

Pairing Approaches in Object Identification

NIK MOHD ZARIFIE BIN HASHIM

Abstract

Object identification has become one of the focussed areas in computer vision. It is widely utilized in various computer vision applications, e.g., robotics, security, mobile apps, transportation, and many more. A conventional definition of object identification, is deciding whether two observations are the same object or not. This thesis redefines this definition by giving a broader alternative context to the existing computer vision applications that could be slotted in such a way to describe the conventional definition more clearly.

Chapter 1 introduces a new definition of image matching for an identity as “object identification” to systematically generalize two, or even more scenarios. To identify objects in a computer vision application, input images are generally matched to another set of output images. These two image sets are generally identified by inferring from both input and output sides as an image pair. Based on this idea, the “pairing concept” is introduced here. To discuss this new concept, existing computer vision, especially object identification tasks are carefully analyzed and redefined. To start with, object identification is categorized into three categories; Instance-to-Class, Instance-to-Value, and Instance-to-Instance. The first category, Instance-to-Class object identification, is familiar and straightforward as in traditional image recognition tasks, e.g., face classification. Next, the second category, Instance-to-Value object identification, provides solutions for different image recognition problem settings, e.g., regression and key-point matching. For these two categories of object identification, since the instances are not paired, but rather straightforwardly identified as one class or value, they are not the main interest of this thesis. Meanwhile, the third category, Instance-to-Instance object identification, is discussed in this thesis through the newly proposed pairing concept between input and output instances. In order to explain the pairing concept, the thesis will focus on Research topic 1: Object pose estimation with incremental viewpoints and Research topic 2: Set-to-set Person

re-identification. The former is an example of pairing images from different viewpoints for optimal multi-view object pose estimation, while the latter is an example of pairing images of people for wide area surveillance, respectively.

Chapter 2 reviews the current work in the discussed fields and comprehensively analyzes the state-of-the-art in this field. Existing works related to object identification for pose estimation and person re-identification, together with object identification applications are introduced.

Chapter 3 introduces an example of the Instance-to-Instance object identification as pairing images from different viewpoints; Research topic 1: object pose estimation with incremental viewpoint. Generally, the aim of the task is estimating an object's pose from a single observation. Traditionally, the estimated pose is a value that is yielded as a result of Instance-to-Value object identification. However, most existing works on single-viewpoint pose estimation face the ambiguity problem, which occurs when an object cannot be fully captured from one viewpoint or occluded. To solve this ambiguity problem, it is essential to select an alternative viewpoint. Averaging the original and current viewpoints with a careful arrangement and decision could infer the best viewpoint among all viewpoints. For this, this thesis introduces an entropy-based score of the object pose ambiguity; by selecting a viewpoint that minimizes this score, the best next-viewpoint is recommended. Evaluation is performed with synthetic object images of several indoor object categories. It demonstrates that the proposed method can properly estimate a pose when facing an ambiguous angular pose for a given object category, which is very important when considering the pose estimation in a categorical level, e.g., comparing mug images from many mug types.

Chapter 4 introduces an example of the Instance-to-Instance object identification as pairing images of people; Research topic 2: simultaneous person re-identification. In general, the aim of this task is identifying all persons in scenes captured from different cameras which could be considered as pairing persons between query and

gallery sets from the pairing perspective. Traditionally, the single query pairing setting in Instance-to-Instance object identification is applied, whereas similarity of persons that appear in two different cameras is compared individually. However, in such a naïve approach, redundant matching or pairing occurs, where one of the persons could be paired with multiple persons, leading to degraded performance. This occurs when the pairing is performed without considering the successfully paired persons. In addition, until recently, a small number of work focussed on non-similar image numbers in the person re-identification task, where the number of images in query and gallery sets are unbalanced. Therefore, this thesis proposes a person re-identification method that challenges two issues; redundant pairing and multiple query pairing for person identification based on object selection and arrangement during the image pairing process. Concretely, the Stable Marriage Algorithm (SMA) is introduced to solve the problem. Evaluation is performed with publicly existing dataset images of pedestrians in a two-camera condition setting. It demonstrates that the proposed method can successfully pair the persons between query and gallery cameras individually or simultaneously, which is essential when facing similar and non-similar numbers of images in query and gallery sets.

Chapter 5 summarizes the thesis and discusses object identification based on Research topics 1 and 2, where their solutions can be considered under one framework. To better understand the implications of pairing approaches in object identification, as the thesis's future outlook, the studies and exploration could address other applications in computer vision under the proposed framework.

Acknowledgments

This dissertation is formally submitted for fulfilling partial requirements for the degree of Doctor of Information Science in Media Science from Nagoya University. This work would not have been possible without the help of many people to whom I owe the sincerest gratitude.

I would first like to thank Prof. Dr. Hiroshi Murase for accepting me into his laboratory twice, beginning as a research student in early 2016 and then starting in October 2016 for my doctoral studies. I want to thank him for his kind help and support throughout my time in Japan.

Here, I would like to thank Associate Prof. Dr. Daisuke Deguchi for giving me many opinions and comments in my study period and being my supervisor for the study.

I want to express my gratitude to Dr. Yasutomo Kawanishi for his supervision and assistance in this research. His input and insight provided the direction for this research, and I am highly grateful for his contributions and support. He also provided me with the chance to do joint research with other students, assisting various research projects around the laboratory. I am grateful for the experiences and knowledge I gained through these opportunities, especially in the Recognition Group activities.

I would also like to thank Prof. Dr. Ichiro Ide, who gave invaluable feedback and assistance for various papers and submissions I wrote while studying at Murase Laboratory. His meaningful feedback, often coming from an entirely different viewpoint than my other supervisors, provided an excellent and helpful addition to improve my work.

I thank Dr. Norimasa Kobori and Ms. Ayako Amma from Woven Planet Group and Toyota Motor Corp. for their feedback and time during the joint research meetings. The research experience from the collaborative university-industry project will be one of the valuable experiences I gained during my doctoral degree study.

Special thanks go to all members of the Murase Laboratory, past and present, who have helped me throughout this journey. The secretaries, Mrs. Hiromi Tanaka and Mrs. Fumiyo Kaba have always been a great help when organizing business trips and official procedures.

I am also grateful to Dr. Mahmud Dwi Sulistiyo for discussing research and future career plans. I would like to thank all of the students, of whom there are too many to name one by one, for their friendship.

I thank the Ministry of Education, Government of Malaysia, and Universiti Teknikal Malaysia Melaka (UTeM) that have supported me in the form of a Skim Latihan Akademik IPTA (SLAI) scholarship throughout my study as a research student (April – September 2016) and Ph.D. student (October 2016 – September 2019.) Here, I would also thank the supports directly or indirectly from my colleague from Fakulti Kejuruteraan Elektronik dan Kejuruteraan Komputer (FKEKK) in completing the doctoral degree study from 2016 to 2022.

Finally, I would like to thank my family. Despite living far away, my mother, Tuan Kismah, has always supported and encouraged my education plans and career goals. I would also like to thank my wife Aishah, my daughters, Nik Dhia Amani, Nik Khayla Zenia, Nik Keisya Zaina, and my son, Nik Rizq Al Haq, for giving me strength and happiness in my whole life. Without them, my life would not be the same.

Thank you ALLAH.

Contents

Abstract	iii
Acknowledgments	vii
Contents	ix
List of Figures	xiii
List of Tables	xv
Abbreviations	xvii
1 Introduction	1
1.1 Pairing	2
1.2 Object Detection Applications in Computer Vision	5
1.2.1 Computer vision for robotic applications	7
1.2.2 Computer vision for security applications	9
1.2.3 Computer vision for mobile applications	11
1.3 Object Identification	12
1.3.1 Pairing instances in object identification	14
1.3.2 Instance-to-Class pairing	16
1.3.3 Instance-to-Value pairing	17
1.3.4 Instance-to-Instance pairing	18
1.3.4.1 Single-query pairing	19
1.3.4.2 Multiple-query pairing	19
1.4 Research Overview	20
1.4.1 How to develop a single-query pairing?	20
1.4.2 How to develop a multiple-query pairing?	22
1.4.3 Research topic 1: Object pose estimation	25
1.4.4 Research topic 2: Person re-identification	26
1.5 Thesis structure	28

2	Related Research	31
2.1	Object Identification for Pose estimation	32
2.1.1	Instance-to-Value pairing: Value estimation for single view-point pose estimation	32
2.1.2	Instance-to-Instance pairing: Viewpoint instance selection for multiple viewpoint object pose estimation	33
2.1.3	Instance-to-Instance pairing: Active vision for object pose estimation	35
2.2	Object Identification for Person Re-identification	36
2.2.1	Single-query pairing: Individual image pairing	37
2.2.2	Multiple-query pairing: Simultaneous image pairing	38
3	Pairing Approach for Object Pose Estimation	39
3.1	Introduction	40
3.2	Details on the Pairing Approach for Object Pose Estimation	41
3.3	Best Next-Viewpoint Recommendation by Selecting Minimum Pose Ambiguity for Category-Level Object Pose Estimation	43
3.3.1	Minimum pose ambiguity selection framework	44
3.3.2	Viewpoint likelihood distribution	46
3.3.3	Pose likelihood distribution	47
3.3.4	Pose estimation	49
3.4	Experiments	50
3.4.1	Dataset	50
3.4.2	Pose estimation method	51
3.4.3	Evaluation criteria	52
3.4.4	Comparative methods	54
3.4.5	Results	55
3.4.5.1	Comparison on Mean Absolute Error (MAE)	55
3.4.5.2	Comparison on Pose Estimation Accuracy (PEA)	57
3.5	Discussion	59
3.5.1	Quantitative evaluation	59
3.5.2	Qualitative evaluation	62
3.5.3	Limitation and further considerations	64
3.6	Summary	65
4	Pairing Approach for Person Re-identification	67
4.1	Introduction	68
4.2	Details on the Pairing Approach for Person Re-identification	69
4.3	Multiple-Query Pairing Approach for Person Re-identification via the Stable Marriage Algorithm	71
4.3.1	Marriage problems	72
4.3.2	Simultaneous image pairing via the Stable Marriage Algorithm	73

4.3.3	Image similarity ranking	74
4.3.4	Simultaneous image pairing implementation for person re-identification	77
4.4	Experiments	78
4.4.1	Dataset	78
4.4.2	Comparative methods	79
4.4.3	Image pairing settings	81
4.4.4	Results	82
4.4.4.1	Comparison on average pairing accuracy	82
4.4.4.2	Comparison on different numbers of images	84
4.5	Discussion	86
4.5.1	Quantitative evaluation	86
4.5.1.1	Comparison of computational time	87
4.5.1.2	Comparison of the selection of image patch	89
4.5.2	Qualitative evaluation	90
4.5.3	Limitation and further consideration	93
4.6	Summary	94
5	Conclusion	95
5.1	Summary	95
5.2	Remaining Challenges and Future Directions	98
5.3	Closing Remarks	100
	Bibliography	103
	Publication list	122

List of Figures

1.4	Example of identification tasks; (a) identification for image sets and (b) identification for generic instances.	13
1.13	Thesis structure	29
3.1	Value estimation for object pose estimation	42
3.2	Idea of the pairing for object pose estimation	43
3.3	Illustration of the idea on the one given viewpoint (top), and two given viewpoints (bottom) for the next viewpoint recommendation	44
3.4	Pose ambiguity distribution when selecting a next viewpoint δ [$^{\circ}$] from the initial viewpoint (Input image I with $\phi = 95^{\circ}$) as shown in Figure 1.10	45
3.5	Viewpoint likelihood distribution $p(\phi I)$ (Input image I with $\phi = 95^{\circ}$)	47
3.6	Pose likelihood distribution $p(\theta I, \delta)$ given two viewpoints (Input image I with $\phi = 95^{\circ}$)	48
3.7	Viewpoint likelihood distribution $p(\phi I)$ (Input image I with $\phi = 95^{\circ}$)	49
3.8	Example of images from the five classes in the ShapeNet dataset [157], used in the experiment	51
3.9	Example of “Mug” images observed from different elevation angles	51
3.10	Original Pose-CyclicR-Net [117]. In the convolution (conv) layers 1 and 2, the light yellow boxes refer to the 2D convolution, the orange boxes to the ReLU layer, and the light red boxes to the maxpooling layer. In the fully connected (fc) layers 3, 4, 5, and 6, the light purple box refers to the dense layer and the dark purple boxes to the ReLU layer	52
3.11	Example of images observed from the estimated viewpoints by the proposed method and comparative methods	55
3.12	Pose Estimation Accuracy by changing the error threshold τ from 0° to 100° (elevation angle = 30°)	58
3.13	Example of one-axis symmetrical object	60
3.14	Image examples for the four viewpoint groups using the “Mug” object category	62
4.1	Image pairing in person re-identification	70

4.2	Overview of the proposed method. The person image is considered as an instance of the Stable Marriage Problem (SMA). The gray circles represent both camera views' person images as an element of each set, while the light blue circles in the blue shaded areas represent preference lists sorted horizontally for each camera view from preferable to unpreferable. SMA matches images from camera view A's image set to camera view B's image set based on the preference list. The blue line represents the matching from person a_1 to b_2 , the green line from a_2 to b_3 , and the red line from a_3 to b_1	72
4.3	Examples of mask images using the VIPeR dataset [48].	75
4.4	Example of image cropping from four different regions of a person.	89
4.5	Examples of unsuccessfully masked images (using Mask R-CNN) for the four datasets used in the evaluation.	91
4.6	Comparison of the matching results between the Greedy Matching (GM; blue rectangular) and the proposed method utilizing SMA (red rectangular). The rectangulars indicate the matched images.	92

List of Tables

1.1	Pairing concept for identification tasks	4
3.1	Comparison of MAE for the five object categories when the elevation angle is 0° by five-fold cross validation	56
3.2	Comparison of MAE for different elevation angles of “Mug” by five-fold cross validation	56
3.3	Comparison using partial - Area Under Curve (pAUC) of Pose Estimation Accuracy (PEA) by changing the error threshold for for the five object categories when the elevation angle is 0° by five-fold cross validation	57
3.4	Comparison using partial- Area Under Curve (pAUC) of Pose Estimation Accuracy (PEA) by changing the error threshold τ from 0° to 100° by five-fold cross validation	58
3.5	Comparison of the overall Mean Absolute Error (MAE) and Partial-Area Under Curve (pAUC) of Pose Eastimation Accucarcy (PEA) for each fold (0° elevation angle)	61
3.6	Comparison of MAE for “Mug” for different initial viewpoint groups (0° elevation angle)	63
3.7	Comparison of MAE for “Mug” when the elevation angle is 0° for viewpoint groups. The value in brackets represent the difference between the initial viewpoint and the pose estimation result	64
4.1	Example of image similarity based on feature similarity	78
4.2	Comparison of pairing methods in matching accuracy on VIPeR[48], CUHK01[49], iLIDS-VID[158], and PRID[50] datasets.	83
4.3	Comparison of matching accuracy with different numbers of images in two camera views settings on the VIPeR dataset [48]	84
4.4	Comparison of matching accuracy with different numbers of images in two camera views settings on the CUHK01 dataset [49]	84
4.5	Comparison of matching accuracy with different numbers of images in two camera views settings on the iLIDS-VID dataset [158]	85
4.6	Comparison of matching accuracy with different numbers of images in two camera views settings on the PRID dataset [50]	85
4.7	Comparison of computation time	86
4.8	Comparison on Different Cropping Regions using SMA + HSV method	90

Abbreviations

2D	2 Dimensional
3D	3 Dimensional
CCTV	Closed-Circuit Tele V ision
CNN	Convolution Neural Network
CPU	Central P rocessing U nit
CUHK01	First Version of Person Re-identification Datasets from Chinese University of H ong K ong
DCNN	Deep Convolution Neural Network
GHz	G iga H ertz
GM	G reedy M atching
HM	H ungarian M atching
HSV	H ue S aturation V alue
iLIDS-VID	Imagery L ibrary for I ntelligent D etection S ystems - V IDeo
LIDAR	L ight D etection A nd R anging
MSDALF	M ask-improved S ymmetry- D riven A ccumulation F eatures
MAE	M ean A bsolute E rror
Mask R-CNN	M ask R eigion-based Convolution Neural Network
pAUC	p artial A rea U nder C urve
PCA	P rincipal C omponent A nalysis
PEA	P ose E stimation A ccuracy
PRID	P erson R e I Dentification
PRiSM	P erson R e- i dentification via S tructured M atching
RAM	R andom A ccess M emory

Re-Id	Re-Identification
ReLU	Rectified Linear Unit
SDALF	Symmetry-Driven Accumulation Features
SMA	Stable Marriage Algorithm
VIPeR	Viewpoint Invariant Pedestrian Recognition
RGB-D	Red Green Blue and Depth

For my late father - Hashim Idris

Chapter 1

Introduction

Today, the innovation and technology surrounding human life keep growing upwards, accelerating highly accurate and reliable systems for all life purposes. Human interactions with technology appliances have changed over time [1]. This interaction change is easy to realize, along with the Industrial Revolution, the transition to new manufacturing processes in United Kingdom, continental Europe, and the United States, from the first Industrial Revolution to the recent Fourth Industrial Revolution (IR 4.0) [2] The positive change in the way we live, is that technology helps us gain quality and safe life. One of the technical fields which deliver good quality in life is computer vision. Computer vision endeavors to extract creative and practical information of an object's visual appearances from signals received from a visual sensor in the form of an an image or a video [3]. From robotic manufacturing to security monitoring and alerting applications, computer vision delivers a vast benefit to people [4]. Recently, it has been considered as one of Artificial Intelligence (AI) solutions to many types of problems [5].

The object identification topics are discussed frequently as a computer vision task such as object detection, object classification, object estimation, object re-identification, and many more. Most of the work in object identification has focussed on extracting better image features until the prevalence of the recent deep learning [6, 7]. Nevertheless, the study on the basic matching or identification between one object to

another object is yet necessary. This thesis is presented as the author’s challenge of looking into how an object is properly matched or identified. The thesis introduces a new definition of “object identification” to identify objects in various aspects in various existing computer vision tasks. Existing computer vision tasks will be redefined under the “pairing” concept; considering an object paired with another object, which will lead the thesis to a new way of observing conventional object identification.

The following Section 1.1 will describe the idea of pairing. Then by utilizing three new categories of object identification tasks under the pairing concept; Instance-to-Class, Instance-to-Value, and Instance-to-Instance, this pairing framework is explained. Since the former two categories have previously been studied well, they will not be studied in detail in this thesis, but they will be explained how they could be a part of the categories under the new concept. Next, Section 1.2 introduces existing computer vision applications, which serves as an introduction to the newly defined pairing concept for object identification tasks. The main contribution, including the overview of the proposed methods is introduced in Section 1.3. Furthermore, general descriptions of the research covering problems and proposed solutions are introduced in Section 1.4. Finally, Section 1.5 presents the structure of this thesis.

1.1 Pairing

To explain the importance of pairing in identification, the matching in an identification task with the pairing concept is introduced and discussed in this Section. Pairing, the term we use in the thesis, finds the best combination of two instances by comparing the similarity of instances or images between compared images. Introducing the pairing concept as the main core of identification could improve the general identification performance, since it carefully pairs the instances based on the needs of each application.

