Human cognition inspired technique for estimation of 2D localization and pose parameters of baggage in uncluttered scenes

Sridhar Sriram

2017

Sridhar Sriram. (2017). Human cognition inspired technique for estimation of 2D localization and pose parameters of baggage in uncluttered scenes. Master's thesis, Nanyang Technological University, Singapore.

<http://hdl.handle.net/10356/69978>

<https://doi.org/10.32657/10356/69978>

**NANYANG
TECHNOLOGICAL
UNIVERSITY**

SINGAPORE

**HUMAN COGNITION INSPIRED TECHNIQUE FOR
ESTIMATION OF 2D LOCALIZATION AND POSE
PARAMETERS OF BAGGAGE IN UNCLUTTERED SCENES**

SRIDHAR SRIRAM

School of Electrical & Electronic Engineering

2017

**HUMAN COGNITION INSPIRED TECHNIQUE FOR
ESTIMATION OF 2D LOCALIZATION AND POSE
PARAMETERS OF BAGGAGE IN UNCLUTTERED SCENES**

SRIDHAR SRIRAM

School of Electrical & Electronic Engineering

**A thesis submitted to the Nanyang Technological University
in partial fulfilment of the requirements for the degree of
Master of Engineering**

Acknowledgement

I would like to thank my supervisor Professor Er Meng Joo for providing me with an opportunity to work closely under his guidance over the course of my Master of Engineering programme. I have thoroughly enjoyed my project meetings and other interactions with him due to his optimistic and friendly personality. His timely words of advice have been invaluable to my progress in research and the successful completion of my degree programme.

I have also had the privilege of learning from and sharing ideas with Ph.D. student Rajasekar Venkatesan and fellow Master of Engineering student Yalavarthi Vijaya Krishna who are a part of his research group.

I would like to thank Mr. Chia Chiu Kiat of the Intelligent Robotics Laboratory for having provided me with all the resources needed to successfully carry out my research.

Last but definitely not the least, I cannot be grateful enough to my parents, sister, brother in law and grandparents for their constant support and encouragement during my studies. They have been my pillars of strength right from my decision to pursue a higher degree abroad and have made me feel at home even as I spend life outside 'home' in the quest of a successful career.

Contents

Acknowledgement.....	2
Abstract	5
List of Figures	6
List of Tables.....	8
Chapter 1 - Introduction	9
Chapter 2 - Object Representation	11
2.1 Object – Centered Representation.....	11
2.2 Viewer-Centered Representation	12
2.3 Vision Techniques.....	13
2.3.1 Object-Centered Computational Models.....	13
2.3.2 Viewer-Centered Computational Models.....	14
2.4 Employed Approach.....	15
Chapter 3–Object Part Saliency & Part Based Attention	17
3.1 Spatial Attention Models.....	18
3.2 Object Attention Models	19
3.3 Part Attention Models	19
3.4 Visual Saliency of Parts	20
3.5 Employment of part saliency and attention to parts in this research.....	22
Chapter 4 – Overview of Part-Based Recognition by Components Model.....	25
4.1 Non-Accidental Properties of Edges	26
4.2 Minima Rule for object decomposition.....	27
Chapter 5- Image Segmentation	30
5.1 Thresholding Techniques	30
5.2 Edge-Based Techniques	32
5.3 Region-Based Techniques.....	33
5.4 Active Contours	34
5.5 Level-Set Formulation.....	35
5.6 Types of Active Contours	36
5.6.1 Edge-based active contours	37
5.6.2 Region-Based Active Contours	37

5.7 Chan-Vese Active Contours.....	38
5.8 Employment of Chan-Vese Active Contours for segmentation in this research.....	40
Chapter 6 – Harris Corner Points	42
6.1 Utilization of corner points in this research	44
Chapter 7–Extreme Learning Machine	46
7.1 Algorithm	47
7.2 Advantages of Extreme Learning Machines	49
7.3 How ELM is employed in this research?	49
Chapter 8–Chamfer Matching.....	51
8.1 Principle of Chamfer Matching.....	51
8.2 Distance Transform.....	52
Chapter 9 - Results & Discussion	56
9.1 Phase 1 – Extraction of Cuboidal Geon Contour	59
9.2 Phase 2 – Estimation of Localization Parameters of Baggage	61
Future Research Directions	65
Conclusion.....	66
Bibliography.....	67
Publications	73
Appendix	74

Abstract

Object localization and its pose estimation in a scene is an extremely important research area in computer vision. In fact, it forms an important first step for performing further manipulation tasks. Several techniques have been proposed in the literature that strive to capture this information. Shape matching is one such technique, as even simple hand drawn contours without the aid of other cues such as color, texture etc. can be used to delineate, identify, localize and compute object pose. In this work, the power and the flexibility of object centered and viewpoint dependent representations of objects from cognitive science are combined to estimate the 2D parameters of baggage in a captured image. Object-centered theories allow us to construe the problem of object representation in terms of its parts. In particular, the recognition by component model represents objects in terms of part primitives referred to as *geons*. The objects are parsed at the points of negative minima of curvature in accordance with the principle of transversality. This concept of object divisibility into parts can be exploited to extract the most visually salient part(s) for pose estimation. In this research, considering that the baggage is a two-part structure comprising of the salient body and the handle, a trained single layer feedforward neural network based on Extreme Learning Machine (ELM) and Harris corner points is used to extract the cuboidal contour of the baggage at fixed scale. The extracted cuboidal contour is then matched with stored cuboidal templates of different poses and aspect ratios using the Chamfer matching technique. This allows for obtaining the best template and the corresponding translation parameters thereby pointing to the integrated approach of object-centered and viewpoint dependent representations for baggage localization and pose estimation. This approach also aligns well with the concept of part saliency, selective attention to and processing of parts from the field of cognitive science in which the top-down influence of the task to be performed (pose estimation) is also taken into consideration. In this research, baggage in uncluttered scenes are considered for the purpose of 2D pose estimation based on which an algorithm is proposed. The success of the algorithm is demonstrated through simulations in MATLAB. Future research in this direction will enable a potential robotics solution for automated baggage handling for which pose estimation is a must.

List of Figures

1. Figure 2-1: Abstracted flow diagram of the baggage pose estimation algorithm.
2. Figure 3-1: a) Main Part.
b) Core Part.
c) Peripheral Part.
3. Figure 3-2: a, c, e) Given Baggage Image.
b, d, f) Extracted cuboidal part contour using our approach that incorporates concepts of part based visual attention as well as saliency of object parts.
4. Figure 4-1: Recognition by Components Theory.
5. Figure 4-2: Two objects on the left. The combined overall object on the right. The dashed contour represents the actual division of object into parts by the human visual system.
6. Figure 4-3: Two negative minima of curvature. Arc indicating object parsing to parts.
7. Figure 5-1: a) Input Image.
b) Segmented Image.
8. Figure 5-2: a) Initial contour represented by a black border.
(b-c) & (d-e) Pairs of given baggage and segmented images using the Chan-Vese active contour approach implemented in MATLAB.
9. Figure 6-1: Demonstration of Harris corner points.
10. Figure 6-2: Explanation of Harris corner points using the sliding window approach.

11. Figure 6-3: a-b) & c-d) Pairs of scene baggage images and corner points extracted on the edges of the segmented images using the Chan-Vese active contour.
12. Figure 8-1: a) Input Binary Image.
b) Euclidean Distance Transform of the Image.
13. Figure 8-2: a) Extracted cuboidal contour from phase 1.
b) Euclidean distance transform of a).
14. Figure 8-3: Computation of Chamfer cost using the pose templates(green) over the distance transform map shown in Figure 8-2 b) with different translation parameters a-c) Fused image showing the target extracted cuboidal geon contour with the translated stored cuboidal geon part contour template. The arrows indicate a large alignment error with a high Chamfer score of 10464, 9074.2 and 10463 respectively d-f) Matching of cuboidal geon template at a different pose with Chamfer scores of 7299.8, 5288.2 and 8439.8.
15. Figure 9-1: Cascading operation of phase 1 and phase 2.
16. Figure 9-2: Pseudocode of the proposed approach.
17. Figure 9-3: a) Example Baggage Image.
b) Image segmentation using Chan–Vese active contour.
c) Detection of Harris Corner Points.
d) Extraction of cuboidal geon contour by finding the convex hull over the corner points which are classified as belonging to the body of the baggage by the trained ELM.
18. Figure 9-4: Variation in the cuboidal contour extraction accuracy for different number of hidden neurons.
19. Figure 9-5: a) Extracted Cuboidal Contour from phase 1.
b)-e) Inaccurate template matches(green) with the cuboidal contour(lavender).
f) Accurate template match yielding the lowest possible score out of stored template database.

List of Tables

1. Table I - Cuboidal Contour Extraction Accuracy.
2. Table II – Pose & Localization Estimation Accuracy.

Chapter 1 - Introduction

Robotics is an important topic for research and studies today due to the increasing need for automating tasks, particularly those that are strenuous or are undesirable for human action. The developed robots need to be equipped with effective sensory and actuation abilities. Vision, being the dominant perceptual modality in humans has much of a role to play in their successful accomplishment of tasks. It is therefore natural that developing and incorporating vision algorithms to create intelligent robotic systems has received a lot of attention by the robotics and computer vision fraternity. Recognizing the significance of vision in robotics, this research work focusses on developing a vision algorithm for estimating the pose of two-part baggage in uncluttered scenes.

Development of vision algorithms for object localization and pose estimation has occupied an important position in computer vision and image processing literature. This follows from the fact that such knowledge forms an important first step in performing manipulation tasks. For instance, in order to successfully grasp an object, one would need to know a priori, the position and orientation of the object in the scene in order to place fingers around it. Typically, the developed algorithms should be feasible to implement and computationally less demanding for use in real time applications. It has also been observed that shape cues are very reliable and sufficient for the purposes of object recognition, localization and pose estimation tasks in humans. Chamfer matching is one such technique which uses shape cues to match template edge maps with a query image edge map by computing the chamfer score. This work uses this technique for estimating the pose of the cuboidal object (baggage) in 2D.

Cues from the object centered and viewer centered representation theories are used in this work. In general, while object centered representation theories propose that objects can be represented in terms of their parts on an object centered reference frame, viewer centered theories propose mental representation of objects in terms of their different viewpoints. These two diverging mental representations of objects have found support in cognitive science and neuropsychological literature. In this work, the ‘best of both

worlds' approach is followed wherein the idea of object divisibility into parts (object centered representation) as well as template matching (viewer centered representation) using Chamfer score computation are combined to estimate the 2D parameters of the two-part baggage in an uncluttered scene. Rather than using the entire object representation for the matching process, the most salient cuboidal contour is extracted thereby pointing to the part based alignment process. This idea of matching parts fits perfectly with the cognitive principles of part saliency and the concept of part based visual attention according to which visual information processing occurs on the basis of parts in which the top down influence of the task to be achieved is taken into consideration.

This work is organized as follows. Chapter 2 reviews the object representation scheme used in this work. Chapter 3 reviews the object part based visual attention model as well as the saliency of different parts of an object as perceived by the human visual system. Chapter 4 reviews the recognition by component model and the minima rule of object parsing drawing inspiration from object centered representation introduced in Chapter 2. Chapter 5 describes the problem of image segmentation and the Chan-Vese active contour model used in this work. Chapter 6 describes the Harris corner points around which visual information is extracted in a spatial window model. Chapter 7 describes the extreme learning machine technique for training the neural network. Chapter 8 describes the template based Chamfer matching technique used in this work building on viewer centered representation of objects from Chapter 2. Chapter 9 describes the specific procedure and the results achieved in this work. A section on future research directions is also included that seeks to focus on the future research scope. The following section concludes this work. Also included in this thesis is the MATLAB implementation of the adopted methodology and the algorithm for 2D pose estimation of baggage in uncluttered scenes under the Appendix section.

Chapter 2 - Object Representation

The concept of object representation has its roots in cognitive science and human inspired studies. The main purpose is to understand how the mental blueprints of objects can be mapped to the current scene containing the object at a particular pose defined by the perspective information. As the topic of representing objects has garnered significant attention, several theories have been proposed in the literature to deal with it. The proposed theories can be categorized mainly into two types – object centered theories, wherein the focus is on a viewpoint invariant description of objects and viewer centered theories, which are ego-centered viewpoint dependent depictions of the object. Naturally, a lot of views of the object need to be stored in the latter case to retrieve the appropriate template corresponding to the current query image. The following paragraphs will review these representations and works pertaining to each representation in greater detail.

2.1 Object – Centered Representation

As mentioned previously, object centered theories are essentially viewpoint invariant. In other words, tasks such as recognition and pose estimation are performed through a structural description of objects such as encoding parts and their spatial relationships. Thus storing an object in multiple viewpoints is not necessary. The theories proposed in literature mainly differ in the way the object structure is brought forward.

[1] presented a scheme of object representation based on 3D generalized cones. Their proposed 3D model was hierarchical as the model parts could be further decomposed into more parts. Recognition, for example, was achieved by matching the scene with the stored 3D model description. [2] presented the description of an object in terms of a limited set of primitives termed as geons (short for geometric ions). These geons were described on the basis of non-accidental properties of edges in an image – collinearity, symmetry, parallelism, curvature and co-termination. As such, their primitives were based on 2D components as opposed to the 3D requirements imposed by [1]. This work focusses on [2] for identifying the cuboidal part in a baggage. However, the recognition by component model by itself does not suggest how the division of an object

into constituent parts can be accomplished. In order to achieve this objective, the minimal rule of division proposed by [3] is to be employed that suggests the division of an object into its constituent parts (geons) at concave points and negative minima of curvature.

Despite the advantages offered by object centered representation, the following are the shortcomings encountered –

- a. Representing objects using merely a small set of primitives may not always account for intra-class variability.
- b. These techniques are slightly difficult to implement using computer vision techniques.
- c. Without the use of rules of decomposition, the way an object can be parsed into parts is ambiguous.

In contrast to the object centered representations, another popular line of thought is the ego-centric or viewer based representations which are defined with respect to the user visual perception of the object. This is discussed in the following paragraphs.

