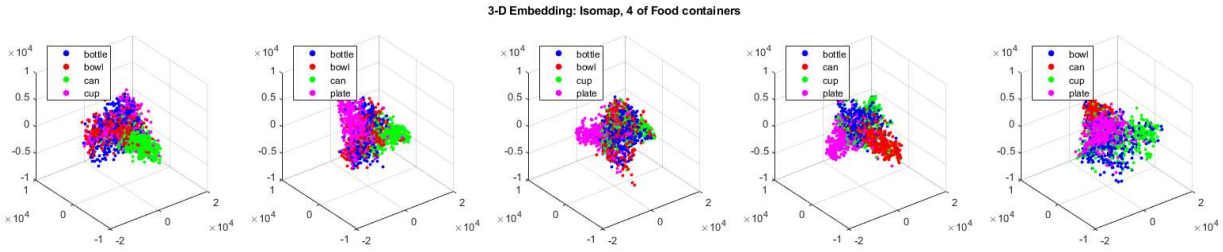


Figure 24: Isomap 3D embedding.

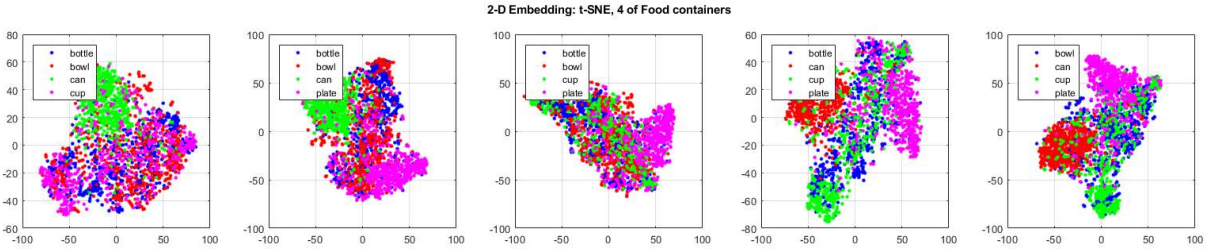


Figure 25: t-SNE 2D embedding.

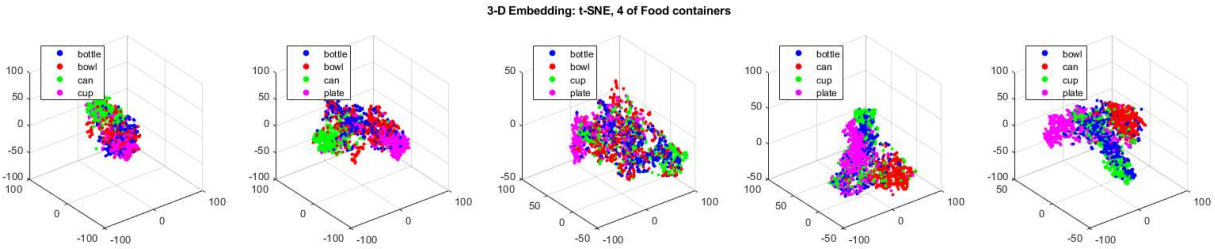


Figure 26: t-SNE 3D embedding.

3.3.1.4 Combinations of 5 concepts

Finally, in this subsection the 5-object case of low dimensional embedding is presented using PCA (Figure 27, 2D (left), 3D (right)). Once again the inter-concept/object overlap in the low dimensional space is extensive.

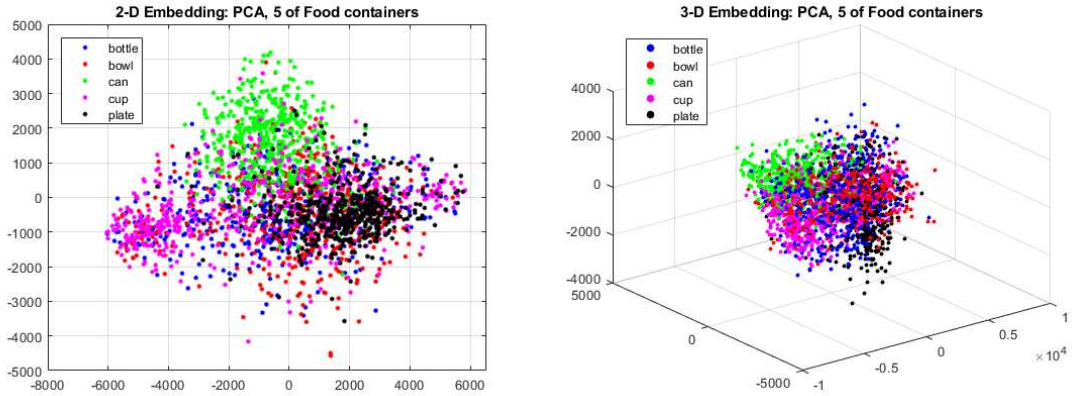


Figure 27: PCA 2D (left) 3D (right) embedding.

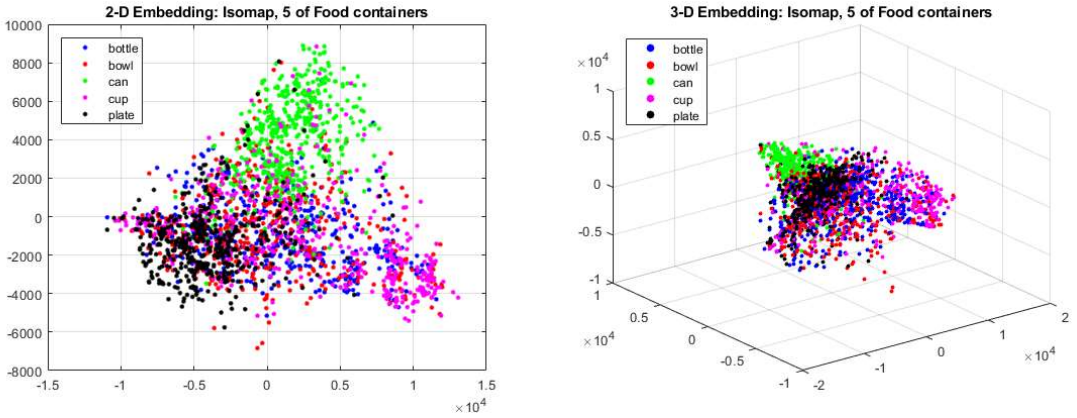


Figure 28: Isomap 2D (left) 3D (right) embedding.

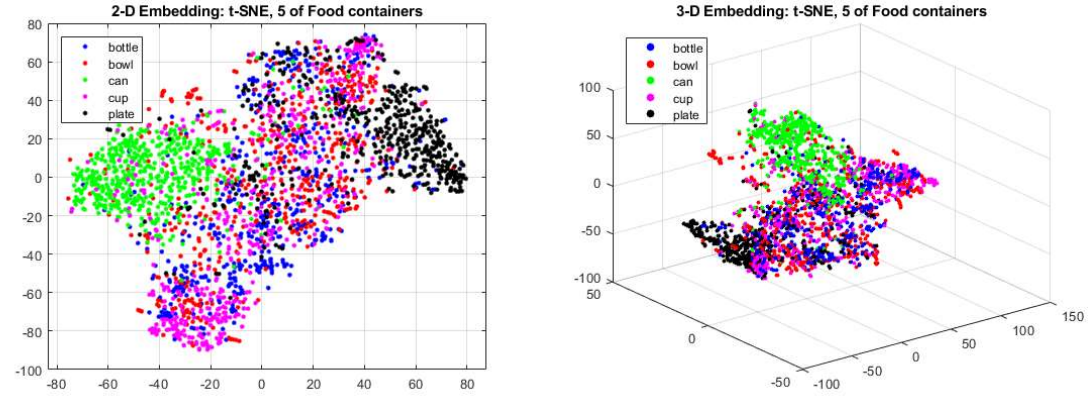


Figure 29: t-SNE 2D (left) 3D (right) embedding.

3.3.2 Other Categories and Overall Evaluation of the Results

The results presented above correspond only to the first category (see Table 3). For the four remaining categories of Table 3 the respective results for 2, 3, 4 and 5 objects/concepts and all DR approaches are presented in the Appendix. The results for the other categories are presented in a similar fashion with the respective presentation of category I. In conclusion, for all categories, the 2-object case exhibits cases, e.g. similar to the *Can* vs. *Plate* example, where the discrimination is plausible with simple linear methods (e.g. simple perceptron network) and the corresponding mapping

to the haptic space would probably lead to promising discrimination by the people with deafblindness. On the other hand the 3, 4, 5-object/concept cases exhibit severe overlap in the low dimensional space.

3.4 Discussion and Future Considerations

The results presented in the previous section correspond to various cases of categories and respective objects. After visual inspection it is obvious that the discrimination of three or more different objects is probably not suitable for transfer onto the haptic space (e.g. different intensities of vibrations, different vibrators etc.). In essence, most of the concepts are overlapped in the low dimensional embedding and thus potential mapping onto the haptic space (e.g. intensities of vibrations etc.) would probably not result in recognizable patterns.

In the case of two-concept embeddings (see e.g. subsection 3.3.5.1) the discrimination is better, e.g. with a linear separation, concepts such as girl-boy, or girl-man, or man-woman are separated using t-SNE with approximately 88% success. Moreover, it should be stressed out that the presented results correspond to a wide variety of objects per category. For instance, for the case of object 'telephone', different types of telephones (in essence the corresponding image) are mapped onto the 2D or 3D space. In case of different images of a specific telephone, that is, a telephone that a person with deafblindness would own in her house, the separation between, e.g. the objects 'telephone' and 'television', would be much more easier.

The haptic space, in terms of the number of vibrators, the modules that would comprise the interface, e.g. the intensity of vibration, the pattern of vibrators etc., will affect the way the different concepts are recognized by the user. Moreover, more sophisticated techniques, like those described in the previous chapter, could improve the discrimination of the different concepts, especially when embedding is realized in more than three dimensions. Nevertheless, the design of the haptic space, i.e. number of vibrations, patterns of vibrator setup etc., will substantially affect the future implementations of the respective algorithms.

4 Semantic Knowledge Representation and Reasoning

In order to facilitate the understanding, sharing and reuse of knowledge between different systems (or even among heterogeneous components within the same system), it is essential to define common vocabularies that represent shared knowledge in a formal way. **Ontologies** constitute the specification of a vocabulary for semantically representing a shared domain of discourse (Gruber, 1993). An ontology semantically models knowledge by defining a set of **classes** (objects, concepts, and other entities) existing in some domain of interest, and their **properties** (attributes, i.e. relationships that hold between them). The expressiveness of the ontology depends on the knowledge representation language used.

This chapter presents the first iteration of the SUITCEYES ontology. Starting with a brief introduction of the relevant background notions, along with an account of various relevant resources that can serve as the basis for our ontology, we then proceed to the core part of the chapter discussing the proposed semantic models in more detail.

4.1 Ontologies and the Semantic Web

The **Semantic Web** is "*a web of data that can be processed directly and indirectly by machines*" (Berners-Lee et al., 2001). It is an extension of the World Wide Web (WWW), in which web resources are augmented with semantics describing their intended meaning in a formal, machine-understandable way. The term was coined by Tim Berners-Lee, the inventor of WWW and director of the World Wide Web Consortium (W3C), which oversees the development of proposed Semantic Web standards. The standards proposed by W3C promote common data formats and exchange protocols on the Web. The Semantic Web is thus regarded as an integrator across different content, information applications, and systems.

Ontologies play a key role in the Semantic Web, providing the machine-interpretable semantic vocabulary and serving as the knowledge representation and exchange vehicle. The **Web Ontology Language** (**OWL**) has emerged as the official W3C recommendation for creating and sharing ontologies on the Web (Bechhofer, 2009).

4.2 The SUITCEYES Ontology

The key aim of the SUITCEYES ontology is **to semantically represent all notions that are pertinent to the project**, serving as the model for **semantically integrating information** coming from the various sensors and analysis components of the system. In this sense, we are primarily interested in semantically representing aspects relevant to the users' context, in order to provide them with enhanced situational awareness and potentially augment their navigation and communication capabilities. This ontology can also serve as the bridge between visually identified concepts and communicated content. Driven by this objective, the following subsections describe the process of designing and implementing the first iteration of the SUITCEYES ontology.

4.2.1 Specification of Ontology Requirements

A key step in designing the ontology is to come up with a set of **Competency Questions** (**CQs**), i.e. queries expressed in natural language that express a pattern for a type of question the ontology

should be able to answer (Grüninger & Fox, 1995). The answerability of CQs, hence, becomes a functional requirement of the ontology.

Towards designing the first iteration of the SUITCEYES ontology, we came up with a list of CQs that the ontology should be able to respond to (see Table 4). To some extent, these CQs are based on discussions with the project's Advisory Board, while (currently ongoing) investigations on end-users' requirements analysis within WP2 will help us revise the CQs and, consequently, the design of the next iteration of the SUITCEYES ontology.

Table 4: Competency Questions (CQs) that drove the design of the SUITCEYES ontology v1.

CQ#	Competency Question	Allowed values / Sample responses
CQ1	*What can the mounted sensors detect?*	*Objects, persons, activities and interconnections between them (e.g. two people playing a game).*
CQ2	*If [X] refers to a detected object/person/activity, which sensor has detected [X]?*	*E.g. the chest-mounted camera.*
CQ3	*If [X] refers to a detected object/person/activity, what is the location (lat/long) where [X] was detected?*	*Latitude/longitude coordinates.*
CQ4	*If [X] refers to an object/person/activity detected by sensor [Y], what is the confidence level of this detection?*	*Any number between 0% and 100%.*
CQ5	*How far from the user is detected object/person/activity [X] located?*	*Distance in meters.*
CQ6	*What object has been detected?*	*E.g. a wall, a door, a glass of water.*
CQ7	*What is the type of detected object [X]?*	*E.g. obstacle, vehicle, electronic device.*
CQ8	*What activity has been detected?*	*E.g. people talking to each other.*
CQ9	*What is the type of detected activity [X]?*	*E.g. communication, movement.*
CQ10	*Is the person detected known to the user or not?*	*Known/Unknown.*
CQ11	*If the detected person is known, who is it?*	*E.g. John Smith, the user's brother.*
CQ12	*If [X] refers to a detected activity, what people are involved in [X]?*	*E.g. two unknown persons.*
CQ13	*If [X] refers to a detected activity, what objects are involved in [X]?*	*E.g. a ball.*
CQ14	*What types of messages are conveyed to the user?*	*Alert/Warning/InfoMessage.*

4.2.2 Relevant Existing Resources

A common practice in ontology engineering involves the **reuse of existing third-party models**; this way we rely on previously used and validated ontologies for developing (parts of) our SUITCEYES ontology. However, since the project's domain is substantially wide, and the user requirements specification within WP2 is still ongoing, we give here a brief account of existing ontologies and models from a variety of domains, which can potentially be of use once the scope of the SUITCEYES ontology is more concrete.

Healthcare

There exist several ontologies, vocabularies and models for the healthcare domain, most of which typically focus on exhaustively modelling the domain of discourse. Two representative examples are SNOMED-CT and ICF. **SNOMED-CT** (Donnelly, 2006) is a systematically organized computer processable collection of medical terms providing codes, terms, synonyms and definitions used in clinical documentation and reporting. The primary purpose of SNOMED-CT is to encode the meanings that are used in health information and to support the effective clinical recording of data with the

aim of improving patient care. The model, featuring more than 340K concepts, includes aspects relevant to clinical findings, symptoms, diagnoses, procedures, body structures, organisms and other aetiologies, substances, pharmaceuticals, devices and specimens.

On the other hand, the **International Classification of Functioning, Disability and Health** (**ICF**) (WHO, 2001) is a classification of the health components of functioning and disability approved by the World Health Organization (WHO). The ICF contains around 1.6K concepts and is structured around two sets of notions: (a) body functions and structure, and (b) domains of activity and participation. Since an individual's functioning and disability occurs in a context, the ICF also includes a list of environmental factors.

Two additional examples are **ADOLENA** (Abilities and Disabilities OntoLogy for ENhancing Accessibility) and **ADOOLES** (Ability and Disability Ontology for Online LEarning and Services). The former model (Keet et al, 2008) is an experimental ontology encompassing abilities, disabilities, devices, and functionalities, and was created as a proof-of-concept for enhancing search capabilities by Ontology-Based Data Access (OBDA). ADOOLES, on the other hand, is an ontology based on ADOLENA for annotating learning resources and represents knowledge in the domains of e-learning and disabilities (Nganji et al., 2012).

Since within SUITCEYES our aim is not to semantically represent the users' disabilities or physical condition, the above ontologies are currently outside of our scope. A potential exception is the ontology created within the Dem@Care FP7 project[2], which ended in 2017, and, in which SUITCEYES partner CERTH participated. The **Dem@Care ontology**[3] is aimed at representing the experimentation protocol towards diagnostic support and assessment of dementia in a controlled environment. The aim of the protocol is to provide a brief overview of the health status of the participants during consultation (cognition, behaviours and function), and to correlate the system (sensor) data with the data collected using typical dementia care assessment tools. Although dementia is not relevant to SUITCEYES, the semantic representation of objects and activities in the Dem@Care ontology is highly relevant to our project and can be adopted and possibly extended, since it involves every-day activities and common objects used in an every-day context.

Internet of Things (IoT)

The predominant ontology for semantically representing IoT concepts is **SSN** (Semantic Sensor Network), along with its lightweight core module called **SOSA** (Sensor, Observation, Sampler, and Actuator)[4]. The SOSA/SSN ontologies can semantically describe sensors, actuators, samplers as well as their observations, actuation, and sampling activities, and currently constitute a W3C recommendation and an OGC implementation. The SOSA/SSN ontologies have been successfully deployed in a wide range of applications and use cases, including satellite imagery, large-scale scientific monitoring, industrial and household infrastructures, social sensing, citizen science, and observation-driven ontology engineering. Within SUITCEYES we are relying on SOSA/SSN for representing sensors and the respective observations.

Two additional relevant ontologies in the IoT domain are **IoT-Lite**, a lightweight ontology to represent IoT resources, entities and services[5] that is largely based on SSN, and **SmartHome**, which is also

[2] http://www.demcare.eu/
[3] http://www.demcare.eu/results/ontologies
[4] https://www.w3.org/TR/vocab-ssn/
[5] https://www.w3.org/Submission/2015/SUBM-iot-lite-20151126/

an SSN extension focusing on representing cover the spatial and temporal aspects of entities involved in smart home settings (Alirezaie et al., 2017). Interestingly, the SmartHome ontology also contains descriptions that provide a basis for the representation of the geometry of entities, as well as the topological and directional relations between any pair of geometrical objects; e.g. the kitchen is connected to the living room, living room is at the right side of the bedroom or the bed is close to the left wall etc. This latter set of representations may potentially prove useful for the SUITCEYES ontology as well, depending on the outcome from the user requirements analysis.

Navigation and Situational Awareness

There have been only a few attempts at producing ontologies for representing aspects relevant to navigation. A prominent ontology in this affair is **OI-space**, which was proposed by Liping & Worboys (2011), with the aim of providing a unified model of indoor and outdoor spaces and making the navigation between and within them seamless. We will consider relying on OI-space, depending on whether we will eventually integrate navigation aspects in our SUITCEYES prototype.

On the other hand, there exist several ontologies dealing with situational awareness, albeit with limited adoption by the respective communities. Such an example is **ESO**, the Event and implied Situation Ontology (Segers et al., 2016), which formalizes the pre-, during-, and post-situations of events and the roles of the entities affected by an event. ESO is a perfect match for predicting situations for SUITCEYES users, given a current status they are in at any given moment, along with the situations they were in during the (recent) past.

The situational awareness ontology within the **SAWA** (Situation Awareness Assistant – Matheus et al., 2005) framework is another related ontology, originating from the military domain. The ontology has embedded rules for deriving relations among objects, but a key drawback hindering its adoption from SUITCEYES is the lack of support for roles. The reason is due to the fact that, in SAWA, events are more powerful, since they also track the evolution of relations. An additional drawback is the ontology's rather fuzzy articulation, with some categories of concepts missing (e.g. non-physical objects), and the existence of some other concepts not being completely justified.

4.2.3 Ontology Conceptualization

This subsection describes the conceptualization of the ontology, in order to satisfy the ontology requirements represented above as CQs. Figure 30 displays an overview of the core ontology classes based on the Grafoo ontology visualization notation (Falco et al., 2014). The yellow rectangles indicate classes, while the green ones indicate data properties (i.e. properties that take a raw data value, like e.g. integers and strings).