The identification problem could be observed in a new **pairing** framework with a variety of problem settings, depending on the given input and output for various

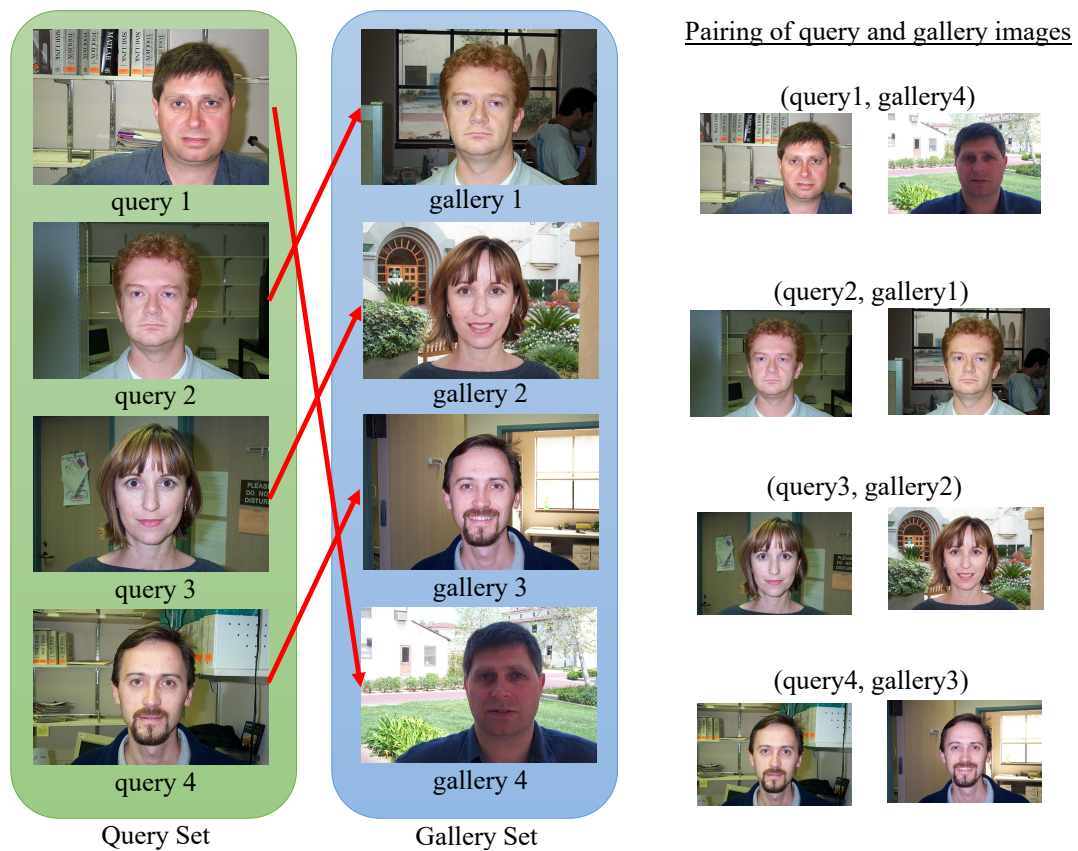


Figure 1.1: Instances in query and gallery sets for face recognition. The query image and gallery image e.g., (query1, gallery4) is defined as paired images under the pairing concept (Images taken from [10]).

computer vision tasks. Such computer vision tasks will be discussed in Section 1.2. For each of those tasks, e.g., surveillance [8], face recognition [9], object identification could be defined as pairing of an query instance ID with gallery instance IDs by binary classification as illustrated in Figure 1.1. As such, we could redefine various object identification tasks in a new manner. Concretely, the object identification tasks are classified into four categories, as illustrated in Table 1.1; classification, regression, single query pairing, and multiple query pairing.

Table 1.1: Pairing concept for identification tasks

Pairing category	Task Category	Applications	
		Security	Robotics
Instance-to-Class	Classification	Face recognition (Object-Class pairing)	Face recognition (Object-Class pairing)
Instance-to-Value	Regression		Single viewpoint pose estimation (Object-Value pairing)
Instance-to-Instance	Single query pairing	Specific person retrieval (Person-Person pairing)	Object image retrieval (Object-Object pairing)
	Multiple query pairing	<i>Person re-identification</i> (Set of <i>Person-Person</i> pairing)	Simultaneous Localization and Mapping (SLAM) (Set of Key-point pairing)

1.2 Object Detection Applications in Computer Vision

Computer vision provides a rich set of information e.g., objects, backgrounds, and site conditions about a scene by taking images or videos, which facilitates the understanding of the complex construction tasks rapidly, accurately, and comprehensively [11]. Computer vision tasks related to images are also described as non-sequential tasks that are generally considered input image to compare their similarity scores with other images [12, 13]. Meanwhile for sequential tasks, this input video will be treated as multiple image frames [14]. The main objective of computer vision is to produce an understanding and useful description of visual scenes and of objects that populate them by performing operations on the signals received from a visual sensor in the form of an image or a video. This thesis focusses on one of the tasks in computer vision; object identification, which analyzes images for several computer vision applications.

Although various methodologies achieved good object identification results in recent works, they could be redefined by considering the object as a new instance to find image pairs in the identification task for various applications. However, the computer vision tasks in a broad view is nearly identical when we consider their applications. For object detection, as one of the computer vision tasks, person re-identification could be implemented via several problem settings in its object identification task [15]. Comparing one person's image with other images in the gallery image set one by one, could be one way to solve the identification task. Then, let's compare all images in the query image set to all gallery image sets simultaneously; this could give us another way to solve the identification task. From these two given scenarios, the security applications via person re-identification tasks could be recategorized by focussing on how objects in each object identification problem is to be paired. On the other hand, for security and mobile applications, the object involved here could be a person as a whole body, thumbprint, or any kind of person's biometrics; Although a person's body is one of the target objects for object identification, the way that person is being compared is not only by pairing the whole person's

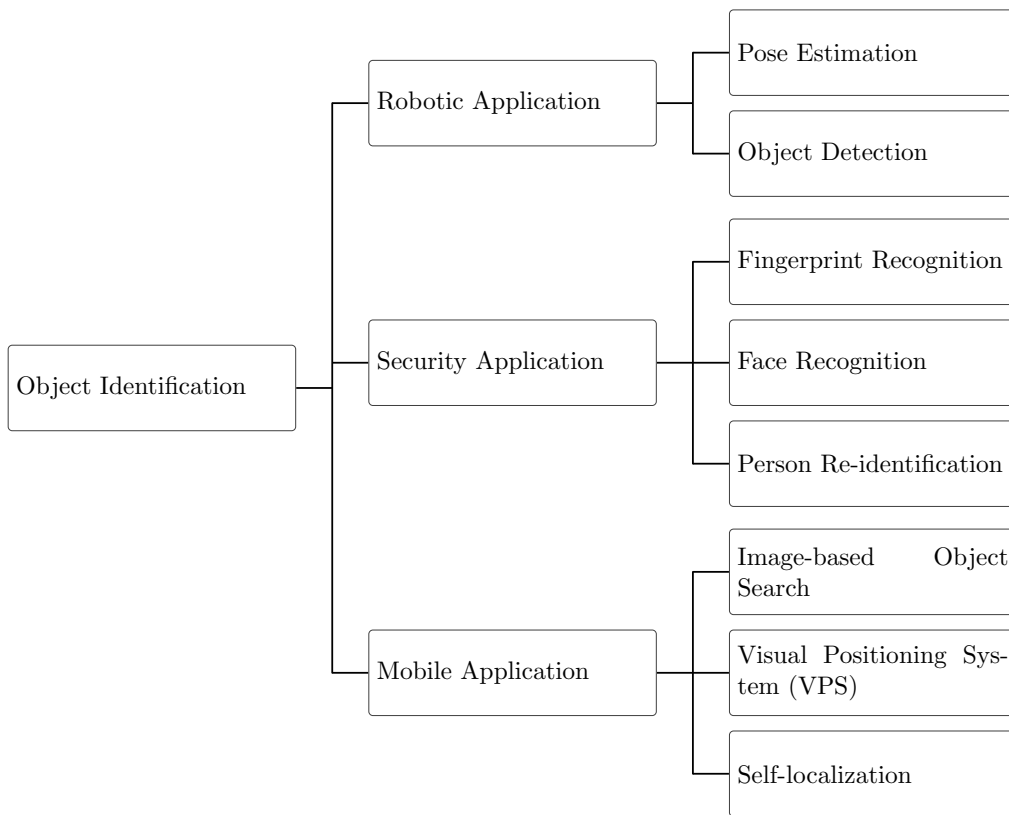


Figure 1.2: Some computer vision applications categorized by applications.

image to image. In this context, object identification could be the target person's walking direction, or the sex.

In this thesis, as mentioned earlier, the target object in one identification task could be redefined into another context than the existing computer vision applications such as robotics, security, mobile, medical, and automotive for finding good matched pairs. To discuss the new categories for object identification in an orderly route, the first three applications, robotic, security, and mobile, are focussed in this thesis as illustrated in Figure 1.2. These three applications are discussed because the understanding of the identification task in each of them could help explaining the new pairing concept. The following Sections will redefine these three application categories under the pairing concept.

This Section discusses how some computer vision applications could be categorized,

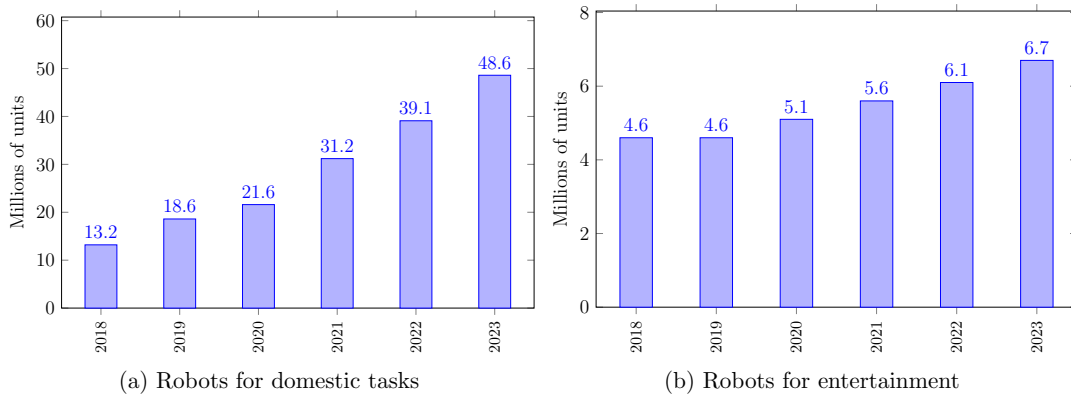


Figure 1.3: Service robots for personal / domestic use. Unit sales 2018 and 2019, potential development 2020-2023 [16].

introduces the pairing concepts to object identification tasks, and discusses the limitations of current computer vision-based techniques.

1.2.1 Computer vision for robotic applications

The demands for robotics applications have recently increased. The upsurging trend of sales statistics for domestic and entertainment purposed robotics are shown in Figure 1.3 [16]. From home to retail, military, and various other sectors, robotics performs various tasks making the work process faster with more efficient and accurate results. To reliably perform these various tasks, a robotics application requires a sound vision of a target object, e.g., estimating the object pose and detecting the presence of an object in a limited camera view here plays like a human eye.

There are several robotics applications that make use of computer vision tasks. Since the new pairing concept introduced in the thesis is related to the application of the human helper robot's pose estimation, pose estimation for robotics applications is discussed here. For robotic applications, this pose estimation task could be discussed in two sub-categories; Pose estimation and Object detection.

For an autonomous robot, it is important to estimate the location of the robot. Feng and Evangelos's work [17] in 1997 is among the early works. They proposed an

object estimation algorithm that prevents the robot from occluding with other objects while moving by matching data points obtained from the previously scanned information. Later, Sim et al. [18] proposed a method for camera pose estimation by parameterizing landmarks on the environment. These two studies analyze the environment information to estimate the object's pose to avoid an obstacle for a moving robot. Warren et al. [19] presented the development of a low-cost sensor platform for use in ground-based visual pose estimation and scene mapping tasks for estimating robot position for SLAM tasks accurately. The implementation of pose estimation from multi-view object recognition with a full pose estimation was also introduced in highly cluttered scenes by Collect et al. [20] from 3D object models.

Shakunaga's work [21] in 1992 is among the early work that proposed object pose estimation using a single camera. He conducted experiments with synthesized and real images and showed the effectiveness of this system for pose estimation and 3D pose tracking. Then, Murino and Foresti [22] proposed a different Hough space embedding 3D information than the traditional 2D Hough space to estimate poses of planar objects in a single gray-level image. Recently, utilizing deep learning, Tatemichi et al. [23] proposed an occlusion-robust pose estimation method of an unknown object instance in an object category from a depth image.

Vision-based object detection is another essential computer vision task for robotics applications. Object detection for robotics applications needs to work under diverse environments in which such applications are required to operate and also strict constraints in terms of run-time, power, and space. Gould et al. [24] integrated the visual and range data for the robot object detection by combining 2D and 3D sensors to enhance the object detection in the cluttered real world. Object detection also utilizes sound in a sound-indicated vision framework for localizing and detecting an object in a robotics application [25]. The demand for real-time object detection later propelled Holz et al. [26] to propose a pipeline for object detection, localization, and verification for robotics depalletizing tasks.

1.2.2 Computer vision for security applications

Compared to traditional security services in human life, implementing computer vision here changes the technology paradigm for human life in maintaining the sustainability of safety and security surrounding us. Computer vision plays the role of helping humans in many ways, including human characteristics/biometrics, people identification, authentication to access people's belonging, and many more. To reliably perform these various security tasks via several types of inputs, image, video, or sound, computer vision can provide a reliable solution, e.g., for keeping the security 24/7 without human help managing a home or kitchen security system, especially at night [27, 28]. Furthermore, accessing email, social media applications or electronic banking has overtaken access from conventional computers, shifting mobile devices into essential tools in our everyday lives [29]. Thus, the security issue in accessing this privacy information is crucial to consider in mobility scenarios. Toward this security in this mobile application, the development of biometric research via facial images and human fingerprint recognition is among the idea to protect the user [30]. With the help of sophisticated computer vision algorithms and techniques, computer vision in security applications has been able to equal or even beat human performance on tasks such as object recognition.

Since each human own individual physical or human behavioral characteristics, which are called biometrics, all such information can be used to digitally identify a person to grant access to systems, devices, or data. Automated fingerprint recognition proposed by Hrechak et al. [31] using structural matching is based on local structural relations among features and an associated automated recognition system. However, the work still struggles with a limited existing fingerprint model. There are many works related to security applications using biometric characteristics [32–34] which expanded their interests here with wavelet feature, combination of image features, and scoring strategy. Lee et al. [35] proposed a recognizable image in their fingerprint recognition, which a fingerprint image with characteristics is sufficient to discriminate an individual from other people. Their work improved the recognition by selecting a valid region from a well-focussed part of the image and estimating the

finger's angle in roll and pitch from the camera plane. Then, the classification task in fingerprints via gait characteristic proposed by Nickel et al. [36] showed a competitive recognition performance. They extracted various image features, e.g., Meland Bark-Frequency Cepstral Coefficients (MFCC), from the measured accelerations and utilized them for training a Support Vector Machine (SVM). Expanding the classification task in fingerprint, Zeng et al. [37] proposed a partial fingerprint recognition algorithm based on deep learning for the recognition of partial fingerprint images. A modern and recent work on fingerprint recognition by Baldi and Chauvin [38] via neural network estimates the probability of two images with the accuracy at 50% in 1990. Utilizing deep learning, their work managed to gain better performance than the existing fingerprint recognition algorithm on the problem of partial fingerprint classification and fingerprint recognition in the public dataset.

As security tasks via a face recognition system, three sub-categories can be considered; face detection, face extraction, and face recognition [39]. Earlier, face detection techniques which consist of the three sub-categories could only handle single or a few well-separated frontal faces in an image with a simple background, while state-of-the-art algorithms can detect many faces and their poses in cluttered scenes [40–43]. Choi et al. [44] demonstrated the effectiveness of the proposed real-time algorithm and the feasibility of its application in this security applications using mobile devices through empirical experiments. Using the Random Forest and SVM in face recognition, Kremic and Subasi [45] managed to gain a comparable recognition accuracy of 97.17%. Deep learning is utilized for face recognition with several strategies, including facial landmark [46], and to the extent of a combination of several neural networks [47].

The third computer vision task for security applications, person re-identification, has also been focussed from the early 1990s [15]. Various approaches in the proposed work for person re-identification focus on retrieving a person of interest across multiple non-overlapping cameras with single-shot images and multiple shot images. Furthermore, the increase of publicly available datasets are a positive factor for the field to have a fast speed of growth [46, 48–50]. Looking into the research trend for

person re-identification from 2008 shows an increment in published paper numbers in reputable conferences such as CVPR and ICCV [15].

In 1997, Huang and Russell [51] proposed a multi-cam tracking method with Bayesian formulation to estimate the posterior of predicting the appearance of objects in one camera given evidence observed in other camera views. The “person re-identification” task was first proposed by Zajdel et al. considering the multi-camera utilization via a latent labeling idea in their work [52]. Farenzena et al. [53] proposed a segmentation model to detect the background and foreground in their person re-identification paper.

1.2.3 Computer vision for mobile applications

Since the new millennium, cameras have become a standard equipment in mobile phones. The rapid improvement of mobile phone’s camera specifications delivers an excellent effect on our society. At the same time, the demand for information access from smartphones and tablets has increased rapidly and has become the mainstream in business and personal environments over the last few years [54]. There are several mobile applications that make use of computer vision tasks; Image-based object search, Visual Positioning System (VPS), and Self-localization.

For finding information based on an object’s visual appearance, Yeh et al. [55] developed a mobile image-based object search system that takes images of objects as queries and finds relevant Web pages by matching them to similar images on the Web. They showed how a shape-based image matching algorithm could be used as the object outline to find similar images on the Web. They organized their proposed SHREC track and built the first 2D scene image-based 3D scene retrieval benchmark by collecting 2D images from ImageNet and 3D scenes from Google 3D Warehouse. For this image-based object search, Minagawa et al. [56] proposed an image-based search system using hierarchical object category recognition algorithm. They significantly improved the standard model’s processing speed with a small decrease in accuracy.

A Visual Positioning System (VPS) creates a new lifestyle for people. The demand for location-based services using mobile devices in indoor spaces without a Global Positioning System (GPS) has increased. Abdul-Rashid et al. [57] utilized a 2D scene image to search relevant 3D scenes from one dataset. Kim et al. [58] proposed method uses a smartphone camera to detect objects through a single image in a web environment and calculates the location of the smartphone to find users in an indoor space.

Self-localization is one of the essential tasks in a mobile application for humans and autonomous robots. It is the basis for orientation and navigation in a spatial environment and mapping tasks. Olson [59] proposed a probabilistic self-localization techniques for mobile robots that are based on a principal of maximum-likelihood estimation. His proposed method compares a map generated at the current robot position to a previously generated map of the environment to probabilistically maximize the agreement between the maps. Later, Arth et al. [60] proposed a system for self-localization on mobile phones using a GPS prior and an online-generated panoramic view of the user's environment. They solved the self-localization task in large-scale outdoor urban scenarios giving robust and accurate results.

1.3 Object Identification

This section will categorize the object identification tasks and then discuss the needs of introducing the pairing concept. To generalize the object identification tasks in various applications, a target is considered as a general "instance" that can be interchanged when needed depending on the application. Accordingly, object identification can be generally defined as pairing an instance with a dedicated identity (ID) from a query set with another instance from a gallery set. The queries are the target while the galleries are candidates of the target matched pair. One of identification problem, which is widely discussed [15, 53], compares the similarity of one instance from the query images with one instance from the gallery images correctly. This concept can be generalized to generic instances as illustrated in Figure 1.4.