2.2 Viewer-Centered Representation

Viewer centered approaches propose that the object is stored as a combination of viewpoint dependent images in memory. The emphasis in this form of representation is therefore on the user. Typically, recognition and pose estimation is made possible by matching the current scene image and the stored prototype templates.

[4] presented a scheme of object representation in which a novel view point of an object is represented by the weighted linear combination of its different stored views. Thus, rather than dealing with an explicit 3D model representation of the object, several 2D images of the objects from differing viewpoints are used which are related to the current scene by a linear description. [5] presented the recognition by alignment model as a two-step process. In the first step, triplets of model and image feature points are identified and the position, scale and the orientation of the object is determined. This is

followed by the computation of the alignment score to estimate the alignment score between the model edges and the image edges.

On the other hand, direct approaches to object recognition also have been proposed. These techniques although similar to [4] in that picture based comparisons are made between the model and the image, are different in that comparisons are made at an individual level without using a specific combination of views. Thus, tasks such as recognition, localization and pose estimation are achieved by means of scoring the candidate matches and choosing the best matching view template based on the score. Several associative models of memory have been proposed in literature based on this direct approach such as [5], [6].

Viewer-centered techniques have some weaknesses as well, such as the requirement of the availability of a large number of stored templates in memory. Nevertheless, they help alleviate some of the problems encountered in object centered representations.

2.3 Vision Techniques

Object centered and viewer centered theories mentioned previously offer descriptions of how familiarity based recognition, localization and pose estimation take place in the biological world including humans. In the domain of computer vision, it is important to be able to utilize these ideas to create effective recognition/pose estimation strategies. Naturally, the two different opinions on object representation have resulted in similar parallel lines of growth in the computer vision fraternity. They are categorized based on the scheme of representation used and are as follows:

2.3.1 Object-Centered Computational Models

[7] presents a method of matching a 3D object model to a 2D image based on corner point correspondence. The challenge then is to identify which features in the model correspond to those in the image. They exploit the fact that there is a unique transformation (translation, rotation and scaling) between the three model vertices and the three image corners. However, to ensure that proper correspondence is indeed achieved, the transformed model has to be overlaid over the image and the alignment of other

boundaries and vertices is cross checked. If not, correspondence between new pairs of model vertices and image corners is conducted. [8] proposed a computational model for object detection and pose estimation analogous to that in [7]. The difference however lies in the fact that they used perceptual groupings based on property of parallelism rather than solitary corner points for the matching process. A transformation step between the model lines and the image edges similar to [7] is used. This approach greatly improves search efficiency by reducing the pitfalls associated with corner based false matches.[9] proposed the ACRONYM algorithm for recognition by volumetric parts. The advantage offered by this method is the opportunity for object parametrization that allowed for variability within the object class during recognition and pose estimation. [10] proposed superquadrics, which can be thought of as deformations parametrized and achieved on a lump of clay for representing volumetric parts of an object. This flexibility in modelling different part primitives has attracted the attention of several research works ([11], [12], [13], [14], [15]).

On the other hand, Biederman's theory of geon primitives [2] for object representation has also attracted significant interest in the computer vision community allowing them to exploit this concept for the purpose of object recognition & pose estimation. Examples of works incorporating geons are [16], [17], [18], [19], [20], [21]. As the focus of this research is on developing a vision system for handling baggage which are primarily cuboidal in nature, the concept of geons offer a convenient way to think, identify and represent cuboids from 2D images and has therefore been used in this work.

2.3.2 Viewer-Centered Computational Models

[22] presented an intelligent approach to the recognition and pose estimation of objects from a multiple objects-multiple views database based on the projection formulation proposed by [23]. Each image is represented as a point in an n -dimensional space by rewriting its pixel values as a row vector ($1 \times n$), where n is the number of pixels in the image. The n -dimensional coordinate system is now replaced by a new one wherein the position of each point (image) can be sufficiently approximated by a smaller number of coordinates. This ensures the reduction in the computational complexity in object recognition. As the camera viewpoint keeps on changing, the object points trace out a

characteristic curve that is different for different objects. The process of recognizing an object then becomes a two-step process. First, project the view into the new coordinate system. Find the closest object specific curve that reveals the object identity. The focus then is on identifying the nearest point on this curve to the current view which reveals the pose of the object in the scene.

2.4 Employed Approach

This work employs the Chamfer matching algorithm[24] in a direct approach to pose estimation of cuboidal objects in a scene wherein predefined pose templates are shape matched with the current scene and a chamfer score is computed for each match that represents the degree of match between the current target scene and the template. The best fitting template is then chosen as the one which yields the lowest chamfer score. This work is also in accordance with the Recognition By Component (RBC) theory proposed by [2] which states that an object can be decomposed into geon primitives. This is done to ensure that the most salient cuboidal geon part contour is extracted for pose estimation thereby reducing the number of unnecessary edge points to be matched. The used shape templates for matching are therefore not the entire object but cuboidal part edge templates at predefined pose and scale. To this end, the approach followed can be treated as one using a combination of both object centered as well as viewer centered representations. The detailed procedure will be provided in a separate chapter later. The simplistic flow diagram of the procedure is given in the following page.

Figure 2-1: Abstracted flow diagram of the baggage pose estimation algorithm.

Chapter 3—Object Part Saliency & Part Based Attention

In this chapter, we introduce the concepts of visual attention and object part saliency as proposed in cognitive science literature. Specifically, a formal definition of attention is provided followed by different models that capture this concept. These models can be broadly divided into three categories, namely – spatial, object based and part based attention theories. In reality, while these models are not mutually exclusive in that neither of them operate in isolation, the part based attention theories offer a convenient way of thinking about the mental representation of objects. This also agrees with many of the object recognition theories proposed in literature such as the RBC theory and therefore forms an important aspect of this research.

Visual Attention: Visual attention refers to the selective processing of information extracted through the imaging modality (‘eyes’ in humans) while ignoring the rest which can include irrelevant/distracting stimuli. Several factors influence where we look and what information is processed. These factors include both bottom-up as well as top-down features. Task dependencies also play a significant role in influencing what information is perceived.

The famous psychologist James has the following to say about visual attention:

“Everyone knows what attention is. It is the taking possession by the mind, in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought. Focalization, concentration, of consciousness are of its essence. It implies withdrawal from some things to deal effectively with the others.

When things are apprehended by the senses, the number of them that can be attended to at once is small.”

This famous statement essentially conveys the principle and the significance of visual attention. Humans are finite creatures who cannot attend to the ocean of information available around them. It is important to focus our mental capacities to focus

on a subset of the sensory input whatever the modality may be. In the case of vision, it refers to interesting artifacts in a scene so that the brain may pay more attention to these aspects than the less interesting ones.

Given this definition/interpretation of visual attention, several theories have been proposed in literature to explain the phenomenon of visual attention. These theories differ in their definition of the region of interest. Traditional models of visual attention are essentially of the view that selection happens in a spatial representation of the visual field. More recently however, research on visual attention has revealed that groupings or objects are the focus of visual perception thereby making objects the fundamental units. This has naturally left many psychologists to wonder if the objects are by themselves the units of attention or if the object can be further divided into smaller units or parts some of which by their visual salience in conjunction with the task at hand, attracts attention. This is the part-centric approach to visual attention.

In the following paragraphs, works that pertain to each of the three categories – spatial, object and part based attention theories are listed. Particular emphasis is given to part based models as it forms the inspiration for this research in which the principle of part based visual attention in conjunction with saliency is used to extract the cuboidal geon contour for pose estimation in an image.

3.1 Spatial Attention Models

According to this approach, attention is directed to particular locations in the visual scene. One of the earliest works in support of the space based attention model was that by [25] in which cues are employed to vary the spatial attention across a scene in the world. Two conditions are examined - valid and invalid. In the valid case, targets appear at the location where the cues are displayed. In the invalid case, targets appear at different locations than the cued locations. It was found that the response times for valid cases was much faster than the invalid trials thereby pointing to the space based nature of visual attention. Treisman's Feature Integration Theory (FIT) [26] is an excellent model that offers a wonderful account of visual attention. According to the theory, the visual scene is divided into a number of low level feature maps in several dimensions. These feature

maps are integrated to form the master map that dictates the saliency of a scene. Other competing models under this technique include Wolfe's guided search model [27], Eriksen's zoom lens model [28],[29], Koch and Ullman's saliency map model [30].

Based on these space-based theories several computational models have also been formulated. [31] presented a model to evaluate how different a point in an image is from its surroundings in terms of its saliency. A winner-take-all neural network is used to govern the selection process analogous to the visual attention. [32] presented a model of visual attention based on selective tuning.

3.2 Object Attention Models

Unlike the spatial model, object attention models propose that objects/perceptual groupings are the focus of visual selection. That is, objects are directly selected as discrete elements rather than contiguous regions of space as in the space based model. All visual features of the attended object are processed at the same instant and much faster than the features of other objects. [33] concluded based on their experiments that observers were more accurate at reporting pairs of features from the same object than two different objects thereby pointing to the object nature of visual attention. This reduction in accuracy in the two object case cannot be attributed to spatial factors as the objects under consideration in this experiment were taken as overlapping with each other and consequently occupied the same position at any given time. Other pioneering research in object based attention can be found in the works of [34], [35], [36]. A detailed review of the object based attention models can be found in [37].

A comparative analysis of the space based and object based attention models can be found in [38]. Of late, however, a new model of attention on the basis of object parts has gained importance. The following section summarizes the part based formulation as it is used in this research.

3.3 Part Attention Models

In the previous section, an attentional mechanism that chooses objects as the units of selection was presented. However, there are increasing evidences in cognitive science

literature such as [1], [2] that the human visual system parses objects and stores them as part representations. Given this understanding, it is natural to think of attention as one of part selection in which one part is selectively attended to at any point in time.

A pioneering work in this direction to investigate part based visual attention was undertaken by [39]. The authors argue that a scene can be decomposed hierarchically into different components. Just as how an object can be selectively attended from among several other objects in a scene, within an object some part is selectively processed from a myriad of object parts at a given time. It must be noted that an object is parsed is at the points of negative minima of curvature or concave cusps. A similar spirit is embodied in this research in which the cuboidal contour of the visually salient object part is extracted for the purpose of pose estimation. Thus, the task to be achieved (that of pose estimation) is also taken into consideration which is a top down influence. The concept of object part visual saliency is explained in the following section as proposed in cognitive science literature.

3.4 Visual Saliency of Parts

[40] introduced the concept of part saliency which is a measure of the usefulness of the part as an index for recognition, pose estimation and other related tasks. This work therefore not only provides further support for a part-centric scheme of object representation but also demonstrates the relative importance of parts. The concept of part salience applies not only to the 2D case but also in general to the 3D objects in a scene. Their method of parsing an object into parts is according to the minima rule in which the object is divided at the points of negative minima of curvature. They introduce the main factors that influence the saliency of a part. They are listed as follows:

1. Relative size of part– Area and Volume.
2. Protrusion.
3. Part boundary strength.

Accordingly, the parts of a 3D object such as baggage are classified as major parts, core parts and peripheral parts. The following points are in general true of object parts -

1. The salience of a part increases in general with increase in area and volume of the part relative to the whole object.
2. The salience of the part increases with increase in the protrusion of the object part measured as the ratio of the sum of its boundaries (excluding part cuts) to the sum of its part cuts.
3. The salience of a part boundary increases with increase in the mean turning angle averaged over the visible portion of the boundary.

The paper also gives the following definitions of the different types of parts that an object can have –

1. Main part – This part is the largest of all the parts of an object. Its saliency is primarily determined by its size and protrusion. An example of a main part is shown in figure 3-1 a.
2. Core part – This is a part which has multiple part cuts. Its saliency is determined mainly by its size. An example of a core part is shown in figure 3-1 b.
3. Peripheral part – A peripheral part is characterized by a single part cut. Its saliency is governed by all three factors – size, protrusion and the part boundary strength. An example of a peripheral part is shown in figure 3-1c.

Although the images shown here are 2D outlines, the idea of different part types can be easily extrapolated to 3D objects captured using 2D images as well and is employed in this research (as baggage is a 3D object which is captured using images).

a)

b)

c)

Fig 3-1: a) Main Part b) Core Part c) Peripheral Part

3.5 Employment of part saliency and attention to parts in this research

In this research, we combine the principles of part centered visual attention proposed by [39] (in which parts of an object are selectively attended) with that of part saliency proposed by [40] to extract the most visually salient cuboidal contour of the baggage ignoring the less salient handle of the baggage for 2D pose estimation. This is achieved using the Extreme Learning Machine (ELM) technique. Figure 3-2 shows the extraction of the visually dominant cuboidal body contour of the baggage in an uncluttered scene given an uncluttered scene containing the baggage thereby also conforming to the part based attention theory that has gained increasing importance in cognitive science literatures.

a)

b)

c)

d)

e)

f)

Figure 3-2: Fig a, c, e) Given baggage image. Fig b, d, f) Extracted approximate cuboidal part contour of respective baggage, using our approach that incorporates concepts of part-based visual attention as well as saliency of object parts.

This extraction of cuboidal geon part contour is therefore significant and advantageous for the following reasons:

1. Only the most salient object part is extracted for further template matching operations and distance map computation thereby eliminating unnecessary edge points belonging to the non-salient parts.
2. Unlike some of the earlier works wherein the entire object is shape matched for tasks such as pose estimation, it is useful to decompose an object into parts and then use the most salient object part for pose estimation which also abides

by the principle of human vision parsing of objects at points of negative minima of curvature as demonstrated in the RBC theory and other cognitive science literature.

Chapter 4 – Overview of Part-Based Recognition by Components Model

In chapter 3, the concept of object part visual attention that suggests that object parts are selectively attended as well as part saliency which proposes that the saliency of an object part is a function of several parameters including area/volume was discussed. In this chapter, the recognition by components model (RBC) of [2] is introduced which is an object centered representation framework that provides further evidence for the mental parsing of the object in a scene by the human visual system. Furthermore, it also provides for the concept of geons which are part primitives obtained by the decomposition of an object. As seen in section 3.5 in chapter 3, the salient cuboidal part contour that is extracted and used for subsequent 2D pose estimation is one of the geon parts [2] listed in this work.