As mentioned in the previous subsection, parts of the SUITCEYES ontology are based on existing third-party vocabularies. This is indicated in the figure by the prefixes in front of some of the class names, which indicate the namespace of the respective third-party ontologies. Therefore, prefix "**dem**" indicated the Dem@Care ontology, prefix "**sosa**" indicates the SSN/SOSA ontology, prefix "**xsd**" corresponds to the XML Schema Definition Language (XSD)[6], while "**foaf**" represents the Friend-Of-A-Friend (FOAF) specification[7]. Classes and properties that have no prefix belong to the core SUITCEYES ontology.

[6] https://www.w3.org/TR/xmlschema11-1/
[7] http://xmlns.com/foaf/spec/

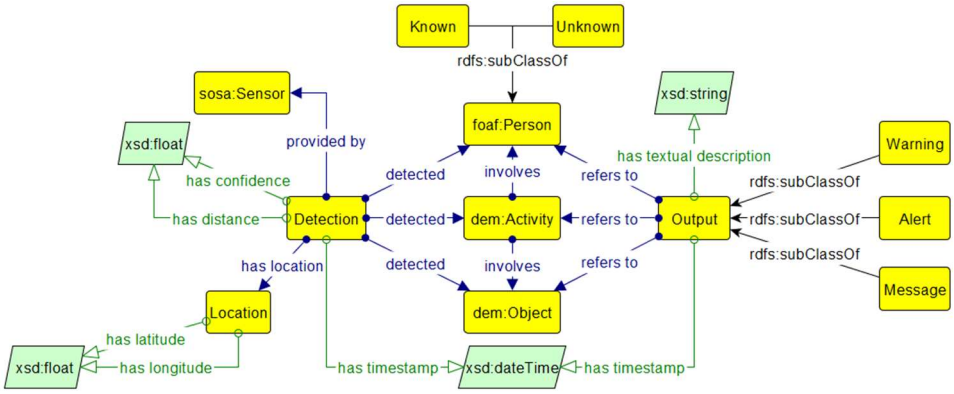


Figure 30: High-level overview of the core classes of the SUITCEYES ontology v1.

Figure 31 illustrates the hierarchies/specializations of objects and activities adopted from the Dem@Care ontology, which will be extended for the next iteration of the ontology, according to the outputs from the end-user requirements analysis taking place in WP2.

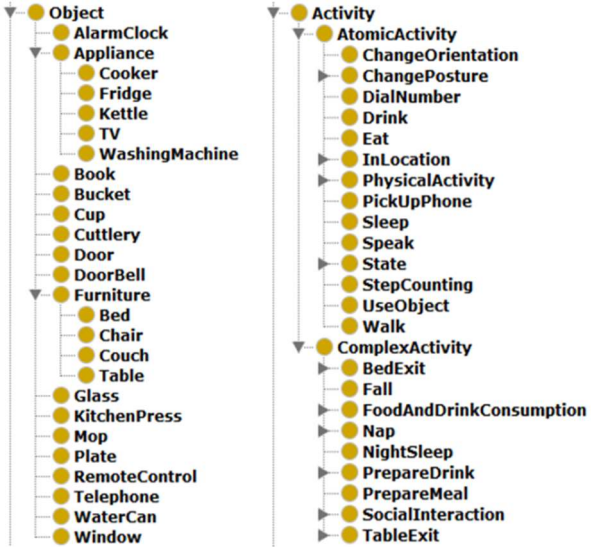


Figure 31: Hierarchies of objects and activities in the Dem@Care ontology, which are adopted in SUITCEYES.

Figure 32 displays a sample instantiation based on the above, where two people have been detected by the HIPI speaking to each other; one of them is "John", a person the user already knows. A respective output message is conveyed to the user with this information.

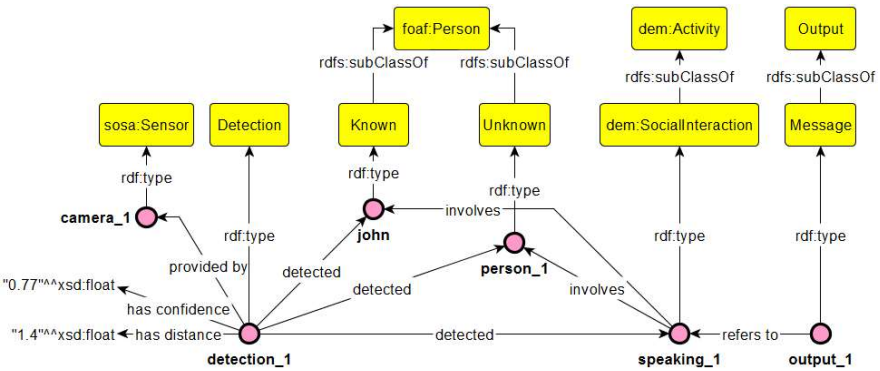


Figure 32: Sample instantiation of detecting two people speaking to each other.

4.2.4 Ontology Formalization and Implementation

The beAWARE ontology is expressed in **OWL 2** (W3C, 2012), a knowledge representation language widely used within the Semantic Web community for developing ontologies. Thus, we capitalize on its wide adoption as well as its formal structure and syntax, based on **Description Logics** (**DL**), a family of knowledge representation formalisms characterised by logically grounded semantics and well-defined reasoning services.

The main building blocks of DL are concepts representing sets of objects (e.g. **Person**), roles representing relationships between objects (e.g. **worksIn**), and individuals representing specific objects (e.g. **Alice**). Starting from atomic concepts, such as **Person**, arbitrary complex concepts can be described through a rich set of constructors that define the conditions on concept membership. For example, the concept ∃**hasFriend.Person** describes those objects that are related through the **hasFriend** role with an object from the concept **Person**; intuitively this corresponds to all those individuals that are friends with at least one person.

For developing and deploying the first iteration of the ontology we relied on the following tools:

- **Protégé-OWL v5.0** (Musen, 2015), a popular ontology development environment;
- **GraphDB**[8] for locally hosting test versions of the ontology;
- **SPARQL** (Harris & Seaborne, 2013) served as the semantic query language for submitting queries to the ontology and running rules on top of the model;
- **YASGUI**[9] for formalizing the SPARQL queries.

4.3 Semantic Reasoning

The term "**semantic reasoning**" refers to the process of deriving facts that are not explicitly expressed in an ontology. Consequently, a "**semantic reasoner**" (also often referred to as "reasoning engine", "rules engine" or simply "reasoner") is a piece of software able to infer logical consequences from a set of asserted facts or axioms in an ontology. A few examples of tasks required from a semantic reasoner are as follows (Obitko, 2007):

- **Satisfiability of a concept**, i.e. to determine whether a description of the concept is not contradictory, namely, whether an individual can exist that would be an instance of the concept;
- **Subsumption of concepts**, i.e. to determine whether concept *C* subsumes concept *D*, namely, whether description of *C* is more general than the description of *D*;
- **Check an individual**, i.e. to check whether the individual is an instance of a concept;
- **Retrieval of individuals**, i.e. to find all individuals that are instances of a concept;
- **Realization of an individual**, i.e. to find all concepts which the individual belongs to, especially the most specific ones.

Within the project, the SUITCEYES ontology will accept input from other modules analysing sensor outputs and will perform semantic reasoning via **SPARQL-based rules**. Rule-based reasoning satisfies the above points, plus additional aspects, like e.g. finding all concepts that satisfy a defined rule, or creating new instances of all concepts that satisfy a defined rule. A set of indicative **semantic reasoning scenarios** that the ontology will address are outlined below – a more complete list of semantic reasoning scenarios will be determined once the user requirements analysis is concluded.

[8] https://ontotext.com/products/graphdb/

[9] http://yasgui.org/

- Determine position and proximity of objects of interest: e.g. *where is my smartphone?*
- Determine position and proximity of locations of interest: e.g. *where is the exit?*
- Determine position and proximity of persons of interest: e.g. *where is my companion?*
- Infer potential risks in user's current situation: e.g. stairs ahead, vehicle approaching, etc.
- Infer types of activities performed by people in the vicinity of the user: e.g. two people in front of each other means that they are probably discussing.
- Determine set of suggested actions in order to achieve something: e.g. get out of the room or issue a ticket on the bus.

Regarding the implementation of the semantic reasoning module, the following parameters are foreseen:

Input: The semantic reasoning does not require any specific input, other than the triggering of the reasoning execution.

Output: The output from semantic reasoning is an ontology file including both the initially asserted and the inferred information.

Execution intervals: Every several minutes or on demand (e.g. whenever new knowledge is inserted into the ontology).

Involved technologies: The relevant RDF Service module will be based on Python 2.x or 3.x, SPARQL, SPARQLWrapper, RDFLib 4.x.x; a REST API will be deployed with a configured public IP/port or domain name.

Critical factors: A valid rule set for the reasoning process (see reasoning scenarios above) is essential for the inference of meaningful knowledge. Hence, it is critical for all involved domain experts (e.g. end user partners) to contribute actively to the task of assembling this rule set.

4.4 Chapter Summary and Future Work

This chapter presented the first iteration of the SUITCEYES ontology. We formulated the competency questions underlying the ontology design, noting that these will be further refined once the ongoing user requirements analysis (WP2) is concluded. We then presented the existing third-party resources we used for implementing parts of the first iteration of the ontology. Finally, we presented the conceptualization of the core aspects of the ontology.

The following directions for improvements are foreseen for the next iteration of the ontology:

- **Extension of the ontology**, in order to cover more extensively aspects that are not adequately covered yet, like e.g. navigational aspects, emotions, etc.
- **Implementation of ontology population techniques** for semantically enriching the ontology with information from external sources, like e.g. DBpedia and WikiData. This thread will be based on previous work of ours (Kontopoulos et al., 2017; Mitzias et al., 2016).
- **Mapping ontology constructs** to other existing models. This has not been implemented yet, since the ontology is still evolving, but such mappings will be integrated in the final iteration of the ontology, in order to establish semantic interoperability with other third-party solutions. This process typically involves the definition of ontology mappings in a separate ontology document that contains the mappings between the SUITCEYES ontology concepts and those of third-party vocabularies. This document facilitates the direct alignment and easy comprehension of terms, data and relations from multiple domains.

- **Extension of the semantic reasoning ruleset** with new and more elaborate rules for providing more meaningful information to the user.

5 Conclusions

This deliverable presents the first version of the tools within WP3 for capturing, translating and semantically representing environmental cues. More specifically, we presented the first versions of the tools for visual analysis (chapter 2), dimensionality reduction (chapter 3) and semantic knowledge representation and reasoning (chapter 4). As already stated, the above will be refined once the end-user requirements analysis process (WP2) is concluded. An additional point to consider in our future work revolves around possible hardware limitations of the implementation platform, which may limit the capacity of our proposed algorithms. These issues will be further explored in the coming months.

With dimensionality reduction playing a key role both for visual analysis, and semantic knowledge representation and reasoning, we have postulated that concept embeddings, constituting geometric or topological vocabularies to encode visual vs. verbal percepts, are a suitable approach in globally ongoing current research in several fields to be tested for SUITCEYES' purposes. This implies that in parallel with collected user needs from WP2, we can depart from visually recognized concepts represented as embeddings, and use them as a test vocabulary for simple interaction between two users with deafblindness (i.e. communication). Such a vocabulary will be scalable and can be tailored to individual needs while offering a robust overlap between perceived and communicated semantic content by haptic means.

References

Ahonen, T., Hadid, A., & Pietikainen, M. (2006). Face description with local binary patterns: Application to face recognition. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2037-2041.

Alirezaie, M., Renoux, J., Köckemann, U., Kristoffersson, A., Karlsson, L., Blomqvist, E., Tsiftes, N., Voigt, T., & Loutfi, A. (2017). An ontology-based context-aware system for smart homes: E-care@home. *Sensors*, *17*(7), 1586.

Andriluka, M., Pishchulin, L., Gehler, P., & Schiele, B. (2014). 2d human pose estimation: New benchmark and state of the art analysis. *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, (pp. 3686-3693).

Avgerinakis, K., & Kompatsiaris, Y. (2016). Demcare action dataset for evaluating dementia patients in a home-based environment. *Impact: The Journal of Innovation Impact*, 6, 83.

Bambach, S., Crandall, D. J., & Yu, C. (2015a). Viewpoint integration for hand-based recognition of social interactions from a first-person view. *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, (pp. 351-354).

Bambach, S., Lee, S., Crandall, D. J., & Yu, C. (2015b). Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. *Computer Vision (ICCV), 2015 IEEE International Conference on*, (pp. 1949-1957).

Bechhofer, S. (2009). OWL: Web ontology language. *Encyclopaedia of Database Systems*, pp. 2008-2009. Springer US.

Belhumeur, P. N., Jacobs, D. W., Kriegman, D. J., & Kumar, N. (2013). Localizing parts of faces using a consensus of exemplars. *IEEE transactions on pattern analysis and machine intelligence*, 35, 2930-2940.

Belongie, S., Malik, J., & Puzicha, J. (2002). *Shape matching and object recognition using shape contexts*. Tech. rep., CALIFORNIA UNIV SAN DIEGO LA JOLLA DEPT OF COMPUTER SCIENCE AND ENGINEERING.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. *Scientific American*, *284*(5), 34-43.

Birdal, A., & Hassanpour, R. (2008). *Region based hand gesture recognition*.

Bolanos, M., & Radeva, P. (2015). Ego-object discovery. *arXiv preprint* arXiv:1504.01639.

Bretzner, L., Laptev, I., & Lindeberg, T. (2002). Hand gesture recognition using multi-scale colour features, hierarchical models and particle filtering. *Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on*, (pp. 423-428).

Cao, Z., Simon, T., Wei, S.-E., & Sheikh, Y. (2017). Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *CVPR*.

Chai, D., & Ngan, K. N. (1998). Locating facial region of a head-and-shoulders color image. fg, (p. 124).

Cootes, T. F., & Taylor, C. J. (1992). Active shape models—'smart snakes'. In *BMVC92* (pp. 266-275). Springer.

Corradini, A. (2001). Dynamic time warping for off-line recognition of a small gesture vocabulary. *Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems, 2001. Proceedings. IEEE ICCV Workshop on*, (pp. 82-89).

Crispim-Junior, C. F., Buso, V., Avgerinakis, K., Meditskos, G., Briassouli, A., Benois-Pineau, J., . . . Bremond, F. (2016). Semantic event fusion of different visual modality concepts for activity recognition. *IEEE transactions on pattern analysis and machine intelligence*, 38, 1598-1611.

Crispim-Junior, C. F., Gómez Uría, A., Strumia, C., Koperski, M., König, A., Negin, F., . . . others. (2017). Online recognition of daily activities by color-depth sensing and knowledge models. *Sensors*, 17, 1528.

Crowley, J., Berard, F., Coutaz, J., & others. (1995). Finger tracking as an input device for augmented reality. *International Workshop on Gesture and Face Recognition*, 1415, p. 195.

Dai, J., Li, Y., He, K., & Sun, J. (2016). R-fcn: Object detection via region-based fully convolutional networks. *Advances in neural information processing systems*, (pp. 379-387).

Dalal, N., Triggs, B., & Schmid, C. (2006). Human detection using oriented histograms of flow and appearance. *European conference on computer vision*, (pp. 428-441).

Damen, D., Doughty, H., Farinella, G. M., Fidler, S., Furnari, A., Kazakos, E., . . . others. (2018). Scaling Egocentric Vision: The EPIC-KITCHENS Dataset. *arXiv preprint* arXiv:1804.02748.

Damen, D., Leelasawassuk, T., & Mayol-Cuevas, W. (2016). You-Do, I-Learn: Egocentric unsupervised discovery of objects and their modes of interaction towards video-based guidance. *Computer Vision and Image Understanding*, 149, 98-112.

Danish Association of the Deafblind. (2012). 103 Haptic Signals – a Reference Book.

Darrell, T. J., Essa, I. A., & Pentland, A. P. (1996). Task-specific gesture analysis in real-time using interpolated views. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18, 1236-1242.

Donnelly, K. (2006). SNOMED-CT: The advanced terminology and coding system for eHealth. *Studies in health technology and informatics*, *121*, 279.

Escalante, H. J., Guyon, I., Athitsos, V., Jangyodsuk, P., & Wan, J. (2017). Principal motion components for one-shot gesture recognition. *Pattern Analysis and Applications*, 20, 167-182.

Espinace, P., Kollar, T., Soto, A., & Roy, N. (2010). Indoor scene recognition through object detection. *Robotics and Automation (ICRA), 2010 IEEE International Conference on*, (pp. 1406-1413).

Falco, R., Gangemi, A., Peroni, S., Shotton, D., & Vitali, F. (2014, May). Modelling OWL ontologies with Graffoo. In *European Semantic Web Conference* (pp. 320-325). Springer, Cham.

Fathi, A., Farhadi, A., & Rehg, J. M. (2011). Understanding egocentric activities. *Computer Vision (ICCV), 2011 IEEE International Conference on*, (pp. 407-414).

Girshick, R. (2015). Fast r-cnn. *Proceedings of the IEEE international conference on computer vision*, (pp. 1440-1448).

Gruber, T. (1993). A translational approach to portable ontologies. *Knowledge Acquisition, 5*(2), 199-229.

Grüninger, M. and Fox, M.S. (1995). Methodology for the design and evaluation of ontologies. In Skuce D (ed) *IJCAI95 Workshop on Basic Ontological Issues in Knowledge Sharing*, pp 6.1–6.10.

Harris, S., Seaborne, A. and Prud'hommeaux, E. (2013), *SPARQL 1.1 query language, W3C recommendation*, 21(10).

He, K., Zhang, X., Ren, S., & Sun, J. (2016, 6). Deep Residual Learning for Image Recognition. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Henriques, J. F., Caseiro, R., Martins, P., & Batista, J. (2015). High-speed tracking with kernelized correlation filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37, 583-596.

Hu, P., & Ramanan, D. (2017). Finding tiny faces. *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, (pp. 1522-1530).

Huang, J., Rathod, V., Sun, C., Zhu, M., Korattikara, A., Fathi, A., . . . others. (2017). Speed/accuracy trade-offs for modern convolutional object detectors. *IEEE CVPR*.

Jiang, H., & Learned-Miller, E. (2017). Face detection with the faster R-CNN. *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*, (pp. 650-657).

Keet, C.M., Alberts, R., Gerber, A., & Chimamiwa, G. (2008). Enhancing web portals with Ontology-based data access: the case study of South Africa's accessibility portal for people with disabilities. *Fifth International Workshop OWL: Experiences and Directions (OWLED'08)*, Karlsruhe, 26-27 October.

Kong, T., Sun, F., Yao, A., Liu, H., Lu, M., & Chen, Y. (2017). Ron: Reverse connection with objectness prior networks for object detection. *IEEE Conference on Computer Vision and Pattern Recognition*, 1, p. 2.

Kontopoulos, E., Mitzias, P., Riga, M., Kompatsiaris, I. (2017). A Domain-Agnostic Tool for Scalable Ontology Population and Enrichment from Diverse Linked Data Sources. In: Kalinichenko, L.A., Manolopoulos, Y., Skvortsov, N.A., and Sukhomlin, V.A. (eds.) *Selected Papers of the XIX International Conference on Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL 2017)*. pp. 184–190. CEUR Workshop Proceedings Vol 2022, Moscow, Russia.

Kroeger, T., Timofte, R., Dai, D., & Van Gool, L. (2016). Fast optical flow using dense inverse search. *European Conference on Computer Vision*, (pp. 471-488).

Kumar, K. P., & Bhavani, R. (2017). Egocentric Activity Recognition using Histogram Oriented Features and Textural Features. *International Journal of Scientific Research in Computer Science*.

Kumar, K. P., & Bhavani, R. (2017). Human activity recognition in egocentric video using PNN, SVM, kNN and SVM+ kNN classifiers. *Cluster Computing*, 1-10.

Li, J., Wang, T., & Zhang, Y. (2011). Face detection using surf cascade. *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, (pp. 2183-2190).

Lin, T.-Y., Dollár, P., Girshick, R. B., He, K., Hariharan, B., & Belongie, S. J. (2017). Feature Pyramid Networks for Object Detection. *CVPR*, 1, p. 4.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., & Berg, A. C. (2016). Ssd: Single shot multibox detector. *European conference on computer vision*, (pp. 21-37).

Matheus, C., Kokar, M., Baclawski, K., Letkowski, J., Call, C., Hinman, M., Salerno, J., Boulware, D. (2005). SAWA: An assistant for higher-level fusion and situation awareness, *Proc. of SPIE Conference on Multisensor, Multisource Information Fusion: Architectures, Algorithms, and Applications*, Orlando, Florida, USA, 2005, 75-85.