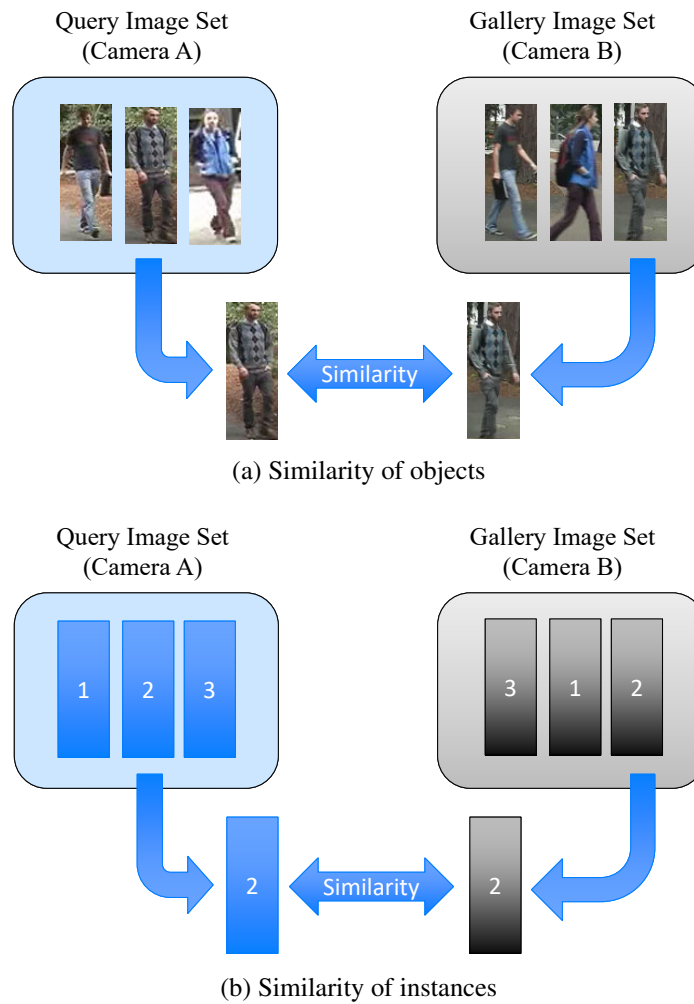


Figure 1.4: Example of identification tasks; (a) identification for image sets and (b) identification for generic instances.

Since 1997, a variety of ideas in identification has been proposed. The majority of the methods used in identification make use of the image features [61–69] for comparing the images, and improvement of these methods have also been challenged [53]. During 2000 and 2010, multi-combination approaches have been proposed in the identification task for making their work competitive. In recently proposed methods, deep learning has started to be utilized.

Notwithstanding, all the conventional methods kept their direction towards obtaining the best similarity by manipulating various features. Zheng et al. changed this paradigm by introducing a matching methodology instead of proposing a new feature. Following their work, this thesis also focusses on the matching and pairing

strategy as the next topic to be discussed in the identification task. Even though it may still be narrow, the scope of identification tasks is widened through the pairing concept, which could be applied in many ways.

In this Section, object identification tasks are categorized and discussed in detail. First, the pairing instances in object identification is discussed. Then, the object identification tasks are categorized into three; Instance-to-Class, Instance-to-Value, and Instance-to-Instance.

1.3.1 Pairing instances in object identification

As mentioned previously pairing is a novel concept for approaching identification-related research topics. For this, identification tasks are categorized based on how the instances are paired. As a matter of fact, issues related to the pairing concept were extensively debated several decades back.

Definition 1.1. Query is defined as a target instance of object identification and grouped in a query set, which later will be paired with the most similar instance from the gallery set.

Definition 1.2. Gallery is defined as a pair instance of object identification and grouped in a gallery set, which will be paired with an instance from the query set.

Definition 1.3. Pairing is defined as the best combination of two instances by comparing the similarity of instances or images between compared images.

As the variety of problem settings in object identification could be set, this query could be treated as single and multiple instances. These single and multiple queries still will infer the same concept of paired instances, e.g. (query1, gallery1). For these single and multiple query to pairing relation, in the thesis, the discussion is directed towards Single-query pairing and Multiple-query pairing.

Single-query pairing is used by the majority of identification researchers, that is, given a query, find the best matching instance with the query from multiple candidates (the gallery set). It finds the most similar one to the query from the gallery set using a pre-defined similarity between the two compared instances, e.g., in one image matching task, a similarity of image features are utilized.

Multiple-query pairing is used when many queries are matched simultaneously. It finds the most similar instance for all queries from the gallery set with a pre-defined similarity between the two compared instances simultaneously [70]. For example, the task for searching the best route for car navigation using greedy matching [71], a distance between two instances is utilized. It may be realized by applying single-query pairing for each query one by one, but the most different point is the utilization of underlying consistency, e.g., an instance for a gallery should be selected only once, or matched pairs follow the same law.

Graph theory is later utilized to match the query to the gallery for expanding the similarity comparison of image contexts to a generalized new image context [72, 73]. Treating the queries and galleries in another context will vary the matching strategy in object identification tasks [74, 75]. Among the graph theory methods known as the graph matching approach, e.g., bipartite graph matching [76], delivers a broad research extension in many applications. Graph matching considers pattern, template, and instance's feature as a node in a graph. As such, all instances involved in identification will be matched by graph matching from the query set to the gallery set.

Beside considering matching in an identification task, utilizing feature similarity-based matching using Scale-Invariant Feature Transform (SIFT) [77], Oriented FAST and rotated BRIEF (ORB) [78], Speeded Up Robust Features (SURF) [79], and Binary Robust Invariant Scalable Keypoints (BRISK) [80] were also conducted by choosing the best target instance ID (object) from its features. Utilizing structure matching, Zheng et al. [64] extended the Bag-of-Words (BoW) concept from the structure matching into the identification task. All these methods are implemented

in various ways of identification tasks. From here, we still can not see systematic research proving that using the pairing approach in identification tasks could give a more reliable matching performance.

1.3.2 Instance-to-Class pairing

In the early days of computer vision research, classification played a broad role in identifying and labelling one target instance to output a ‘class.’ The class, here, defines a concept as a set of instances, which has been grouped based on the oracle. Later, classification became widely used in computer vision tasks to recognize the class that a target instance belongs. In order to consider the conventional classification through the pairing concept in this thesis, it is defined as “Instance-to-Class”, since it can be considered as finding a pair of the target instance and its class. For Instance-to-Class object identification, in the thesis, traditional classification is re-defined as a pairing problem between the two, instance and class. The instance is represented by the query for identification, which could be any instance type, face image, fingerprint, or others.

We could consider this Instance-to-Class object identification with binary classification; the class is identified and labelled from a given input, e.g., two-animal classification, for labelling the input’s most suitable class. Such binary classification is conducted by separating the given input into binary classes, e.g., 0/1 or yes/no. This is broadly used in daily life, e.g., to differentiate two types of fruits [81], classification of milk conditions, either stale or fresh [82]. Some of the methods commonly used for binary classification are Logistic Regression (LogReg) [83, 84], Linear Discriminant Analysis (LDA) [85, 86], and Support Vector Machines (SVM) [87, 88].

However, later it was extended to multi-class classification by introducing k -Nearest Neighbors (k NN) [89, 90], Naïve Bayes (NB) [91, 92], Random Forest [93, 94], Ensemble Models [95, 96], and Neural Network [97, 98]. The extension from binary

to multi-class pairing could ease various Instance-to-Class pairings in object identification tasks. Character recognition is one of the examples of multi-class Instance-to-Class applications, in which a query is set as an instance, and galleries are set as classes of the multi-class pairing. For this, the MNIST dataset [99] is a well-known dataset utilized for multi-class classification conducted via the pairwise coupling idea [100] and SVM classifier [101, 102].

1.3.3 Instance-to-Value pairing

Regression is well-known as a task that answers the most suitable single value for the input. In order to consider the conventional regression through the pairing concept proposed in this thesis, it is redefined as “Instance-to-Value” as one of the categories of object identification. For a simple example, there is a relation between the healthiness of a person and some features, body height as an instance and weight as an estimation value. We can consider the regression as an instance-to-value pairing where the set of features is an instance, and the healthiness is a value. Regression analysis has been widely used not only in economics and health science, but also in computer vision tasks.

For example, let’s consider a robotics application to discuss the Instance-to-Value pairing in object identification, focussing on object pose estimation. The single viewpoint pose estimation task estimates the most suitable pose between 0° and 360° for a given object, object image from the viewpoint as an instance and estimated pose as a value.

Among the first research on the single viewpoint pose estimation task, Murase et al. [103] proposed the parametric eigenspace method to estimate the pose from a single viewpoint. They developed both object recognition and pose estimation algorithms based on the parametric eigenspace representation. An input is mapped onto the object eigenspace as a point. The closest point on the manifold is output as the estimated pose. For this Instance-to-Value pairing, the input image is considered the

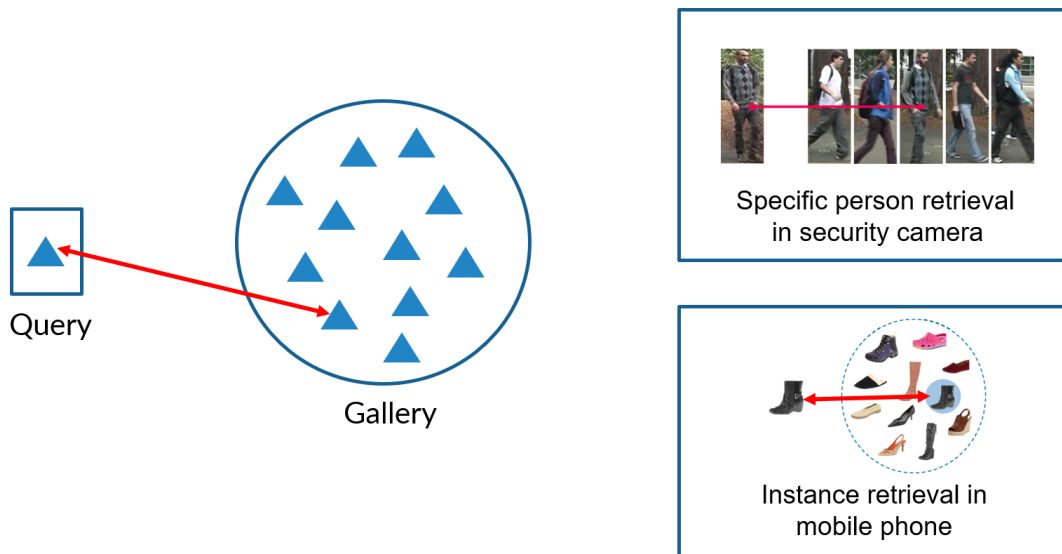


Figure 1.5: Single-query pairing. Two examples that consider person and object as instances.

instance, while the angle regressed through the object eigenspace is regarded as the value.

1.3.4 Instance-to-Instance pairing

The previous two categories represented the well-known classification and regression. Newly in this thesis, a new category, “Instance-to-Instance” pairing is defined as an object identification task. The “Instance-to-Instance” pairing is possible if the problem is set for a dedicated computer vision application with a specification of the instances; What is the instance’s characteristic, and how are the instances matched in one identification task. Here, this instance is defined as an actual identical object. By changing the setting of the identification task, we could divide the Instance-to-Instance pairing into two sub-categories; single-query pairing, and multiple-query pairing.

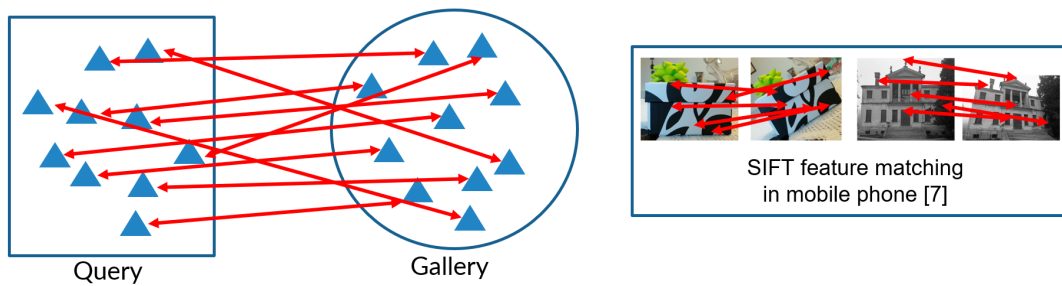


Figure 1.6: Multiple-query pairing. Example that considers a whole image as a set of instances.

1.3.4.1 Single-query pairing

An instance for single-query pairing could be defined in many ways according to the application, especially in the three main computer vision applications discussed earlier in Section 1.2. Single query pairing is defined as pairing an instance as a query with an instance from the gallery set. As examples, there are specific person retrieval for security applications, and object image retrieval for mobile applications. In such applications, person and object are considered instances, and from a query instance, a pair is selected from the many available instances from the gallery images as illustrated in Figure 1.5.

1.3.4.2 Multiple-query pairing

Compared to the single-query pairing, multiple-query pairing involves multiple instances in the query sets, later paired to multiple instances in the gallery sets. Here, multiple-query pairing defines an instance as the part of the whole image, such as an image patch, edges or representative points from an image. Not only comparing the similarity of these instances but also consistency of them could help the multiple-query pairing gain more information than the single-query pairing. When a simultaneous key-point matching problem is considered as in Figure 1.6, both the similarity comparison and consistency of a matching provides additional information for this multiple-query pairing.

1.4 Research Overview

The previous section has explained the object identification task through the pairing concept, including the rapid development of technologies and algorithms in the computer vision field. Based on the discussions there, we have seen that the Instance-to-Class and Instance-to-Value pairing approaches are among the basic ones that can provide a good understanding in identifying an instance to the desired output. Furthermore, this thesis addresses the demand for proposing a better pairing approach to perform an object identification task comprehensively. Concretely, this thesis contributes to developing a new pairing approach in the robotics and security fields.

The core idea of this research is to quantify the Instance-to-Instance pairing performance accurately. For this, in the Instance-to-Instance pairing category, two subcategories must be deeply discussed; single query pairing and multiple-query pairing. We could consider a similar target object as our instance in the identification task for these two Instance-to-Instance pairing categories, but they manipulate the instances differently. Therefore, we need to consider each of these pairing approaches with an individual target object.

1.4.1 How to develop a single-query pairing?

The single query pairing requires deep attention from a robotic perspective for robotics applications since the current Instance-to-Value pairing, also known as regression, has limited the robot's capability to estimate an object's pose with a better estimation result. With only Instance-to-Value pairing, the pose estimation value is difficult to re-estimate the object pose because the estimated value here requires another pose to realize a new pose estimation for an ambiguous scenario. Considering existing research on pose estimation, most of them focus on finding the next viewpoint from multiple viewpoints, whereby the best next-viewpoint is still not discussed. This re-definition of object identification delivers us an opportunity to expand and provide a contribution. Since the person and object were chosen as the instance in the existing

applications for single query pairing, to fill the gap in the pose estimation application, a new single-query pairing is introduced in this thesis using the object pose's images as a pair.

Based on this problem statement, many researchers focus on object pose estimation from multiple viewpoints. For example, Zeng et al. [104], Erkent et al. [105], Collet and Srinivasa [20], Kanazaki et al. [106], and Vikstén et al. [107], propose pose estimation methods using observations from multiple viewpoints. However, these methods only focus on pose estimation of a given image and do not consider the selection of viewpoints. Bajcsy et al. [108] discussed the idea of active perception, initially in the context of the sensor planning problem. Since controlling the camera view is the most crucial issue in recognizing objects, recent researches focus on predicting the best next-view [109–111]. Recent active perception works [112, 113] propose the best next-viewpoint prediction for multiple target object pose estimation based on Hough Forest [134]. However, these methods focus on a target object from unknown shapes. Still, there is a case that pose estimation methods should be robust to the shape variation within the target object category; category-level object pose estimation is desirable.

To begin this task, the traditional single viewpoint pose estimation is extended to implement the single query pairing for an indoor target object. For a general pose estimation, observing the object pose from an image is not easy; there is a case that the object pose is ambiguous from the observation location, as illustrated in Figure 1.7. On the other hand, since robots have embodiment, they can move to other locations. From this new location, a different observation could be obtained. The question is how to select the next viewpoint given an initial viewpoint for object pose estimation from multiple viewpoints. Not to forget, the scale that will be utilized for scaling the possible viewpoint choices also requires deep attention before estimating the object pose. The way we see the viewpoint is also crucial as the viewpoint could be manipulated as we have hundreds of viewpoint choices that could be paired.

Here, in this thesis, for the single-query pairing, Research topic 1 will introduce the

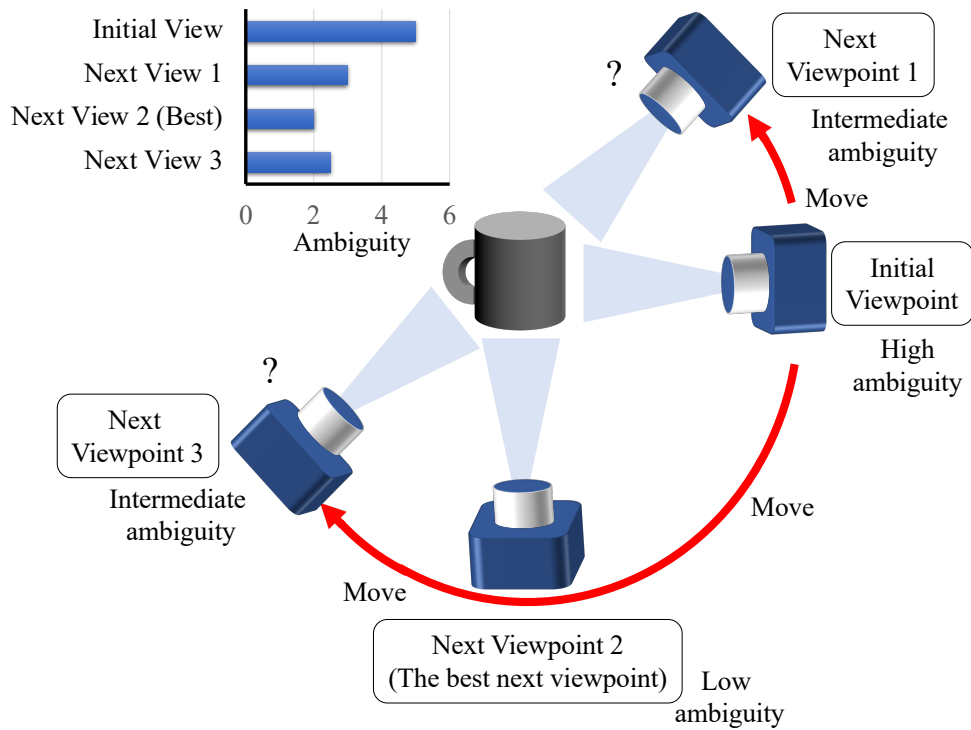


Figure 1.7: Difficulty in object pose estimation when the viewpoint is ambiguous.

idea of pairing two poses as instances which will be catered to the limitation of using the single viewpoint object pose estimation.

1.4.2 How to develop a multiple-query pairing?

The multiple-query pairing requires extended study on pedestrian recognition for security applications since the current single-query pairing also known as specific person retrieval, has a limitation to perform in a complex environment setting. For existing security applications using the single-image pairing, a simultaneous image pairing setting could not be implemented, especially at a railway station or an airport where a lot of people are present in a single image. Since the multiple-query pairing suits this problem setting in single-image pairing, person re-identification is considered here as an security application for the object identification task.