RBC theory [2] states that the human vision system perceives count objects by decomposing them into parts which are referred to as geons. That is, there are standard geons that serve as the primitive components, combinations and different spatial arrangements of which would produce the objects we see and recognize. The geons presented are cylinders, cones, wedges etc., which are indivisible and are based on the five non-accidental properties of edges – curvature, collinearity, symmetry, parallelism, and cotermination.

The process works as follows. Given a retinal image of an object, the human visual system parses the object into different parts based on the principle of concavities, a term referred to as transversality in [3] in their minima rule for partitioning silhouettes. The result of this parsing is that we end up with convex volumes (geons), all of them in specified spatial relationships which form the basis for object recognition. An example is shown in figure 4-1 which depicts the lower order components (geons or primitives) on the left hand side and the final objects formed by their proper spatial positioning on the right hand side. It can be seen in the figure how the different part constituents (or geons) can be combined to form the objects of interest. Specifically, for the purposes of this

research, it can be seen from the figure how the handle and the cuboid (the two geon primitives) combine to form the baggage object. Note also the differences in saliency of these two geons which constitute the baggage. In this research however, we follow the reverse approach in which the baggage in an uncluttered scene forms the starting point for subsequent extraction of the visually more salient cuboidal contour for 2D pose estimation / localization thereby ignoring the less salient handle geon of the baggage.

Figure 4-1: Recognition by Component Theory

It must be emphasized at this stage that this research is not a faithful implementation of the recognition by components model. It however uses the idea that a given object can be divided into its parts at the points of negative minima of curvature and concave cusps. By using this concept, it is possible to extract the most salient cuboidal contour for obtaining the estimate of its location and orientation in a scene while ignoring unnecessary edge points corresponding to non-salient parts of the baggage (the handle in this case) during the matching process. It therefore leads us to the concept of a part based matching process for estimation of pose parameters.

4.1 Non-Accidental Properties of Edges

The following properties of edges hold when projected from the 3D world into a 2D image scene. These properties are therefore termed as ‘non-accidental’ because they remain fairly unchanged by changes in viewpoint.

- a. Collinearity – A straight line in the two-dimensional world corresponds to a straight line in the real world. That is, it is highly unlikely for other features

such as curved segment to project such a straight segment in the image and that its likely cause is a straight segment in the 3D world.

- b. Curvature - A curved segment in an image is a result of the projection of a similar feature in the three-dimensional world and not due to features such as straight lines.
- c. Symmetry – If the image is symmetrical, then the object producing the image in the real world is also symmetrical.
- d. Parallelism – Two non-intersecting and parallel lines in the real world also exhibit similar characteristics when projected on to the two-dimensional image.
- e. Co-termination – Edges which co-terminate in the real world also co-terminate in captured image.

These properties enable identification of specific geons from a myriad of possibilities given an image of the object.

4.2 Minima Rule for object decomposition

Although the RBC theory introduces the concept of object division into its constituent parts (geons), it is not self-sufficient as it does not suggest where the object has to be exactly parsed so as to obtain these primitives. It relies on the results of [3] to provide the points of deep concavity at which to segment an object.

The premise behind the theory of establishing specific points or regions of object parsing as proposed by [3] is that of the regularities of nature. In other words, the choice of parts defined by a human visual system is not arbitrary but follows a specific pattern or rule. This is pretty obvious as otherwise there would have to be an infinite storage of parts all of which are unique to each individual which would result in lack of commonness and wastage of memory. It is therefore that a more general uniformity be in place that can account for the mental parsing of objects. This leads the discussion into the principle of transversality.

Transversality – The principle of transversality states that “when two objects are made to interpenetrate, they always intersect each other in contour of concavity of their tangent planes.” (as stated by [3]).

Let us consider the figure 4-2 to understand the concept even better. Consider the silhouettes of the two separate 3-D object parts in the left. These object parts are combined to form the overall silhouette shown in the figure on the right. By the principle of transversality, the new object parts according to the parsing by the human visual system is given by the dotted line in the figure.

Figure 4-2: Two objects on the left. The combined overall object on the right. The dashed contour represents the actual division of object into parts by the human visual system. (Image Courtesy:

A similar spirit of object parsing can be extended to baggage as well. Assuming the baggage to be comprising of two parts – cuboidal geon and the handle geon, the baggage is parsed at line indicated by the dashed contour which represents the intersection of the handle and cuboidal geon. An example of baggage parsing is shown in Figure 4-3 considering a binarized image for easier representation and visualization. The two points of negative minima of curvature are also marked in the figure.

Fig 4-3: Two negative minima of curvature. Arc indicating object parsing to parts.

In order to account for a more generic object division into parts the minima rule was proposed. The minima rule can be formally summarized as “Parse an object into its components at the points of negative minima of curvature or concave cusps”. This concept is implemented in this research through the extraction of a window of information around each Harris corner point (to be discussed in Chapter 6) which then serves as the input to a trained neural network for extracting the salient cuboidal contour (in a two-part baggage comprising of the handle and the body). This is subsequently used for the computation of the Euclidean distance transform map for the purpose of pose estimation and localization in a 2D image.

Chapter 5- Image Segmentation

The goal of image segmentation is to group/cluster pixels of an image into different categories based on some criterion. In doing so, it is also possible to separate the foreground or region(s) of interest from the background. However, it is important to note that the process of segmentation although fundamental is not a trivial one. An example of image segmentation is shown in figure 5-1. Several techniques have been proposed in literature to deal with the issue of segmentation. These techniques can be broadly classified as – threshold based, edge based and region based techniques. Strictly speaking, edge based and region based techniques are duals of each other. Detecting edges in the image will help identify the region bounded by the edges whereas identifying the regions within an image will invariably provide the edges of the region. In the following sections the threshold based, edge based and region based segmentation techniques will be introduced which is succeeded by a comprehensive explanation of active contours and the Chan-Vese active contour specifically used in this work.

Figure 5-1: a) Input Image b) Segmented Image

(Image Courtesy: matlabtricks.com)

5.1 Thresholding Techniques

It is the simplest approach to image segmentation in which the pixels are allocated into different categories based on the range of values held by the pixels. This is made

possible by choosing a value of the threshold and comparing the value of each pixel with reference to the threshold.

Let I be a grayscale image, and t denotes the threshold chosen. Also, f is the function mapping the pixel location to grayscale intensity value. Then, for the case of a binary partition problem, a pixel at coordinates (i, j) is allocated to category 1 if $f(i, j) \leq t$, otherwise it belongs to category 2. Similarly, multi-category partition can also be achieved by having multiple thresholds.

Different strategies have been proposed in literature over the choice of the threshold(s). Although it is possible to choose the thresholds manually on a trial and error basis, automatic methods of selection have been proposed in literature. The proposed algorithms in literature can be further classified as – 1. Histogram based thresholding for intensity images 2. Manual and automatic classifiers for multi-variate images 3. Context based thresholding algorithms which incorporate neighborhood values as well as histogram of pixel values information.

Under the histogram based thresholding, [41] proposed inter means algorithm for obtaining the threshold by iteratively adjusting it. The threshold is computed as the mean of the mean computations of the two regions of pixel values starting with an initial arbitrary chosen value. [42] proposed the minimum error algorithm which is an extension and generalization of the inter-means algorithm. For multi-variate images, the technique of linear discrimination based on the Mahalanobis distance is employed when the variance is assumed equal. In the event this is not so, the problem becomes one of quadratic discrimination. This is however a manual method of classification as it relies on a training set and is therefore a supervised mode of learning. An automatic model of pixel segmentation is also possible through the use of the k-means clustering algorithm which is an example of an unsupervised learning mechanism and is an automatic version of the linear discriminant problem. Context based algorithms take the neighborhood of the pixel in question while performing the segmentation process. Under this approach, a manual post classification smoothing by using a majority filter that reassigns the pixel category as the most common group under a square window of predefined size with the pixel (i, j) as the center was proposed. An automatic version of the post classification smoothing was

then subsequently proposed to iteratively tune the threshold until it does not change over successive iterations. This proposed algorithm was also compared with the iterated conditional modes algorithm.

It must be emphasized however that thresholding mechanisms for segmentation are often ineffective, particularly, when automatic means are employed. Manual intervention is still necessary which reduces the efficiency and effectiveness of the thresholding process. This calls for the study into a new category of edge detection based segmentation techniques.

5.2 Edge-Based Techniques

Edge refers to the demarcation between two regions with distinct properties. Edges are important because they significantly reduce the data to be processed without any loss of significant information. They are characterized by local intensity changes. They are typically used in many object recognition, pose estimation and higher level image analysis algorithms. The assumption made by these techniques is that the constituent sub-regions are uniform so that the discontinuities actually signal the transition from one sub-region to another. When this does not hold, region based segmentation techniques provide more accurate results than the edge based counterparts.

As mentioned before, employing edge based segmentation techniques refers to the application of methods that can delineate the contours of objects. Generally, edge filters are used to obtain these demarcations such as the Prewitt edge detector, Sobel filter, Canny edge detector, Laplacian of Gaussian, Robinson filter, Marr-Hildreth filter that convolve an intensity image with masks of a pre-defined size in the horizontal and the vertical direction. There are no hard criteria for evaluating the performance of the proposed edge detection algorithms in literature but is in fact application dependent.

The advantages offered by these edge based approaches are the low computational cost involved and the simplicity of implementation. However, despite these benefits offered, edge based techniques suffer from two major shortcomings:

1. Edges can be detected only if they are characterized by rapid changes in intensity.

2. Edge detection techniques are highly susceptible to noise and therefore require smoothing. However, smoothing the image also blurs out the edges which could induce problems while using the edge detection algorithms.

5.3 Region-Based Techniques

An alternative way of looking at segmentation is to look for specific regions or classes within an image. This is possible by grouping pixels into clusters in an image based on certain shared characteristics such as color, texture, intensity etc. A region is therefore a group of connected pixels with similar properties. The grouped pixels also exhibit spatial uniformity in that pixels in a group form a single connected component. The concept of a region is important as it can be used to represent objects in a scene. Consequently, for correct image interpretation, it is imperative to partition an image into regions that represent useful entities or objects of interest. Generally, this is achieved on intensity images in which each pixel has an associated gray value.

The following principles are employed in the division of an image into regions –

- a. Similarity – The variance or difference in the gray values of the pixels.
- b. Location – Whether the pixels are compactly packed by using a distance measure such as the Euclidean distance between the pixels.

Typically, the region based segmentation techniques employ one of the following strategies in clustering pixels together into separate groups -

- a. Merge – Merging techniques group pixels based on some inclusion criteria. For example, seeded region growing technique[43] starts with an initial number of pixels (seeds) and culminates in the grouping of the image space into fairly homogenous regions based on some pixel value properties.
- b. Split – Split techniques work on the basis of dividing the image into different regions.
- c. Split and Merge – Split and merge techniques combine the power and flexibility of both the split and merge operations to create effective

segmentation of images. Examples of such split and merge algorithms can be found in works of [44],[45],[46].

Of late, active contours have found particular interest as a segmentation tool and will be discussed in detail. As we shall see active contours can be both edge based or region based. Owing to the advantages offered, it is also the principle underlying the image segmentation framework used in this research. Specifically, the Chan-Vese active contours that operate on the principle of minimizing intra-group variance of image intensity will be employed for segmentation in this work. It can therefore be categorized as a region based segmentation technique.

5.4 Active Contours

Active contours offer an effective way of delineating an object outline in a 2D scene image. They offer a novel way of looking at the segmentation problem. Specifically, they formulate the problem as one of constraint optimization. The concept of active contours originated with the advent of ‘Snakes’, an energy minimizing model proposed by [47]. It falls under the category of boundary (edge) based active contours. Ever since, research in active contours has increased significantly to the point that it is being increasingly used in object detection, motion tracking, image segmentation etc. Since their inception, active contours have followed one of the following two paths – snakes or level sets and have employed one of the following two forces for contour evolution – edges or regions. Edge based active contours use image gradients in the stopping function to attract the contour to region edges and to inhibit further contour evolution. As such, the object to be segmented is characterized by the properties of its contour only. On the other hand, region based active contours follow the statistical information of image intensities in each sub-region instead of explicitly finding geometric boundaries. This statistical information is also called as the descriptors and can refer to the mean or variance computation which is used in the optimization function.

As mentioned before, the goal of snakes is to minimize an energy function which is the sum total of the internal and the external energy of the contour, in order to locate the edge/boundary of the object.

Mathematically, the contour C with arc length s can be expressed as:

$$C(s) = \{x(s), y(s)\}, 0 \leq s \leq L \quad (5.1)$$

Where L is the length of the contour.

The total energy possessed by the contour can be defined as:

$$E = E_{int} + E_{ext} \quad (5.2)$$

$$\text{Where } E_{int} = \int_0^L \alpha |C'(s)|^2 + \beta |C''(s)|^2 ds \text{ and } E_{ext} = \int_0^L E_{img}(C(s)) ds$$

Here, α controls the tension of the contour, β controls the rigidity of the contour. $E_{img}(C(s))$ is the function responsible for attracting the contour to the edges. This edge attraction function is usually defined on the image gradients.

The snakes were an important starting point. However, they suffered from the following shortcomings:

- a. The classic snakes provide an accurate location of the edges only if the initial contour is given sufficiently near the edges because they make use of only the local information along the contour.
- b. Snakes cannot split to multiple boundaries or merge from multiple initial contours.
- c. Snakes and other boundary based techniques in general, work only if the object and the background differ significantly in contrast or image intensity so that the stopping function tends to zero.

Besides the regular parametrized versions of the snakes, implicit models of contour evolution based on level sets have also been proposed in literature. The following section briefly describes the concept of level sets.

5.5 Level-Set Formulation

An alternative method of formulating active contours is based on level sets. Level set based active contours are also called geometric active contours where as the ones based on the snake model discussed previously are also referred to as the parametric active contours. It must be emphasized that the change in representational strategy from snakes to level sets does not imply that the guiding forces that propagate contour motion vary. The same forces can be chosen to guide the contour evolution.