McCandless, T., & Grauman, K. (2013). Object-Centric Spatio-Temporal Pyramids for Egocentric Activity Recognition. *BMVC*, 2, p. 3.

Meditskos, G., Plans, P.-M., Stavropoulos, T. G., Benois-Pineau, J., Buso, V., & Kompatsiaris, I. (2018). Multimodal activity recognition from egocentric vision, semantic enrichment and lifelogging applications for the care of dementia. *Journal of Visual Communication and Image Representation*, 51, 169-190.

Mitra, S., & Acharya, T. (2007). Gesture recognition: A survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 37, 311-324.

Mitzias, P., Riga, M., Kontopoulos, E., Stavropoulos, T.G., Andreadis, S., Meditskos, G., Kompatsiaris, I. (2016). User-Driven Ontology Population from Linked Data Sources. In: *7th Int. Conf. on Knowledge Engineering and the Semantic Web (KESW 2016)*. pp. 31–41. Springer International Publishing, Prague, Czech Republic.

Molchanov, P., Gupta, S., Kim, K., & Kautz, J. (2015). Hand gesture recognition with 3D convolutional neural networks. *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, (pp. 1-7).

Musen, M.A. (2015). The Protégé project: A look back and a look forward. *AI Matters. Association of Computing Machinery Specific Interest Group in Artificial Intelligence*, 1(4), June. DOI: 10.1145/2557001.25757003.

Najibi, M., Samangouei, P., Chellappa, R., & Davis, L. S. (2017). SSH: Single Stage Headless Face Detector. *ICCV*, (pp. 4885-4894).

Nganji, J.T., Brayshaw, M., & Tompsett, B. (2012). Ontology–driven disability-aware e-learning personalisation with ontodaps. *Campus-Wide Information Systems 30*(1), 17–34.

Obitko, M. (2007). *Translations between Ontologies in Multi-Agent Systems*, Ph.D. dissertation, Faculty of Electrical Engineering, Czech Technical University in Prague.

Pandey, M., & Lazebnik, S. (2011). Scene recognition and weakly supervised object localization with deformable part-based models.

Parizi, S. N., Oberlin, J. G., & Felzenszwalb, P. F. (2012). Reconfigurable models for scene recognition. *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, (pp. 2775-2782).

Pigou, L., Van Den Oord, A., Dieleman, S., Van Herreweghe, M., & Dambre, J. (2018). Beyond temporal pooling: Recurrence and temporal convolutions for gesture recognition in video. *International Journal of Computer Vision*, 126, 430-439.

Pirsiavash, H., & Ramanan, D. (2012). Detecting activities of daily living in first-person camera views. *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, (pp. 2847-2854).

Poveda-Villalón, M., Gómez-Pérez, A., & Suárez-Figueroa, M. C. (2014). Oops!(ontology pitfall scanner!): An online tool for ontology evaluation. *International Journal on Semantic Web and Information Systems (IJSWIS)*, *10*(2), 7-34.

Quattoni, A., & Torralba, A. (2009). Recognizing indoor scenes. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, (pp. 413-420).

Quattoni, A., Collins, M., & Darrell, T. (2008). Transfer learning for image classification with sparse prototype representations. *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, (pp. 1-8).

Ranjan, R., Patel, V. M., & Chellappa, R. (2017). Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Rautaray, S. S., & Agrawal, A. (2015). Vision based hand gesture recognition for human computer interaction: a survey. *Artificial Intelligence Review*, 43, 1-54.

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. *Proceedings of the IEEE conference on computer vision and pattern recognition*, (pp. 779-788).

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, (pp. 91-99).

Ren, S., He, K., Girshick, R., Zhang, X., & Sun, J. (2017). Object detection networks on convolutional feature maps. *IEEE transactions on pattern analysis and machine intelligence*, 39, 1476-1481.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., . . . others. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115, 211-252.

Saxe, D., & Foulds, R. (1996). Toward robust skin identification in video images. fg, (p. 379).

Segers R., M. Rospocher, P. Vossen, E. Laparra, G. Rigau, A. Minard. *The Event and Implied Situation Ontology: Application and Evaluation.* In: Proceedings of the 10th edition of the Language Resources and Evaluation Conference (LREC2016), Portoroz, Slovenia, May 23-28 2016.

Shu, C., Ding, X., & Fang, C. (2011). Histogram of the oriented gradient for face recognition. *Tsinghua Science and Technology*, 16, 216-224.

Sigal, L., Sclaroff, S., & Athitsos, V. (2004). Skin color-based video segmentation under time-varying illumination. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26, 862-877.

Silberman, N., Hoiem, D., Kohli, P., & Fergus, R. (2012). Indoor segmentation and support inference from rgbd images. *European Conference on Computer Vision*, (pp. 746-760).

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint* arXiv:1409.1556.

Starner, T. E. (1995). *Visual Recognition of American Sign Language Using Hidden Markov Models*. Tech. rep., Massachusetts Inst Of Tech Cambridge Dept Of Brain And Cognitive Sciences.

Sun, X., Wu, P., & Hoi, S. C. (2018). Face detection using deep learning: An improved faster RCNN approach. *Neurocomputing*, 299, 42-50.

Sun, Y., Wang, X., & Tang, X. (2013). Deep convolutional network cascade for facial point detection. *Proceedings of the IEEE conference on computer vision and pattern recognition*, (pp. 3476-3483).

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., . . . Rabinovich, A. (2015, 6). Going Deeper With Convolutions. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tenenbaum, J. B., de Silva, V., and Langford, J. C. (2000). A Global Geometric Framework for Nonlinear Dimensionality Reduction. Science, vol. 290, no. 5500, pp. 2319–23, 2000.

Uijlings, J. R., Van De Sande, K. E., Gevers, T., & Smeulders, A. W. (2013). Selective search for object recognition. *International journal of computer vision*, 104, 154-171.

Vaca-Castano, G., Das, S., Sousa, J. P., Lobo, N. D., & Shah, M. (2017). Improved scene identification and object detection on egocentric vision of daily activities. *Computer Vision and Image Understanding*, 156, 92-103.

Van De Sande, K., Gevers, T., & Snoek, C. (2010). Evaluating color descriptors for object and scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 32, 1582-1596.

Van Der Maaten, L. J. P., and Hinton, G. E. Visualizing high-dimensional data using t-sne. J. Mach. Learn. Res., vol. 9, pp. 2579–2605, 2008.

Van Der Maaten, L. J. P., Postma E. O., and H. J. Van Den Herik. (2009). Dimensionality Reduction: A Comparative Review. J. Mach. Learn. Res., vol. 10, pp. 1–41, 2009.

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, 1, pp. I--I.

W3C. (2012). *OWL 2 Web Ontology Language Document Overview (Second Edition)*. W3C Recommendation 11 December 2012, available online: http://www.w3.org/TR/owl2-overview/

Wang, H., Kläser, A., Schmid, C., & Liu, C.-L. (2011). Action recognition by dense trajectories. *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, (pp. 3169-3176).

Wang, X., Gao, L., Song, J., Zhen, X., Sebe, N., & Shen, H. T. (2018). Deep appearance and motion learning for egocentric activity recognition. *Neurocomputing*, 275, 438-447.

Wang, X., Shrivastava, A., & Gupta, A. (2017). A-fast-rcnn: Hard positive generation via adversary for object detection. *IEEE Conference on Computer Vision and Pattern Recognition*.

Wolf, L., Hassner, T., & Maoz, I. (2011). Face recognition in unconstrained videos with matched background similarity. *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, (pp. 529-534).

World Health Organization. (2001). *International classification of functioning, disability and health: ICF*. Geneva: World Health Organization.

Yan, Y., Ricci, E., Liu, G., & Sebe, N. (2015). Egocentric daily activity recognition via multitask clustering. *IEEE Transactions on Image Processing*, 24, 2984-2995.

Yang, J., Lu, W., & Waibel, A. (1998). Skin-color modeling and adaptation. *Asian Conference on Computer Vision*, (pp. 687-694).

Yang, L., & Worboys, M. (2011). A navigation ontology for outdoor-indoor space: (work-in-progress). *In Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Indoor Spatial Awareness (ISA '11)*. ACM, New York, NY, USA, 31-34.

Yang, S., Luo, P., Loy, C. C., & Tang, X. (2016). WIDER FACE: A Face Detection Benchmark. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yang, S., Luo, P., Loy, C.-C., & Tang, X. (2015). From facial parts responses to face detection: A deep learning approach. *Proceedings of the IEEE International Conference on Computer Vision*, (pp. 3676-3684).

Zhang, K., Zhang, Z., Li, Z., & Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23, 1499-1503.

Zhou, B., Khosla, A., Lapedriza., A., Oliva, A., & Torralba, A. (2016). Learning Deep Features for Discriminative Localization. *CVPR*.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. (2017). Places: A 10 million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

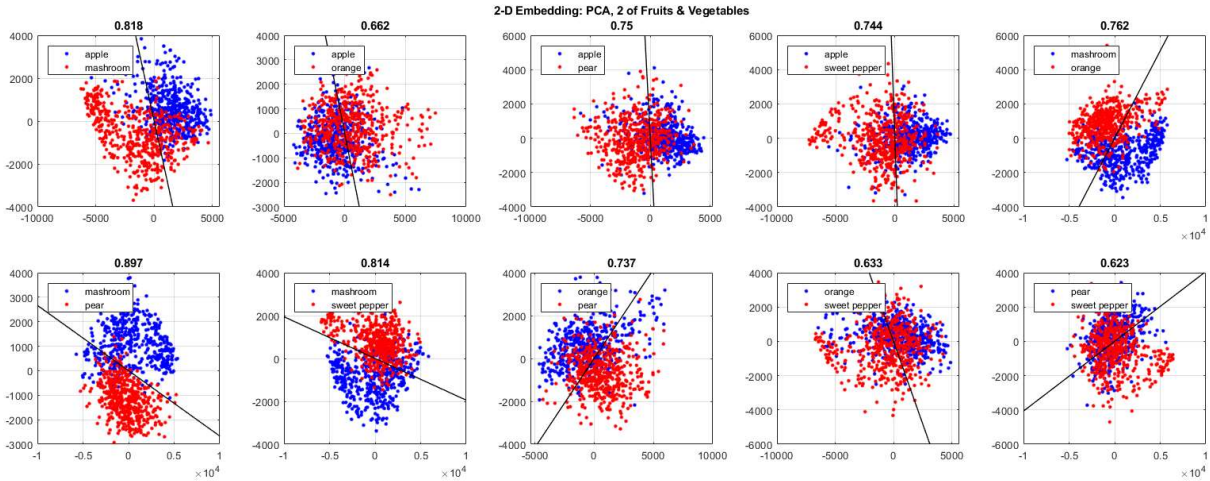
Zhou, Y., Ni, B., Hong, R., Yang, X., & Tian, Q. (2016). Cascaded interactional targeting network for egocentric video analysis. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (pp. 1904-1913).

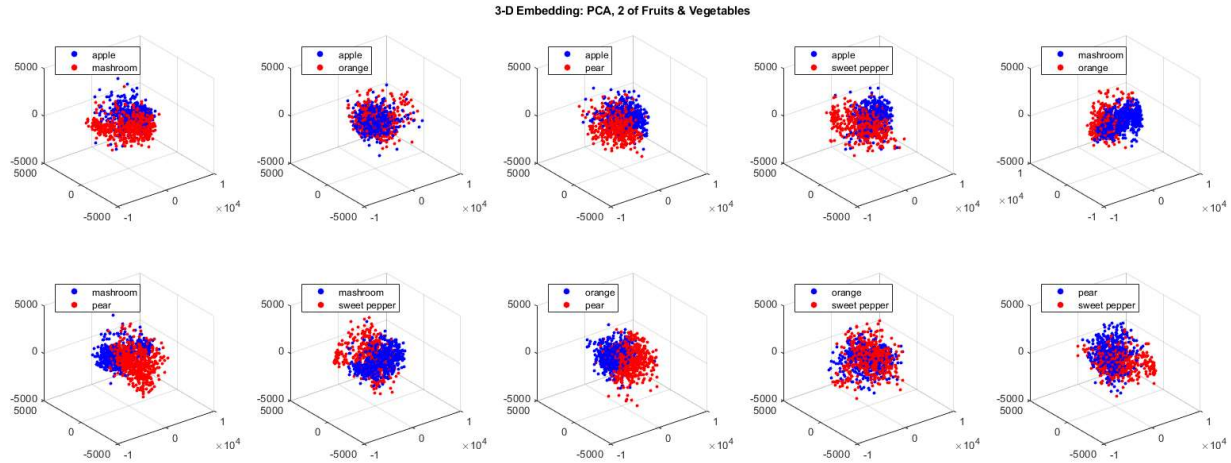
Zhu, X., & Ramanan, D. (2012). Face detection, pose estimation, and landmark localization in the wild. *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, (pp. 2879-2886).