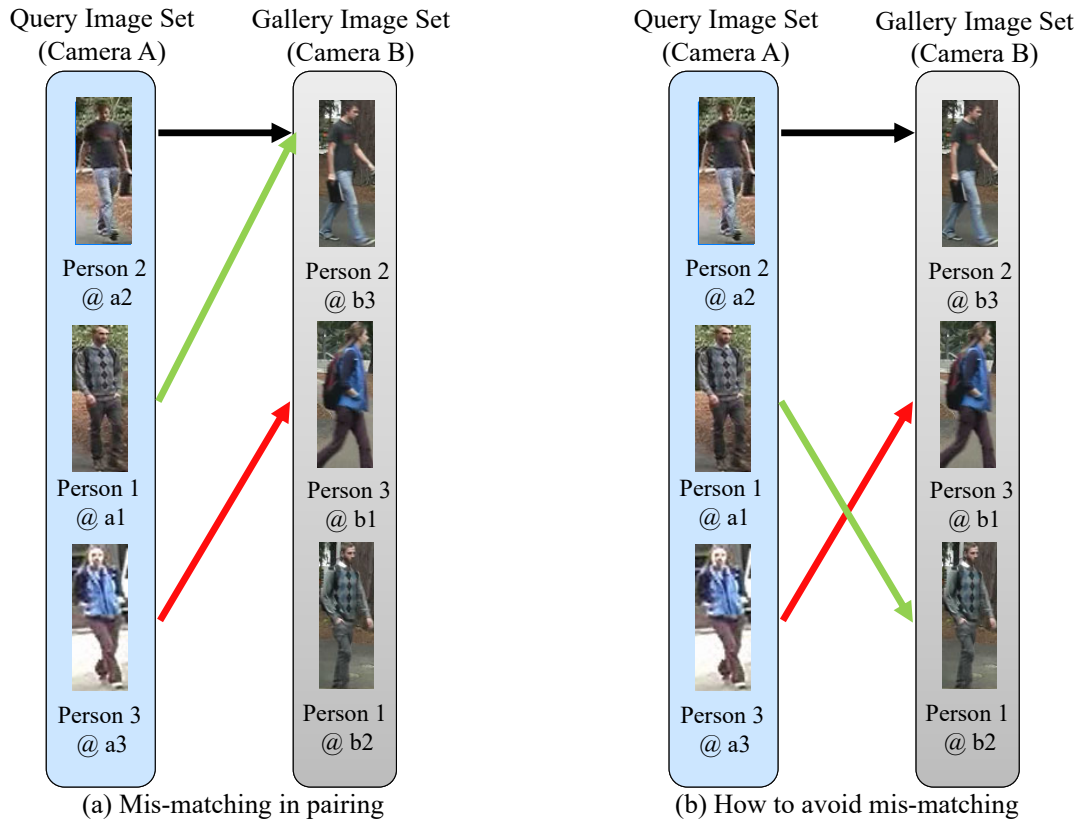


Figure 1.8: Difficulty in person re-identification via single image pairing when (a) a mismatched image pair occurred, and (b) simultaneous image pairing.

Based on this problem statement, the majority of the traditional single pairing, discussed in the previous Section is not suitable for this condition [61–69]. This traditional single pairing or in other words, individual pairing scheme pairs images independently based on the image similarity between the query and the gallery images. The independent image pairing allows the previously wrongly matched gallery image with a query image to be matched again to another query image without any restriction. This wrongly matched image pair leads to a redundant pairing where one gallery image is matched with multiple query images. In this case, the remaining query image which has not been paired will be discarded. These discarded images will remain unmatched images; Person 1 in gallery image set, until the end loop of the identification task as illustrated in Figure 1.8 (a). Because of this, redundant pairing may lead to a decrease in the matching rate and cause a tracking failure.

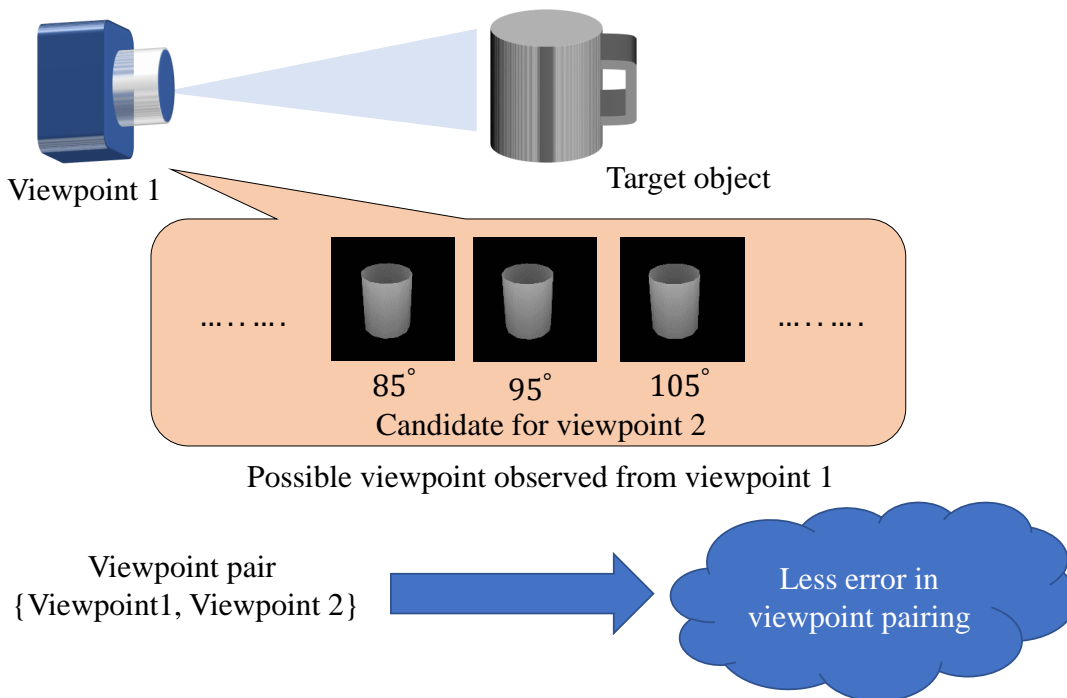


Figure 1.9: Goal of Research topic 1.

This simultaneous image pairing is similar to the concept of two sets of an element matching in a graph [114] as illustrated in Figure 1.8 (b). Recently, Zhang et al. proposed a method named Person Re-identification via Structured Matching (PRiSM). They deployed a weighted structured matching method [115] to the person re-identification problem by considering it as an instance of structured matching. The question is how to carefully select the simultaneous image pairing for this person re-identification task. The pairing performance also requires a deep analysis for a reliable proposed method and the simultaneous pairing approach.

Here, in this thesis, for the multiple-query pairing, Research topic 2 will introduce the idea of pairing two whole-person images as instances, which will solve the disadvantages of the mismatched image pair and redundant matching in the person re-identification task. Solving these problems would allow analyzing the instance pairing from a diverse angle.

1.4.3 Research topic 1: Object pose estimation

For single-query pairing, many studies exist related to the baseline single viewpoint pose estimation technique. There are previous studies involving pose estimation from a single viewpoint, such as template matching [116], parametric eigenspace method [103], embedding templates into a pose manifold [117], deep learning [118], and many more.

A recent work, RotationNet [106] proposed the idea of using multi-view images of an object as input and jointly estimates its pose and object category. This pose estimation also has attracted many researchers and practitioners to establish various approaches [20, 104, 105, 107, 119] estimating the object pose from multiple viewpoints. However, selecting which of the next viewpoints should be chosen as the best for the given initial viewpoint is still not profoundly discussed. The question here is how to choose the best next viewpoint among the good available viewpoints in the list. Besides aiming for the best next-viewpoint, the problem of having ambiguity could not be avoided. The ambiguity occurs when the pose estimation has a given opposite viewpoint image of the target object. Here, to estimate the object pose, several output results will be inferred, which later could be wrongly estimated.

Research topic 1 proposes a novel approach to choose the best next-viewpoint among the candidates so that the proposed method can choose the best viewpoint from the many viewpoint angles as illustrated in Figure 1.9. For this research evaluation's target object, category-level object images are utilized in both the training and testing phase. In this research, network modeling is not the aim but rather how the best next-viewpoint is chosen. Pose likelihood for one viewpoint, and two viewpoints are statistically analyzed by indirectly plotting and analyzing the object pose distribution using comparative methods as illustrated in Figure 1.10. A complete explanation for this Research topic 1 is provided in Chapter 3.

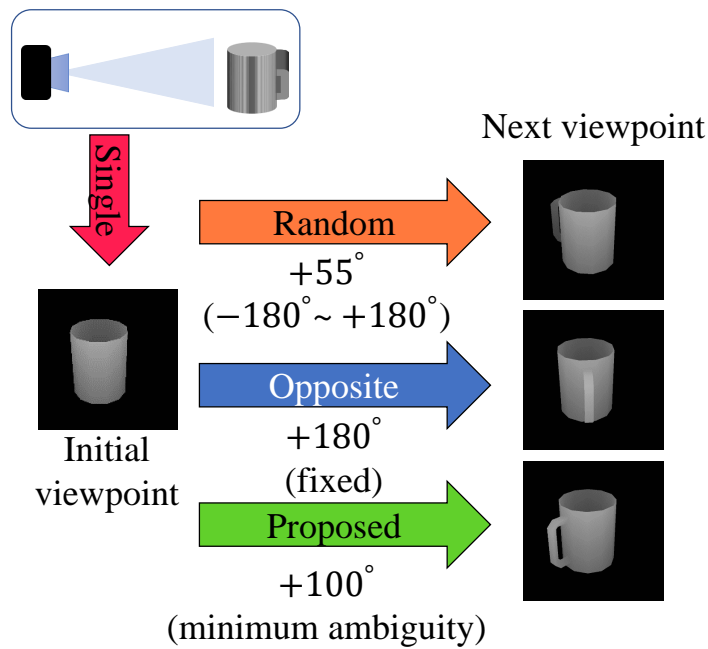


Figure 1.10: Overview of Research topic 1

1.4.4 Research topic 2: Person re-identification

In the multiple-query pairing, a pairing approach that considers pairing between multiple instances is proposed. Traditionally the problem related to multiple -instances to multiple-instances in person pair selection has not gained sufficient attention; Instance pairing has been approached by many researchers, by the conventional single-pairing approach as illustrated in Figure 1.11. The main issue of person re-identification is how to make sure that a wrong pairing does not occur in the identification as illustrated in Figure 1.12. The Greedy matching and Hungarian Matching [153] are among the image pairing strategies used in conventional methods.

Besides, these methods focus on feature comparison, metric learning, and deep learning which allow the matched image to be matched again with others. This thesis introduces the multiple-query pairing, which is similar to the simultaneous image pairing in the person re-identification problem. Wrong matching could be avoided by introducing the Stable Marriage Algorithm (SMA) [120], with an assumption that each instance could be paired only once. The effect of pairing simultaneously could

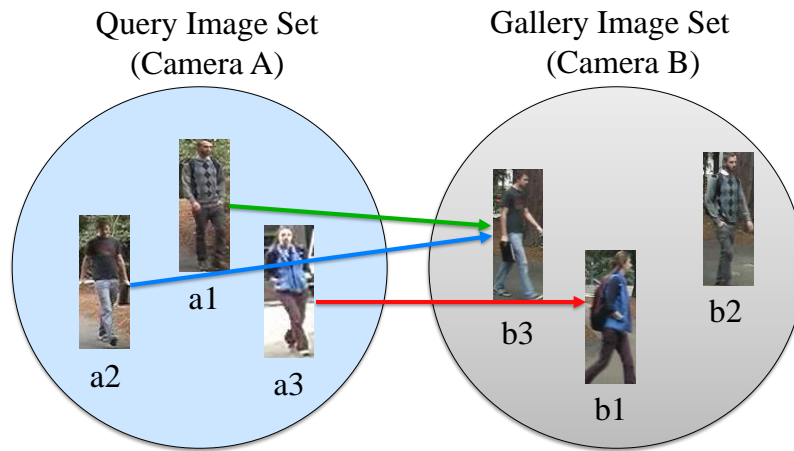


Figure 1.11: Overview of Research topic 2

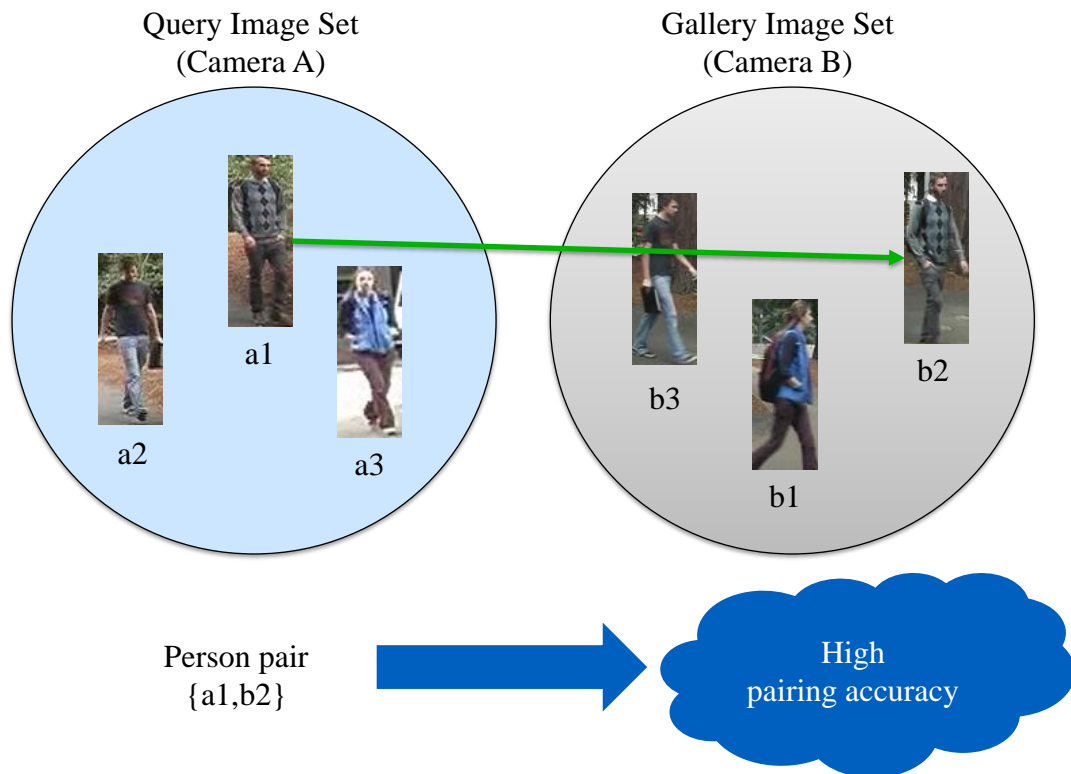


Figure 1.12: Goal of Research topic 2 introduced in this thesis.

avoid the unpaired instances and wrong pairing. In addition, to improve the similarity comparison, which occurs before the instance pairing, a new image feature called MSDALF is proposed. The quantitative pairing performance and the computational time are also among the target of improvement in this Research topic 2. A complete explanation for this Research topic 2 is provided in Chapter 4.

1.5 Thesis structure

This thesis contains five chapters. The relationships between the different chapters of this thesis are visualized in Figure 1.13.

This Chapter 1 discussed the motivation of this doctoral research and gave an overview on the background of the object identification task and their categorization: Single-query pairing and multiple-query pairing. Two Research topics were introduced to study performance of the pairing concept applied to different applications with different target objects.

Chapter 2 reviews existing work in the discussed two computer vision applications; object pose estimation and person re-identification, thoroughly, by giving a comprehensive analysis of the state-of-the-art on these Research topics.

Chapter 3 discusses Research topic 1 using category-level indoor object image as the target to analyze the best next-viewpoint for object pose estimation by minimizing the ambiguity problem. Here, the object pose estimation for two viewpoints is considered object pose-to-object pose or the proposed Instance-to-Instance pairing for security applications.

Chapter 4 discusses Research topic 2 using an offline person image as the target to analyze the high matching rate for person re-identification by simultaneous image pairing. Here, the person in the query gallery image with the person in the gallery image is considered a whole person-to-whole person or the proposed Instance-to-Instance pairing for security application.

Lastly, Chapter 5 concludes this thesis by summarizing the research contributions and results found through these studies. In addition, future research directions, remaining challenges, and applications that can be built from the results will be discussed.

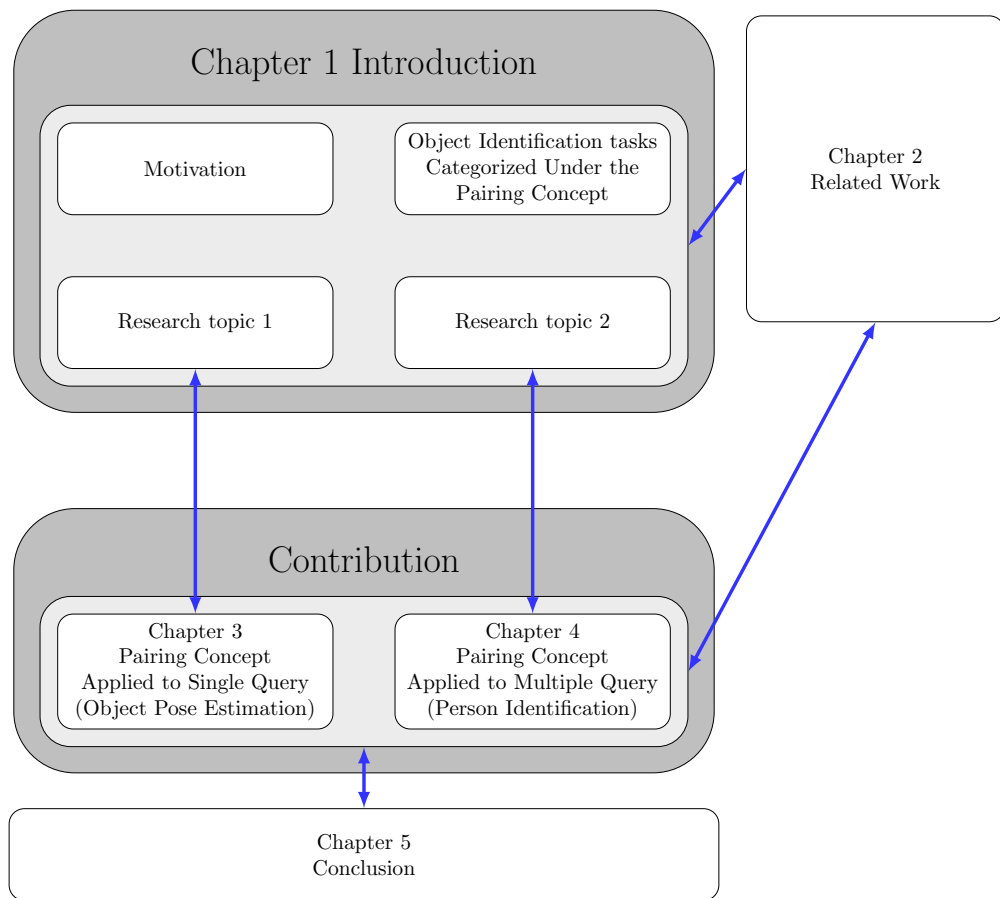


Figure 1.13: Thesis structure.

Chapter 2

Related Research

In Chapter 1, the third pairing concept was newly introduced to object identification; Instance-to-Instance pairing, which are categorized into single-query pairing, and multiple-query pairing. These two pairing approaches have also been utilized in previous work for quantifying the identification task in various computer vision applications as illustrated in Table 1.1. This chapter gives an overview of existing literature related to both Research topics 1 and 2, discussing the identification tasks that are indirectly related to the pairing in the related object pose estimation and person re-identification tasks.

Section 2.1 will discuss existing literatures regarding object pose estimation. This includes over-viewing work and some ideas toward single viewpoint, multiple viewpoint, and active vision object pose estimation. Section 2.2 will discuss existing literature regarding person re-identification. This includes single query pairing for individual image pairing and multiple query pairing for simultaneous image pairing.

2.1 Object Identification for Pose estimation

In the past, many researchers proposed methods to tackle various difficulties in 3D object pose estimation. Here, the pose estimation methods can be categorized into the following two pairing approaches; single viewpoint pose estimation is categorized into Instance-to-Value pairing, multiple viewpoints pose estimation and active vision object pose estimation are categorized into single query pairing of Instance-to-Instance pairing. This Section introduces each method.

2.1.1 Instance-to-Value pairing: Value estimation for single viewpoint pose estimation

Value estimation in object identification can be categorized into Instance-to-Value pairing approaches. The instance is paired to a dedicated estimation pose value by utilizing a variety of pose estimation methods.

The template matching approach is one of the earliest pose estimation methods from a single viewpoint [116]. This approach utilizes many templates of the target object captured from various viewpoints beforehand, and the pose estimation result is taken from the best matched template. Later, to reduce the number of templates in single viewpoint pose estimation setting, Murase and Nayer [103] proposed the Parametric Eigenspace method. This method represents an object's pose variation on a manifold in a low-dimensional subspace obtained by Principal Component Analysis (PCA). By interpolating the pose (template) of the target object on the manifold, they achieved accurate object pose estimation even with few templates. Since PCA focuses on the appearance variation of templates, some poses with similar appearances may be mapped to similar points in the low-dimensional subspace, which is difficult to distinguish. This diminishes the accuracy of the pose estimation.