Level set description for active contours was introduced by [48]. According to the level set theory, a contour is represented implicitly as a smooth (usually Lipschitz) function on the domain of the image, $\phi(x, y)$. In particular, a specific level of the function (usually the zero level) is taken as the contour C . This is in contrast to the snakes wherein the contour points serve as representatives of the contour. To this end, level sets can be regarded as a computational trick that minimizes the hassles associated with the method of snakes by handling any topological complexities and changes naturally.

Mathematically, a level set function is expressed as:

$$\begin{aligned}\phi(x, y) &> 0 \quad \forall (x, y) \in \Omega \\ \phi(x, y) &< 0 \quad \forall (x, y) \in \Omega \\ \phi(x, y) &= 0 \quad \forall (x, y) \in \partial\Omega = C\end{aligned}$$

The contour C (also referred to as the interface or level curve) which is to be evolved is therefore represented in terms of level sets as:

$$C = \{(x, y) \mid \phi(x, y) = 0\} \quad \forall (x, y) \in \partial\Omega \quad (5.3)$$

It is this interface information which needs to be captured by computing the set for which the function ϕ vanishes. The evolution of the contour C is analogous to the evolution of the level set function $\phi(x, y)$ i.e. $\partial C / \partial t = \partial \phi(x, y) / \partial t$. Several contour evolution algorithms have been proposed in literature. The reader is referred to additional material such as papers and online resources for a detailed understanding of level sets.

The Chan-Vese algorithm based on level-sets is explained in a separate section in this chapter as it is implemented in MATLAB and forms an important part of image segmentation employed in this research.

5.6 Types of Active Contours

The proposed active contour approaches for segmenting images in literature can be broadly classified into two categories:

1. Edge based active contours.
2. Region based active contours.

5.6.1 Edge-based active contours

As the name suggests, the basic principle behind the edge based active contours rely on the edge detection stopping functions to guide and stop the contour evolution process. That is, the contour movement stops when the stopping function (a function of the image gradient at a point) vanishes. This is possible by choosing the stopping function to be inversely proportional to the gradient at a point in the image. A large gradient at a point signals the presence of an edge and therefore the contour movement is halted. Examples of edge based active contours proposed in literature include [49],[50].

Despite the advantages offered, the following are some of the general limitations imposed by the application of edge based active contours to the problem of segmentation:

- a. They are not effective when used for detecting blurry boundaries which implies low contrast between the object and the background.
- b. They are sensitive to local minima and noise in the image.
- c. In practice, the gradient based stopping function is never zero and the curve may evolve beyond the actual boundary of the object.
- d. Prior knowledge is required while positioning initial contours in the edge based active contour approach.

5.6.2 Region-Based Active Contours

More recent work has focused on region based techniques such as the work on region flows inspired by the region based competition introduced in [51]. Other important works assume that the image regions to be separated i.e. the foreground and the background are of constant intensity such as [52], [53], [54], [55] of which [52] will be used in this work.

In general, region based techniques offer the following main advantages over the edge based active contour models:

- a. The initial contour in the region based approaches can be placed at any arbitrary position in the image.
- b. They do not suffer from local minima problems.
- c. Less manual intervention and knowledge of the scene image is required than edge contour techniques.
- d. Highly robust when compared to edge detection active contour techniques.

5.7 Chan-Vese Active Contours

Chan-Vese active contours [52] is a region based segmentation technique that debunks the idea of edge based stopping function for contour evolution. They instead develop an active contour model that can detect objects in an image that are not necessarily defined by a gradient. Their proposed model is based on the Mumford Shah segmentation[56] and level sets, in which an energy function is minimized. It uses variational calculus to evolve the level set function. Variational methods work by seeking a level set that minimizes some function. In the case of the Chan-Vese approach, it is referred to as the “fitting energy” functional.

To understand the approach, let us consider the case of minimizing the fitting energy in an image comprising of the object in a background. The two regions to be segmented are assumed to be of piecewise constant intensities of distinct values u_0^i and u_0^o . The contour is represented by C and c_1, c_2 represent the averages of the images within and outside the contour C .

The fitting energy of the curve C is then given by:

$$F_1(C) + F_2(C) = \int_{inside\ C} |u_0 - c_1|^2 dx dy + \int_{outside\ C} |u_0 - c_2|^2 dx dy \quad (5.4)$$

It can also be seen that the boundary of the object is the minimizer of the fitting energy associated with the curve. In other words,

$$\text{Inf}_C \{F_1(C) + F_2(C)\} \approx 0 \approx F_1(C) + F_2(C)$$

The above concept is extended in developing the segmentation model by adding regularizing terms such as the length of C and / or the area enclosed by C . The fitting energy now can be expressed as:

$$F(C, c_1, c_2) = \mu \cdot \text{Length}(C) + v \cdot \text{Area}(C) + \lambda_1 \int_{\text{inside } C} |u_0 - c_1|^2 dx dy + \lambda_2 \int_{\text{outside } C} |u_0 - c_2|^2 dx dy \quad (5.5)$$

where $\mu > 0$, $v \geq 0$, $\lambda_1, \lambda_2 > 0$ are the assigned weights.

Generally, in most of the computations v is taken to be 0 and $\lambda_1 = \lambda_2$. Area weight v , for instance, is used only in the event we want our curve to move only inwards. The problem then becomes one of minimization over the fitting energy mentioned in (5.5).

The Chan-Vese active contour model can be efficiently formulated and solved using the level set method introduced by [30]. Since the evolving curve is represented by the zero level of the Lipschitz function ϕ , we can replace C and rewrite (5.5) in terms of the variable ϕ as,

$$F(\phi, c_1, c_2) = \mu \cdot \text{Length}\{\phi=0\} + v \cdot \text{Area}\{\phi \geq 0\} + \lambda_1 \int_{\phi \geq 0} |u_0 - c_1|^2 dx dy + \lambda_2 \int_{\phi < 0} |u_0 - c_2|^2 dx dy. \quad (5.6)$$

Using the definition of Heaviside function H ,

$$H(x) = \begin{cases} 1, & \text{if } x \geq 0 \\ 0, & \text{if } x < 0 \end{cases} \quad (5.7)$$

and the Dirac measure δ in one dimension centered at 0 given by,

$$\delta(x) = \frac{dH(x)}{dx} \quad (5.8)$$

the fitting energy function given by (5.6) can be further expressed as,

$$F(\phi, c_1, c_2) = \mu \int_{\Omega} \delta(\phi) |\nabla \phi| dx dy + v \int_{\Omega} H(\phi) dx dy + \lambda_1 \int_{\Omega} |u_0 - c_1|^2 H(\phi) dx dy + \lambda_2 \int_{\Omega} |u_0 - c_2|^2 (1 - H(\phi)) dx dy \quad (5.9)$$

With ϕ fixed and minimizing the energy $F(\phi, c_1, c_2)$ with respect to the constants c_1 and c_2 , we obtain the values of these constants c_1 and c_2 as :

$$c_1(\phi) = \frac{\int_{\Omega} u_0 H(\phi) dx dy}{\int_{\Omega} H(\phi(x, y)) dx dy}, \quad c_2(\phi) = \frac{\int_{\Omega} u_0 (1 - H(\phi)) dx dy}{\int_{\Omega} (1 - H(\phi(x, y))) dx dy} \quad (5.10)$$

In a similar way, keeping c_1 and c_2 fixed, and minimizing the energy with respect to ϕ , we obtain the Euler Langrange equation for ϕ given by,

$$\begin{cases} \frac{\partial \varphi}{\partial t} = \delta(\varphi) \left[\mu \operatorname{div} \left(\frac{\nabla \varphi}{|\nabla \varphi|} \right) - v - \lambda_1 (u_0 - c_1)^2 + \lambda_2 (u_0 - c_2)^2 \right] & \text{in } \Omega \\ \frac{\delta(\varphi)}{|\nabla \varphi|} \frac{\partial \varphi}{\partial n} = 0 & \text{on } \Omega \end{cases} \quad (5.11)$$

For practical reasons, however, regularized versions of H and δ are used. Specifically, second approximations are used instead of first approximations as proposed in [57] for the following two reasons:

- a. The energy function is non-convex and the first approximation may not produce a global minimizer as it has a small compact support.
- b. The interior contour is automatically detected.

5.8 Employment of Chan-Vese Active Contours for segmentation in this research

In this research, the MATLAB implementation of the Chan-Vese active contour is achieved. As in the case of any active contour, an initial contour is to be chosen based on which further contour evolution takes place until the final segmentation of the object (baggage) is achieved. The initial contour is chosen from near the image frame and stored in the mask variable as shown in division A in the appendix section. The contour therefore moves inwards subject to the maximum number of iterations specified and latches onto the baggage in the scene resulting in a binarized segmented image. An example is shown in Fig 5-2.

a)

b)

c)

d)

e)

Fig 5-2: a) Initial contour represented by a black border (b-c) &(d-e) Pairs of given baggage and segmented images using the Chan-Vese active contours implemented using MATLAB.

Chapter 6 – Harris Corner Points

In this chapter, we discuss the Harris corner points[58] around which visual information is extracted in this work. Specifically, the process of segmenting and binarizing an uncluttered scene image through the Chan-Vese active contour approach discussed in the previous chapter is followed by the identification of salient points (Harris) on the baggage edges which are characterized by drastic intensity changes.

Consider an example image shown in the figure 6-1. As we shall see, the regions marked by the red circles are characterized by drastic intensity changes. This concept can be best understood by using the window approach as seen in Figure 6-2. By simple observation we find that moving the first square window in a random direction would result in a drastic change in intensity. This is the corner by definition. Moving the second square shaped window however, would result in a drastic change in intensity only along the direction perpendicular to the line. This characterizes an edge. Lastly, moving the third square by a small amount does not result in an intensity change. This is therefore a region of homogenous intensity.

Figure 6-1: Demonstration of Harris Corner Points

This change in intensity as a result of the shift by an amount $[u, v]$ can be represented mathematically as:

$$E(u, v) = \sum_{x,y} w(x, y) [I(x+u, y+v) - I(x, y)]^2 \quad (6.1)$$

where $w(x, y)$ is called the window function.

For small shifts $[u, v]$ the energy function $E(u, v)$ can be approximated as follows:

$$E(u, v) \approx [u \ v] M [u \ v]^T \quad (6.2)$$

Figure 6-2: Explanation of Harris corner points using the sliding window approach.

where M is a 2×2 matrix defined by:

$$M = \sum_{x,y} w(x, y) \begin{bmatrix} I_x^2 & I_x I_y \\ I_x I_y & I_y^2 \end{bmatrix} \quad (6.3)$$

Using Eigen value analysis where λ_1 and λ_2 are the Eigen values of the matrix M defined above we can draw the following conclusions:

- If λ_1 and λ_2 are large and $\lambda_1 \sim \lambda_2$, then the function E increases drastically and the point is a corner.
- If $\lambda_1 \gg \lambda_2$ or $\lambda_2 \gg \lambda_1$ then E increases along a particular direction and therefore defines an edge.
- If $\lambda_1 \sim \lambda_2 \approx 0$ then E does not change significantly anywhere and therefore represents a flat region.

In an alternative representation, a response function R which determines if a point is a corner or not can be defined as follows:

$$R = \det M - k (\text{trace } M)^2 \quad (6.4)$$

$$\text{where } \det M = \lambda_1 \lambda_2, \text{ trace } M = \lambda_1 + \lambda_2$$

It can be seen that the value of R depends solely on the Eigen values of the matrix. The following conclusions can be reached after computing the value of R :

- A large positive value of R indicates the presence of a corner in the image.
- R is negative with large magnitude if the feature in the image is an edge.
- The value of R is a very small one in case the image region is a flat one.

The proposed technique of obtaining the corner points is the Harris corner method. The following are some of the properties exhibited by this technique:

- a. The detector exhibits rotation invariance.
- b. It does not obey scale invariance.
- c. It is partially invariant to affine transformation.

6.1 Utilization of corner points in this research

In this work, points are extracted on the edges of the binarized baggage image (after Chan-Vese segmentation) to be fed into the trained Extreme Learning Machine(ELM) for extracting the cuboidal part contour, which is a convex hull over the classified points for the purpose of estimating the 2D pose parameters. Figure 6-3 shows an example of the extraction of points on the edges of the binarized segmented baggage images.

a)

b)

c)

d)

Figure 6-3: a-b) & c-d) Pairs of scene baggage images and corner points extracted on the edges of the segmented images using the Chan-Vese active contour.

In the following chapter, a review of the ELM technique for training Single Layer Feedforward Neural Network (SLFN) for the purpose of classifying corner points as belonging to the cuboidal body contour based on the window of extracted visual information is discussed. It can be seen in the following chapter and in the chapter on results and discussion that employing a trained ELM this way to extract the most salient cuboidal geon contour of a given two-part baggage can yield good accuracy percentages which also fulfills the part-centric representation/division of objects by the human visual system as explained in the chapter 2, 3 and 4.

Chapter 7—Extreme Learning Machine

Ever since their inception, feedforward neural networks have been extensively used in several applications which include both regression and classification problems. This is due to their ability to achieve complex non-linear mappings from the input to the output through a learning mechanism. In the case of neural networks, learning refers to the technique of choosing appropriate values of the weights and biases. As important as the neural networks are, conventional learning algorithms have typically focused on minimizing the training error. There is little or no emphasis on the training time needed to achieve optimal performance not only on the training set but also on the test set. As such, depending on the dataset and the scope and the objectives of the task, training a neural network can typically take hours or even a few days with no guarantee on the optimal performance as the solution can easily converge to local minima. This is particularly true of gradient based approaches to neural network training.

An equally important problem posed by conventional learning techniques is that they assume that all the parameters of the network need to be tuned. This line of understanding not only adds to the training time but also creates a dependency between the different layers of parameters. Furthermore, the speed of convergence of weights is dictated by the learning step also referred to as the learning rate.