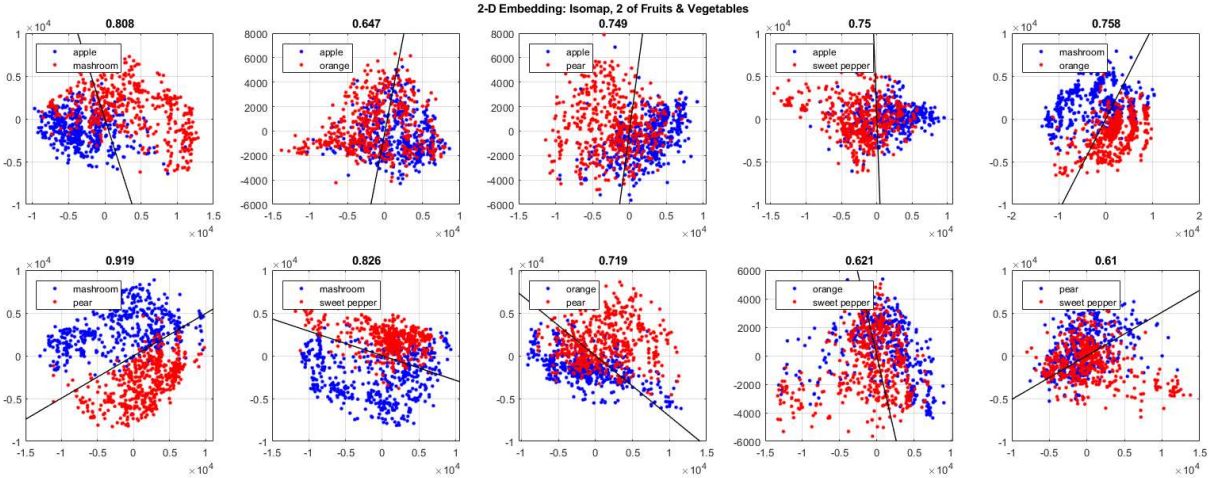
Appendix

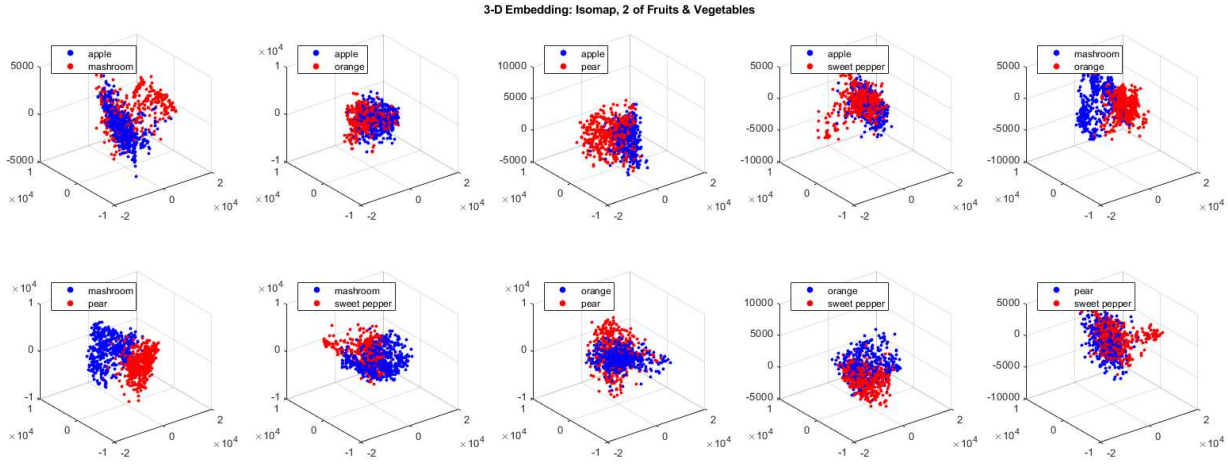
Category II

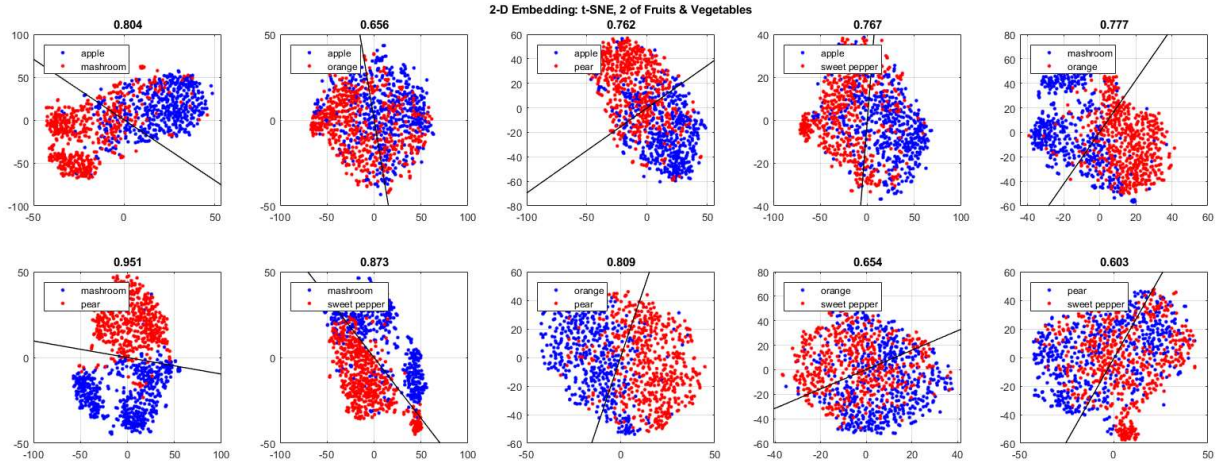
Combinations of 2 concepts

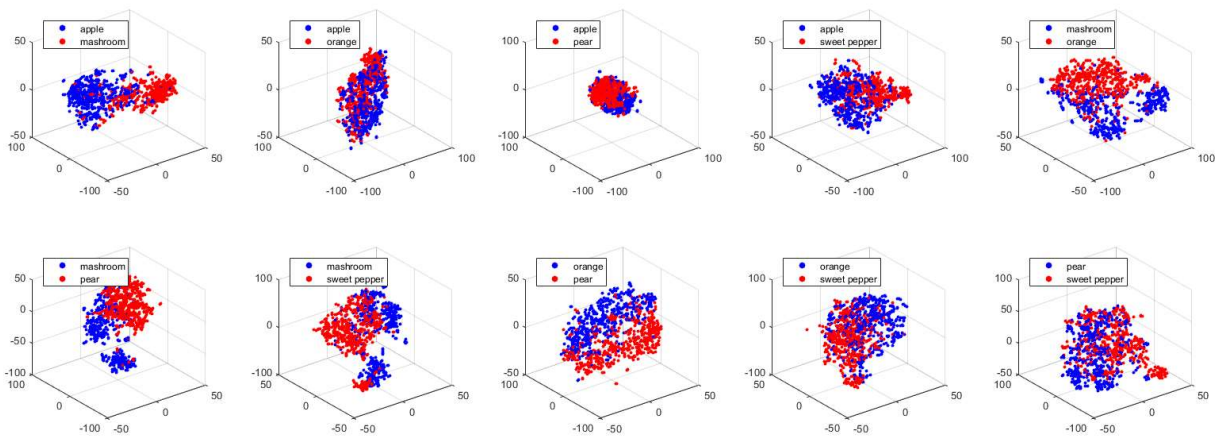




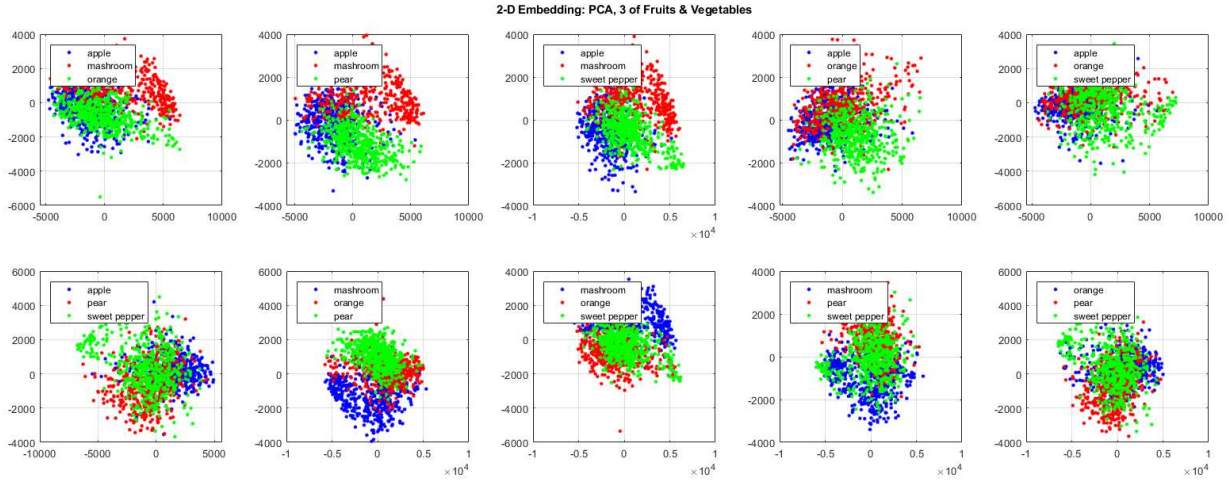


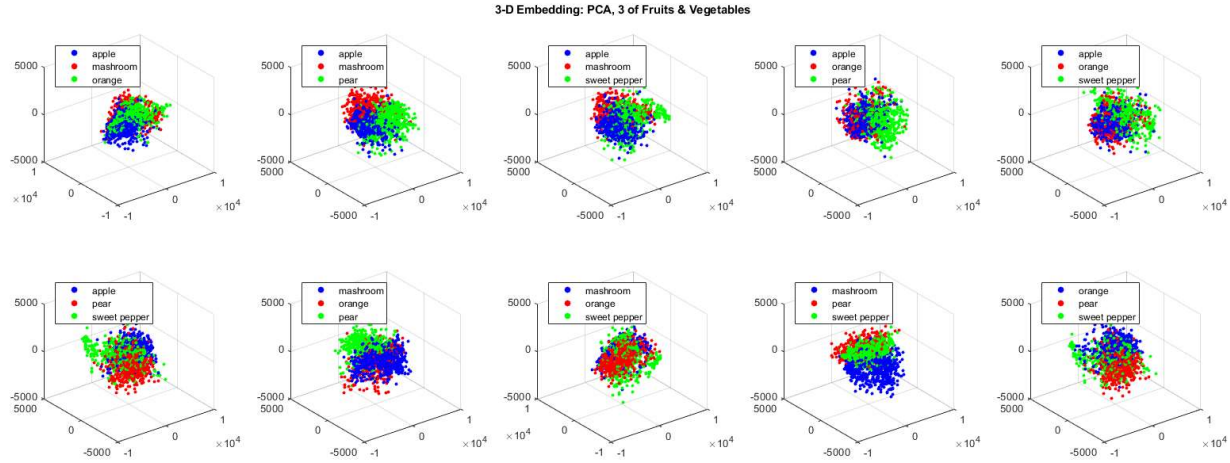

3-D Embedding: Isomap, 2 of Fruits & Vegetables
apple
mashroom
apple
orange
apple
pear
apple
sweet pepper
mashroom
orange
mashroom
pear
mashroom
sweet pepper
orange
pear
orange
sweet pepper
pear
sweet pepper

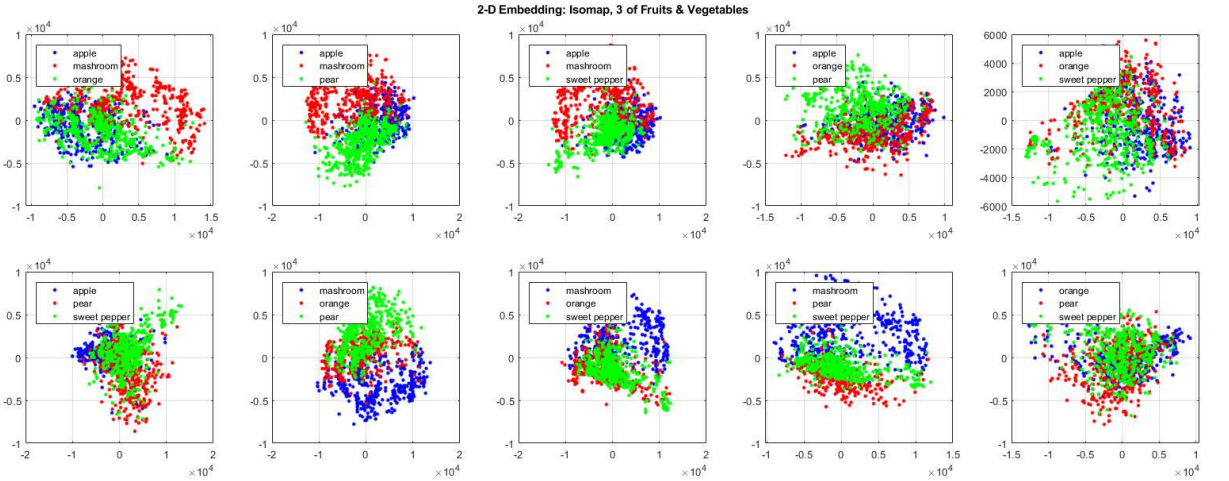

2-D Embedding: t-SNE, 2 of Fruits & Vegetables
0.804
0.656
0.762
0.767
0.777
0.951
0.873
0.809
0.654
0.603
apple
mashroom
apple
orange
apple
pear
apple
sweet pepper
mashroom
orange
mashroom
pear
mashroom
sweet pepper
orange
pear
orange
sweet pepper
pear
sweet pepper


3-D Embedding: t-SNE, 2 of Fruits & Vegetables
apple
mashroom
apple
orange
apple
pear
apple
sweet pepper
mashroom
orange
mashroom
pear
mashroom
sweet pepper
orange
pear
orange
sweet pepper
pear
sweet pepper

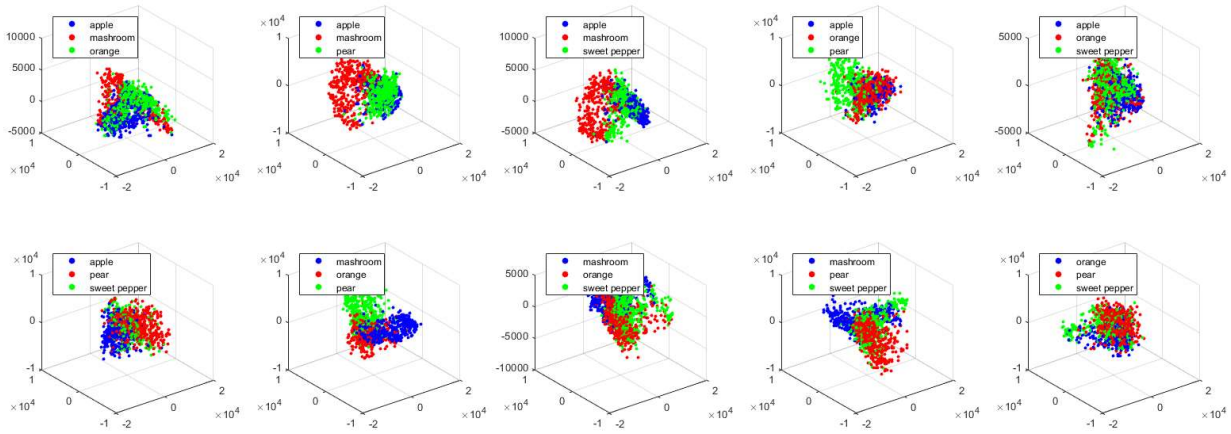
Combinations of 3 concepts



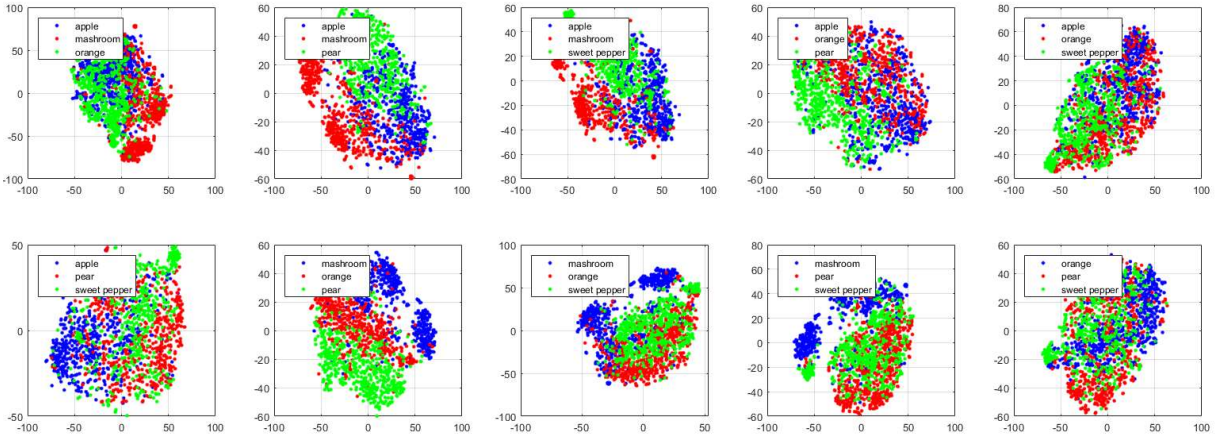




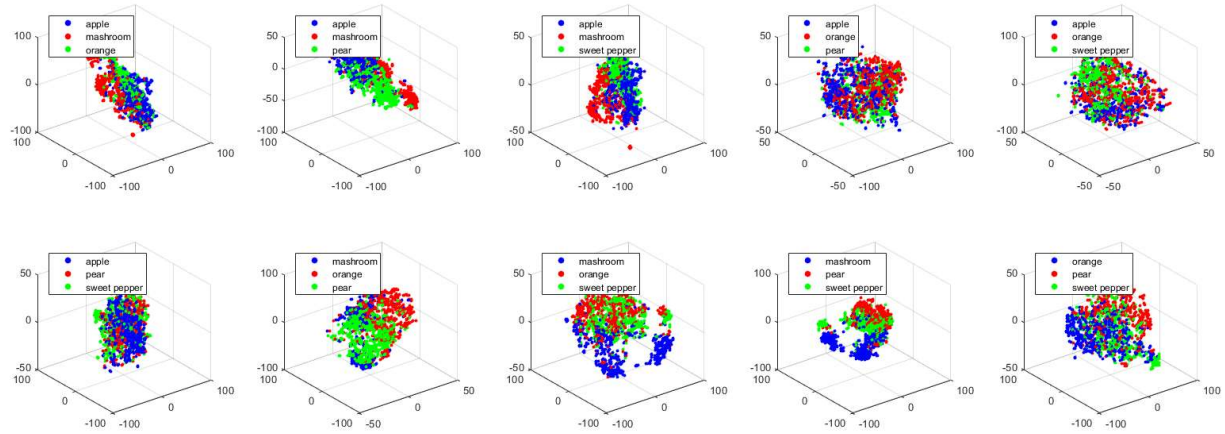
3-D Embedding: Isomap, 3 of Fruits & Vegetables



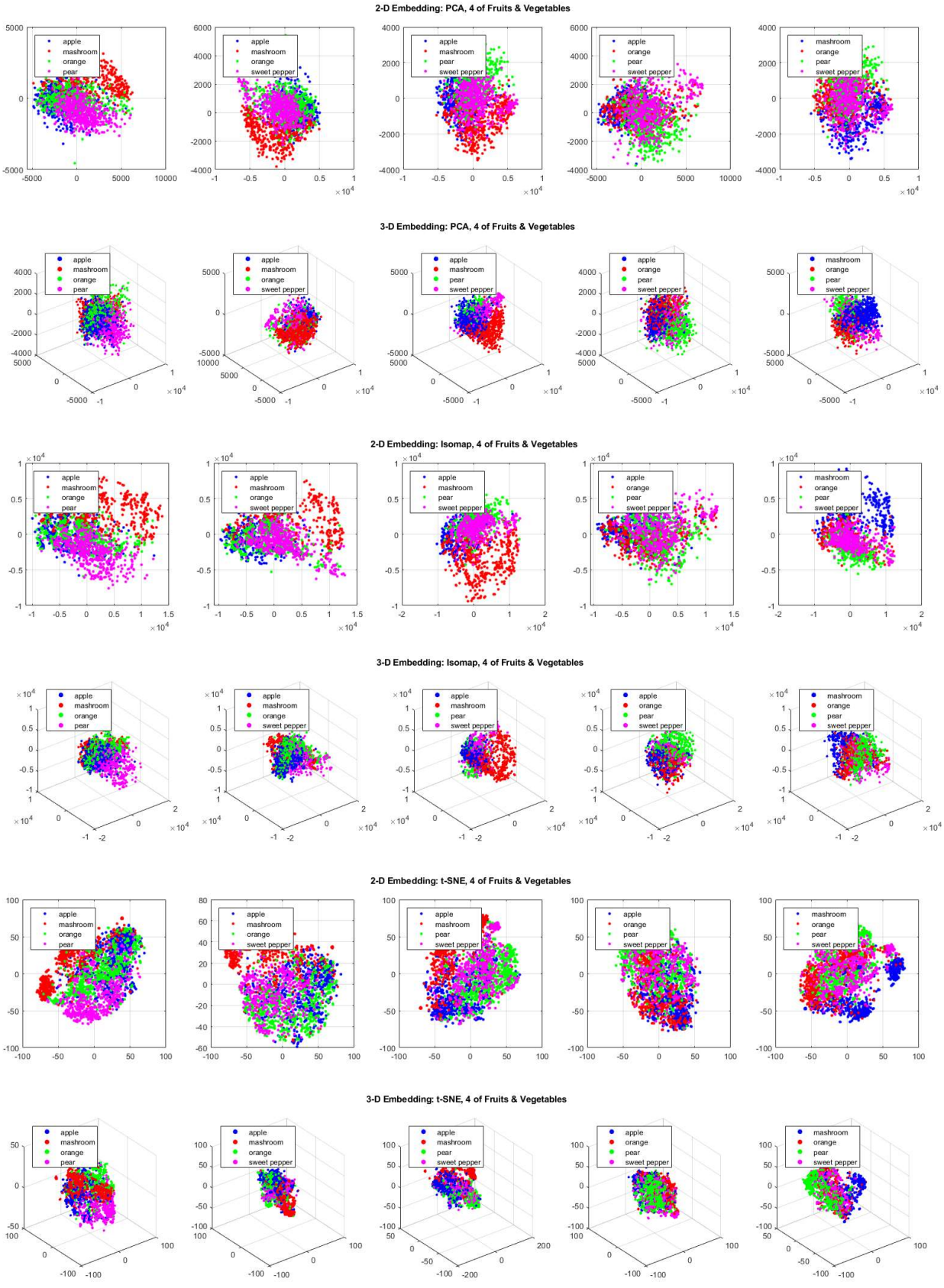
2-D Embedding: t-SNE, 3 of Fruits & Vegetables



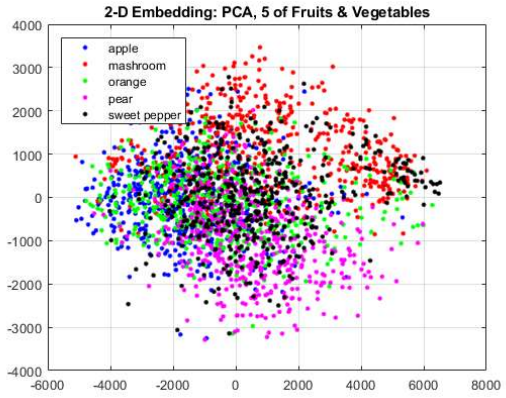
3-D Embedding: t-SNE, 3 of Fruits & Vegetables