Recently, Ninomiya et al. [117] proposed a supervised feature extraction method for embedding templates into a pose manifold. They focused on Deep Convolutional

Neural Networks (DCNNs) [118], which is one of the deep-learning models, as a supervised learning model for manifold embedding. They modified DCNNs for object pose estimation, named Pose-CyclicR-Net, which can accurately handle object rotation by describing the rotation angle using trigonometric functions. By introducing the Pose-CyclicR-Net-based manifold embedding, which is called Deep Manifold Embedding, the method estimates the object's pose from a single viewpoint. Hashimoto et al. [121] proposed a method by utilizing Convolutional Neural Network (CNN) for probabilistic 6D object pose estimation from a color image in their work. Unlike other methods which analyze a single data point as their inference, their proposed network passes the information necessary to estimate the complete likelihood distributions of 6D object poses. This work also caters to the issues discussed in the thesis: the ambiguity of object appearance in the image in a principled manner. Tatemichi et al. [23] proposed an occlusion-robust pose estimation method of an unknown object instance in an object category from a depth image.

A recent work by Hashimoto et al. proposed a novel method using a CNN for probabilistic 6D object pose estimation from color images [121]. This recent work does not only capture the ambiguity of object appearance in the image in a principled manner, but also outputs the probability distribution of 6D object poses. It enables the results to be fused with other sensing modalities using well-established probabilistic inference techniques.

In general, object pose estimation from a single viewpoint faces the problem of inaccurate pose estimation due to the ambiguity issue, namely, an object may have some poses which look similar and hard to be distinguished.

2.1.2 Instance-to-Instance pairing: Viewpoint instance selection for multiple viewpoint object pose estimation

To avoid the pose ambiguity issue, the main problem in a single viewpoint pose estimation, several methods focus on object pose estimation from multiple viewpoints.

Multiple viewpoints are considered in two approaches for estimating the object pose: multi-view and incremental view. The multi-view approach estimates the poses from multiple physical viewpoints, e.g., viewpoints from many cameras. On the other hand, the incremental view approach considers an additional angle from the original viewpoint to another better viewpoint. Here the next viewpoint is generally used to describe another viewpoint from the original viewpoint or the best among a selection of next viewpoints. There are versatile methods of multi-view and incremental view approach, but in the end, the main target is the same as finding the best pose estimation from more than a single viewpoint.

The majority of the existing works in pose estimation focussed on the former multi-view approach. Collet and Srinivasa [20] proposed a multi-view object pose estimation method based on multi-step optimization and global refinement. Erkent et al. [105] tackled object pose estimation in cluttered scenes. The method is a multi-view approach based on probabilistic modelling and appearance-based pose estimation. Vikstén et al. [107] proposed a method combining several pose estimation algorithms and information from several viewpoints. Zeng et al. [104] proposed a self-supervised approach for object pose estimation in the Amazon Picking Challenge [119] scenario. Kanezaki et al. [106] proposed the RotationNet, which takes multi-view images of an object as input and jointly estimates its pose object category. Ashutosh et al. [122] later presented a classification-based strategy for selecting the next best view selection. They presented how a supervisory signal can be plausibly obtained for the object pose estimation task. The proposed strategy is end-to-end trainable and endeavors to achieve the best possible 3D reconstruction quality with a pair of passively acquired 2D views via several experiments on synthetic and real images.

For the incremental view approach, among the earlier works, Detry and Piater [123] proposed a 3D probabilistic object-surface model and a mechanism for probabilistically integrating unregistered 2.5D views into the model, and for segmenting model instances in cluttered scenes. Later, Teney and Piater [124] utilized a pair or triple views, for estimating the changes of appearance as a function and then matched them

with those already stored in the model. The incremental view approach has also been studied for object detection by Konno et al. [125]. They proposed a method to predict the pose of an object when integrating the object detection scores. There are also some applications related to a specific robotics field which utilizes the pairing in order to improve the single query pair in the identification task. Ni et al. [126] proposed an approach, which is an improved image matching algorithm based on feature points, to enhance the anti-interference ability of the algorithm.

While the multi-view approach only considers pose estimation from multiple view-points, the majority of the incremental view approach focusses on object detection. Therefore, these methods require several cameras to capture all the available view-points and simultaneously need additional hardware settings and plenty of space for the camera movement area to estimate an object's pose. Unlike others, in this thesis, we propose an idea of estimating the best next-viewpoint via the incremental view approach, which will decrease the pose estimation ambiguity for an arbitrary object instance in a specific object category.

2.1.3 Instance-to-Instance pairing: Active vision for object pose estimation

Since the pose estimation for an indoor object is relevant and closed to the robotic field, some of the related active vision works for the object pose estimation are introduced for comparison. Formerly, the term “active vision” stands for strategies of a robot for sensor placement or configuration to complete either general purpose or specific tasks. Because active vision is a real-time approach for estimating the object pose, the discussion is essential for expanding the proposed method in the future. Although this active vision approach may be interpreted as a kind of multi-view approach introduced in the previous Section, here, it is discussed separately in this Section.

Bajcsy et al. [108] discussed the idea of active perception, initially in the context of the sensor planning problem. Wilkes and Tsotsos [127] boosted this idea by introducing a concept of “active object recognition” and a proposal for the concept’s solution. They proposed and described a method for selecting an additional viewpoint when an object is uncertain to be identified with the ability to vary the viewpoint accordingly to the current object’s interpretation status. Pose determination and visual sensor configuration are the two concerns in active vision research. Later, active vision research became to play an essential role in robot vision [128–133].

Since controlling the camera view is the most important issue to recognize objects, recent researches focus on predicting the best next-viewpoint [109–111]. This trend can also be found in the object pose estimation task. Dumanoglou et al. [112] and Sock et al. [113] proposed the best next-viewpoint prediction methods for multiple object pose estimation based on Hough Forest [134]. We can expect that this approach, an active vision method applied to multiple object pose estimation, will be the next exciting topic whereby the impact of the versatile and fast development of robot vision research demands the rapid changes in object surrounding environment. This idea will allow us to support an application in which various instances in a specific object category need to be considered as the target object. However, these methods could not be applied for the category-level object pose estimation since they are designed only for instance-level object pose estimation.

2.2 Object Identification for Person Re-identification

Traditionally, an image in one object identification task will be identified from one image set to another, inferring a similarity between the two images. Pairing an instance to another instance, in which the instance is considered a person image, is one of the computer vision applications related to object identification, also known as re-identification. In other words, we can describe the object identification here as a case of an image being paired with another image to define their similarity. As the

pairing is the priority and the main issue of this thesis, the pairing between instances is considered a close situation to the traditional re-identification task.

In the past, many researchers proposed various methods to improve the person re-identification performance. Illumination, human pose, and occlusion are common difficulties discussed in conventional person re-identification research. This section briefly introduces some of the representative methods. Here, existing re-identification methods are categorized into the following two; single-query pairing for individual image pairing and multiple-query pairing for simultaneous image pairing.

2.2.1 Single-query pairing: Individual image pairing

The majority of the existing person re-identification methods focus on pairing the person images as individual instances. The idea of the individual image pairing scheme is that each query image is paired with an image in the gallery set one-by-one at one time without sharing the existing image pair. This scheme matches all the images in query images with gallery images without any replacement of the other matching results. Pairing without any replacement leads to redundant pairings where a gallery image is paired several times to the query images.

The majority of the previous work utilizes the variety of image features for image matching [53, 61–69, 135–140], which is related to color characteristic, image patching strategies, and bag-of-words. They perform the pairing of instances by first extracting the image feature and then pairing the similar instance in terms of feature distances.

Farenzena et al. proposed a feature aggregation method that focussed on the symmetry of a human body. Recent methods in person re-identification tend to focus on metric learning. By utilizing metric learning, better similarity metrics are obtained [141–149].

Ahmed et al. [150] initiated person re-identification with Deep Learning by proposing a CNN-based method. Liu et al [151] proposed a triplet-loss based CNN model to find useful features and metrics. A similar approach also has been introduced by Zhang et al. [152]. Applying Deep Learning in person re-identification, however, struggles with the image labeling problem in the training process. The Deep Learning approach will also consume more time especially for manual labeling of training data.

2.2.2 Multiple-query pairing: Simultaneous image pairing

This kind of pairing scheme for person re-identification matches all given pedestrian images simultaneously, considering all pairing results are listed at once during the identification task. It is usually formulated as a bipartite graph matching problem. A bipartite graph consists of given two sets of vertices and edges which connect the two vertices. The bipartite graph matching problem finds a set of edges of a bipartite graph which maximizes the total scores corresponding to the edges. For person re-identification, the score of an edge can be considered as a similarity score between two images, when a person is considered as a vertex in V of a graph $G = (V, E)$.

The most naïve approach to find a solution for this is greedy matching. Its algorithm starts from giving scores of connecting two vertices in different sets, and then adds edges to the vertice pair whose vertices are not connected to any vertice yet in descending order. Another well-known solution is Hungarian matching [153]. Despite being the choice in a particular field, many work consider the Hungarian matching as a tool for optimization at the back end [154, 155].

Ye et al. [156] proposed another idea in simultaneous image matching. They initiated the idea of structured learning in graph matching for person re-identification. This idea was extended and improved by Zhang et al. [115] named Person Re-identification via Structured Matching (PRiSM).

Chapter 3

Pairing Approach for Object Pose Estimation

In this Chapter, the pairing approach is introduced in object identification, which can be utilized in various identification applications such as robotics, security, and mobile. Particularly for the robotics application, predicting the best next-viewpoint in object pose estimation is effective to improve the pose understanding. Therefore, Research topic 1, the Best Next-Viewpoint prediction task is introduced in this Chapter as an example of the single-query pairing for object identification to meet the increasing demand on multiple viewpoint object estimation.

This chapter is structured as follows. Section 3.1 describes the general background of the single query pairing in object identification and its relation to object pose estimation. Discussion on how the instances are considered in the object pose estimation context is discussed in Section 3.2. Section 3.3 elaborates the proposed framework made to obtain a good pose estimation using the best next-viewpoint gained from multiple viewpoints. The approach to the problem is by investigating the pose estimation error between the given input images with several comparative methods. Next, evaluation through experiments and their results and analysis are reported in Section 3.4. Additional experiments for further discussion are also provided in Section 3.5 before the chapter is closed with a summary in Section 3.6.

3.1 Introduction

Single query pairing as an Instance-to-Instance pairing and its example have been introduced in Chapter 1, including purposes and applications. However, there is still a paucity of direct discussion on the pairing activities in the object pose estimation in identifying the similarity of a given target object. To obtain a better understanding of the pairing approach in the pose estimation application, single-query pairing for object pose estimation that demands an acceptable estimation arrangement toward the identification setting is crucial. This leads to the demand for realizing the single-query pairing for the object identification task by utilizing the identification of the object pose estimation task. The pairing approach in identification was introduced indirectly via the usage of classification and regression in previous works.

As previously explained, this Chapter focusses on the pairing approach for one of the object identification tasks. As discussed in Chapter 2, especially in Section 2.1, various existing works have been proposed for object pose estimation [20, 103–113, 116–119, 127–134]. The results from these methods are targeted to yield the best estimation for the object pose with a single viewpoint object pose estimation. The newly proposed idea for the multiple viewpoint object pose estimation via minimizing the pose ambiguity could improve and infer comparable best next-viewpoint results compared to the existing works. The proposed pairing approach with ambiguity minimization yields outputs compatible with the goal of the Object Pose Estimation.

The best next-viewpoint is recommended as the viewpoint where the pose ambiguity is the smallest given observations from the initial and the next-viewpoint candidates. However, since the estimated initial viewpoint is also ambiguous, it is also considered as a latent variable to keep all object pose possibilities from the initial viewpoint. This chapter focusses on the essential part of the problem setting and idea, so the proposed method is explained and evaluated by limiting it to a single axis rotation. However, the extension to 3D rotation could be straightforwardly explainable. With the utilization of object images rendered from the public 3D object

dataset, ShapeNet [157], evaluation is performed to assess the proposed method's effectiveness.

In summary, the contributions in this chapter are as follows:

1. Introduction of the single-query pairing to object pose estimation, as an Instance-to-Instance pairing approach in object identification.
2. Definition of a new metric called "*pose ambiguity*" to examine the ambiguity for the pose estimation task.
3. Introduction of a new standard for finding the best next viewpoint for category-level object pose estimation.
4. The proposed method outperforms two other naïve viewpoint recommendation methods in several pose estimation analysis, and also it achieves a better result than the result from a single viewpoint.

3.2 Details on the Pairing Approach for Object Pose Estimation

Chapter 1 introduced single-query pairing as one of the pairing categories for object identification. To see and perceive the pairing approach in the object pose estimation, we should define an instance. To begin this discussion, instance, value, and pairing are introduced and explained based on the object pose estimation tasks in both Instance-to-Value and Instance-to-Instance pairing categories.

In object pose estimation, an instance is initially determined as a whole captured image for a target object. The Instance-to-Value pairing approach yields the estimated pose as a value from the given image, e.g., 1° , 45° , or 270° as illustrated in Figure 3.1. However, this Instance-to-Value pairing approach could perform in a straightforward setting by inferring a value after estimating a pose from one given image, but not in

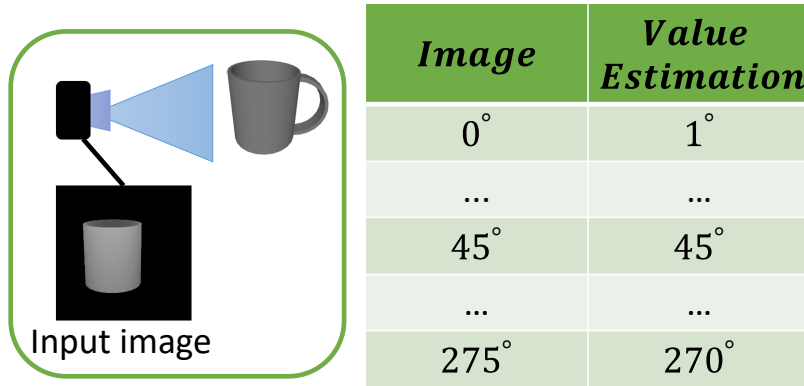


Figure 3.1: Value estimation for object pose estimation

a difficult one. This difficult condition setting could occur when a part of the object is hidden or occluded due to the camera angle. This scenario will decrease the estimation performance as the estimation could not identify the distinct part of a given image which is hidden from the camera view. For object estimation of an occluded object, for example, delivers difficulty. This difficulty could be the same if we face an object that could not or partially be observed well from one viewpoint. As the unseen observation is the main priority of this research, the probability of an object pose that could not be estimated initially is considered in the proposed method. This probability is utilized to estimate the best viewpoint from all 360 degrees.

Since the previous single viewpoint pose estimation utilized the Instance-to-Value pairing, it was difficult to estimate the pose in a challenging environment setting. In this Chapter, this approach, difficult cases such as occluded viewpoint is handled by Instance-to-Instance pairing. In contrast to the Instance-to-Value pairing approach, the Instance-to-Instance pairing approach defines instance as the given image viewpoint in degrees. The pairing term here describes a pairing between a given image pose and the estimated pose.

To explain how the instance is defined in the object estimation task, Figure 3.2 illustrates how image viewpoints are considered an instance in both query and gallery image sets in object identification. Here, we pair the query instance as a viewpoint from the given object image with the best selection of viewpoint from the gallery instances.

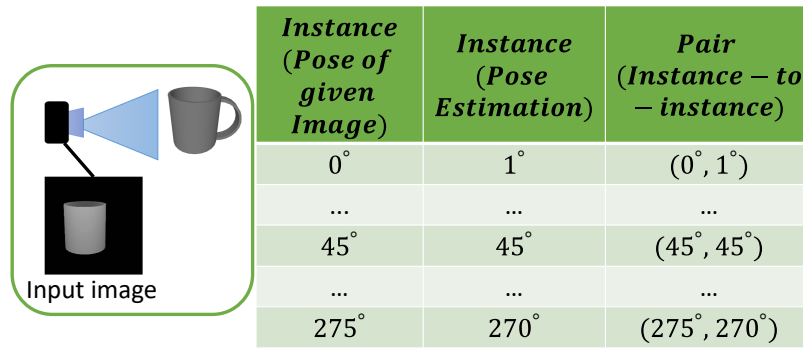


Figure 3.2: Idea of the pairing for object pose estimation

3.3 Best Next-Viewpoint Recommendation by Selecting Minimum Pose Ambiguity for Category-Level Object Pose Estimation

To discuss how the best next-viewpoint is found for object pose estimation, in this Chapter, a novel next-viewpoint recommendation method is proposed based on the selection of the minimum pose ambiguity. A metric called “*pose ambiguity*” is defined given two different viewpoints; initial viewpoint ϕ and rotation angle δ to the next viewpoint from ϕ . These are considered as parameters.

Here, rotation angle δ to the next viewpoint from the initial viewpoint ϕ is used as the parameter. Since the initial viewpoint from the current observation I may be ambiguous, by handling the current viewpoint as a latent variable, the pose ambiguity function is decomposed into “pose ambiguity under given two viewpoints” and “viewpoint ambiguity under a given observation” as illustrated in Figure 3.3. The current viewpoint is defined as the query instance, which will then be paired with the gallery instance to gain the best estimation for one given target object image. The idea to select the best next-viewpoint from the possible viewpoints based on the probability is a critical element for the proposed idea. Figure 3.4 illustrates the angle distribution for the “pose ambiguity” $A(\delta; I)$. Here, the minimum value of A infers the best next-viewpoint as δ [°]. In the end, the pose estimation from two

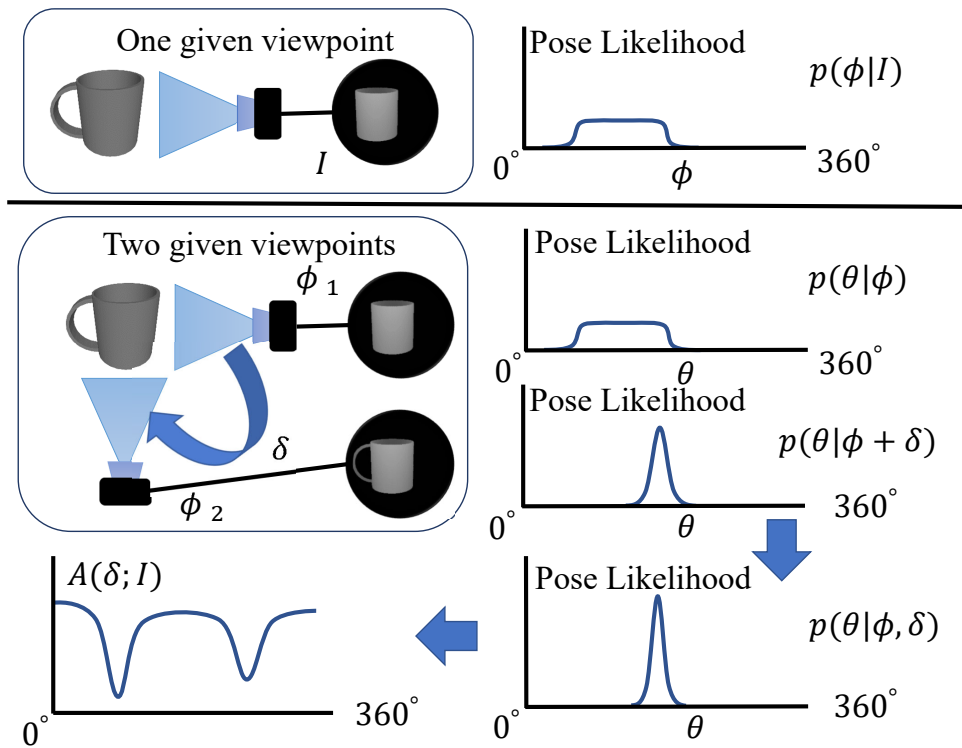


Figure 3.3: Illustration of the idea on the one given viewpoint (top), and two given viewpoints (bottom) for the next viewpoint recommendation

viewpoints is calculated by averaging them. Details of the process are introduced in the following Sections.