In this chapter, a review of the new scheme of training neural networks employed in our research is provided. The discussed technique is different from other learning strategies such as the backpropagation learning algorithm which propose that all the parameters of the neural network need to be tuned. In contrast, the concept of Extreme Learning Machine(ELM)[59] debunks this notion by randomly initializing the hidden node parameters and then only solving for the output weights using a matrix representation. The developed algorithm tries to break away from the conventional neural network learning techniques and moves towards human brain inspired learning strategies. In doing so, ELM achieves better generalization performance and greater speed (roughly thousands of times faster) over other comparative learning algorithms.

ELM can be defined as the suite of learning techniques that can be used to train feed-forward neural networks (both single layer and multi-layer neural networks). This training scheme is therefore important as feedforward neural networks are used extensively in several applications due to their ability to: 1. Establish complex non-linear mapping between the input samples and the output values, 2. Provide abstract representations for natural and artificial phenomena that are difficult to model by hand. ELM strives to produce excellent generalization performance over its peers by conforming to [60] and tries to avoid the pitfalls of gradient based learning techniques. In the following sections we discuss the ELM theory as applied to Single Layer Feedforward Neural Networks (SLFN) and show how learning can be effectively achieved without having to deal with training all the parameters as previously assumed.

7.1 Algorithm

[61], [63] introduced the concept of ELM as an alternative to conventional methods of training feedforward neural networks. They also rigorously prove that the input nodes and the hidden biases can be randomly initialized if only that the choice of the activation function is such that it is infinitely differentiable. The activation functions include sigmoid, sine, Radial Basis Function (RBF) etc.

Consider a set of N data samples (x_i, t_i) where $x_i = [x_{i1}, x_{i2}, x_{i3}, x_{i4} \dots x_{in}]^T \in \mathbb{R}^n$ is the input feature vector and $t_i = [t_{i1}, t_{i2}, \dots, t_{im}]^T \in \mathbb{R}^m$ is the output vector. N' and $g(x)$ represents the number of hidden neurons and the hidden layer activation function respectively. We can mathematically model it as:

$$\sum_{i=1}^{N'} \beta_i g_i(x_j) = \sum_{i=1}^{N'} \beta_i g(w_i x_j + b_i) = o_j, \quad j = 1, 2, \dots, N \quad (7.1)$$

where $w_i = [w_{i1}, w_{i2}, \dots, w_{in}]^T$ is the vector of weights connecting the i^{th} hidden node and the input nodes, $\beta_i = [\beta_{i1}, \beta_{i2}, \dots, \beta_{im}]^T$ is the vector of output weights connecting the i^{th} hidden node and the output nodes and b_i is the threshold of the hidden node i . The output nodes are chosen to be linear in this technique.

Standard SLFNs can approximate these samples with zero error mean. (7.1) can therefore be rewritten as :

$$\sum_{i=1}^{N'} \beta_i g(w_i x_j + b_i) = t_j, j = 1, \dots, N. \quad (7.2)$$

Using matrix notation, we can express it more compactly as:

$$H\beta = T \quad (7.3)$$

where H is the hidden layer output matrix, β is the matrix of output weights and T is the matrix of outputs.

Equation 7.3 can be solved by using inverse operation as:

$$\beta' = H^* T \quad (7.4)$$

where H^* is the Moore Penrose generalized inverse of the hidden layer output matrix H .

Unlike regular inverse computation, Moore Penrose generalized inverse or simply referred to as the ‘pseudoinverse’ can be used to compute the least square fit to the data in the event that the matrix H is not square or singular.

It should be emphasized that in the proposed methods, the input weights w_i and the hidden layer biases b_i are randomly initialized only at the start of the learning phase and the hidden layer output matrix H is computed only once throughout, using these parameter values.

The solution proposed in (7.4) is one of the possible least square solution to the data which tries to reach the minima of the training error as under:

$$\|H\beta' - T\| = \|HH^*T - T\| = \min \|H\beta - T\| \quad (7.5)$$

However, in addition to minimizing the training error, ELM for training neural networks also obtains the minimum norm of weights among all the least square solutions thereby improving generalization. In other words,

$$\|\beta'\| = \|H^*T\| \leq \|\beta\| \quad \forall \beta \in \{\beta: \|H\beta - T\| \leq \|Hz - T\|, \forall z \in \mathbb{R}^{N' \times N}\} \quad (7.6)$$

In the next section, we list the advantages offered by the extreme learning machines over the conventional gradient based learning methods of training neural network parameters. In addition to the listed advantages, [62] compares the ELM algorithm with the existing variants of the Support Vector Machines (SVM). They show that ELM can achieve similar or better generalization performance for binary

classification tasks compared to SVM and is much simpler, scalable and computationally less demanding than SVM. It is due to these plethora of advantages that we have employed ELM in our work with sigmoid as the activation function.

7.2 Advantages of Extreme Learning Machines

Compared to the gradient based learning techniques such as the backpropagation algorithm and its variants such as the R-prop backpropagation technique, ELM offers the following main advantages:

- a. The iterative nature of the gradient based algorithms make them highly reliant on the parametric value of the learning rate α which decides the size of the learning step - a high value of α could make the algorithm unstable and result in divergence of the algorithm whereas a small value of α will result in a high convergence time of the algorithm. ELM by virtue of being non-gradient based does not require the selection of an optimal value of α .
- b. The gradient based techniques could result in sub optimal solutions as they suffer the problem of local minima. ELM avoids this problem.
- c. The principle underlying the operation of gradient based techniques is that all the parameters of the neural network need to be tuned. This does not align well with the biological principle of learning. ELM tries to bridge the barriers between artificial and biological learning by removing the need to tune input to hidden node parameters and biases.
- d. Unlike gradient based approaches, ELM based techniques are computationally inexpensive and less time consuming.
- e. ELM offers better generalization performance in addition to minimizing the training error for the given dataset.

7.3 How ELM is employed in this research?

In the previous chapter, the extraction of Harris corner points on the edges of the two-part baggage as well as the concept of part saliency & visual attention in Chapter 3

was discussed. In this research, a computational model based on a trained ELM that uses the information around each point is utilized to extract the salient cuboidal contour of the baggage in a scene which can then be utilized as the starting point for distance transform map computation and subsequent template matching for the purpose of 2D pose estimation of baggage.

This extraction of the salient cuboidal geon contour using ELM therefore emphasizes the following key points thereby building on the studies on human vision and cognition to develop a feasible computational system:

1. Not all parts of baggage are salient from the point of view of pose estimation and therefore can be selectively attended thereby underlining the role of task (to be achieved) in the selection of relevant visual features or parts of an object in a scene.
2. Rather than matching the entire object with stored templates for pose estimation, it is often useful and sufficient to match the most salient geon part primitive of the object in a scene with stored primitives of the type at different poses and orientations to estimate the best pose of the object in a scene as in the case of baggage in this research.
3. The extraction of the baggage body contour using the proposed ELM technique is in accordance with the concept of parts divisible at the points of negative minima of curvature proposed in cognitive science literature on human visual system.

Chapter 8—Chamfer Matching

Humans can effectively identify and estimate the pose of most objects, if not all, by using simple shape cues such as silhouettes, object edges and outlines without the need for other cues such as color, texture etc. Similar techniques for object recognition and pose estimation have been employed in the computer vision literature. Template based matching is one such concept that has emerged out of studies in shape matching techniques and viewpoint dependent object representation studies in cognitive science. In this research, we will be using Chamfer matching for estimating the location and the orientation of the cuboidal object in the 2D image plane. Distance transform representation is used to ease the cost computation associated with each match. The stored templates for matching should also implicitly encode object scale as it affects cost computation and as such, the scale is considered a constant. The following sections will briefly describe the method used in this research.

8.1 Principle of Chamfer Matching

The premise behind Chamfer Matching[24] is the usage of shape descriptors (contours or silhouettes) as symbolic maps for parametric matching with the sensed image. This is in contrast to the conventional matching techniques that use image patches for establishing correspondence.

Chamfer matching can be depicted as a minimization problem. Given a template image edge map $U=\{u_i\}$ and a query image edge map $V=\{v_j\}$, the chamfer distance between the two edge maps is defined as the average distance of each pixel u_i from its nearest edge in V . Mathematically, we can write it as:

$$d_{CM}(U, V) = \frac{1}{n} \sum \min_{v_j \in V} |u_i - v_j| \quad (8.1)$$

where $n = |U|$

If W is a warping function defined on the image plane and parametrized by s , where $s = (\theta, tx, ty)$, then its action on image points can be defined as:

$$W(x; s) = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} t_x \\ t_y \end{bmatrix} \quad (8.2)$$

The best parameters for the mentioned problem is given by:

$$s' = \arg \min_{s \in SE(2)} d_{CM}(W(U; s), V) \quad (8.3)$$

Chamfer matching provides a smooth measure of fitness, can tolerate small rotations and is easy to implement. Changes in orientation of the object is encoded as an implicit measure while matching. The implementation aspect is further simplified by computing the matching cost using a distance transform[64] image. Using this mathematical trick simplifies the computational demands by making calculation of (8.1) possible in linear time. This concept is used in this research.

Equation 8.1 can therefore be rewritten in terms of distance transform as:

$$d_{CM}(U, V) = \frac{1}{n} \sum_{u_i \in U} DT_v(u_i) \quad (8.4)$$

Besides the ease of implementation, there is another advantage that the use of Chamfer matching has to offer. As already mentioned before, by virtue of being a template based matching procedure, it can be used to implement the direct approach to viewpoint dependent mental representation of objects or object parts in humans thereby aligning well with cognitive science and psychology based studies. In the following section, we briefly explain the concept of distance transform.

8.2 Distance Transform

Distance transform, also referred to as the distance field allows for linear time algorithms for solving a class of problems in which the objective is to minimize the cost incurred. Distance transform is normally applied only to binary images. Edge images are a particular sub-class of binary images where application of distance transform would be particularly useful. The outcome of the distance transform is a gray-level image that is similar to the input image but is in fact a representative of the distance of each point from the closest boundary point.

A distance transform image therefore specifies the distance of each pixel from its nearest non-zero pixel. For a binary image, edges are typically used as the features to guide the computation of the distance transform. In general, however, a distance transform

is a map of distances of each point to the nearest point of a particular type. They are generally, used for data in a grid such as pixels in 2D or voxels in 3D. Different distance metrics have been used in practice such as Euclidean, Cityblock, Chessboard etc. for computing the distance transform of images. Euclidean is by far the most familiar and popular because of its property of rotational invariance. This research also employs Euclidean distance metric for computing the distance transform image and the implementation is carried out in MATLAB.

Consider the example of a grid shown in figure 8-1 a. The grid represents a binary image with pixel values 0 and 1 as shown in the figure. For simplicity of understanding, let us assume that the image is a 3×3 matrix with one edge pixel denoted by the value 1. Although the specifics of the algorithm used to compute the distance transform vary, figure 8-1 b shows the distance transform matrix that should be obtained by using the Euclidean distance metric in which each pixel holds the computed Euclidean distance to its nearest edge pixel. Figure 8-2 shows the image obtained by the application of the Euclidean distance transform on the extracted cuboidal contour of the baggage after phase 1. It can be seen that in figure 8-2 b the regions near the contour as seen in figure 8-2 a are black indicating low distance. As the distance increases from the edges it reaches close to white indicating a large distance. Figure 8-3 shows a few examples of the computation of the Chamfer score for three different translations each of two given pose templates based on the distance transform map obtained in Figure 8-2 b. As can be seen, a higher Chamfer score while translating indicates a greater mismatch error while a lower score implies a shift towards the contours of interest. It is also evident that the value of the Chamfer score varies as a function of not only the pose but also the aspect ratio thereby allowing for accurate localization of baggage.

Applying the definition of Euclidean distance in conjunction with the concept of distance transform, the Euclidean distance transform can be written as:

$$D(x, y) = \sqrt{\min((x - i)^2 + (y - j)^2)} \quad \forall 0 \leq i < m, 0 \leq j < n, I(i, j) = 1 \quad (8.5)$$

where (x, y) is the particular pixel under consideration, (i, j) are the coordinates of the pixels of the image I of size $m \times n$.

0	0	0
0	1	0
0	0	0

a)

1.4142	1	1.4142
1	0	1
1.4142	1	1.4142

b)

Figure 8-1: a) Input Binary Image b) Euclidean Distance Transformation of the Image

Figure 8-2: a) Extracted cuboidal geon contour from phase 1 b) Euclidean distance transform map of a).

Figure 8-3: Computation of Chamfer cost using the pose templates (green) over the distance transform map shown in Figure 8-2 b) with different translation parameters a-c) Fused image showing the target extracted cuboidal geon contour with the translated stored cuboidal geon part contour template. The arrows indicate a large alignment error with a high Chamfer score of 10464, 9074.2 and 10463 respectively d-f) Matching of cuboidal geon template at a different pose with Chamfer scores of 7299.8, 5288.2 and 8439.8.

Chapter 9 - Results & Discussion

In this chapter the results of our experiments in estimating the pose (localization and orientation) is presented. For simplicity of explanation and understanding, the work is divided into two phases. Phase 1 is concerned with the extraction of the salient cuboidal contour of the baggage which serves as the input edge map for the next phase. Phase 2 is concerned with the estimation of the 2D localization parameters of the baggage in an uncluttered scene relative to the best fitting edge template which is ascertained using Chamfer matching. To this end, the two phases can be seen as a cascading or serial operation in which the success of the first phase affects the performance of the second phase. Figure 9-1 shows the cascading effect of the two phases while representing the sequence of operations that are carried out under each phase. Figure 9-2 represents the pseudo code of the implemented algorithm the particulars of which are given in the Appendix.

All the experiments and simulations were conducted on a 64-bit intel Core i7 PC running Window 7. The proposed algorithm was implemented in MATLAB and was augmented by the Image Processing Toolbox. A set of 100 cuboidal templates with different aspect ratios and poses were created in Microsoft Paint for the purpose of matching.

Figure 9-1 - Cascading operation of phase 1 and phase 2

% Pseudocode to obtain the localization and pose estimates of single baggage in uncluttered scene images.

% Phase 1 – Extraction of the cuboidal contour of the two-part baggage containing the salient body and the non-salient handle for the computation of the distance transform image.

% Training phase.