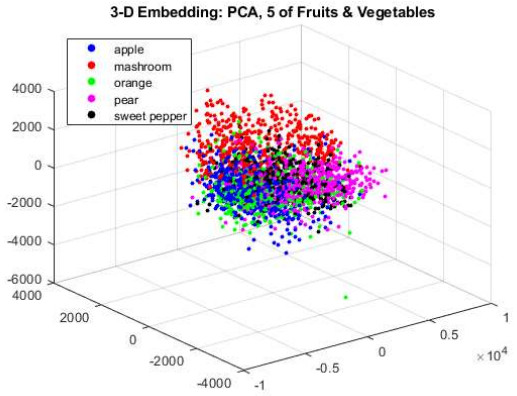


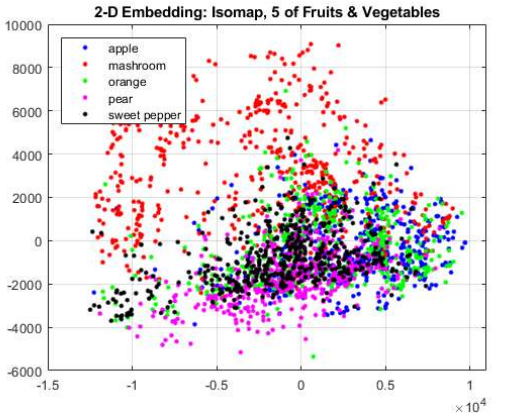
Combinations of 4 concepts

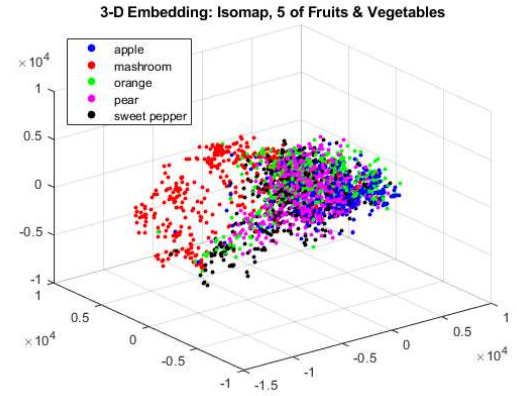


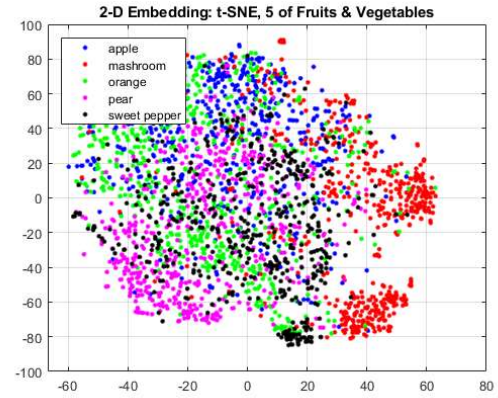
Combinations of 5 concepts

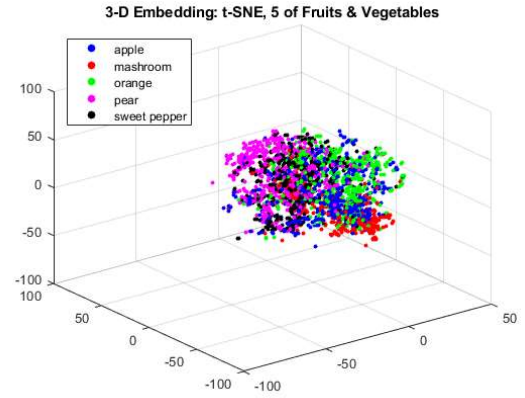






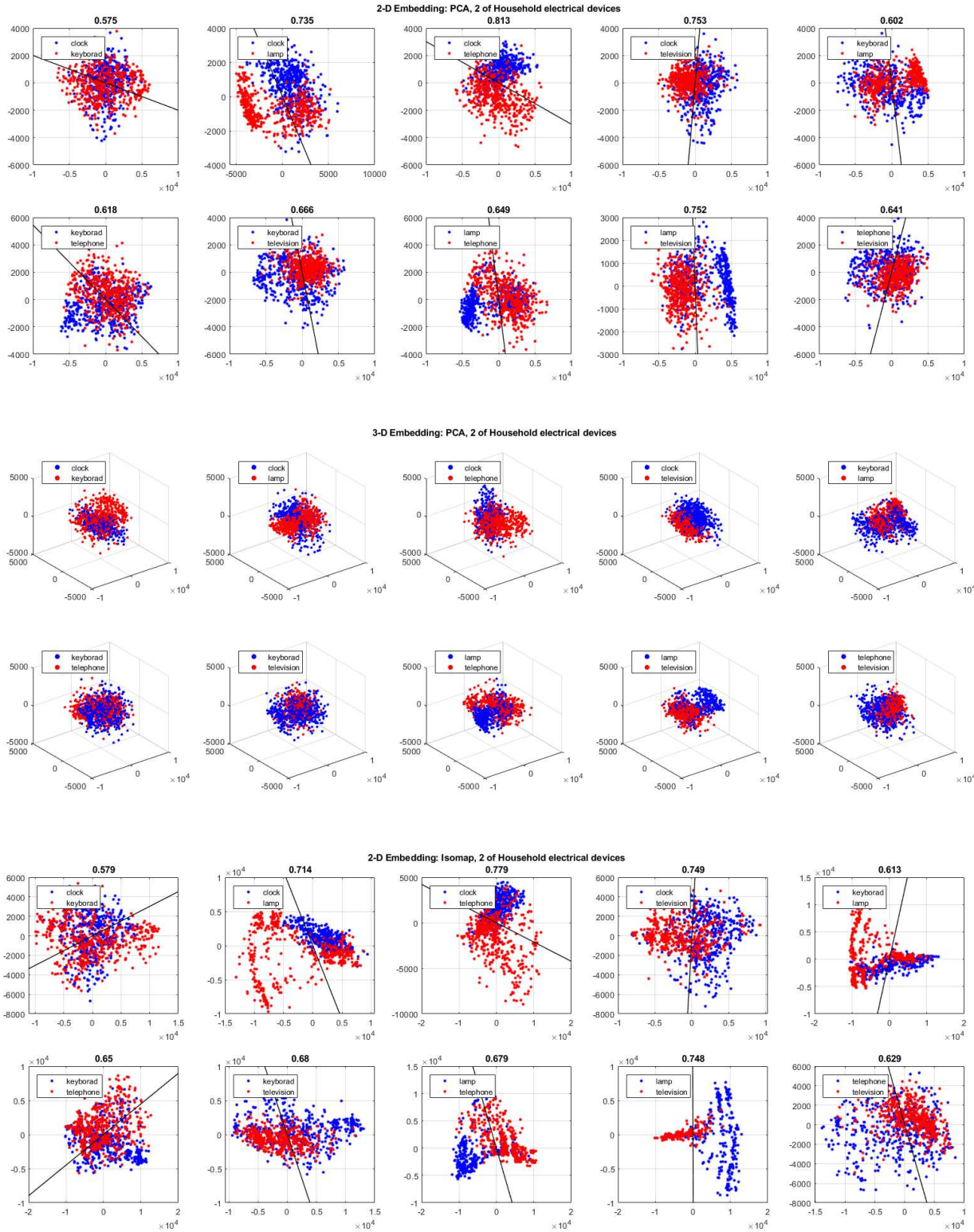


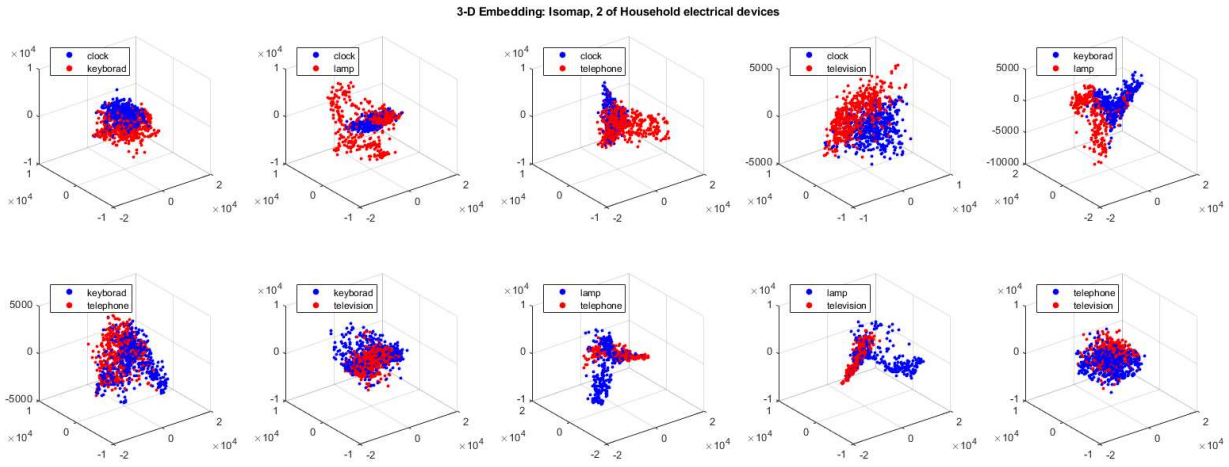


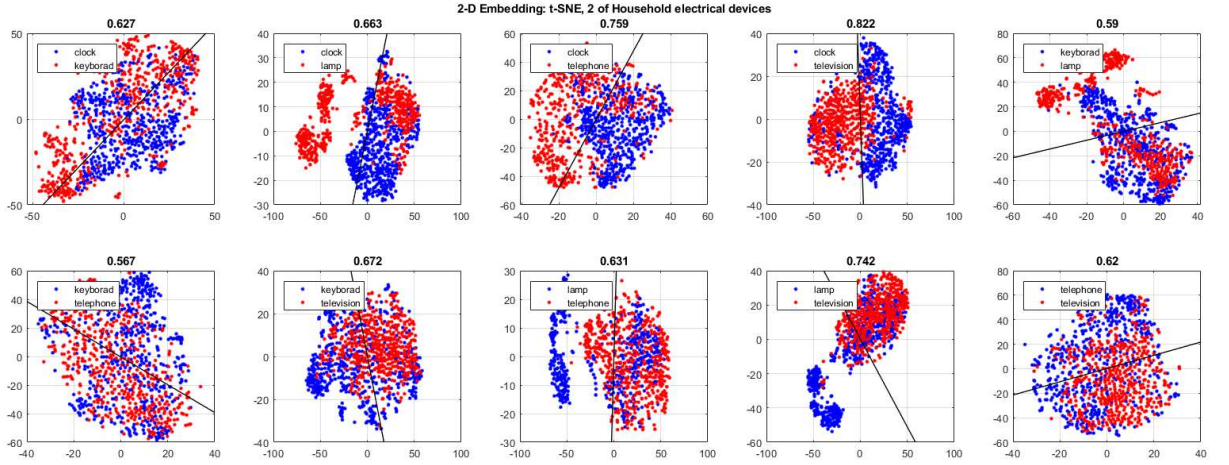


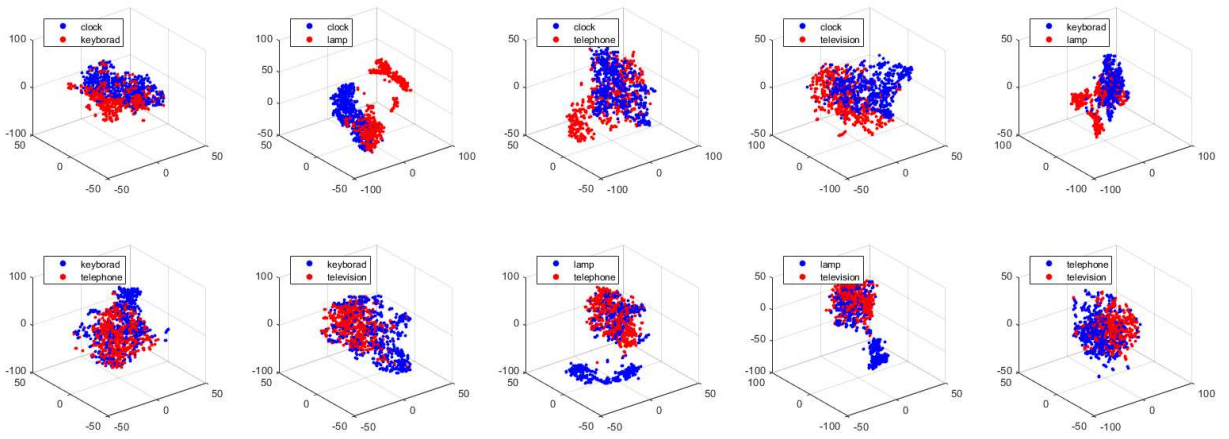
Category III

Combinations of 2 concepts

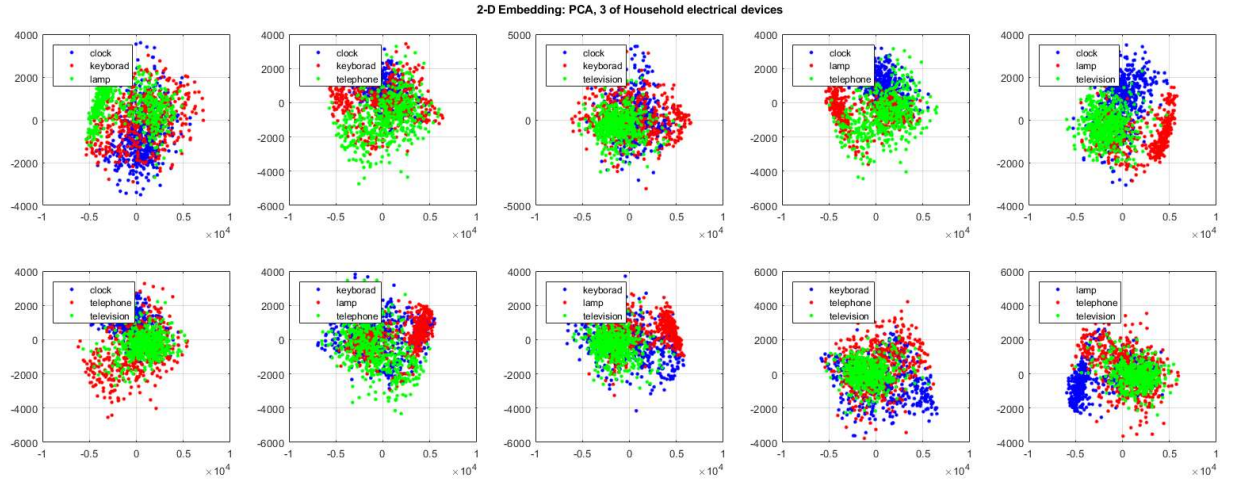


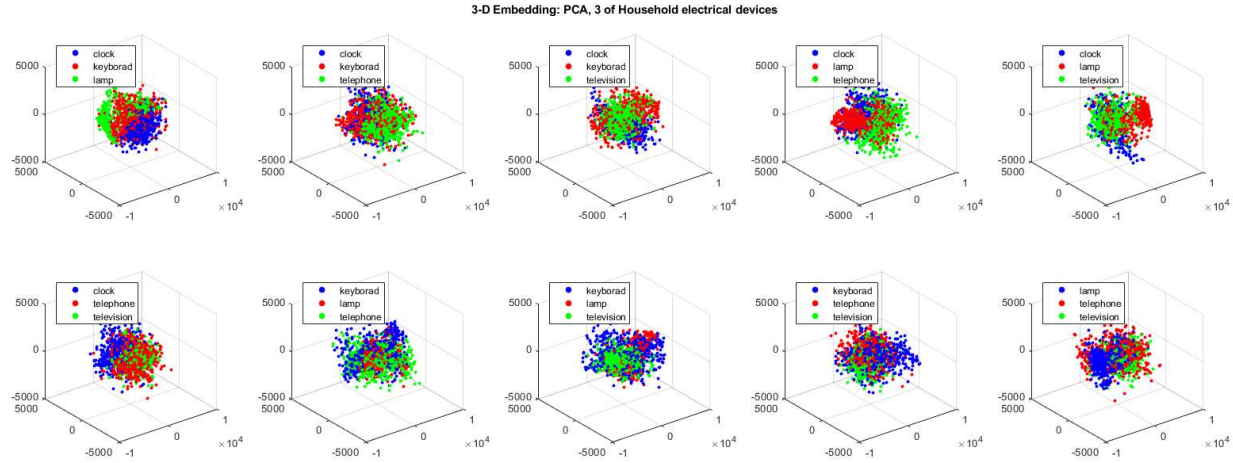

3-D Embedding: Isomap, 2 of Household electrical devices
clock
keyborad
clock
lamp
clock
telephone
clock
television
keyborad
lamp
keyborad
telephone
keyborad
television
lamp
telephone
lamp
television
telephone
television

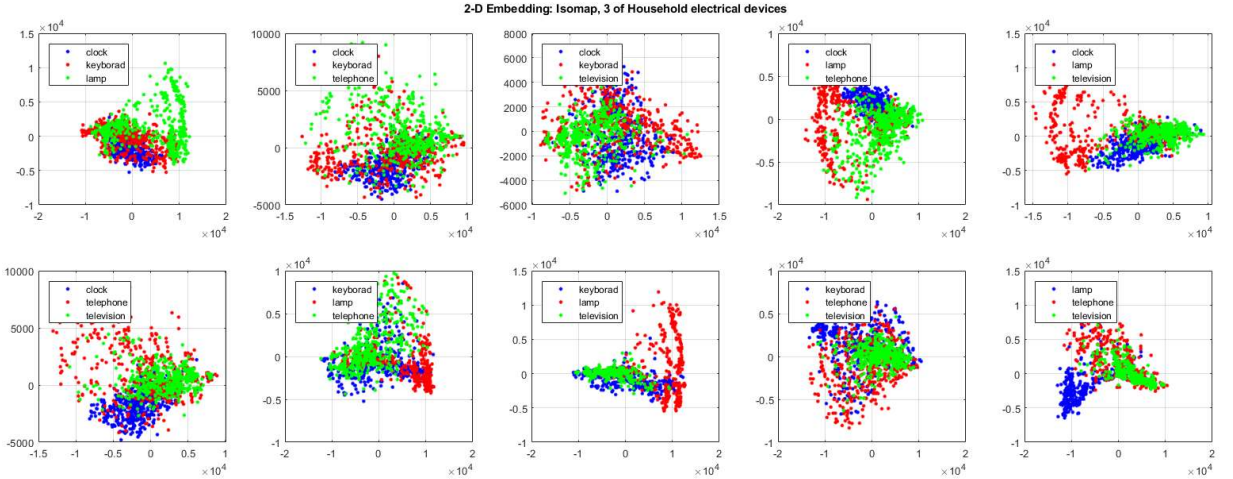

2-D Embedding: t-SNE, 2 of Household electrical devices
0.627
clock
keyborad
0.663
clock
lamp
0.759
clock
telephone
0.822
clock
television
0.59
keyborad
lamp
0.567
keyborad
telephone
0.672
keyborad
television
0.631
lamp
telephone
0.742
lamp
television
0.62
telephone
television


3-D Embedding: t-SNE, 2 of Household electrical devices
clock
keyborad
clock
lamp
clock
telephone
clock
television
keyborad
lamp
keyborad
telephone
keyborad
television
lamp
telephone
lamp
television
telephone
television

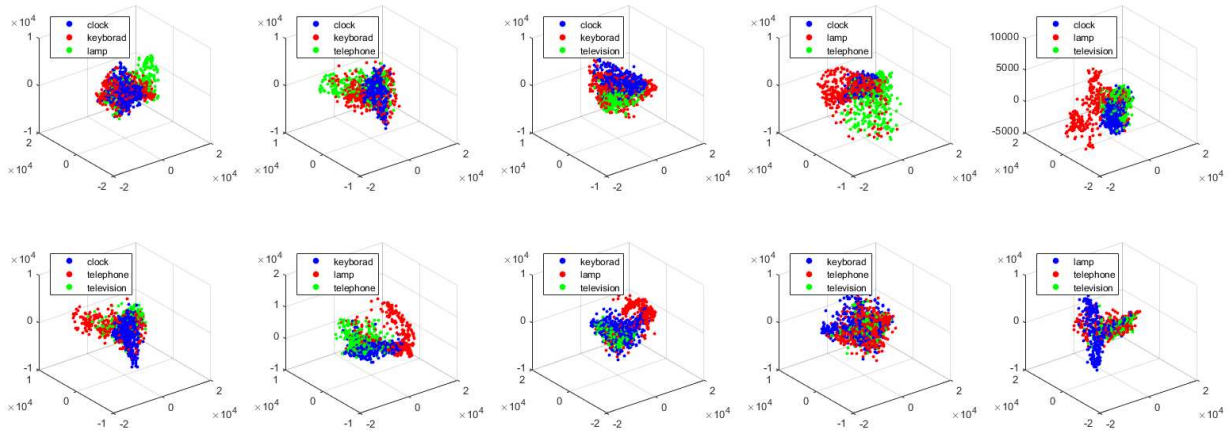
Combinations of 3 concepts



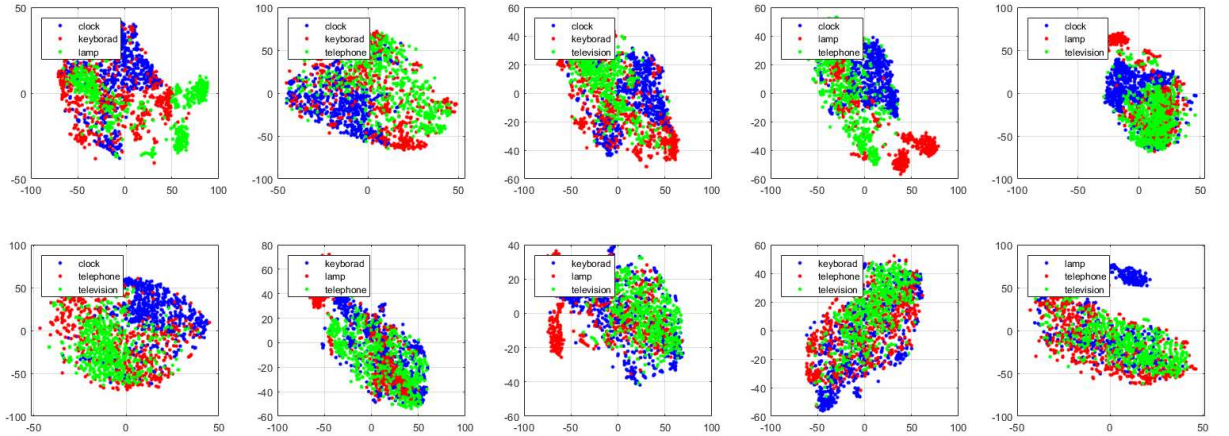




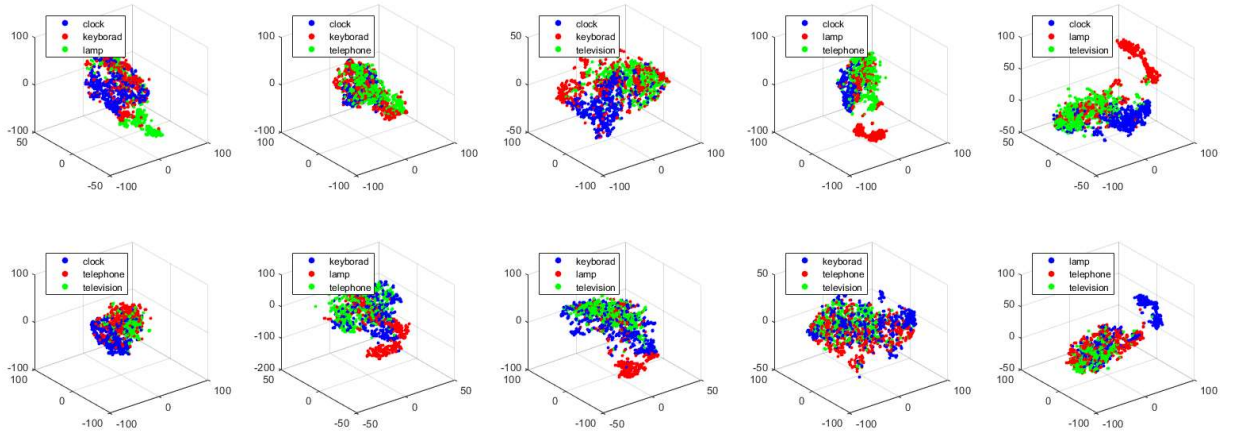
3-D Embedding: Isomap, 3 of Household electrical devices