3.3.1 Minimum pose ambiguity selection framework

This framework will measure the pose ambiguity in a quantitative way. First of all, what is pose ambiguity? Here, it is defined as the difficulty to estimate the pose of an object from a viewpoint. If the possibility of the object pose θ is widely distributed, the result can be considered as ambiguous. Therefore, the pose ambiguity $A(\delta; I)$ is defined based on pose likelihood distribution $p(\theta|I, \delta)$ given the initial observation I and rotation angle δ as illustrated in Figure 3.4. Here, a mapping functional G is introduced from the pose likelihood distribution to the pose ambiguity. For example, G can be defined by the Entropy of $p(\theta|I, \delta)$ as

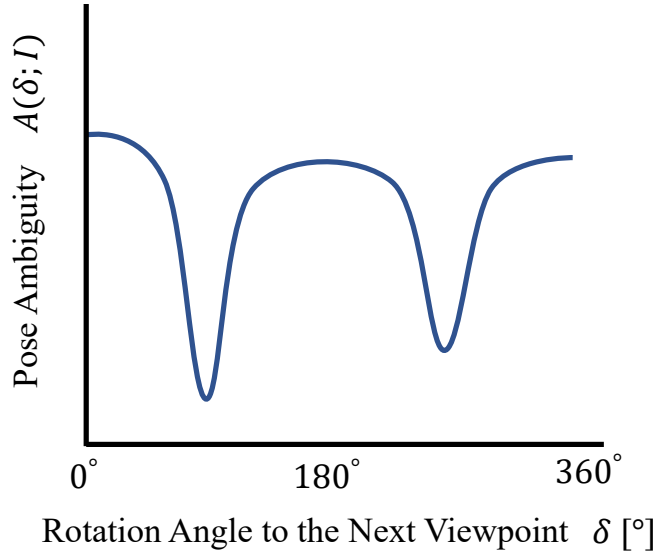


Figure 3.4: Pose ambiguity distribution when selecting a next viewpoint δ [°] from the initial viewpoint (Input image I with $\phi = 95^\circ$) as shown in Figure 1.10

$$A(\delta; I) = G(p(\theta|I, \delta)) = - \int p(\theta|I, \delta) \log p(\theta|I, \delta) d\theta. \quad (3.1)$$

Here, the pose likelihood distribution is evaluated under an image observed from the initial viewpoint, and then the rotation angle to the best next-viewpoint is yielded. Therefore, the pose likelihood distribution is defined as a conditional distribution $p(\theta|I, \delta)$ when an image I from the current viewpoint and a rotation angle δ are given.

The minimum value of the pose ambiguity in G will tell us the best next viewpoint for accurate pose estimation using the two viewpoints. By using the formulation, we find the best viewpoint by finding the minimum entropy as

$$\widehat{\delta} = \arg \min_{\delta} G(p(\theta|I, \delta)). \quad (3.2)$$

To handle the ambiguity of the initial viewpoint, the pose likelihood distribution is further decomposed as follows:

$$p(\theta|I, \delta) = \int p(\theta|\phi, \delta)p(\phi|I)d\phi. \quad (3.3)$$

The first term $p(\theta|\phi, \delta)$ indicates the pose likelihood distribution under two given viewpoints ϕ and $\phi + \delta$, and the remaining $p(\phi|I)$ indicates the viewpoint likelihood under a given observation. In the following sections; more details on the two distributions are explained.

3.3.2 Viewpoint likelihood distribution

Since the absolute viewpoint of an observation is difficult to obtain, the viewpoint likelihood distribution can be considered as a relative pose estimation from the initial viewpoint. In the ideal case, if we have a discrete pose classifier in hand for the pose estimation, we may obtain not only the estimation result (pose) but also the likelihood for all possible poses. On the other hand, if we take a regression-based approach for the pose estimation, such as Pose-CyclicR-Net proposed by Ninomiya et al. [117], we may only obtain an estimation result such as

$$\phi = f(I), \quad (3.4)$$

where I represents a given image and f the pose estimator.

For such a regression-based pose estimator, how can we obtain the viewpoint likelihood distribution? Since we have many images I_i of various objects in a class, by applying pose estimation to those corresponding images, we can obtain many pose estimation results ϕ_i . From these results and corresponding ground-truth poses, we can obtain a huge number of pairs of an estimation result and its ground truth.

By applying density estimation to these data, we can obtain a conditional distribution as $p(\phi|f(i)) = p(\phi_{\text{gt}}|\phi_{\text{est}})$, where ϕ_{gt} represents the ground truth and ϕ_{est} the

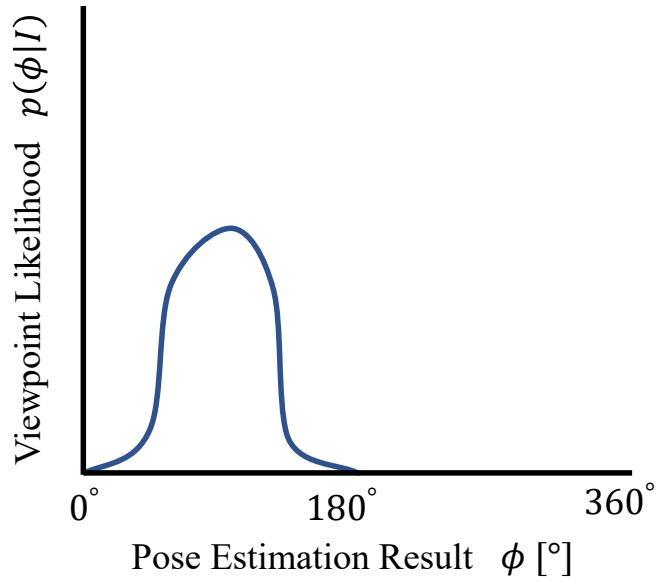


Figure 3.5: Viewpoint likelihood distribution $p(\phi|I)$ (Input image I with $\phi = 95^\circ$)

estimation result. By using this conditional distribution, we can obtain the viewpoint likelihood distribution as,

$$p(\phi|I) = p(\phi|f(I)) \quad (3.5)$$

for a regression-based object pose estimator. This viewpoint likelihood distribution is illustrated in Figure 3.5, which shows the likelihood of pose estimation results from an observation I , the initial viewpoint. Since it is difficult to estimate the pose from an observation which is captured from the viewpoint as shown in Figure 3.4, the pose likelihood is widely distributed from about 45° to 220° .

3.3.3 Pose likelihood distribution

Here the pose likelihood distribution given two viewpoints ϕ and $\phi + \delta$ is explained, where ϕ represents the initial viewpoint and δ the rotation angle to the next viewpoint. The likelihood represents how likely the estimated poses are given the two

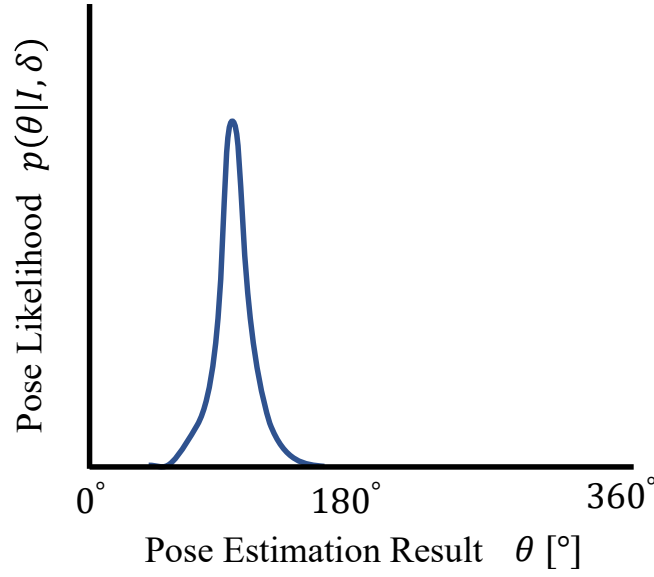


Figure 3.6: Pose likelihood distribution $p(\theta|I, \delta)$ given two viewpoints (Input image I with $\phi = 95^\circ$)

viewpoints. The pose likelihood distribution given two viewpoints is illustrated in Figure 3.6. Here, the likelihood distribution is simply decomposed into two pose likelihoods as

$$p(\theta|\phi, \delta) = p(\theta|\phi)p(\theta|\phi + \delta), \quad (3.6)$$

where $p(\theta|\phi)$ and $p(\theta|\phi + \delta)$ denote the pose likelihood distributions given a viewpoint ϕ and $\phi + \delta$, respectively. This equation holds by assuming $p(\theta)$, which is the pose likelihood without any information, follows a uniform distribution. Each likelihood distribution given a viewpoint can also be calculated by applying density estimation for the pairs of a pose estimation result and the ground truth similar to the method described in Section 3.3.2.

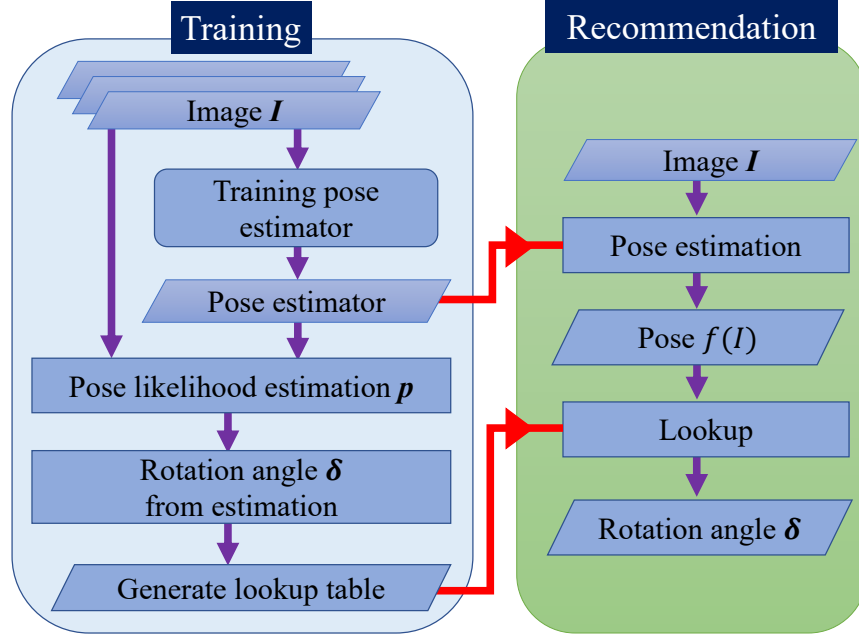


Figure 3.7: Viewpoint likelihood distribution $p(\phi|I)$ (Input image I with $\phi = 95^\circ$)

3.3.4 Pose estimation

After finding the best δ by using Equation (3.2), finally we can estimate the object pose from two viewpoints; the initial viewpoint ϕ and the next viewpoint $\phi + \widehat{\delta}$. Here, I_1 is the image observed from the initial viewpoint ϕ . After rotating $\widehat{\delta}$, we obtain I_2 , which is the image observed from the next viewpoint.

We estimate the pose θ_e from these two viewpoints as the average of pose estimation results ϕ_1 and ϕ_2 from I_1 and I_2 , respectively, as

$$\theta_e = \frac{\phi_1 + \phi_2 - \widehat{\delta}}{2}, \quad (3.7)$$

where $\phi_1 = f(I_1)$ is the pose estimation from the initial viewpoint and $\phi_2 = f(I_2)$ that from the next viewpoint. Since $\widehat{\delta}$ is selected in terms of the minimum pose ambiguity given an initial viewpoint and the rotation angle $\widehat{\delta}$, the averaged pose θ_e will be optimal.

The next viewpoint recommendation idea, which consists of training and recommendation phases, is shown in Figure 3.7. In practice, the pose likelihood distribution $p(\theta|I, \delta)$ is implemented by a lookup table. This lookup table is pre-computed in the training phase. In the recommendation phase, after the current pose is estimated from the initial viewpoint $p(\phi|I)$, we can obtain $\hat{\delta}$ by referring to the lookup table.

3.4 Experiments

3.4.1 Dataset

To show the effectiveness of the proposed viewpoint recommendation method, a simulation-based evaluation is performed. For the simulation, 125 3D models in five object categories; “Airplane”, “Car”, “Chair”, “Mug”, and “Toilet”, were selected from the ShapeNet dataset [157]. The five categories of indoor objects were selected from the versatile characteristic for each of them. The “Airplane” and “Car” are considered and defined as an indoor object reflecting the kid’s toys. Concretely, a 3D model was put in a virtual environment and observed using a virtual depth sensor. By rotating the sensor around the z-axis of the 3D model, 360 depth images in the range of $[0^\circ, 360^\circ)$ were obtained for each model as shown in Figure 3.8. To focus on the essential part of the proposed algorithm, the pose was estimated in the single axis rotation setting. Additionally, in the simulation, the elevation angle of the virtual sensor was changed as 0° , 15° , 30° , 45° , 60° , and 75° which was elevated upright from the z-plane as shown in Figure 3.9.

In total, 125 objects were observed from each elevation angle with a total of 45,000 images. The rendered images were used for both training and testing in the evaluation. These 125 objects in a category from the synthetic datasets were divided into five folds for evaluating the proposed pose estimation method compared to other methods in a five-fold cross-validation setup. For each fold, images of 25 objects were used for testing and the remaining objects for training the model.

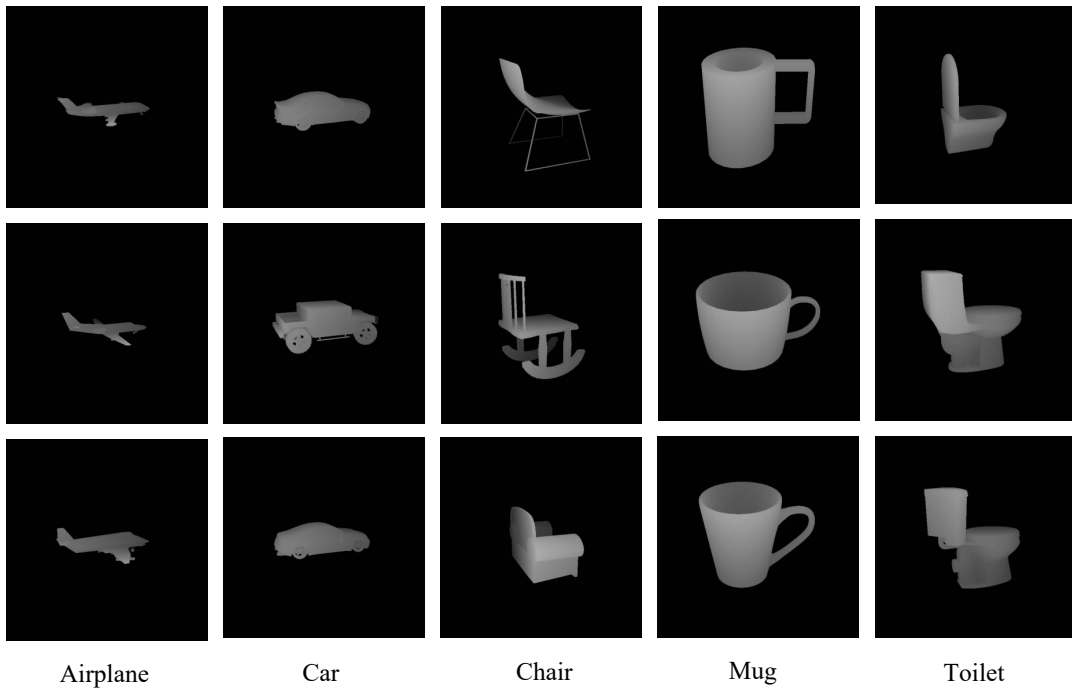


Figure 3.8: Example of images from the five classes in the ShapeNet dataset [157], used in the experiment

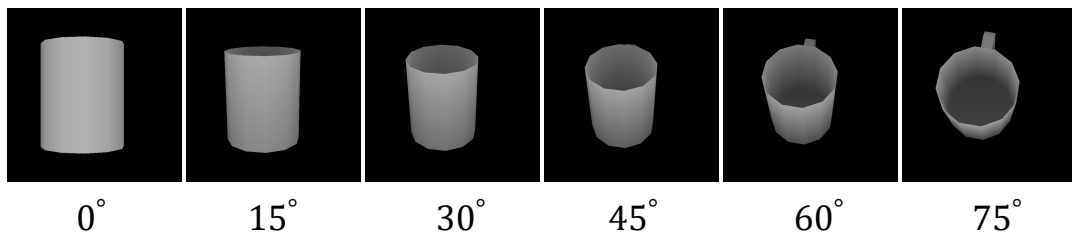


Figure 3.9: Example of “Mug” images observed from different elevation angles

3.4.2 Pose estimation method

In the proposed method, any regression-based pose estimation can be used. Since this part is not the core of the proposed method, a network architecture similar to the Pose-CyclicR-Net proposed by Ninomiya et al. [117] was selected as the pose estimator. Their original network architecture is shown in Figure 3.10. Since the object pose variation was limited to a single axis rotation, the network output was modified to a pair of trigonometric functions ($\cos \theta$, $\sin \theta$) instead of the original quaternion. The pose estimator was trained using the training images.

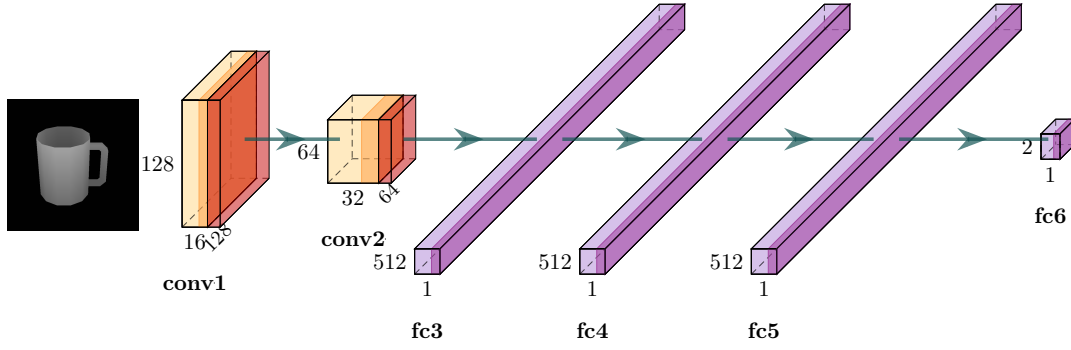


Figure 3.10: Original Pose-CyclicR-Net [117]. In the convolution (conv) layers 1 and 2, the light yellow boxes refer to the 2D convolution, the orange boxes to the ReLu layer, and the light red boxes to the maxpooling layer. In the fully connected (fc) layers 3, 4, 5, and 6, the light purple box refers to the dense layer and the dark purple boxes to the ReLU layer

3.4.3 Evaluation criteria

The evaluation on the appropriateness of the recommended viewpoints for the pose estimation is performed by using several criteria.

One criterion is the Mean Absolute Error (MAE) of the pose estimation results with the ground truth. The pose estimation results are obtained by using a pair of the initial viewpoint and the recommended viewpoint. By considering the circularity of angles, the error can be calculated as

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^N d(\theta_e^i, \theta_g^i), \quad (3.8)$$

where N represents the number of images, θ_e^i and θ_g^i the pose estimation result and the ground truth, respectively. $d(\theta_e^i, \theta_g^i)$ is the absolute difference of the poses considering the circularity defined as

$$d|\theta_e - \theta_g| = \begin{cases} |\theta_e - \theta_g| & \text{if } |\theta_e - \theta_g| > 180^\circ, \\ 180^\circ - |\theta_e - \theta_g| & \text{otherwise.} \end{cases} \quad (3.9)$$

The other criterion is Pose Estimation Accuracy (PEA), which is defined as

$$\text{PEA}(\tau) = \frac{1}{N} \sum_{i=1}^N F(d(\theta_e^i, \theta_g^i) < \tau), \quad (3.10)$$

where τ represents a threshold error which reflects the difference of pose estimation result θ_e^i and the ground truth θ_g^i , $F(\cdot)$ is a function which returns 1 if the condition in the function holds and 0 vice versa.