1. NUMBER_TRAINING_IMAGES = OPEN_TRAINING_BAGGAGE_IMAGES.
2. FOR IMAGE IN NUMBER_TRAINING_IMAGES
3. BW = SEGMENT_CHAN-VESE_ACTIVE_CONTOUR(IMAGE). % Segment the baggage image using the Chan-Vese active contours.
4. POINTS = HARRIS(BW). % Edge decomposition by extracting the points(Harris) on the edges.
5. IMG_WINDOW = EXTRACT_WINDOW(POINTS). %Extracting a window of 50×50 around each corner point to distinguish points lying on the body of the baggage from those on the handle.
6. REDUCED_WINDOW = DIMENSIONALITY_REDUCTION(IMG_WINDOW). % Reducing the dimensionality of the 50×50 data to 10×10 input vector to be fed into the Extreme Learning Machine(ELM) for training.
7. END
8. ELM_TRAIN(REDUCED_WINDOW). %Train the ELM to obtain the weights using the data obtained above.

% Testing Phase for the extraction of the cuboidal contour of actual baggage in a scene using the trained ELM.

9. NUMBER_TEST_IMAGES = OPEN_TEST_BAGGAGE_IMAGES.
10. FOR IMAGE IN NUMBER_TEST_IMAGES
11. BW_TEST = SEGMENT_CHAN-VESE_ACTIVE_CONTOUR(IMAGE). % Segment the image using the Chan-Vese active contour to obtain the binarized image.
12. POINTS_TEST = HARRIS(BW_TEST). % Edge decomposition by extracting the points (Harris) on the edges for the extraction of the visually salient cuboidal contour of the baggage for pose estimation.
13. IMG_WINDOW_TEST = EXTRACT_WINDOW(POINTS_TEST). %Extracting a window of 50×50 around each corner point to distinguish points lying on the body of the baggage from those on the handle.
14. REDUCED_WINDOW_TEST = DIMENSIONALITY_REDUCTION(IMG_WINDOW_TEST).
%Dimensionality reduction of the data from 50×50 TO 10×10 input vector for the trained ELM classification. Classification is used to extract only those edge (Harris) points lying on the salient cuboidal body of the baggage ignoring the non-salient handle of the baggage and consequently the unnecessary edge points for the matching process.
15. PART_CONTOURS = CONVEX_HULL (POINTS CLASSIFIED AS BELONGING TO THE BODY OF THE BAGGAGE).
16. END
17. CORRECT_CONTOURS = NUMBER_OF_CORRECTLY_EXTRACTED_CONTOURS(PART_CONTOURS).
18. PERCENTAGE_ACCURACY = CORRECT_CONTOURS / NUMBER_TEST_IMAGES.

% Phase 2 – Estimation of 2D Pose Parameters of the baggage using the scene-specific extracted cuboidal contour and stored pose templates in an uncluttered scene.

1. NUMBER_STORED_TEMPLATES = OPEN_STORED_CUBOIDAL_EDGE_TEMPLATES.
2. BW_DISTS = DISTANCE_TRANSFORM (CORRECTLY EXTRACTED CONTOURS).
3. FOR BW_DIST IN BW_DISTS
4. FOR TEMPLATE IN NUMBER_STORED_TEMPLATES
5. FOR TRANSLATION IN TRANSLATIONS
6. %Translate each pose template over BW_DIST and compute the Chamfer cost. Chamfer cost is measured as the accrued cost obtained by summing the Euclidean distance costs under the edge pixels of each template for every translation.
7. SCORES = RECORD(CHAMFER_COST).
8. END
9. MIN_SCORE = RECORD_LOWEST_SCORE(SCORES). % Record the lowest score of all the translations for each template and note the corresponding translation parameters.
10. END
11. LOWEST_SCORE = RECORD_LOWEST_SCORE(MIN_SCORE). % Apply a second minimization across the lowest scores of all the templates to obtain the best fitting template and the corresponding translation parameters. Pose and orientation is implicitly encoded in the template itself.
12. END

Figure 9-2 – Pseudocode of the proposed approach.

9.1 Phase 1 – Extraction of Cuboidal Geon Contour

In this phase, the cuboidal part contour (salient part boundary) of the baggage is identified and extracted for the purpose of pose estimation in phase 2. This implies the elimination of those edges that do not belong to the body of the baggage and instead lie on the handle of the baggage which is not salient from the point of view of pose estimation. In other words, the object will need to be parsed at the points of negative minima of curvature in accordance with [3] leading to part primitives (body and the handle) termed geons as proposed by [2]. Only the body (cuboidal) contour of the baggage is used for localization and pose estimation through the part based matching process in phase 2.

A set of 50 baggage images at a fixed scale and varying poses was chosen during the training process. Chan – Vese active contour was then used to segment the training baggage images to obtain the binarized image representation. By the principle of edge decomposition, Harris corner points were then identified on the edges of the segmented baggage. Around each point a filter window of 50×50 was extracted, vectorized and fed into the extreme learning machine as a training example. A total of 1500 such training examples serve as the batch training input. Each training example is associated with a label that identifies the corner point as either belonging to the body or the handle. The ELM then learns the function mapping the input to the output. A similar procedure is then followed on the test set of 60 baggage images. Upon segmentation and extraction of window information of the specified size around each corner point, the previously trained ELM is then used to classify each Harris corner point. A convex hull over the corner points classified as belonging to the body of the baggage is then computed to obtain the cuboidal contour of the baggage. Figure 9-3 shows an example of extraction of the cuboidal contour of the baggage starting with a given baggage image. Figure 9-4 shows the pictorial representation of Table I which represents the cuboidal contour extraction accuracy for a subset of the varying number of hidden neurons chosen for our ELM framework. Five trials of experiments were performed for each chosen number of hidden neurons and the best accuracy rates are reported.

It must be emphasized from Table I and Fig.9-4 that setting the number of hidden neurons to 120 provided the highest testing accuracy by allowing the correct extraction of

cuboidal geon contour in 54 of the 60 test images with differing baggage poses and aspect ratios leading to an accuracy of 90% on the test set. It can therefore be seen that incorporating object centered representation theory analogous to the recognition by component model and the part based visual attention model is made possible using a trained ELM based classifier to extract the most salient cuboidal baggage contour.

Figure 9-3 - a) Example Baggage Image b) Image segmentation using Chan–Vese active contour c) Detection of Harris corner points d) Extraction of cuboidal geon contour by finding the convex hull over the corner points which are classified as belonging to the body of the baggage by the trained ELM.

Table I - Cuboidal Contour Extraction Accuracy

No. of Hidden Neurons	No. of Test Images	No. of Correctly Extracted Contours	Percentage Accuracy
80	60	44	73.33
100	60	51	85
120	60	54	90
140	60	50	83.33
160	60	41	68

180	60	43	71.67
200	60	40	66.67

Figure 9-4 – Variation in the cuboidal contour extraction accuracy for different number of hidden neurons.

9.2 Phase 2 – Estimation of Localization Parameters of Baggage

In this phase, the cuboidal contours which were extracted in the previous phase are utilized to obtain information regarding the location and pose of the baggage. This is achieved through implicitly encoded pose templates and then employing the Chamfer shape matching technique to estimate the localization and pose information of the baggage in the scene. This aligns well with the viewpoint dependent object representation theories which suggest that objects are stored as a series of 2D views that elicit tasks such as recognition and pose estimation. In this work, however, edge maps of the baggage body (part edges) rather than the baggage itself (object) are stored as it is sufficient to extract the localization information. Figure 9-5 shows the example of the matching process of one

of the 54 correctly extracted cuboidal contours in phase 1 with a few of the 100 stored templates.

The matching process is conducted by translating each template by 20 pixels in either dimension over the distance transform image of the extracted cuboidal contour. At each translation of the template the Chamfer score as explained in [24] (Chapter 8) is recorded. A minimization is applied across the obtained scores of the translation process that yields the best translation parameters for a given template. This process is repeated for all the templates in the database. Finally, a second minimization is then applied across the lowest scores for each template which yields the best fitting template and the corresponding translation parameters.

Mathematically, if t_x and t_y represent the translation parameters of the best fitting template T in the set of stored templates Γ and ϕ denotes the space of possible translations, then the localization process can be represented as:

$$(T, t_x, t_y) = \arg \min_r (\min d_{CM}(W(U; \phi), V)) \quad (9.1)$$

where $U \in \Gamma$ and W is the warping function.

Figure 9-5: a) Extracted cuboidal contour from phase 1 b) – e) Inaccurate template matches (green) with the cuboidal contour (lavender) f) Accurate template match yielding the lowest possible score out of the stored template database.

The translation parameters of the best fitting cuboidal template are obtained as the displacement of the template relative to the original position in either dimension (x and y) in the image plane required to obtain the lowest Chamfer score while matching with the extracted cuboidal contour in the query image. It must also be emphasized that the pose, orientation and information concerning the aspect ratio are intrinsically encoded in the stored templates and therefore form an inherent component of the matching process.

By following this procedure, it was observed that out of the 54 query images in which the cuboidal contours were correctly extracted, 48 instances yielded correct results during the template matching process implying that the closest matching template from among the stored 100 templates was found to yield the lowest Chamfer score. The results obtained are best summarized in Table II. False matches in the remaining 6 cases arise when the lowest score yielding template does not correspond well with the scene cuboidal contour. This is because Chamfer matching only tries to minimize the distance between two image edge maps with little regard to the orientation of the edge segments. This inaccuracy could however be reduced by incorporating the edge orientation information into the Chamfer cost function as in the case of directional Chamfer matching. It is also important to note that the accuracy of the method in recognizing the correct matching template increases with the increase in the number of stored templates albeit at the expense of an increased computational cost as the matching is a one-on-one process. Careful consideration therefore needs to be given to both the issues and there is a need to balance the requirement of precision in estimation of localization parameters and the

computational complexity involved. Nevertheless, an optimum number of pose templates is preferred. This is because the proposed equation in (9.1) returns the lowest score yielding template which may still be a suboptimal solution to the pose estimation problem if there is a scarcity of stored templates in the database. It is also important to keep the scale of the image fixed as the Chamfer cost computation is not invariant to scale changes.

Table II – Pose & Localization Estimation Accuracy

Number of Hidden Neurons as determined from Phase 1	120
Number of Test Images	60
Number of correctly extracted cuboidal contours in Phase I	54
Number of cuboidal contours with correct template matches in Phase II	48
Accuracy in Phase II	88.89%
Overall Accuracy	80%

Future Research Directions

The following aspects will be looked into and investigated in the futuristic development, modification and implementation of the algorithm:

1. To evaluate the performance of the proposed algorithm in the cluttered scenes containing not one but multiple baggage and to make necessary changes as appropriate.
2. To incorporate edge orientation information in the proposed algorithm as in the case of directional Chamfer matching and to evaluate the accuracy in the estimation of pose parameters as compared to Chamfer matching used in this work. The addition of this edge orientation information can be a useful one especially when operating on 2D images wherein the lack of availability of the depth information as in the case of 3D can lead to false matches yielding the lowest scores using the conventional Chamfer matching technique in this work.
3. To explore the availability of scale invariant techniques that achieve the combined principles of object-centered and viewer-centered theories in the problem of pose estimation.
4. Practical demonstration of the proposed algorithm on an actual robotic hardware platform.
5. Integration of the developed pose estimation technique with effective grasping algorithms that will allow us to create full-fledged robotic systems for handling baggage in uncluttered scenes.

Conclusion

In this work, a computer vision algorithm for identifying the 2D localization parameters of the two-part baggage in an uncluttered scene is developed. This algorithm is based on the Chamfer shape matching algorithm that combines the power and the flexibility of both object centered and viewer centered theories of object representation. In particular, the recognition by component model of object parsing at the points of negative minima of curvature to parts (geons) of interest and the template matching technique based on viewpoint representation of object part contours is utilized in this work. The premise behind this idea is that an object can be decomposed into part primitives as in this case wherein the baggage was parsed into the body and the handle. The pose of the object can then be construed as a problem of estimating the pose of its salient or dominant part as in this case in which the body of the baggage is sufficient for pose estimation. This concept of object decomposition into parts followed by shape matching the most salient part with predefined templates helps decouple the problem of recognition from that of pose and viewpoint estimation. This work therefore strives to drive the emphasis on the application of our understanding of the cognitive science theories in the development of computer vision algorithms as it represents the dominant perceptual modality of focus, as we move away from traditional robotics to cognitive robotics. Future research along these lines will help develop a potential robotics solution for automation of baggage handling and for grasping applications in general for which pose estimation is a must.

Bibliography

- [1] D. Marr and H. K. Nishihara, "Representation and Recognition of the Spatial Organization of Three-Dimensional Shapes," *Proceedings of the Royal Society of London B: Biological Sciences*, vol. 200, pp. 269-294, 1978-02-23 00:00:00 1978.
- [2] I. Biederman, "Recognition-by-components: a theory of human image understanding," *Psychological review*, vol. 94, p. 115, 1987.
- [3] D. D. Hoffman and W. A. Richards, "Parts of recognition," *Cognition*, vol. 18, pp. 65-96, 12// 1984.
- [4] S. Ullman, "Three-dimensional object recognition based on the combination of views," *Cognition*, vol. 67, pp. 21-44, 7/17/ 1998.
- [5] D. J. Willshaw, O. P. Buneman, and H. C. Longuet-Higgins, "Non-Holographic Associative Memory," *Nature*, vol. 222, pp. 960-962, 06/07/print 1969.
- [6] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," *Proceedings of the National Academy of Sciences*, vol. 79, pp. 2554-2558, April 1, 1982 1982.
- [7] D. P. Huttenlocher and S. Ullman, "Recognizing solid objects by alignment with an image," *International Journal of Computer Vision*, vol. 5, pp. 195-212.
- [8] D. G. Lowe, *Perceptual Organization and Visual Recognition*: Kluwer Academic Publishers, 1985.
- [9] R. A. Brooks, "Model-Based Three-Dimensional Interpretations of Two-Dimensional Images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 5, pp. 140-150, 1983.
- [10] A. P. Pentland, "Perceptual organization and the representation of natural form," *Artif. Intell.*, vol. 28, pp. 293-331, 1986.
- [11] S. J. Dickinson and D. Metaxas, "Integrating qualitative and quantitative shape recovery," *International Journal of Computer Vision*, vol. 13, pp. 311-330.
- [12] S. J. Dickinson, D. Metaxas, and A. Pentland, "The role of model-based segmentation in the recovery of volumetric parts from range data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, pp. 259-267, 1997.