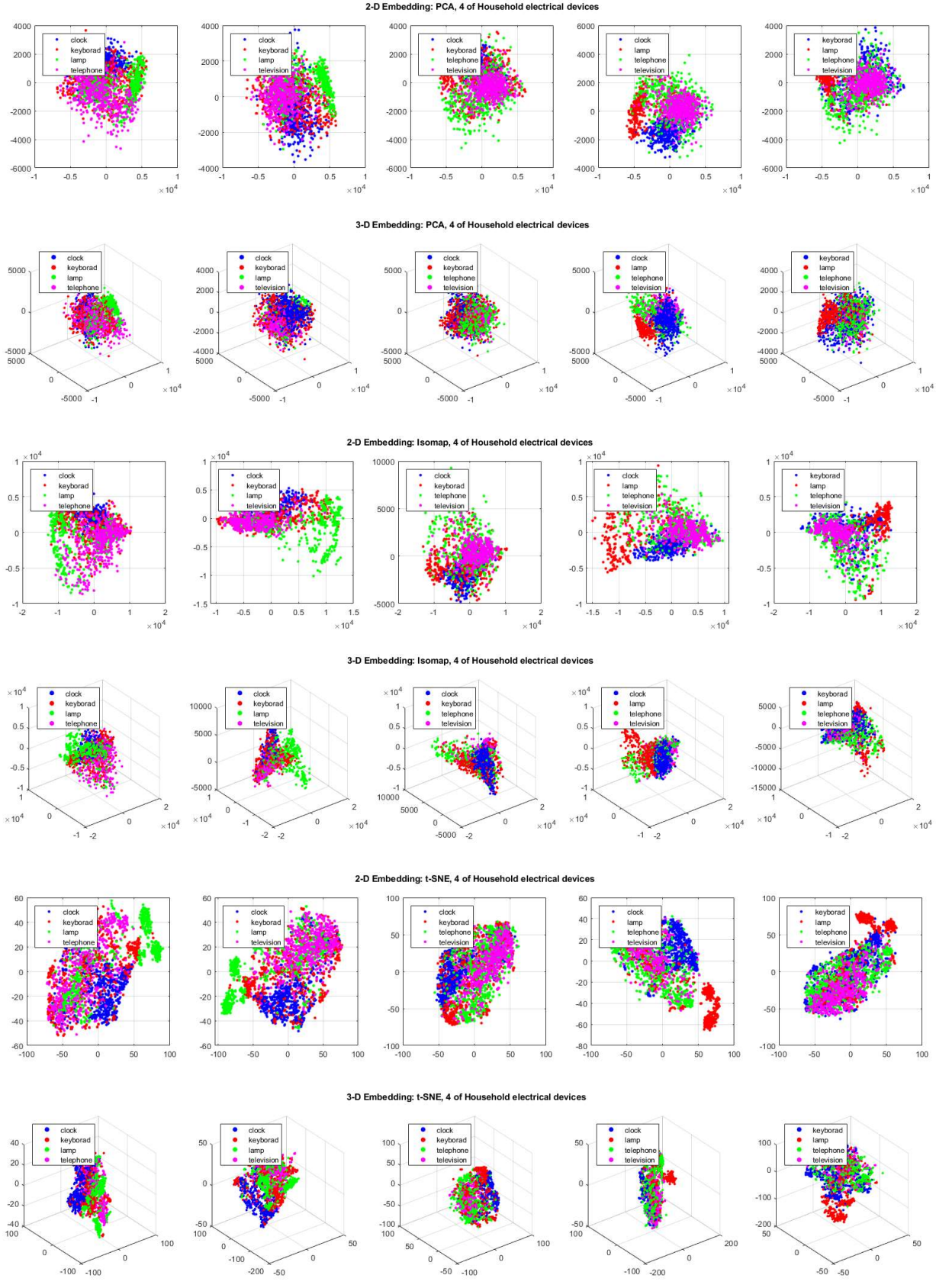
2-D Embedding: t-SNE, 3 of Household electrical devices



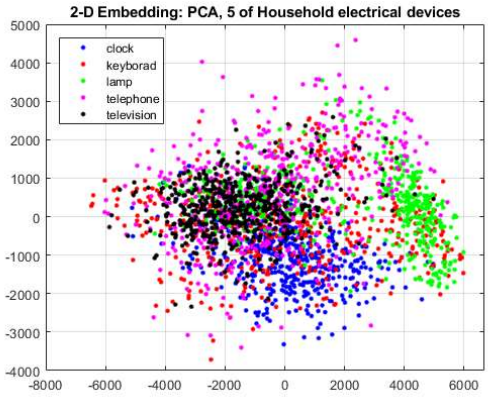
3-D Embedding: t-SNE, 3 of Household electrical devices

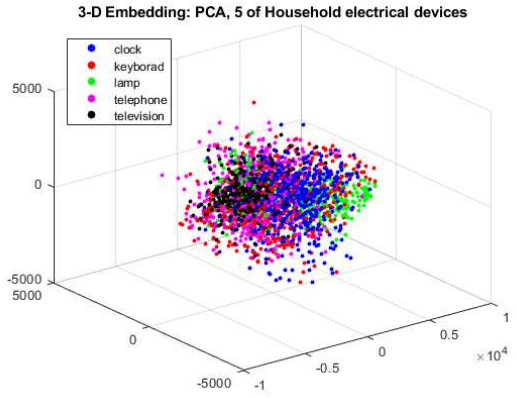


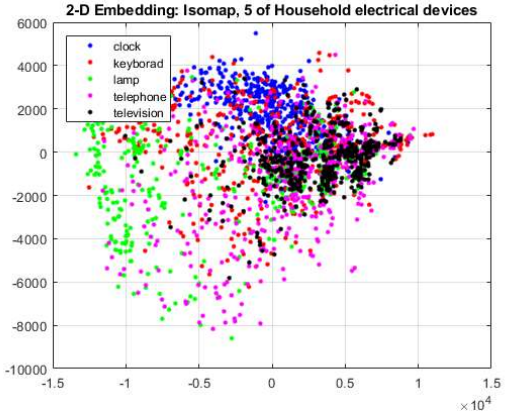
Combinations of 4 concepts

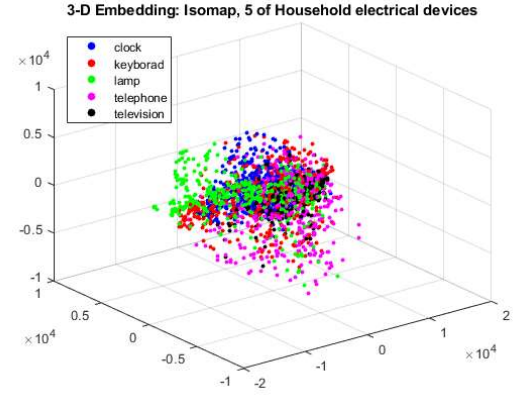


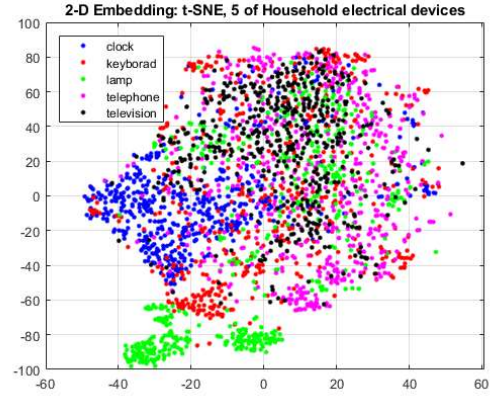
Combinations of 5 concepts

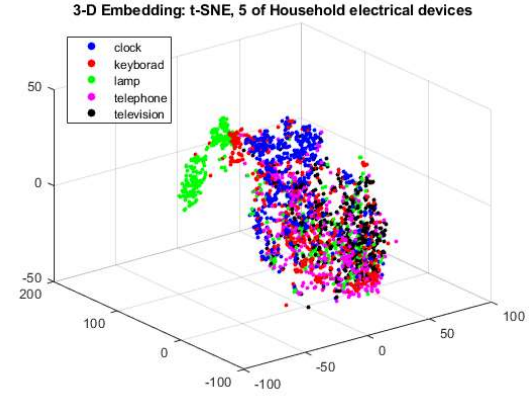






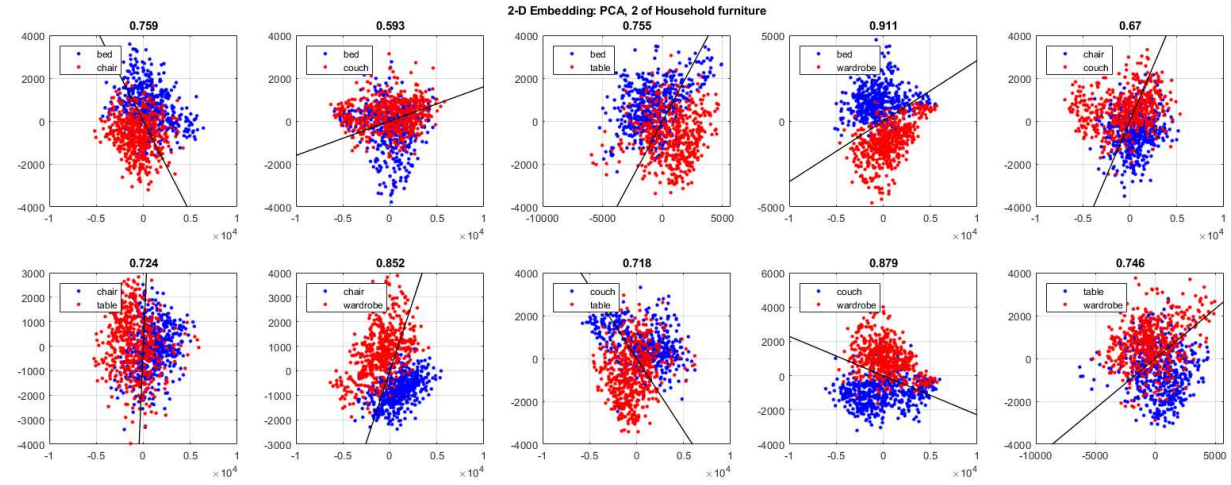


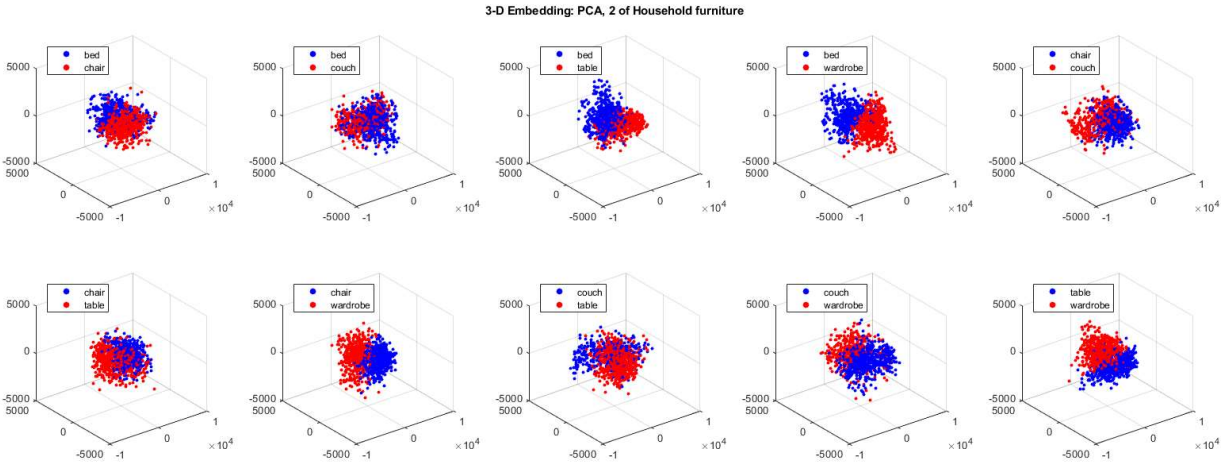


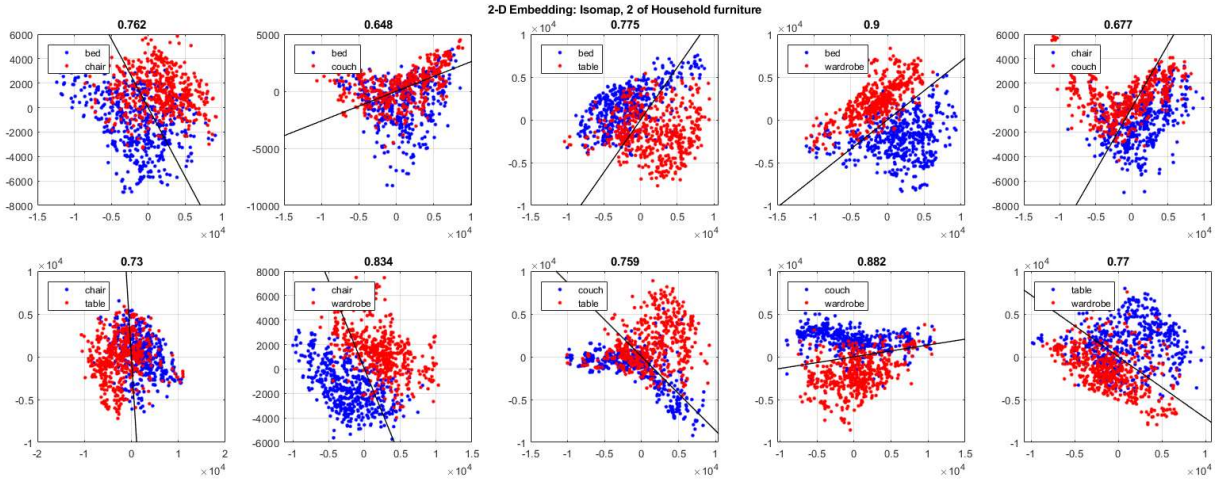


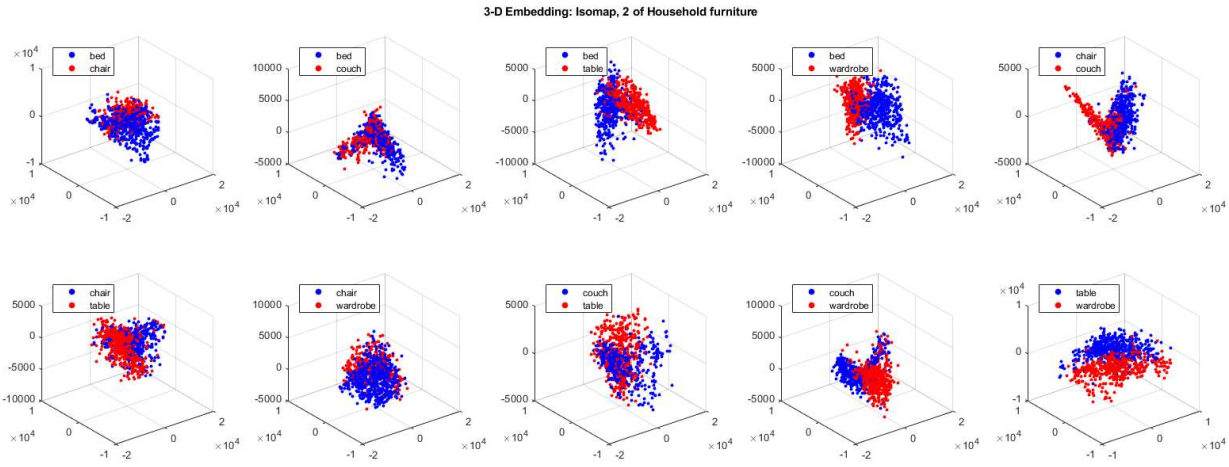
Category IV

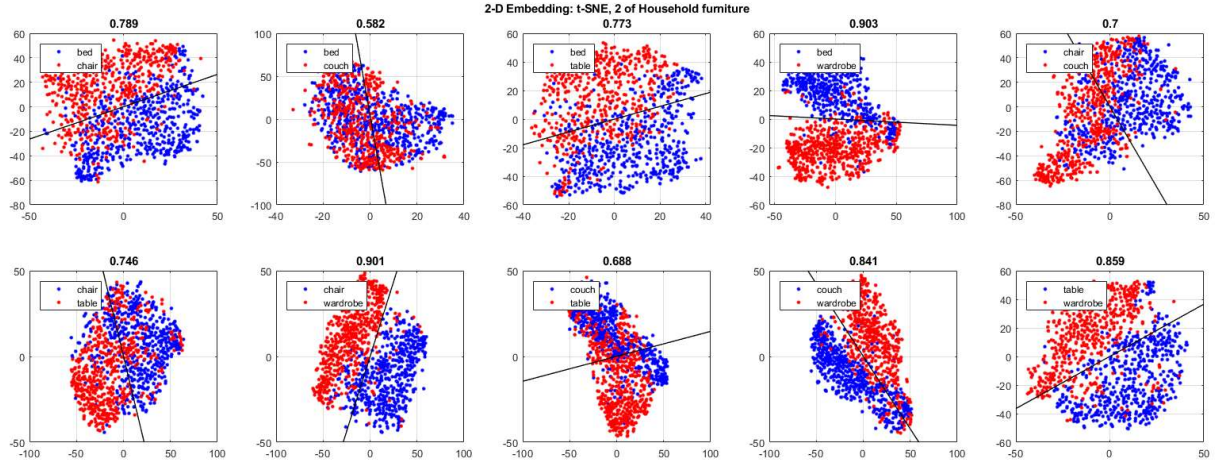
Combinations of 2 concepts

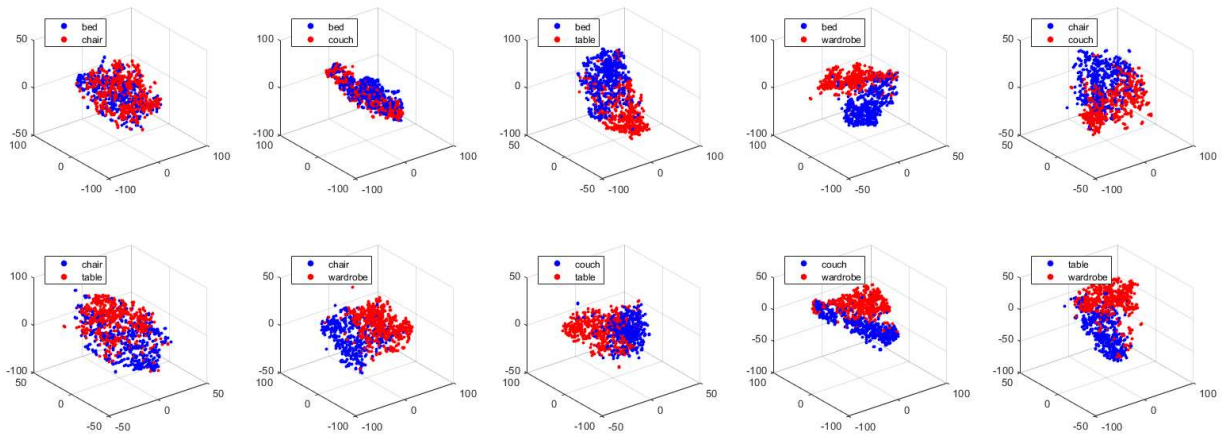




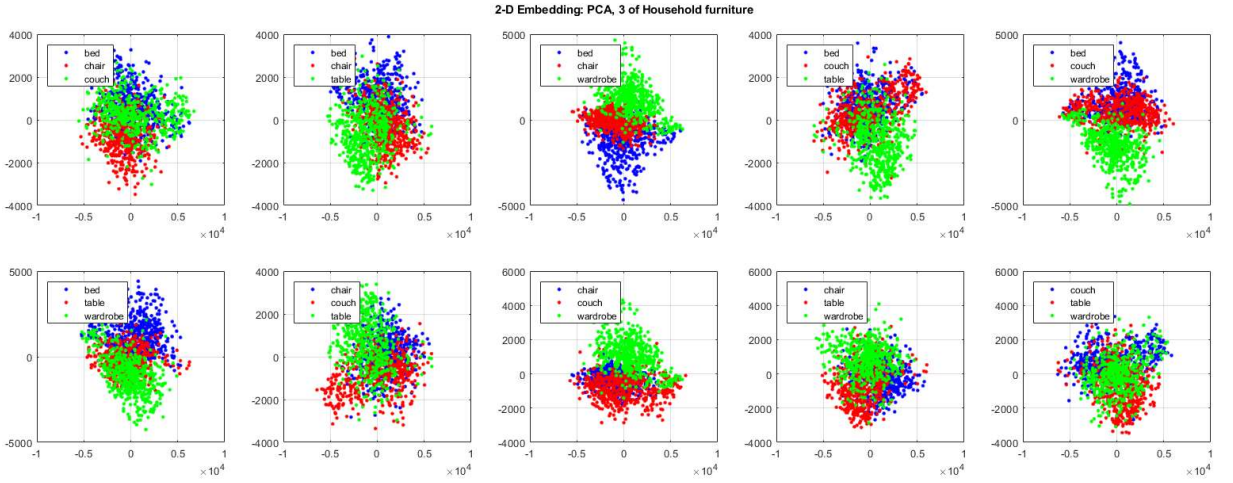


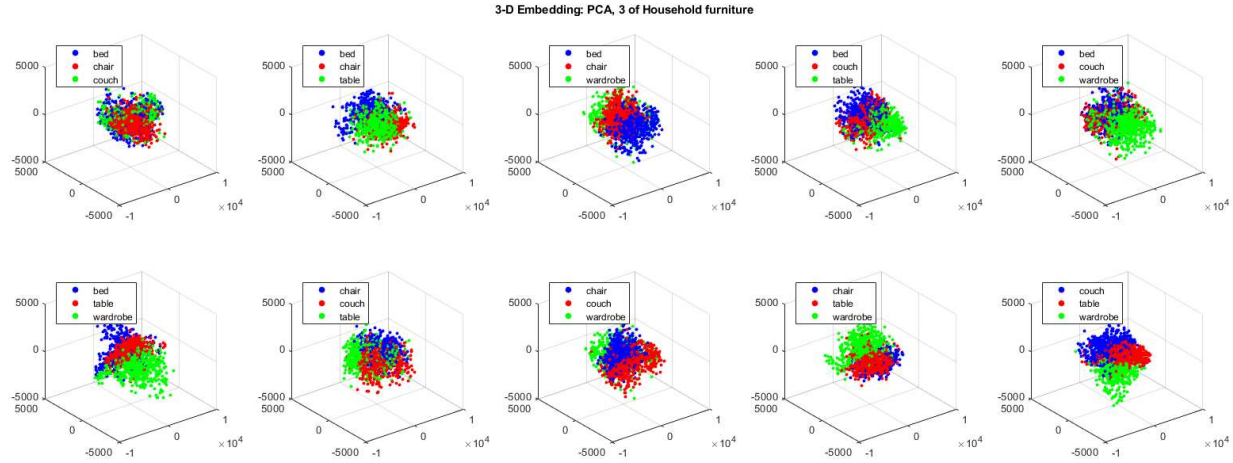

3-D Embedding: Isomap, 2 of Household furniture
bed
chair
bed
couch
bed
table
bed
wardrobe
chair
couch
chair
table
chair
wardrobe
couch
table
couch
wardrobe
table
wardrobe