Standard deviation over the five-fold cross-validation is also evaluated. This criterion is defined as

$$\sigma = \sqrt{\frac{1}{K} \sum_{i=1}^K (\text{MAE}_k - \overline{\text{MAE}})^2}, \quad (3.11)$$

where MAE_k is the MAE for each fold, and $\overline{\text{MAE}}$ represents the average of them.

3.4.4 Comparative methods

The pose estimation results by the proposed method and several other baseline methods are compared. To the best of my knowledge, there is no existing method that could be directly compared with the proposed method as the category-level best next-viewpoint estimation task has just been initiated by this work. Thus as a baseline, pose estimation from a single viewpoint, which just applies a Pose-CyclicR-Net [117]-like network to the input image from the initial viewpoint is used.

Several viewpoint recommendation methods are also compared.

1. Single Viewpoint Object Pose Estimation method: A traditional viewpoint recommendation method utilized by a majority of previous work. This method considers pose from only a single viewpoint.
2. Adopted from Sock et al.'s work [113], Random Viewpoint Object Pose Estimation method, this method considers pose estimation from two viewpoints and random: The next viewpoint is randomly selected from all around 360°. Because of the randomness, ten viewpoints are tried randomly and the estimation results are averaged in the evaluation.
3. Also adopted from Sock et al.'s work [113], Opposite Viewpoint Object Pose Estimation method, this method considers pose estimation from two viewpoints and opposite: The next viewpoint is selected from the angle opposite to the current viewpoint. This is the furthest point from the initial point, so Sock et al. call it the "Furthest" in their paper.

Figure 3.11 shows examples of the viewpoint recommendation methods for a given "Mug" image. The images shown in Figure 3.11 are observations from the recommended viewpoints.

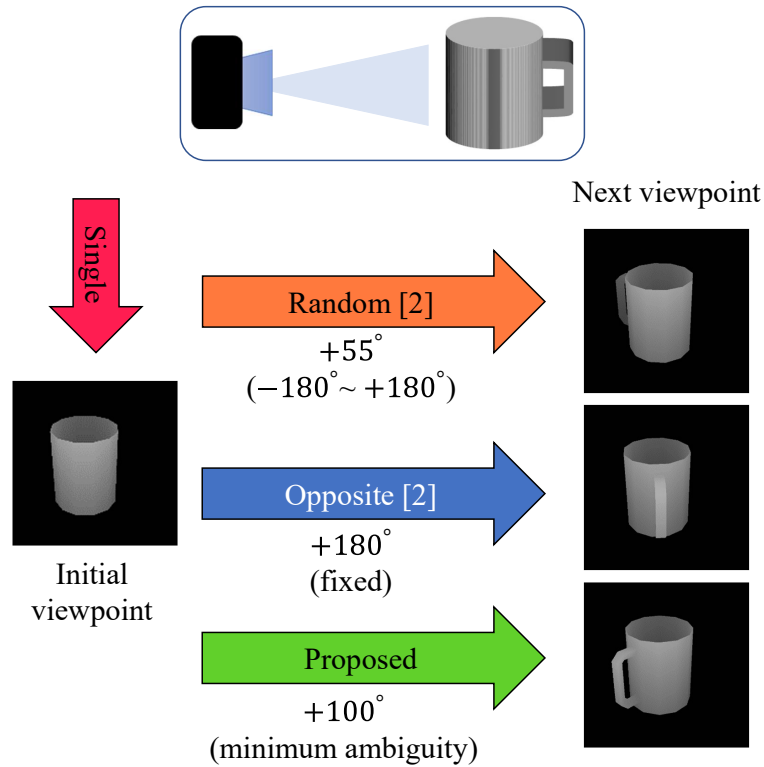


Figure 3.11: Example of images observed from the estimated viewpoints by the proposed method and comparative methods

3.4.5 Results

3.4.5.1 Comparison on Mean Absolute Error (MAE)

Pose estimation error is analyzed with various object categories; “Airplane” “Car”, “Chair”, “Mug”, and “Toilet”. Table 3.1 tabulates the MAE comparison between all five object categories when the elevation angle is 0° . We can see that the proposed method almost outperforms the comparative methods for different target objects. Only in the “Chair” category, it ranks the third after “Opposite” and “Random”. Overall, this result shows that the proposed method is capable to produce a good estimation for various categories.

Then, to investigate the relation between the elevation angle with the pose estimation method, “Mug” was chosen as the target object category and the results are shown

Table 3.1: Comparison of MAE for the five object categories when the elevation angle is 0° by five-fold cross validation

Object Category	Single	Random	Opposite	Proposed
“Airplane”	12.86°	11.76°	11.79°	11.45°
“Car”	8.73°	7.75°	8.12°	7.09°
“Chair”	10.88°	8.24°	7.30°	8.77°
“Mug”	13.07°	11.34°	11.03°	9.58°
“Toilet”	10.12°	8.54°	7.96°	7.66°

Table 3.2: Comparison of MAE for different elevation angles of “Mug” by five-fold cross validation

Elevation Angle	Single	Random	Opposite	Proposed
0°	12.55°	10.67°	10.14°	9.34°
15°	12.63°	10.79°	10.36°	9.51°
30°	12.11°	10.25°	9.97°	8.81°
45°	8.98°	7.46°	7.12°	6.76°
60°	6.05°	5.04°	5.07°	4.68°
75°	4.80°	4.01°	3.83°	4.44°

in Table 3.2. For most of the elevation angles, the proposed method almost outperformed the comparative methods. The “Opposite” and the “Random” methods outperformed the proposed method only in the case of 75° elevation angle. For this elevation angle, the proposed method was not the best but still could be considered comparable to the comparative methods. However, since in the proposed work, the main priority is the pose estimation from 0° to 45° considering the ambiguity problem, these cases become less critical.

These results clearly show that the proposed method is effective and gives a better way (next viewpoint) for object pose estimation. The proposed method successfully managed to reduce the pose ambiguity in the difficult observation which has been mentioned earlier in Figure 3.3. We can see that estimating an object’s pose from

Table 3.3: Comparison using partial - Area Under Curve (pAUC) of Pose Estimation Accuracy (PEA) by changing the error threshold for for the five object categories when the elevation angle is 0° by five-fold cross validation

Object Category	Single	Random	Opposite	Proposed
“Airplane”	89.01	89.01	89.03	89.65
“Car”	92.35	92.00	91.52	92.58
“Chair”	88.76	91.27	92.19	90.72
“Mug”	87.12	88.35	88.46	90.24
“Toilet”	90.72	91.20	91.55	92.23

two viewpoints yields a better result than that from a single viewpoint. By comparing with the other pose recommendation methods, the proposed method achieves better results by carefully selecting the best viewpoint for object pose estimation.

3.4.5.2 Comparison on Pose Estimation Accuracy (PEA)

In general, smaller error is the main priority for pose estimation analysis with the comparative methods. Thus, the Pose Estimation Accuracy (PEA) is analyzed by changing the error threshold τ in Equation (3.10) in the case of elevation angle 0° . To see the performance of the proposed method with the various categories, partial-Area Under Curve (pAUC) is calculated and summarized in Table 3.3. The proposed method achieved the highest pAUC than the comparative methods for each of the four object categories except for the “Chair”.

Using “Mug” as the target object category, an evaluation for PEA is conducted with the elevation angles, from 0° to 75° . In five out of six elevation angles, the proposed method achieved the most accurate results compared to the other methods as illustrated in Table 3.4. For the remaining elevation angle, 75° , the proposed method was not the best but could be considered comparable to the comparative methods. This is because object poses are easily distinguishable from the initial viewpoints. In these cases, viewpoint selection becomes less important.

Table 3.4: Comparison using partial- Area Under Curve (pAUC) of Pose Estimation Accuracy (PEA) by changing the error threshold τ from 0° to 100° by five-fold cross validation

Elevation Angle	Single	Random	Opposite	Proposed
0°	87.73	89.09	89.52	90.68
15°	87.79	88.91	89.16	90.53
30°	88.23	89.44	89.74	91.11
45°	91.01	92.13	92.50	92.95
60°	93.68	94.49	94.44	94.92
75°	94.85	95.51	95.64	95.14

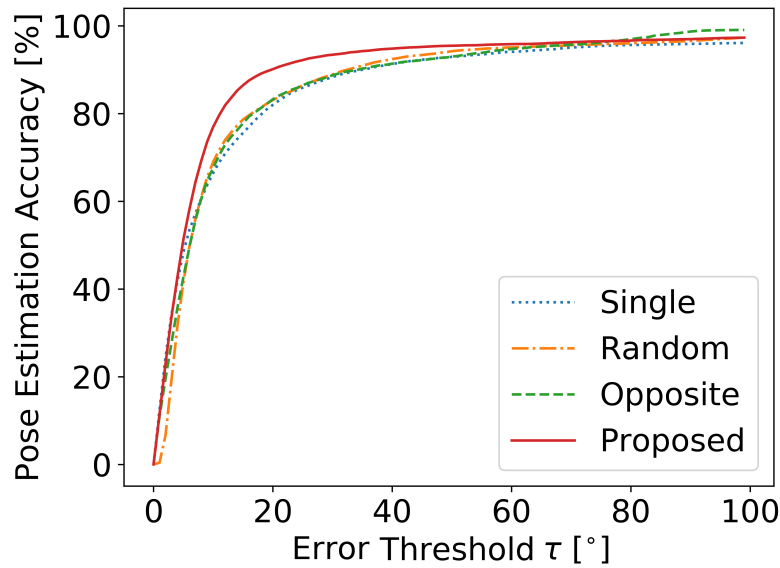


Figure 3.12: Pose Estimation Accuracy by changing the error threshold τ from 0° to 100° (elevation angle = 30°)

The relation between the pose estimation accuracy and the error threshold is plotted in Figure 3.12. Here, the pose estimation accuracy is plotted by changing the error threshold τ in Equation (3.10) in the case of elevation angle 30° (testing fold number 2). We can see that the proposed method outperformed all the comparative methods when the error threshold is within $0^\circ \leq \tau < 70^\circ$. When $70^\circ \leq \tau < 100^\circ$, the “Opposite” method outperformed the proposed method. However, a large error threshold value will not critically influence the pose estimation accuracy, so we can consider

the results when $\tau \geq 70^\circ$ are not significant for this purpose.

3.5 Discussion

This section discusses more details about the experimental results presented in the previous section. The comparisons of results between methods and discussion on the relative relation of the indoor object type image with the object's characteristics that can be recognized from the newly proposed ambiguity minimization pose estimation are provided here.

The methods' performances are analyzed quantitatively and qualitatively to validate the proposed method's effectiveness to affirm the experimental goals.

3.5.1 Quantitative evaluation

To discuss the quantitative evaluation of the results obtained in Section 3.4, the Mean Absolute Error (MAE) and partial Area Under Curve (pAUC) are analyzed statistically. A box plot graph of MAE is illustrated in Figure 3.13. This graph indicates the minimum, lower quartile, median, upper quartile, and the maximum values.

The median values of the MAE of the proposed method show the lowest value for "Airplane" and "Mug". According to the "Airplane" and "Mug" statistical results, they show that not only the "Mug", which has only a part of the one-axis symmetrical object, delivers a good result in the analysis, but, "Airplane" as a normal one-axis symmetrical object also achieved a promising performance. The one-axis symmetrical objects here, "Airplane", "Car", "Chair", "Mug", and "Toilet" are object categories considered to be symmetrical in one-axis or one plane. The "Mug" has the minimum part from the whole object, which indicates the symmetrical characteristic compared to the other four object categories. An earlier assumption was that, the unique characteristic of an object could yield a good result. However, this experiment shows that the proposed method could perform with a versatile object type.

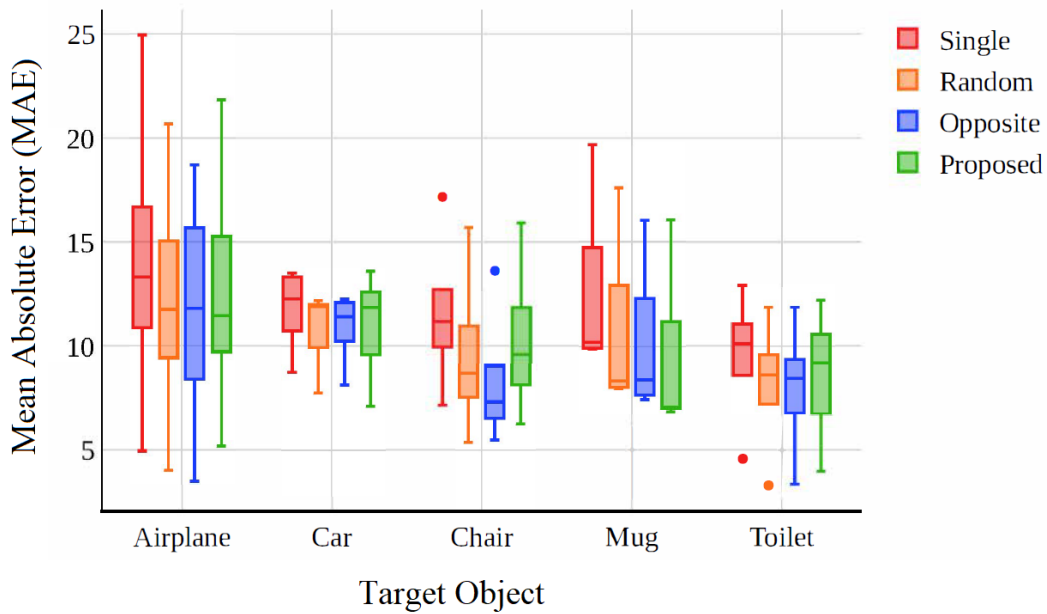


Figure 3.13: Example of one-axis symmetrical object

The difference between one-axis symmetrical object examples for the MAE analysis by using the three comparative methods and the proposed method are illustrated in Figure 3.13.

Furthermore, from all the object type analyses, the absence of outliers which are indicated by dots in Figure 3.13 shows that the proposed method delivers a promising approach to estimating the pose from an ambiguous viewpoint. Here the dots presented an outlier for the “Chair” with “Single” and “Opposite” and “Toilet” with “Single” and “Random” methods, in several MAE data from this analysis.

For “Car” and “Toilet”, the proposed method performs competitively with the “Opposite” method. However, for “Chair”, the proposed method ranked the third.

By focussing on the elevation angle at 0° with “Mug” images, MAE and pAUCs of PEA for each fold is illustrated in Table 3.5. The standard deviation is used to ensure agreement between the theoretical prediction and the proposed pose estimation method’s performance.

Table 3.5: Comparison of the overall Mean Absolute Error (MAE) and Partial- Area Under Curve (pAUC) of Pose Estimation Accuracy (PEA) for each fold (0° elevation angle)

Testing Fold Number	Single		Random		Opposite		Proposed	
	MAE	pAUC of PEA	MAE	pAUC of PEA	MAE	pAUC of PEA	MAE	pAUC of PEA
Fold 1	19.68°	82.23	17.59°	82.91	16.03°	84.29	16.06°	85.53
Fold 2	13.07°	87.12	11.34°	88.35	11.03°	88.46	9.58°	90.24
Fold 3	10.20°	89.36	8.05°	91.45	7.76°	91.73	6.85°	92.59
Fold 4	9.92°	90.09	8.00°	91.59	7.45°	92.04	7.10°	92.59
Fold 5	9.88°	89.87	8.35°	91.15	8.41°	91.09	7.09°	92.42
Average	12.55°	87.23	10.67°	89.09	10.14°	89.52	9.34°	90.68
σ	4.20°	3.29	4.11°	3.70	3.59°	3.25	3.92°	3.05

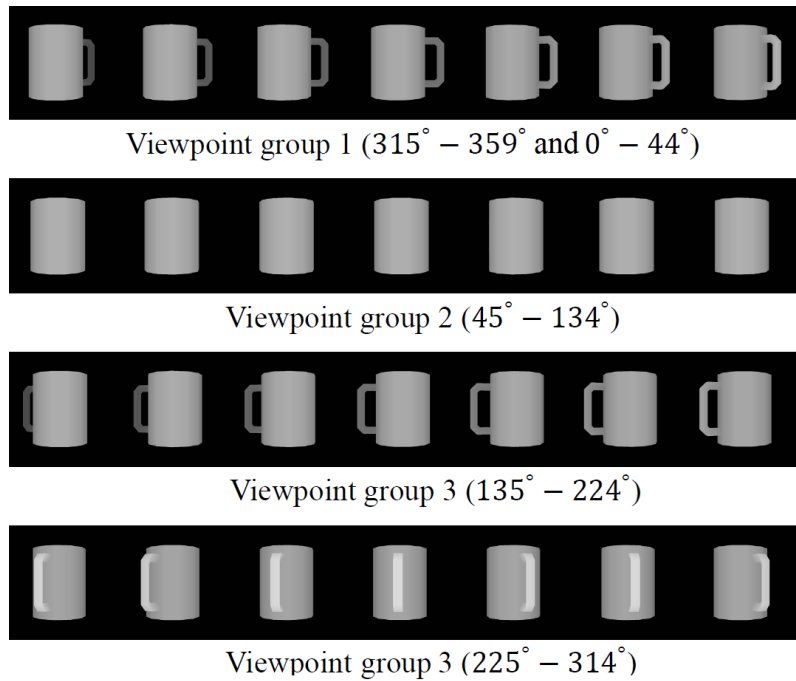


Figure 3.14: Image examples for the four viewpoint groups using the “Mug” object category

By looking at each of the standard deviation value for MAE and the pAUCs of PEA, the pose estimation for the five object categories are further being analyzed and compared.

The proposed method’s standard deviation is 3.92, which shows that it is the second stable among all the methods. For the pAUC of PEA analysis, for the fold numbers 2 to 5, the proposed method gained the highest estimation accuracy. The standard deviation of the proposed method for the pAUC of PEA for all folds, is 3.05 which also shows that it is the most stable method among all the methods.

3.5.2 Qualitative evaluation

Furthermore, qualitative evaluation is investigated to confirm the advantages of the proposed method by grouping the initial viewpoint into four groups as shown in Figure 3.14; initial viewpoint group 1 ($0^\circ - 44^\circ$ and $315^\circ - 359^\circ$), group 2 ($45^\circ - 134^\circ$), group 3 ($135^\circ - 224^\circ$), and group 4 ($225^\circ - 314^\circ$).

Table 3.6: Comparison of MAE for “Mug” for different initial viewpoint groups (0° elevation angle)





Initial Viewpoint Group	Single	Random	Opposite	Proposed
1	9.03°	7.31°	6.30°	6.00°
2	17.66°	11.43°	9.23°	9.97°
3	7.13°	6.77°	6.30°	6.01°
4	6.97°	6.71°	9.22°	5.41°

The four group selection is sufficient to analyze four types of initial viewpoints, since all the viewpoint possibilities in the range of $[0^\circ, 360^\circ)$ slightly differ in few degrees of the representative viewpoint which are 0° , 90° , 180° , and 270° .

In Table 3.6, estimation results of different initial viewpoint groups are shown. We can see that the proposed method outperforms the comparative methods in most groups. For group 2, although the proposed method ranks the second, it still presents its competitiveness to the “Opposite” method with a difference of 0.74° .