- [13] F. P. Ferrie, J. Lagarde, and P. Whaite, "Darboux frames, snakes, and superquadrics: geometry from the bottom up," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, pp. 771-784, 1993.
- [14] A. Leonardis, F. Solina, and A. Macerl, "A direct recovery of superquadric models in range images using recover-and-select paradigm," in *Computer Vision — ECCV '94: Third European Conference on Computer Vision Stockholm, Sweden, May 2–6, 1994 Proceedings, Volume I*, J.-O. Eklundh, Ed., ed Berlin, Heidelberg: Springer Berlin Heidelberg, 1994, pp. 309-318.
- [15] D. Metaxas and D. Terzopoulos, "Shape and nonrigid motion estimation through physics-based synthesis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, pp. 580-591, 1993.
- [16] R. Bergevin and M. D. Levine, "Extraction of line drawing features for object recognition," in *Pattern Recognition, 1990. Proceedings., 10th International Conference on*, 1990, pp. 496-501 vol.1.
- [17] R. Bergevin and M. D. Levine, "Part decomposition of objects from single view line drawings," *CVGIP: Image Understanding*, vol. 55, pp. 73-83, 1992/01/01 1992.
- [18] R. Bergevin and M. D. Levine, "Generic object recognition: building and matching coarse descriptions from line drawings," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, pp. 19-36, 1993.
- [19] I. Biederman, J. E. Hummel, P. C. Gerhardstein, and E. E. Cooper, "From image edges to geons to viewpoint-invariant object models: a neural net implementation," 1992, pp. 570-578.
- [20] S. J. Dickinson, A. P. Pentland, and A. Rosenfeld, "A Representation for Qualitative 3-D Object Recognition Integrating Object-Centered and Viewer-Centered Models," in *Science of Vision*, K. N. Leibovic, Ed., ed New York, NY: Springer New York, 1990, pp. 398-421.
- [21] J. K. Aggarwal, B. C. Vemuri, N. S. Raja, and A. K. Jain, "Range Image Understanding Recognizing geons from superquadrics fitted to range data," *Image and Vision Computing*, vol. 10, pp. 179-190, 1992/04/01 1992.

- [22] H. Murase and S. K. Nayar, "Visual learning and recognition of 3-d objects from appearance," *International Journal of Computer Vision*, vol. 14, pp. 5-24.
- [23] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. Cognitive Neuroscience*, vol. 3, pp. 71-86, 1991.
- [24] H. G. Barrow, J. M. Tenenbaum, R. C. Bolles, and H. C. Wolf, "Parametric correspondence and chamfer matching: two new techniques for image matching," presented at the Proceedings of the 5th international joint conference on Artificial intelligence - Volume 2, Cambridge, USA, 1977.
- [25] M. I. Posner, "Orienting of attention," *Quarterly journal of experimental psychology*, vol. 32, pp. 3-25, 1980.
- [26] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive psychology*, vol. 12, pp. 97-136, 1980.
- [27] J. M. Wolfe, "Guided search 2.0 a revised model of visual search," *Psychonomic bulletin & review*, vol. 1, pp. 202-238, 1994.
- [28] C. W. Eriksen and Y.-Y. Yeh, "Allocation of attention in the visual field," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 11, p. 583, 1985.
- [29] C. W. Eriksen and J. D. S. James, "Visual attention within and around the field of focal attention: A zoom lens model," *Perception & psychophysics*, vol. 40, pp. 225-240, 1986.
- [30] C. Koch and S. Ullman, "Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry," in *Matters of Intelligence: Conceptual Structures in Cognitive Neuroscience*, L. M. Vaina, Ed., ed Dordrecht: Springer Netherlands, 1987, pp. 115-141.
- [31] L. Itti, C. Koch, and E. Niebur, "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, pp. 1254-1259, 1998.
- [32] J. K. Tsotsos, S. M. Culhane, W. Y. Kei Wai, Y. Lai, N. Davis, and F. Nuflo, "Modeling visual attention via selective tuning," *Artificial Intelligence*, vol. 78, pp. 507-545, 1995/10/01 1995.

- [33] J. Duncan, "Selective attention and the organization of visual information," *J Exp Psychol Gen*, vol. 113, pp. 501-17, Dec 1984.
- [34] J. Duncan and G. W. Humphreys, "Visual search and stimulus similarity," *Psychol Rev*, vol. 96, pp. 433-58, Jul 1989.
- [35] S. Grossberg and R. D. Raizada, "Contrast-sensitive perceptual grouping and object-based attention in the laminar circuits of primary visual cortex," *Vision Res*, vol. 40, pp. 1413-32, 2000.
- [36] D. Heinke and G. W. Humphreys, "SAIM: A model of visual attention and neglect," in *Artificial Neural Networks — ICANN'97: 7th International Conference Lausanne, Switzerland, October 8–10, 1997 Proceedings*, W. Gerstner, A. Germond, M. Hasler, and J.-D. Nicoud, Eds., ed Berlin, Heidelberg: Springer Berlin Heidelberg, 1997, pp. 913-918.
- [37] B. J. Scholl, "Objects and attention: the state of the art," *Cognition*, vol. 80, pp. 1-46, Jun 2001.
- [38] M. C. Mozer and S. P. Vecera, "Space-and object-based attention," in *Neurobiology of Attention*, L. Itti, G. Rees, and J. K. Tsotsos, Eds., ed: Academic Press, 2005, pp. 130--134.
- [39] S. P. Vecera, M. Behrmann, and J. McGoldrick, "Selective attention to the parts of an object," *Psychonomic Bulletin & Review*, vol. 7, pp. 301-308, 2000.
- [40] D. D. Hoffman and M. Singh, "Salience of visual parts," *Cognition*, vol. 63, pp. 29-78, 4/1/ 1997.
- [41] "Picture Thresholding Using an Iterative Selection Method," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 8, pp. 630-632, 1978.
- [42] J. Kittler and J. Illingworth, "Minimum error thresholding," *Pattern Recognition*, vol. 19, pp. 41-47, 1986/01/01 1986.
- [43] R. Adams and L. Bischof, "Seeded region growing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, pp. 641-647, 1994.
- [44] F. Cheevasuvit, H. Maitre, and D. Vidal-Madjar, "A robust method for picture segmentation based on split-and-merge procedure," *Comput. Vision Graph. Image Process.*, vol. 34, pp. 268-281, 1986.

- [45] S.-Y. Chen, W.-C. Lin, and C.-T. Chen, "Split-and-merge image segmentation based on localized feature analysis and statistical tests," *CVGIP: Graphical Models and Image Processing*, vol. 53, pp. 457-475, 1991/09/01 1991.
- [46] S. L. Horowitz and T. Pavlidis, "Picture Segmentation by a directed split-and-merge procedure," *Proceedings of the 2nd International Joint Conference on Pattern Recognition, Copenhagen, Denmark*, pp. 424-433, // 1974.
- [47] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *International Journal of Computer Vision*, vol. 1, pp. 321-331.
- [48] S. Osher and J. A. Sethian, "Fronts propagating with curvature-dependent speed: Algorithms based on Hamilton-Jacobi formulations," *Journal of Computational Physics*, vol. 79, pp. 12-49, 11// 1988.
- [49] S. Kichenassamy, A. Kumar, P. Olver, A. Tannenbaum, and A. Yezzi, "Conformal curvature flows: From phase transitions to active vision," *Archive for Rational Mechanics and Analysis*, vol. 134, pp. 275-301.
- [50] V. Caselles, R. Kimmel, and G. Sapiro, "Geodesic Active Contours," *International Journal of Computer Vision*, vol. 22, pp. 61-79.
- [51] S. C. Zhu, T. S. Lee, and A. L. Yuille, "Region competition: unifying snakes, region growing, energy/Bayes/MDL for multi-band image segmentation," in *Computer Vision, 1995. Proceedings., Fifth International Conference on*, 1995, pp. 416-423.
- [52] T. F. Chan and L. A. Vese, "Active contours without edges," *Image Processing, IEEE Transactions on*, vol. 10, pp. 266-277, 2001.
- [53] A. Yezzi, A. Tsai, and A. Willsky, "A statistical approach to snakes for bimodal and trimodal imagery," in *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, 1999, pp. 898-903 vol.2.
- [54] M. Rousson and R. Deriche, "A Variational Framework for Active and Adaptive Segmentation of Vector Valued Images," presented at the Proceedings of the Workshop on Motion and Video Computing, 2002.
- [55] A. Yezzi Jr, A. Tsai, and A. Willsky, "A Fully Global Approach to Image Segmentation via Coupled Curve Evolution Equations," *Journal of Visual Communication and Image Representation*, vol. 13, pp. 195-216, 3// 2002.

- [56] D. Mumford and J. Shah, "Optimal approximations by piecewise smooth functions and associated variational problems," *Communications on pure and applied mathematics*, vol. 42, pp. 577-685, 1989.
- [57] H.-K. Zhao, T. Chan, B. Merriman, and S. Osher, "A Variational Level Set Approach to Multiphase Motion," *Journal of Computational Physics*, vol. 127, pp. 179-195, 8// 1996.
- [58] C. Harris and M. Stephens, "A Combined Corner and Edge Detection," in *Proceedings of The Fourth Alvey Vision Conference*, 1988, pp. 147-151.
- [59] G.-B. Huang, "What are Extreme Learning Machines? Filling the Gap Between Frank Rosenblatt's Dream and John von Neumann's Puzzle," *Cognitive Computation*, vol. 7, pp. 263-278, 2015/06/01 2015.
- [60] P. L. Bartlett, "The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network," *Information Theory, IEEE Transactions on*, vol. 44, pp. 525-536, 1998.
- [61] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: theory and applications," *Neurocomputing*, vol. 70, pp. 489-501, 2006.
- [62] H. Guang-Bin, Z. Hongming, D. Xiaojian, and Z. Rui, "Extreme Learning Machine for Regression and Multiclass Classification," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 42, pp. 513-529, 2012.
- [63] H. Guang-Bin, Z. Qin-Yu, and S. Chee-Kheong, "Extreme learning machine: a new learning scheme of feedforward neural networks," in *Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on*, 2004, pp. 985-990 vol.2.
- [64] P. Felzenszwalb and D. Huttenlocher, "Distance transforms of sampled functions," Cornell University 2004.

Publications

1. Meng Joo Er, Sriram Sridhar and Ning Wang, “A Novel Vision System for Baggage Handling”, Presented and published in the proceedings of the IEEE International Conference on Intelligent Control, Power and Instrumentation 2016 (ICICPI 2016) in October 2016.

Appendix

In this section, the MATLAB computer program for implementing the algorithm that incorporates both object-centered and viewer-centered representations for the purpose of matching the extracted cuboidal geon contour with different cuboidal pose templates to obtain the localization information is given. Where necessary, to enhance the reader understanding, comments are added to explain the functionality of the code. Also, the code is divided into snippets based on the specific functionality / purpose which it serves.

A. Creation of a training set of vectorized binary image patches around each of the Harris corner points after segmentation of the baggage in uncluttered scene to obtain the visually salient cuboidal part contour for subsequent Chamfer matching.

```
% Code snippet to obtain the training set of binarized image patches
around each of the corner points for determination of the visually
salient part primitive cuboidal contour for template based Chamfer
matching.
global im;
im = imread('C:\Users\sriram005\Downloads\baggage2.jpg');
im=rgb2gray(im);
mask = zeros(size(im));
mask(2:end-2,2:end-2)= 1;
im = activecontour (im, mask,2000,'Chan-Vese','SmoothFactor',4);
% Computes the active contour of the object to be grasped with a
maximum of 2000 iterations.
im = padarray(im,[50 50]);
im = imgaussfilt(uint8(im),8)
im = logical(im)
imshow(im)
hold on
b=corner(im);
plot(b(:,1),b(:,2),'*')
dataset = ExtractingKeypoints(im)

% A total of 50 baggage images are chosen for the purpose of training
with 30 training samples from each image - 15 belonging to the
cuboidal geon contour of interest (positive sample '1') and 15
belonging to the handle of the baggage (negative sample '0')
function dataset = ExtractingKeypoints(im)
global im
train= zeros(1,100);
for i =1:15
    [x,y] = ginput(1);
    g= floor(x)
    h= floor(y)
    data = Object_Activecontour(floor(y),floor(x));
    train = [train;data];
```

```

end
for i = 1:15
    [x,y] = ginput(1)
    g= floor(x)
    h= floor(y)
    data = Object_Activecontour(floor(y),floor(x));
    train = [train;data];
end
output = [1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0]
dataset = [train(2:end,:) output]

function data = Object_Activecontour(p,q)
window = zeros(10);
global im
window = im(p-49:p+50,q-49:q+50); % Extracting the image patch of size
100×100 around the Harris corner points which lie on both the visually
salient baggage cuboid and the handle.
window = imresize(window,[10 10]); % Reducing the dimensionality of
the data from 100×100 image pixels to 10×10 which serves as input to the
ELM during the training phase.
reshaped = reshape(window,[1,100]); % Using the reshape function to
vectorized the input data to be fed into the ELM.
data = reshaped;

```

B. Extreme Learning Machine (ELM) for training on the image samples for classifying the corner points as either belonging to the visually dominant cuboidal part of the baggage or the handle.(Code credits – Mr. Qin-Yu Zhu and Dr. Guang-Bin Huang, Nanyang Technological University, Singapore).