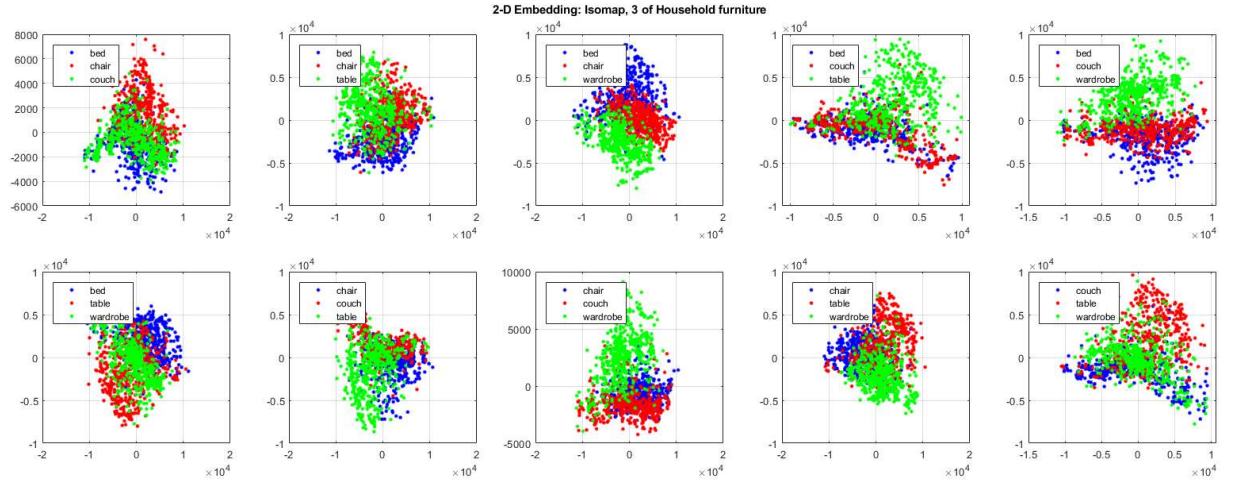

2-D Embedding: t-SNE, 2 of Household furniture
0.789
0.582
0.773
0.903
0.7
0.746
0.901
0.688
0.841
0.859
bed
chair
bed
couch
bed
table
bed
wardrobe
chair
couch
chair
table
chair
wardrobe
couch
table
couch
wardrobe
table
wardrobe


3-D Embedding: t-SNE, 2 of Household furniture
bed
chair
bed
couch
bed
table
bed
wardrobe
chair
couch
chair
table
chair
wardrobe
couch
table
couch
wardrobe
table
wardrobe

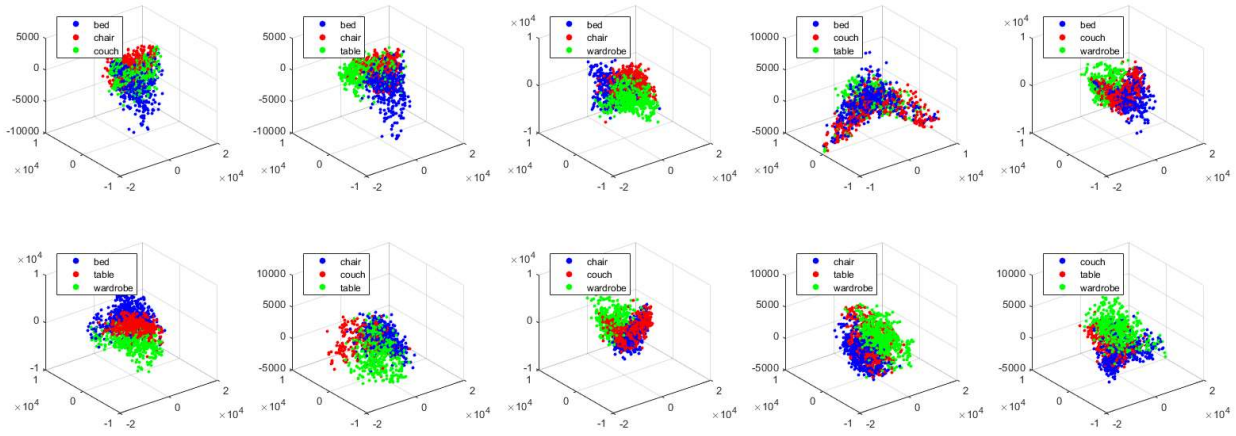
Combinations of 3 concepts



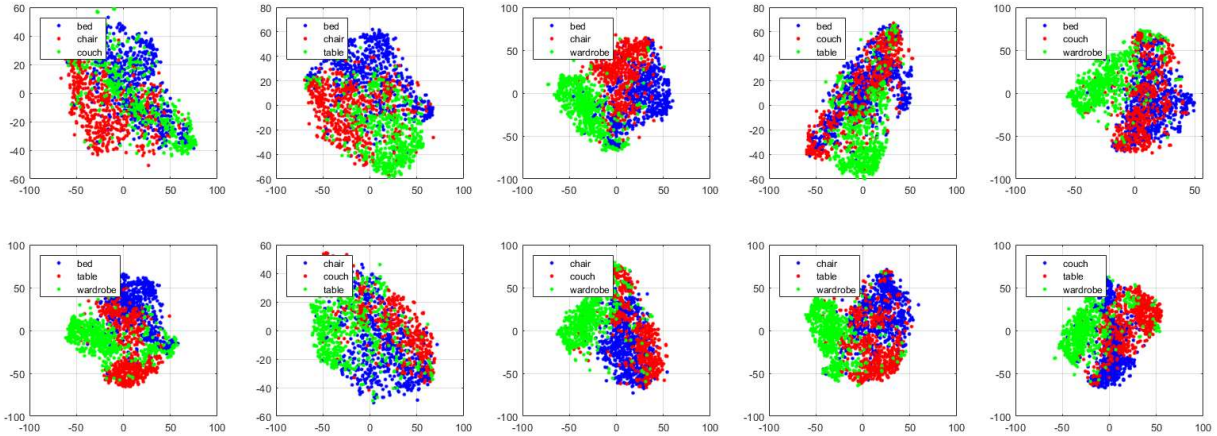




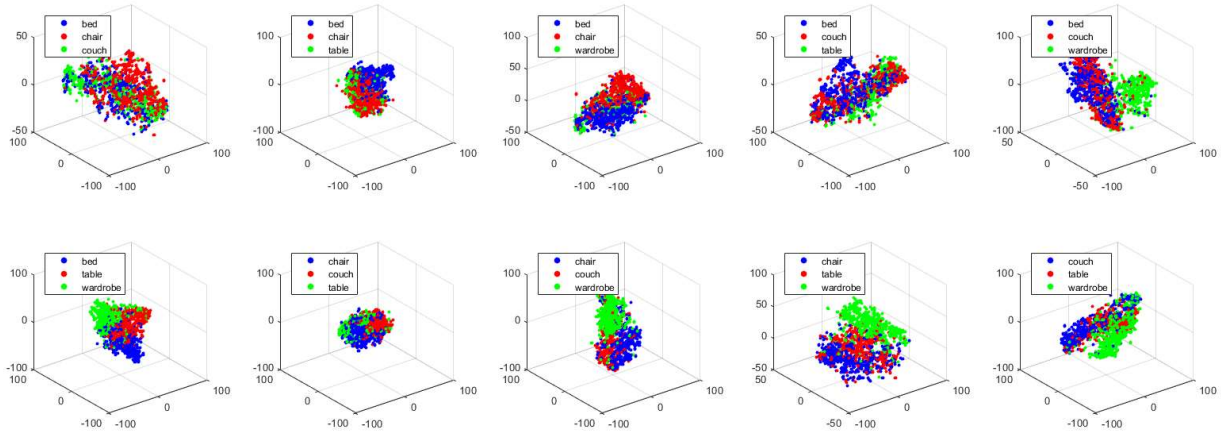
3-D Embedding: Isomap, 3 of Household furniture



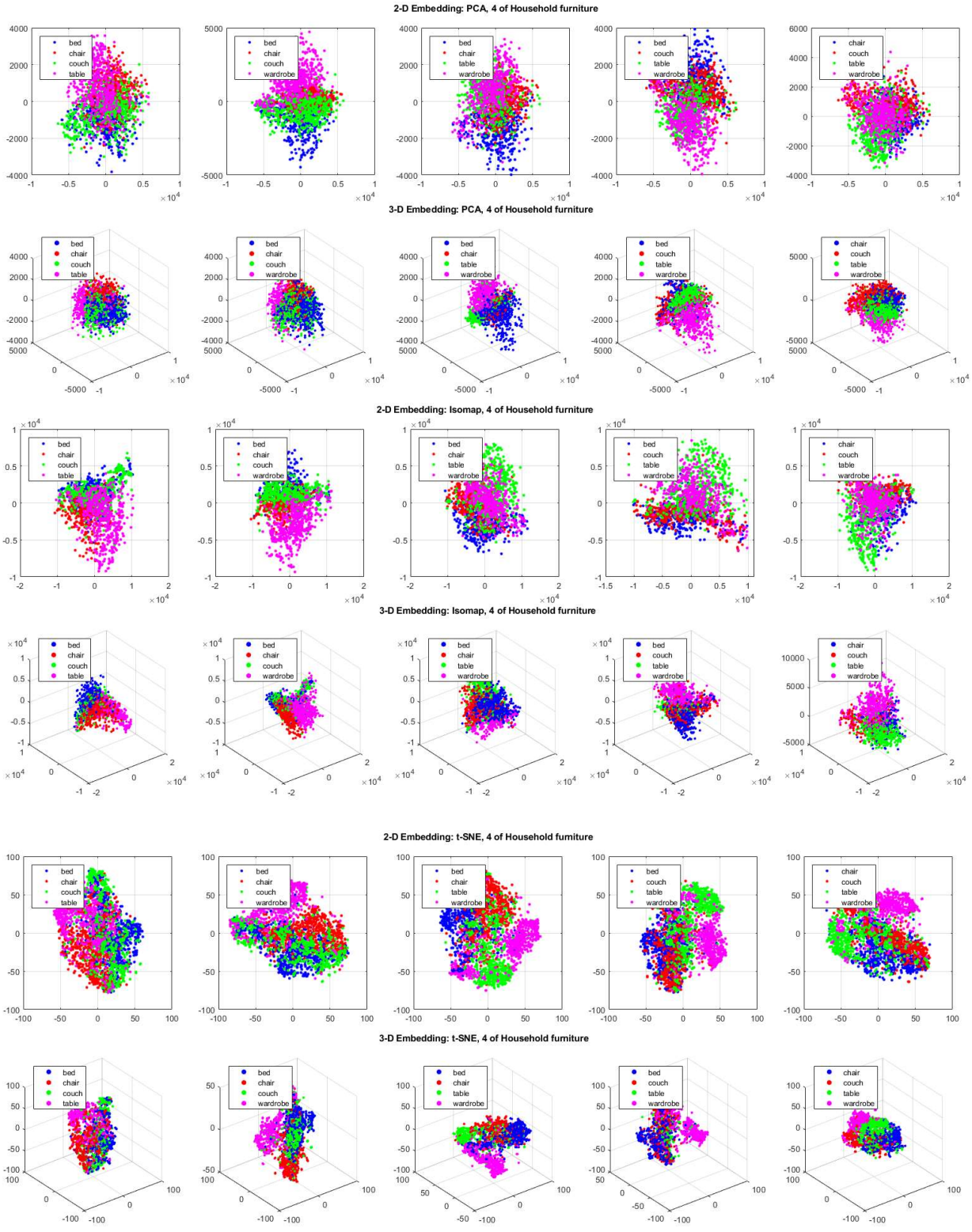
2-D Embedding: t-SNE, 3 of Household furniture



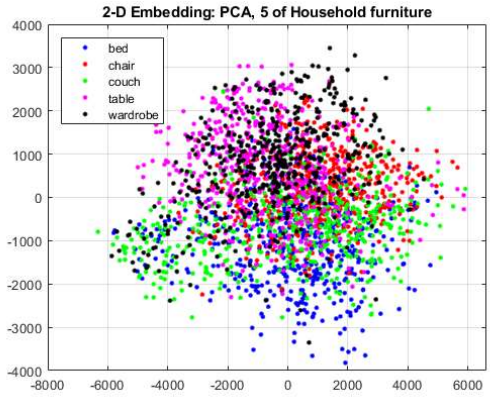
3-D Embedding: t-SNE, 3 of Household furniture