% An ELM is trained for identification and classification of baggage body edge points from the points on the handle for the extraction of the visually salient cuboidal contour of the baggage.

```

function [TrainingTime,TrainingAccuracy] = elm_train
(TrainingData_File, Elm_Type, NumberofHiddenNeurons,
ActivationFunction)
% Usage: elm_train(TrainingData_File, Elm_Type,
NumberofHiddenNeurons, ActivationFunction)
% OR: [TrainingTime, TrainingAccuracy] =
elm_train(TrainingData_File, Elm_Type, NumberofHiddenNeurons,
ActivationFunction)
%
% Input:
% TrainingData_File - Filename of training data set
% Elm_Type - 0 for regression; 1 for (both binary and
multi-classes) classification
% NumberofHiddenNeurons - Number of hidden neurons assigned to the
ELM
% ActivationFunction - Type of activation function:
% 'sig' for Sigmoidal function
% 'sin' for Sine function
% 'hardlim' for Hardlim function
%
% Output:

```

```

% TrainingTime          - Time (seconds) spent on training ELM
% TrainingAccuracy      - Training accuracy:
%                        RMSE for regression or correct
classification rate for classification
%
% MULTI-CLASSE CLASSIFICATION: NUMBER OF OUTPUT NEURONS WILL BE
AUTOMATICALLY SET EQUAL TO NUMBER OF CLASSES
% FOR EXAMPLE, if there are 7 classes in all, there will have 7
output
% neurons; neuron 5 has the highest output means input belongs to 5-
th class
%
% Sample1 regression: [TrainingTime, TrainingAccuracy,
TestingAccuracy] = elm_train('sinc_train', 0, 20, 'sig')
% Sample2 classification: elm_train('diabetes_train', 1, 20, 'sig')
%
% Authors:      MR QIN-YU ZHU AND DR GUANG-BIN HUANG
%              NANYANG TECHNOLOGICAL UNIVERSITY, SINGAPORE
%              EMAIL:      EGBHUANG@NTU.EDU.SG; GBHUANG@IEEE.ORG
%              WEBSITE:
http://www.ntu.edu.sg/eee/icis/cv/egbhuang.htm
%              DATE:      APRIL 2004

%%%%%%%%%%%% Macro definition
REGRESSION=0;
CLASSIFIER=1;

%%%%%%%%%%%% Load training dataset
train_data=xlsread(TrainingData_File);
T=train_data(:,1)';
P=train_data(:,2:size(train_data,2))';
% clear train_data; % Release raw
training data array

NumberofTrainingData=size(P,2);
NumberofInputNeurons=size(P,1);

if Elm_Type~=REGRESSION
    %%%%%%%%%%%%% Preprocessing the data of classification
    sorted_target=sort(T,2);
    label=zeros(1,1); % Find and save
in 'label' class label from training and testing data sets
    label(1,1)=sorted_target(1,1);
    j=1;
    for i = 2:NumberofTrainingData
        if sorted_target(1,i) ~= label(1,j)
            j=j+1;
            label(1,j) = sorted_target(1,i);
        end
    end
    number_class=j;
    NumberofOutputNeurons=number_class

    %%%%%%%%%%%%% Processing the targets of training
    temp_T=zeros(NumberofOutputNeurons, NumberofTrainingData);
    for i = 1:NumberofTrainingData

```

```

        for j = 1:number_class
            if label(1,j) == T(1,i)
                break;
            end
        end
        temp_T(j,i)=1;
    end
    T=temp_T*2-1;
end % end if of
Elm_Type

%%%%%%%%%% Calculate weights & biases
start_time_train=cputime;

%%%%%%%%%% Random generate input weights InputWeight (w_i) and biases
BiasofHiddenNeurons (b_i) of hidden neurons
InputWeight=rand(NumberofHiddenNeurons,NumberofInputNeurons)*2-1;
BiasofHiddenNeurons=rand(NumberofHiddenNeurons,1);
tempH=InputWeight*P;
clear P; % Release input of
training data
ind=ones(1, NumberofTrainingData);
BiasMatrix=BiasofHiddenNeurons(:,ind); % Extend the bias
matrix BiasofHiddenNeurons to match the demention of H
tempH=tempH+BiasMatrix;

%%%%%%%%%% Calculate hidden neuron output matrix H
switch lower(ActivationFunction)
    case {'sig','sigmoid'}
        %%%%%%%%% Sigmoid
        H = 1 ./ (1 + exp(-tempH));
    case {'sin','sine'}
        %%%%%%%%% Sine
        H = sin(tempH);
    case {'hardlim'}
        %%%%%%%%% Hard Limit
        H = hardlim(tempH);
        %%%%%%%%% More activation functions can be added here
end
clear tempH; % Release the
temporary array for calculation of hidden neuron output matrix H

%%%%%%%%%% Calculate output weights OutputWeight (beta_i)
OutputWeight=pinv(H') * T';
end_time_train=cputime;
TrainingTime=end_time_train-start_time_train % Calculate CPU
time (seconds) spent for training ELM

%%%%%%%%%% Calculate the training accuracy % Y: the actual
Y=(H' * OutputWeight)';
output of the training data
if Elm_Type == REGRESSION % Calculate
    TrainingAccuracy=sqrt(mse(T - Y))
    training accuracy (RMSE) for regression case
    output=Y;
end

```

```

clear H;

if Elm_Type == CLASSIFIER
    %%%%%%%%% Calculate training & testing classification accuracy
    MissClassificationRate_Training=0;

    for i = 1 : size(T, 2)
        [x, label_index_expected]=max(T(:,i));
        [x, label_index_actual]=max(Y(:,i));
        output(i)=label(label_index_actual);
        if label_index_actual~=label_index_expected

MissClassificationRate_Training=MissClassificationRate_Training+1;
            end
        end
        TrainingAccuracy=1-
MissClassificationRate_Training/NumberOfTrainingData
    end

if Elm_Type~=REGRESSION
    save('elm_model', 'NumberOfInputNeurons', 'NumberOfOutputNeurons',
'InputWeight', 'BiasofHiddenNeurons', 'OutputWeight',
'ActivationFunction', 'label', 'Elm_Type');
else
    save('elm_model', 'InputWeight', 'BiasofHiddenNeurons',
'OutputWeight', 'ActivationFunction', 'Elm_Type');
end

```

C. Trained ELM for classifying corner points (for test baggage) as belonging to the cuboidal body or the handle of the baggage and to compute the convex hull over only those points that are classified as belonging to the visually salient cuboidal body of the baggage thereby implementing a computational model of the part based object representation theory. (Original Code credits – Mr. Qin-Yu Zhu and Dr. Guang-Bin Huang. Modified to meet the needs of this research/application).

% This trained ELM is then used to operate on the image patches around the extracted corner points after segmentation using the Chan-Vese active contours using the code similar to A. After classifying the points as belonging to the body / handle of the baggage a convex hull over all the points on the baggage body edge is computed to obtain the salient cuboidal contour of the baggage. This therefore conforms to the part based object representation model as well as captures the intuition of part saliency in which the top-down influence of the task to be performed (that of pose estimation) is taken into consideration.

```

function [TestingTime, TestingAccuracy] = elm_predict(TestingData_File)

% Usage: elm_predict(TestingData_File)
% OR:    [TestingTime, TestingAccuracy] = elm_predict(TestingData_File)
%
% Input:
% TestingData_File      - Filename of testing data set

```



```

%
% Output:
% TestingTime          - Time (seconds) spent on predicting ALL testing
data
% TestingAccuracy      - Testing accuracy:
%                      RMSE for regression or correct
classification rate for classification
%
% MULTI-CLASSE CLASSIFICATION: NUMBER OF OUTPUT NEURONS WILL BE
AUTOMATICALLY SET EQUAL TO NUMBER OF CLASSES
% FOR EXAMPLE, if there are 7 classes in all, there will have 7 output
neurons; neuron 5 has the highest output means input belongs to 5-th
class
%
% Sample1 regression: [TestingTime, TestingAccuracy] =
elm_predict('sinc_test')
% Sample2 classification: elm_predict('diabetes_test')
%
%      %%%      Authors:      MR QIN-YU ZHU AND DR GUANG-BIN HUANG
%      %%%      NANYANG TECHNOLOGICAL UNIVERSITY, SINGAPORE
%      %%%      EMAIL:       EGBHUANG@NTU.EDU.SG; GBHUANG@IEEE.ORG
%      %%%      WEBSITE:     http://www.ntu.edu.sg/eee/icis/cv/egbhuang.htm
%      %%%      DATE:       APRIL 2004

%%%%%%%%%%%% Macro definition
REGRESSION=0;
CLASSIFIER=1;
global im;
global shape;

%%%%%%%%%%%% Load testing dataset
test_data=xlsread(TestingData_File);
% TV.T=test_data(:,1)';
TV.P=test_data(:,1:(size(test_data,2)-2))';
%clear test_data; % Release raw
testing_data_array
size(TV.P);
NumberofTestingData=size(TV.P,2);

load elm_model.mat;

% if Elm_Type~=REGRESSION
%
%      %%%%%%%%% Processing the targets of testing
%      temp_TV_T=zeros(NumberofOutputNeurons, NumberofTestingData);
%      for i = 1: NumberofTestingData
%          for j = 1: size(label,2)
%              if label(1,j) == TV.T(1,i)
%                  break;
%              end
%          end
%          temp_TV_T(j,i)=1;
%      end
%      TV.T=temp_TV_T*2-1;
%

```

```

% end                                     % end if of
Elm_Type

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%% Calculate the output of testing input
start_time_test=cputime;
tempH_test=InputWeight*TV.P;
clear TV.P;                               % Release input of testing data
ind=ones(1,NumberOfTestingData);
BiasMatrix=BiasofHiddenNeurons(:,ind);    % Extend the bias
matrix BiasofHiddenNeurons to match the dimention of H
tempH_test=tempH_test + BiasMatrix;
switch lower(ActivationFunction)
    case {'sig','sigmoid'}
        %%%%%%%%% Sigmoid
        H_test = 1 ./ (1 + exp(-tempH_test));
    case {'sin','sine'}
        %%%%%%%%% Sine
        H_test = sin(tempH_test);
    case {'hardlim'}
        %%%%%%%%% Hard Limit
        H_test = hardlim(tempH_test);
        %%%%%%%%% More activation functions can be added here
end
t=H_test';
q=OutputWeight;
size(t)
size(q)
TY=(H_test' * OutputWeight)'              % TY: the actual
output of the testing data
end_time_test=cputime;
TestingTime=end_time_test-start_time_test; % Calculate CPU
time (seconds) spent by ELM predicting the whole testing data

if Elm_Type == REGRESSION
    TestingAccuracy=sqrt(mse(TV.T - TY))    % Calculate
testing accuracy (RMSE) for regression case
    output=TY;
end

if Elm_Type == CLASSIFIER
    %%%%%%%%% Calculate training & testing classification accuracy
    % MissClassificationRate_Testing=0;
    figure()
    imshow(im)
    hold on
    axis on
    A=zeros(size(im));
    coordx= 0;
    coordy= 0;
    for i = 1 : size(TV.P, 2)
        % [x, label_index_expected]=max(TV.T(:,i));
        [x, label_index_actual]=max(TY(:,i));
        output(i)=label(label_index_actual);
        % label_index_actual
        % if label_index_actual== label_index_expected
        % if label_index_actual ~=1

```

```

        plot(test_data(i,101),test_data(i,102),'r*');
        hold on
        A(test_data(i,102),test_data(i,101))=1;
        coordx =[coordx, test_data(i,101)];
        coordy= [coordy, test_data(i,102)];
    else
    end
    % else label_index_actual~=label_index_expected
    %     t=test_data(i,101)
    %     s=test_data(i,102)
    %
    MissClassificationRate_Testing=MissClassificationRate_Testing+1;
    % end
end

% TestingAccuracy=1-
MissClassificationRate_Testing/NumberofTestingData;
end
figure(2)
imshow(A)
axis on

revcoordx = coordx(2:end)';
revcoordy = coordy(2:end)';
combcoords = [revcoordx;revcoordy];
sort(combcoords);
k = convhull (revcoordx, revcoordy); % Computation of the convex hull
over the corner points classified as belonging to the cuboidal
contour(body) of the baggage.
figure(3)
imshow(A)
hold on
plot(revcoordx(k), revcoordy(k))
F = getframe (figure(3));
X = frame2im(F);
iptsetpref('ImshowBorder','tight');
set(gca,'position',[0 0 1 1],'units','normalized')
p = min(revcoordx)
q = max(revcoordx)
r = min(revcoordy)
s = max(revcoordy)
plot((p+q)/2,(r+s)/2,'*')
hold off
figure(4)
X= rgb2gray(X);
SIZ = size(X);
I = find(X>0);
X(I)= 255;
X= im2bw(X); % Binarizing the image obtained by joining the contour
points on the baggage body.
shape = X;
imshow(X);
save('elm_output','output');

```

D. Estimation of best fitting template from the database of stored cuboidal part templates using the edge alignment based Chamfer matching technique. These templates differ not only in pose but also aspect ratios so that the best fitting

template obtained through our approach yields the best localization information and the corresponding translation parameters. In general, the accuracy of pose estimation increases with increase in number of templates in the database by avoiding the false match. The problem of baggage pose estimation is therefore reconstructed as the problem of estimating the pose of its salient cuboidal part ignoring other non-salient part descriptors thereby simplifying the matching process.

```
% This code is used to ascertain the best fitting template from among
the database of stored cuboidal templates. Each template is
translated over the cuboidal part contour image and the Chamfer score
is recorded for each translation. A minimizer of scores is applied
across the translations for a given template that yields the best
parameter estimates for that template. A second minimizer is applied
across all the templates in the database to obtain the best fitting
template and the corresponding translation parameters.
global shape;
image = im2bw(imread('C:\Users\sriram005\Downloads\shape41.jpg'));
image = imresize(image,size(shape));
D = bwdist(shape);
sum((D(find(image == 1)))));
[r c] = find(image==1);
image(r,c)= 0;
for i=1: size(r,1)
    image(r(i,1)-20,c(i,1)-20)=1; % Translating each database template
cuboidal part shape over the computed distance transform image of the
extracted cuboidal contour of the baggage;
end
sum((D(find(image == 1)))) % Record the Chamfer score after each
translation.
figure(3)
c= imfuse(image,shape); % Show the extracted baggage cuboidal contour
fused with the edge template in one image for better understanding.
imshow(c)
```