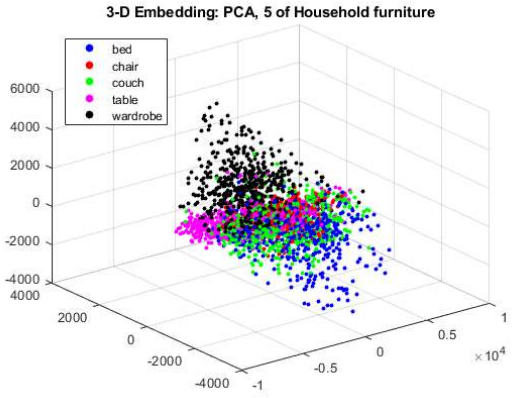


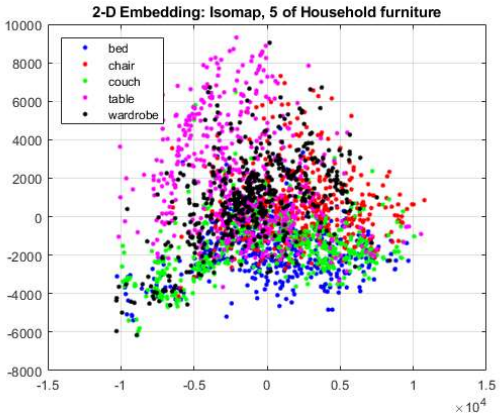
Combinations of 4 concepts

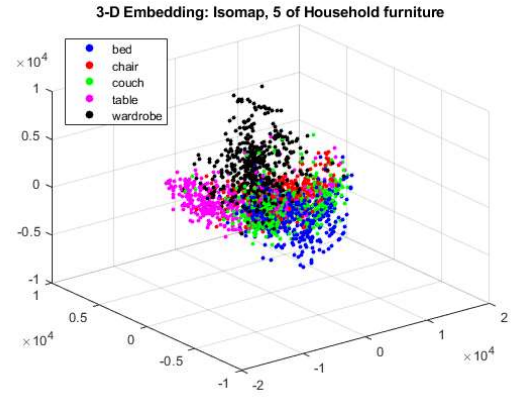


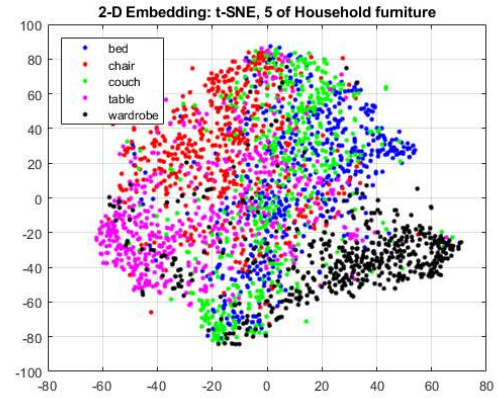
Combinations of 5 concepts

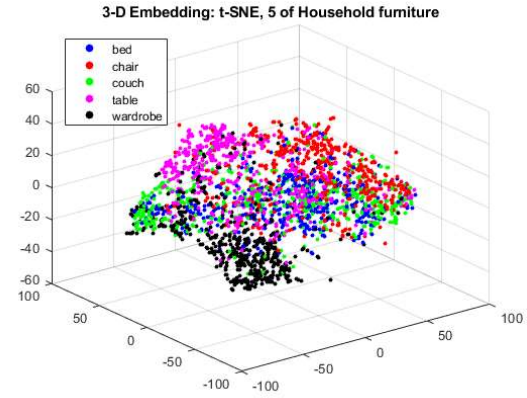






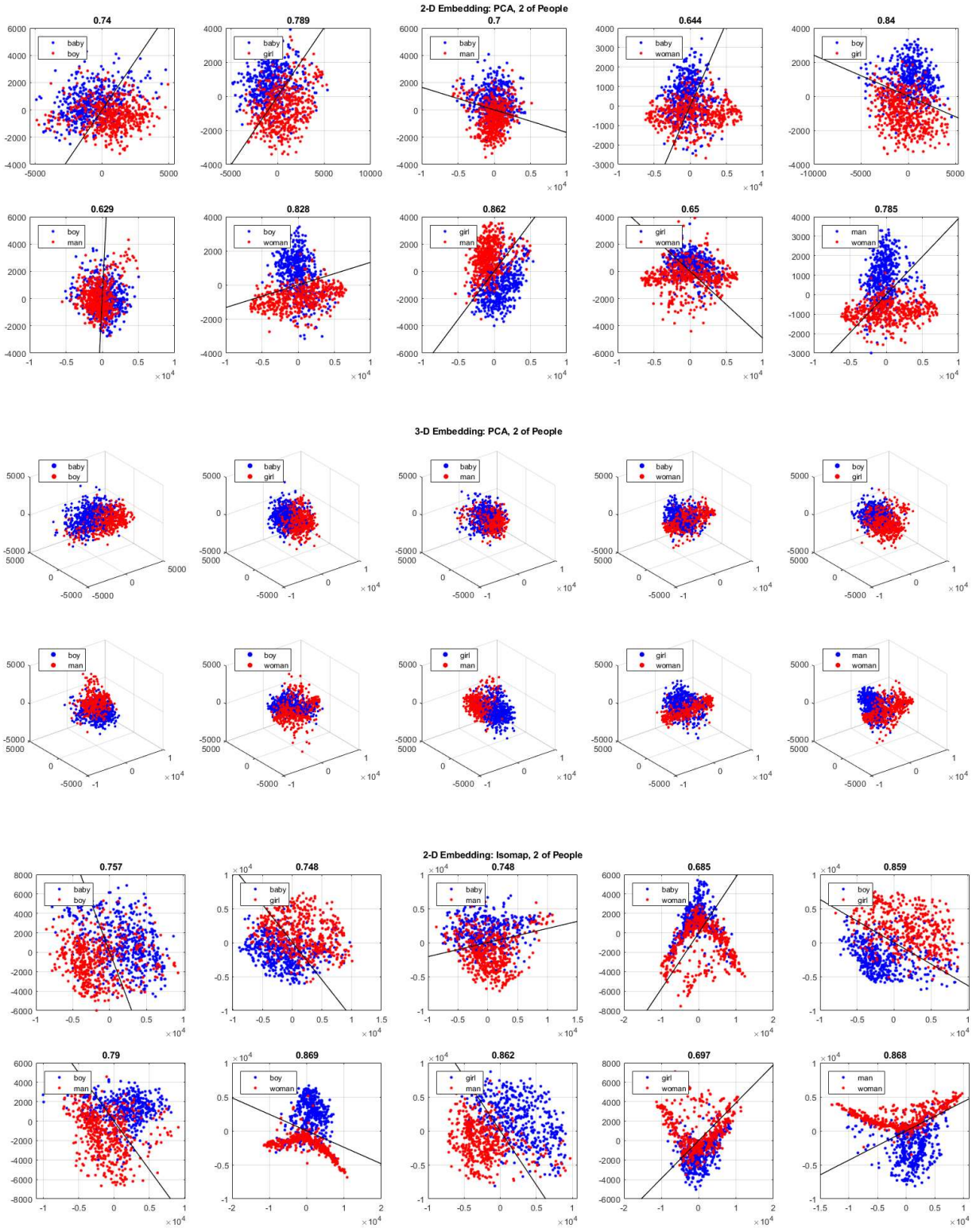


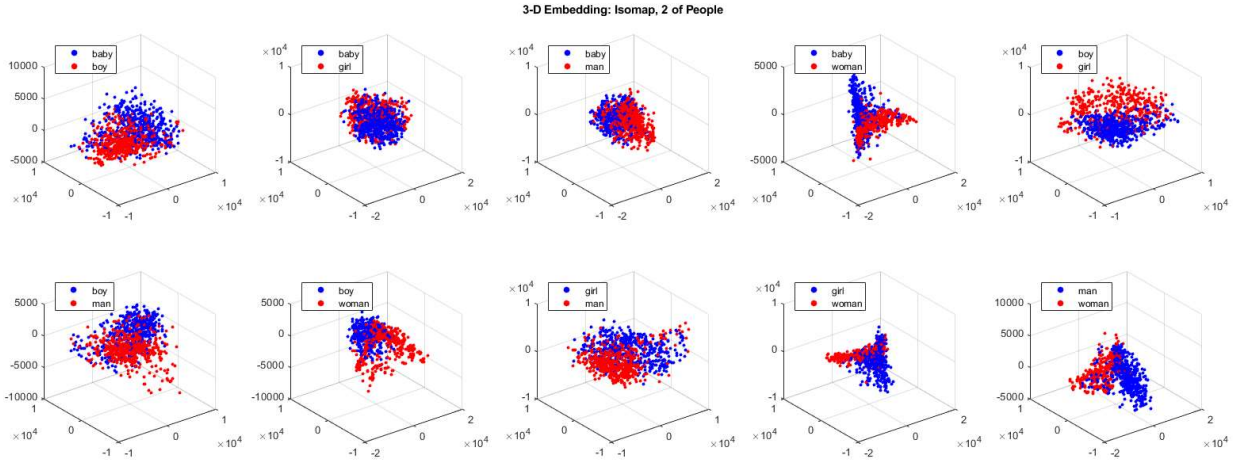


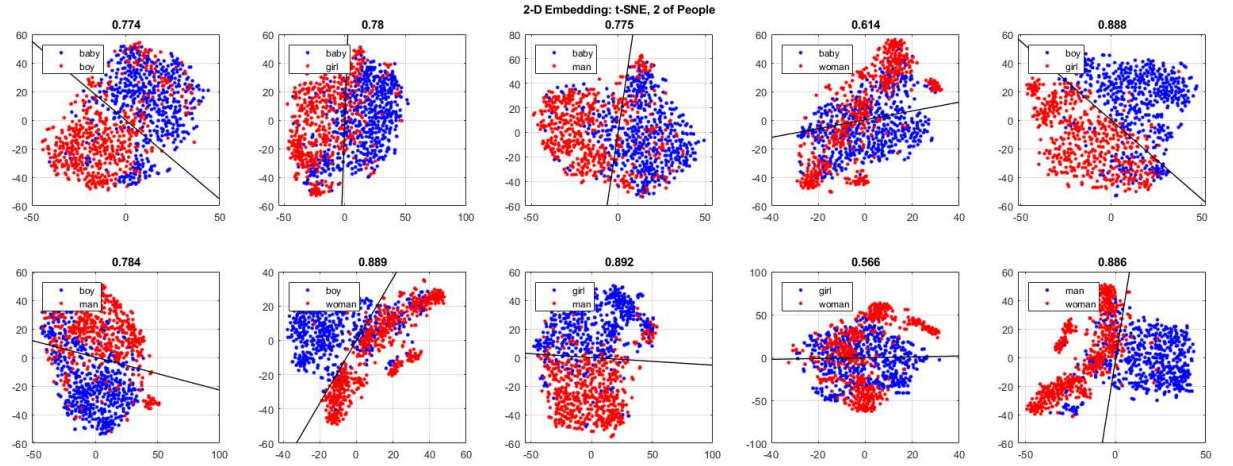


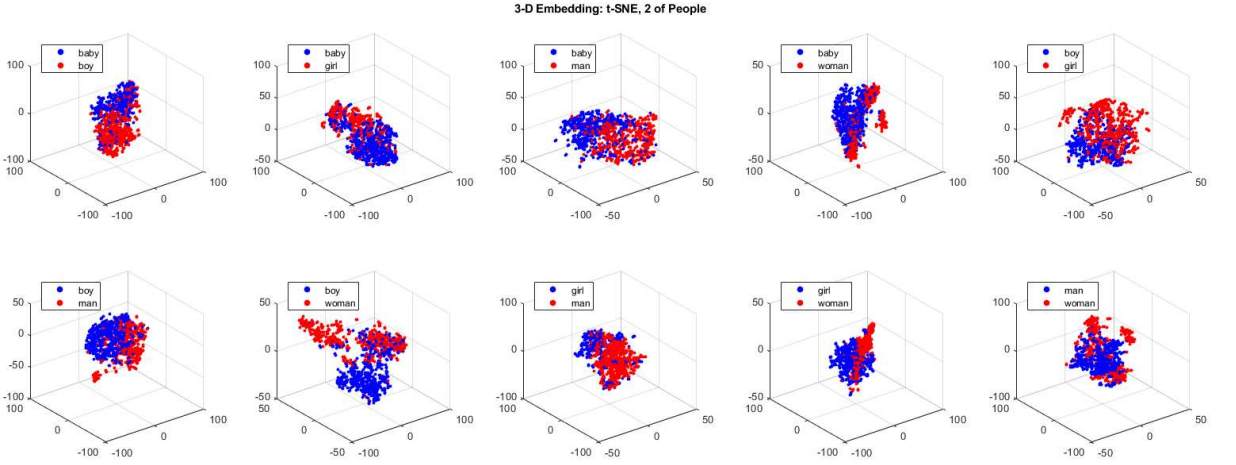
Category V

Combinations of 2 concepts

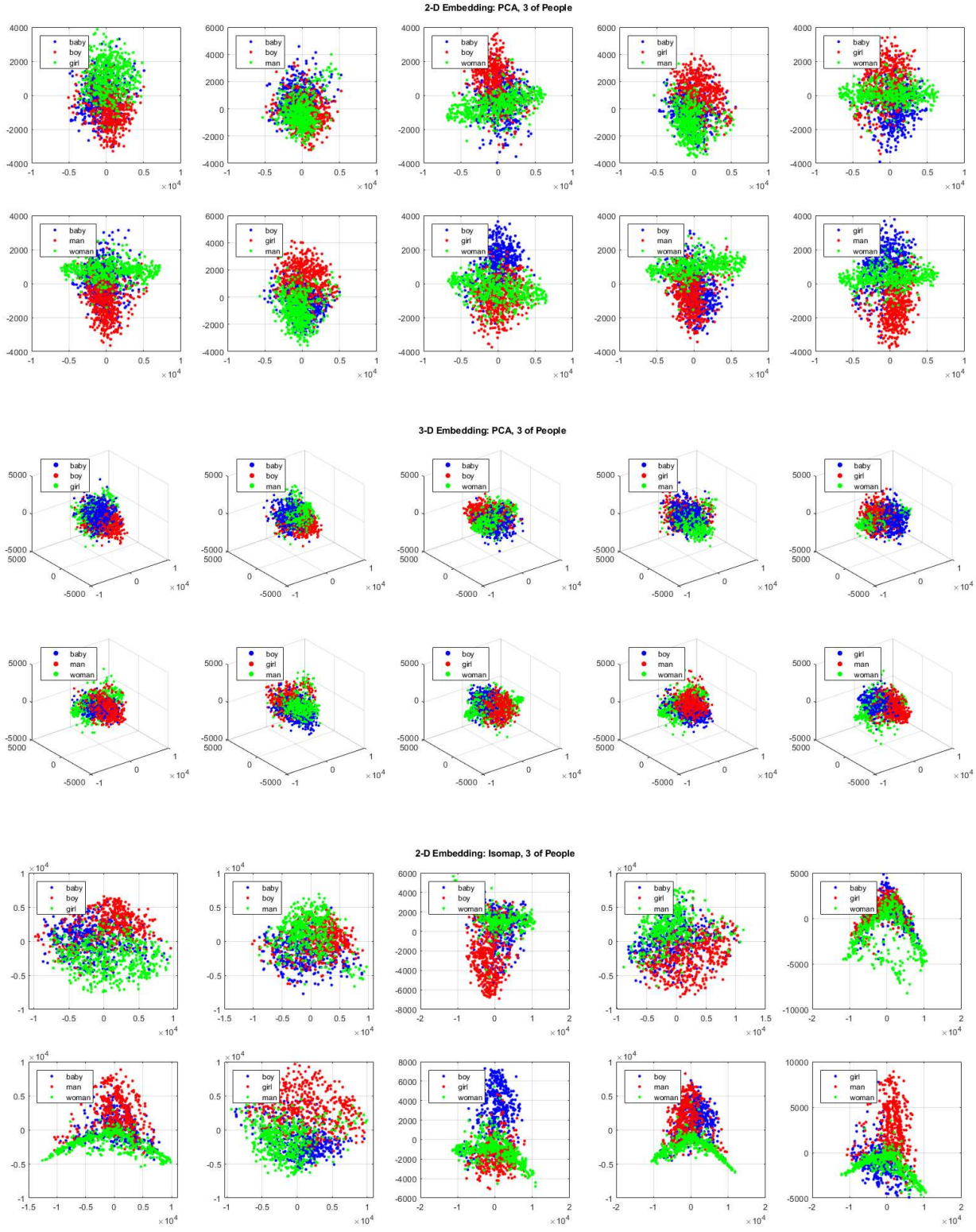



3-D Embedding: Isomap, 2 of People
baby boy
baby girl
baby man
baby woman
boy girl
boy man
boy woman
girl man
girl woman
man woman

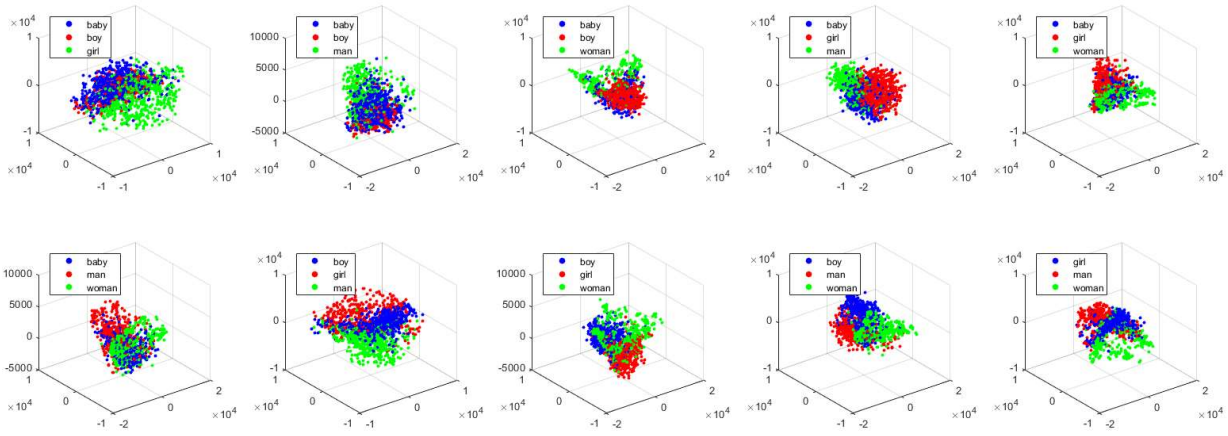

2-D Embedding: t-SNE, 2 of People
0.774
0.78
0.775
0.614
0.888
0.784
0.889
0.892
0.566
0.886
baby boy
baby girl
baby man
baby woman
boy girl
boy man
boy woman
girl man
girl woman
man woman


3-D Embedding: t-SNE, 2 of People
baby boy
baby girl
baby man
baby woman
boy girl
boy man
boy woman
girl man
girl woman
man woman

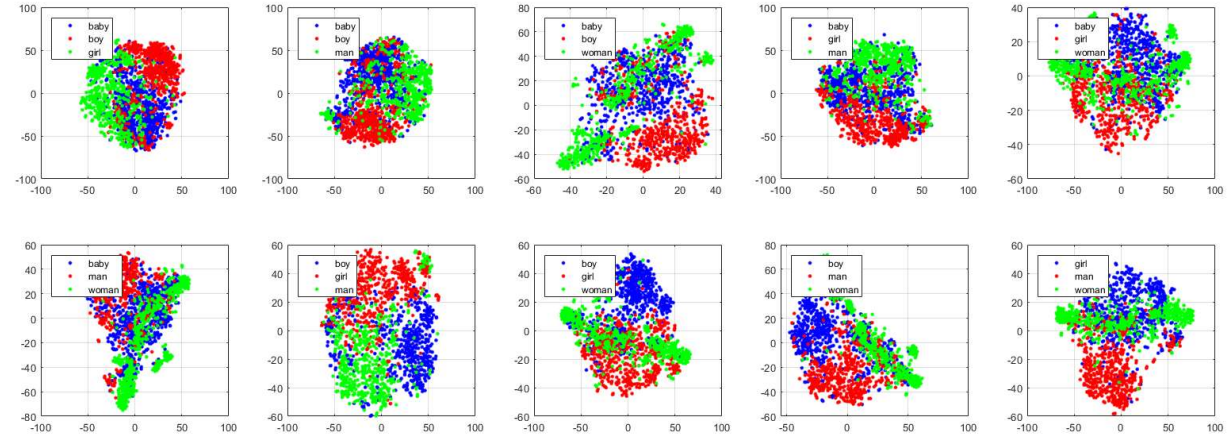
Combinations of 3 concepts



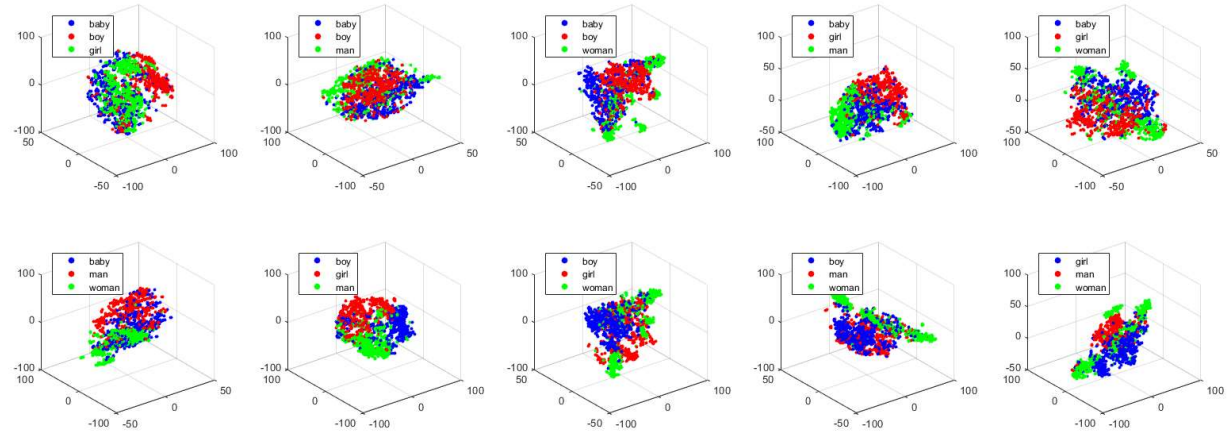
3-D Embedding: Isomap, 3 of People



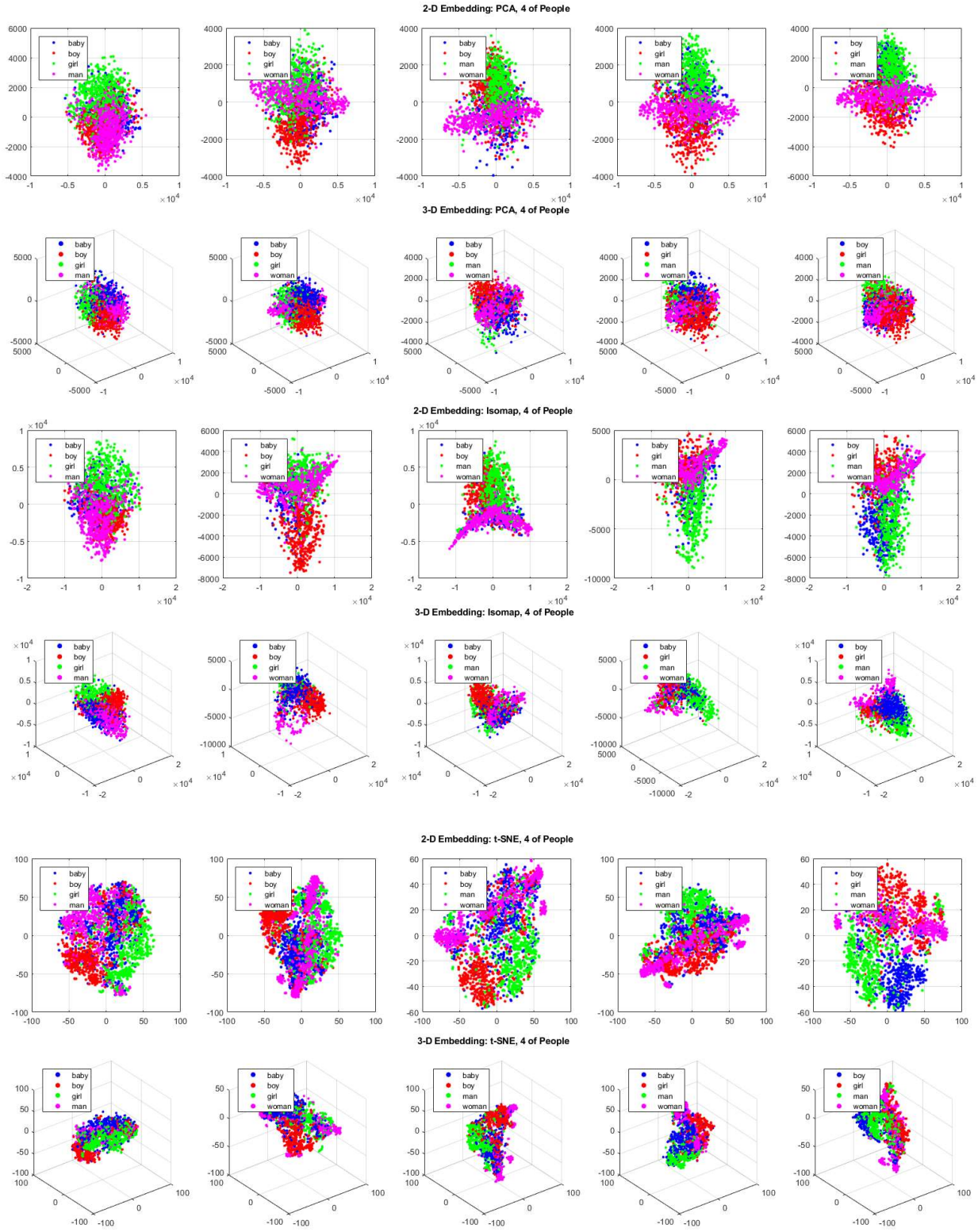
2-D Embedding: t-SNE, 3 of People



3-D Embedding: t-SNE, 3 of People



Combinations of 4 concepts



Combinations of 5 concepts