For a more qualitative study, the initial viewpoint image is also compared with the best next-viewpoint using the “Mug” images. Since the proposed method could suggest and select the best next-viewpoint, even if we had an ambiguous image as the initial viewpoint, the proposed method could still estimate the object’s pose accurately. For the visual purpose of the qualitative analysis, four initial viewpoints were randomly selected as the example object images. Each of the image represents the four initial viewpoint groups. Table 3.7 provides the output examples from a less ambiguous initial viewpoint for the proposed method and comparative methods. We can see that the proposed method achieves better pose estimation results than the comparative methods.

Table 3.7: Comparison of MAE for “Mug” when the elevation angle is 0° for viewpoint groups. The value in brackets represent the difference between the initial viewpoint and the pose estimation result

Image	Single	Random	Opposite	Proposed
 0.0°	342.0° (18.0°)	356.0° (3.5°)	357.5° (2.5°)	0.0° (0.0°)
 30.0°	40.0° (10.0°)	39.5° (9.5°)	41.5° (11.5°)	32.0° (2.0°)
 210.0°	219.0° (9.0°)	223.0° (13.0°)	220.5° (10.5°)	207.5° (2.5°)
 330.0°	298.0° (32.0°)	292.0° (38.0°)	298.5° (31.5°)	326.5° (3.5°)

3.5.3 Limitation and further considerations

Regarding the proposed multiple viewpoint pose estimation method via minimization of pose ambiguity with the utilization of the best next-viewpoint, there are some limitations and considerations necessary for its development in the future.

- The proposed method is designed for the multiple-viewpoint object pose estimation task with indoor object images and with the single-dimensional axis rotation for the image capturing the viewpoint. It is easily-expandable for the multi-dimensional axis rotation environment setting. It is expected that there will be a more straightforward strategy to reduce the time for training all images for more than one rotation axis, up to three-axes, in an acceptable time for gaining a network model.

- The proposed method uses a simple averaging for the two pose estimation results ϕ_1 and ϕ_2 , and considers the rotation angle $\widehat{\delta}$ as shown in Equation (3.7). However, increasing the number of dimensional axis rotations requires modification of the pose estimation formula. There is room for further improvement in revising the equation, which will be aligned with the increasing dimensional axis rotation, e.g., yaw and pitch, instead of the roll used here.
- Utilizing the simulated rendered images is one of the limitations. Careful handling of real captured data with heavy noise is one of future work. Comparing the effectiveness of simulated images and real captured images in pose estimation will also be one of future research prospects.
- Since the object categories utilized in the evaluation analysis is limited, there is still room for improvement using a wide variety of object category images to ensure the versatility of the proposed method in various environment settings.

3.6 Summary

This chapter introduced a solution for Research topic 1, which corresponds to the improvement on selecting the best next viewpoint given an initial viewpoint for object pose estimation from multiple viewpoints. The limitations on single viewpoint pose estimation for finding the best estimation pose which could not select the best viewpoint was first discussed here. Considering the initial viewpoint as the latent variable for exploring the best next-viewpoint performed extensively to pair the best next-viewpoint for a given current viewpoint. Several experiments were performed to evaluate the quantitative and qualitative performance for the pose ambiguity minimization method. This pose ambiguity minimization method was demonstrated to be adequate to estimate the ambiguous viewpoint for a given object category-level. Through this example, the single query pairing approach showed its effectiveness in estimating a good estimation pair by looking at the multiple viewpoint pose estimation task as one of the object identification problems.

Chapter 4

Pairing Approach for Person Re-identification

As presented and discussed in Chapter 1, the main problem raised in the thesis is how to develop and discern the suitable methods for the pairing approaches in versatile object identification applications. In the previous Chapter, single-query pairing was discussed and proved through the two viewpoints object pose estimation for an indoor object category-level. In this chapter, as Research topic 2, multiple-query pairing is discussed and proved through the person re-identification task. For making the person re-identification task suit with the multiple-query pairing approach, the simultaneous image pairing task is introduced. Therefore, Simultaneous Multiple Query Pairing is introduced as one of the multiple-query pairing approaches for object identification in this Chapter to meet the increasing demand on the person re-identification task.

This chapter is structured as follows. First, Section 4.1 describes the general background of ideas of multiple-query pairing in person re-identification. Next, how the instances are regulated and considered to work accordingly with the person re-identification context is discussed in Section 4.2. Then, Section 4.3 elaborates the proposed framework to gain a good pairing strategy for the person re-identification

task via a simultaneous image pairing approach. In Section 4.4 the problem is evaluated by investigating the effectiveness of the pairing strategy for given images with several comparative methods and by changing the image number for both gallery and query data. Additional experiments for further discussion are also provided in Section 4.5 before finally the chapter is closed with a summary in Section 4.6.

4.1 Introduction

This chapter discusses the actual pairing process for person re-identification for successfully pairing similar person identities from the query to the gallery set of dataset images. Multiple-query pairing for person re-identification demands a good person pairing strategy.

As discussed in Chapter 2, especially in Section 2.2, various methods have been proposed for the person re-identification task from the basic idea of image pairing by image feature to the recent deep learning methods [53, 61–69, 115, 135–152, 156]. Most of the proposed works including Zheng et al.’s work [115] utilize the individual pairing strategy to match the query to gallery images. They show good re-identification accuracy, but with the introduction of the simultaneous image pairing in both query and gallery sets, it could be improved by restricting the matched image from re-matching among all the remaining person images in the query image set.

To find the best pairing between the query images and the gallery images aligned with multiple query pairing, the pairing requires to occur simultaneously, whereby, compared to the traditional image pairing in the previous works conducted separately with a single pairing strategy. Concretely, single pairing methods could yield unmatched images as illustrated in Figure 1.11, but simultaneous person pairing could solve this issue. This chapter focuses on the essential part of the problem setting and the idea of simultaneous person pairing; The method is explained and evaluated by limiting it to two viewpoints involved in the analysis. However, extension

to a more than two viewpoints is possible if we considered a different research setting and viewpoints involved. With the utilization of several public datasets, Viewpoint Invariant Pedestrian Recognition (VIPeR) dataset [48], CUHK01 dataset [49], iLIDS-VID dataset [158], and PRID dataset [50], several assessments are made on the proposed method's effectiveness.

In summary, the contributions in this chapter are as follows:

1. Introduction of a new multiple query pairing approach in person re-identification, considering the Instance-to-Instance pairing approach in the person pairing task.
2. Introduction of a simultaneous person matching scheme for person re-identification based on the Stable Marriage Algorithm (SMA) [120] previously known in Economics.
3. Analysis on a comparison of the image matching methods in cases where the image sets contain equal or non-equal numbers of images.

4.2 Details on the Pairing Approach for Person Re-identification

Chapter 1 introduced multiple-query pairing as one of the pairing categories for object identification. To apply the pairing concept to person re-identification, we should define the instances in the task. In person re-identification, a person is defined as an object for the identification task. During the identification, generally, an object is identified from query camera view to gallery camera view based on the similarity of their identity (ID). Thus, the instance is defined as a person image in this context. Since traditional person re-identification methods took the individual pairing approach based on their re-identification's matching idea, they have difficulty yielding good matching results in a challenging setting. In this thesis, instead of taking the individual pairing approach, a simultaneous pairing approach is considered.

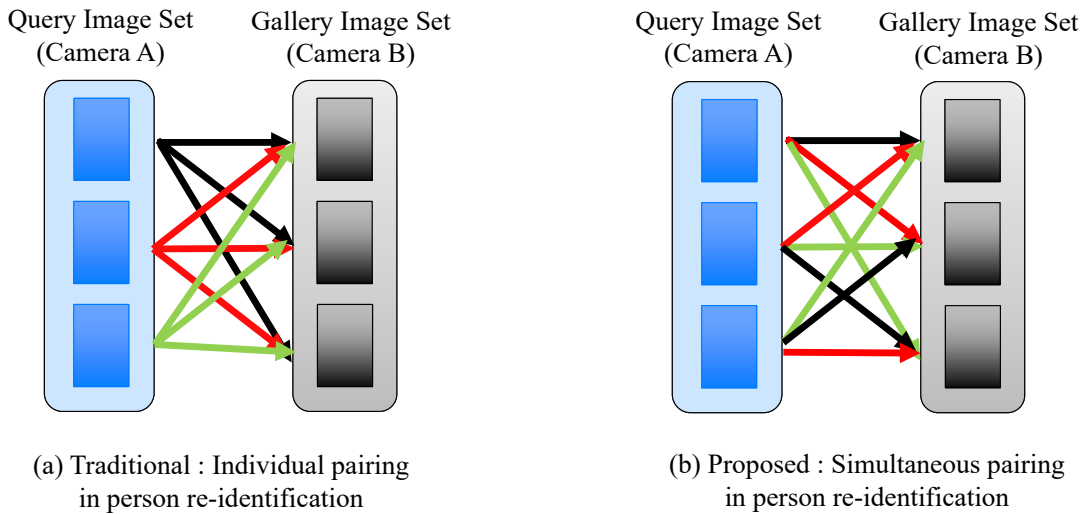


Figure 4.1: Image pairing in person re-identification

Single query pairing, also categorized as specific person retrieval, illustrated in Table 1.1, is defined as an Instance-to-Instance pairing approach that could perform in the traditional person re-identification setting but not in a complex environment setting. This difficult environment may occur in the person's re-identification problem when a repeated matching of images between query and gallery sets is allowed on an event and also with non-similar image numbers between query and gallery sets. A problem also occurs when unmatched images remain at the end of the matching. Ensuring that the person images are not repeatedly matched during the re-identification would solve the first problem in single image pairing person re-identification. To avoid the second problem, the remaining unmatched images are forced to be paired with at least one person image. Here, the instances are paired simultaneously, depending on the image numbers given in the dataset for all the available query images. Expanding the single query pairing into simultaneous image pairing could avoid these repeatedly matched image pairs and unmatched images for producing a good re-identification accuracy for the person re-identification task.

To explain how the instance works in person re-identification, Figure 4.1 illustrates how images are considered an instance in both query and gallery image sets. Black, red and green lines in Figure 4.1 (a) indicate that each image in the query image set is paired with an image in the gallery image set one by one. On the other hand,

the three colored lines in Figure 4.1 (b) show all images in the query set are paired simultaneously with the images in the gallery set.

4.3 Multiple-Query Pairing Approach for Person Re-identification via the Stable Marriage Algorithm

To perform object identification with simultaneous person re-identification, the instance-to-instance pairing is considered as a multiple-query pairing for all the instances in both query and gallery sets simultaneously.

Here, we assume that a camera view detects every person in the scene without redundancy. A specific pairing approach is essential to ensure that the image in the query has a pair instance in the gallery. Considering the person re-identification problem as an instance of the marriage problem, a pairing method is proposed in this chapter that finds a stable pairing of query and gallery image sets. To solve the image matching in person re-identification, the Stable Marriage Algorithm (SMA) [120] is introduced for the person re-identification task. SMA was widely utilized in Economics in early 21st century. To the extent of my knowledge, the introduction of SMA in the identification task has not been challenged. Therefore, the implementation of person re-identification is initiated by re-strategizing the pairing approach. Most conventional person re-identification methods consider only the analysis on the similarities of images but do not consider the similarity ranking from both sides. Meanwhile, the proposed method with the SMA considers the similarity ranking, which provides a more robust matching.

Thanks to the algorithm, a bipartite graph matching based person re-identification method is realized. The overview of the proposed method is illustrated in Figure 4.2. Here, the gist of person re-identification is followed where one query image set is paired with another gallery image set. The core idea of the pairing approach is the middle part of this whole picture. For pairing images on the left-hand side; query

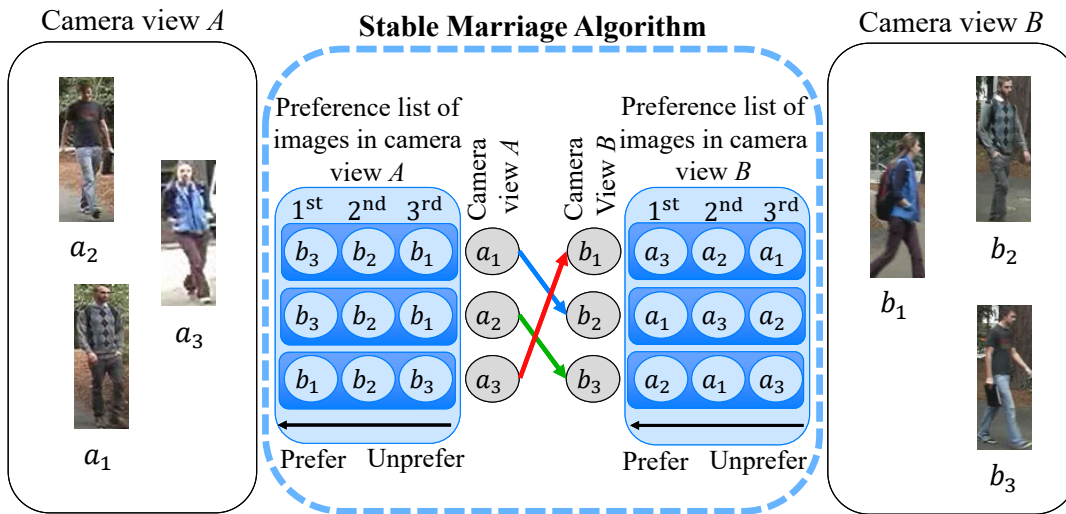


Figure 4.2: Overview of the proposed method. The person image is considered as an instance of the Stable Marriage Problem (SMA). The gray circles represent both camera views' person images as an element of each set, while the light blue circles in the blue shaded areas represent preference lists sorted horizontally for each camera view from preferable to unpreferable. SMA matches images from camera view A's image set to camera view B's image set based on the preference list. The blue line represents the matching from person a_1 to b_2 , the green line from a_2 to b_3 , and the red line from a_3 to b_1 .

image set, the images will be carefully analyzed based on their preference or, in this context, the similarity rank.

In this Section, since the instance pairing part is the core issue in this thesis, the pairing process is explained first. After that, the details of feature extraction and similarity distance calculation are explained in the following sections. In this thesis, since the image feature selection is not the primary focus, simple image features are used and sophisticated features proposed in state-of-the-art person re-identification methods are used for the comparison study.

4.3.1 Marriage problems

The marriage problem was initially proposed by Gale and Shapley [120] a few decades ago in Economics. The concept, which described a matching solution for elements in two sets of elements given a ranking of preference for each element, was

successfully implemented in a real-life situation; doctor-hospital assignment in the United States of America in 1962 [159]. The marriage problem is composed of the following two definitions, assuming that there are men 1, 2, ..., and women 1, 2, ..., each with a preference list of the opposite sex to propose marriage to.

Definition 4.1. An unstable matching of marriage is defined as two persons; man 1 and man 2, are assigned to woman 1 and woman 2, respectively, although man 2 prefers woman 1 to woman 2 and woman 1 prefers man 2 to man 1.

Definition 4.2. The marriage problem is defined as *stable* if every person is matched with a partner.

SMA gives stable pairing of the instances by rematching an existing pair if it showed less rank preference by the instance's partner. Thus, if there are the same numbers of men and women, each of them will have a partner based on his/her preference list. SMA is known to have a polynomial solution, and it also is one example of bipartite graph matching.

4.3.2 Simultaneous image pairing via the Stable Marriage Algorithm

For person re-identification, the similarity rank of a query image to all its gallery images could be considered a similar state with an element with their preference list in the SMA. The similarity image rank represents quantitatively how close a query image is to other gallery images. For this purpose, the idea of simultaneous image pairing is proposed here, by considering the similarity image rank as the preference list in SMA.

The idea of having a similarity rank of the images before the pairing is illustrated in Algorithm 1. SMA uses two preference ranking lists for Camera View A to B and Camera View B to A. In other words, the pairing with SMA for two camera views is considered from both camera view's preference rankings. Meanwhile, the traditional

Algorithm 1 Simultaneous Image Matching via the Stable Marriage Algorithm for Person Re-identification

1: **Inputs:**

$\mathcal{I}^A \leftarrow$ Query for N Images , $\mathcal{I}^B \leftarrow$ Gallery for N Images

Phase 1 - Preference List as Similarity Ranking

2: **for** $i \leftarrow 1$ to N **do**

3: Rank(I_i^A) \leftarrow sort($[I_1^B, I_2^B, \dots]$) by $s(I_i^A, I_j^B)$ in descending order

4: Rank(I_i^B) \leftarrow sort($[I_1^A, I_2^A, \dots]$) by $s(I_i^B, I_j^A)$ in descending order

5: **end for**

Phase 2 - Simultaneous Image Matching

$\forall I_i^A \in \mathcal{I}^A$ and $\forall I_i^B \in \mathcal{I}^B$ as *free*

6: **while** \exists free I_i^A which still has an I_i^B to be matched with **do**

7: $I_i^B \leftarrow$ first rank in I_i^A 's list and I_i^A has not yet been matched

8: **if** I_i^B is *free* **then:**

9: (I_i^A, I_i^B) becomes matched

10: **else:**

11: \exists pair (I_k^A, I_i^B) already matched

12: **if** I_i^B is more similar to I_i^A than I_k^A **then:**

13: I_k^A becomes free

14: (I_i^A, I_i^B) becomes matched

15: **else:**

16: (I_k^A, I_i^B) is kept matched

17: **end if**

18: **end if**

19: **end while**

individual image pairing considers only the ranking for Camera View A to B with several criteria.

By utilizing the proposed simultaneous image matching, preference rankings are considered for both Camera A to B and Camera B to A. Compared to traditional Greedy Matching and Hungarian Matching [153] methods, the proposed method will not only sort the preference ranking and choose the best image in the rank as the final matched image, but also confirms the opposite preference ranking which is a requirement for SMA. In the end, this yields a robust image pairing.

4.3.3 Image similarity ranking

In the proposed method, two kinds of image features are focussed; one is the traditional HSV color histogram, and the other is the modified image feature, named

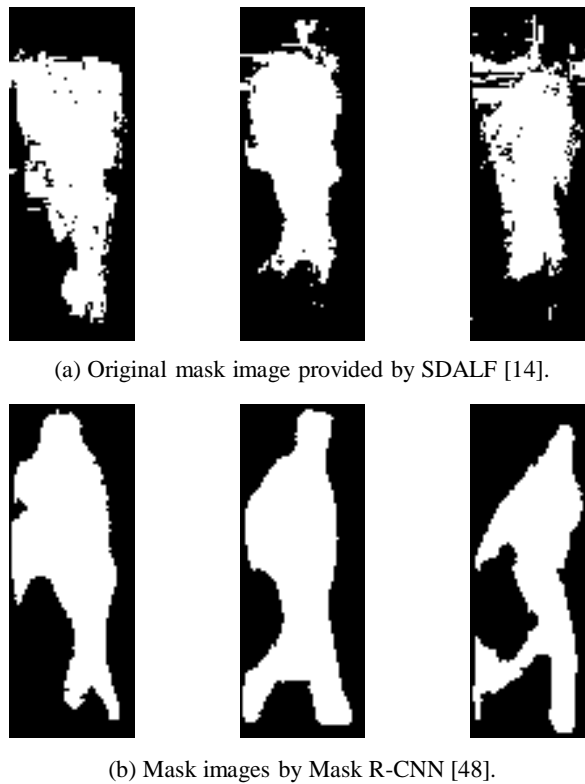


Figure 4.3: Examples of mask images using the VIPeR dataset [48].

Mask-improved Symmetry-Driven Accumulation of Local Features (SDALF) to calculate the similarity between images in both query and gallery images. For the HSV color histogram, the histogram intersection method which is the same as that in [160, 161], is used for image comparison to calculate the distance between the images. For HSV color histogram, since it is commonly used to extract a feature from an image, it is considered as the baseline image feature.

Meanwhile, Mask-improved SDALF is an extension of SDALF, a well-known image feature for person re-identification. While the traditional SDALF utilizes a person image mask generated by Stel Component Analysis [162] to calculate image features, the proposed Mask-improved SDALF estimates the person image mask by using Mask R-CNN [163]. These masked images of the VIPeR dataset utilizing the Mask-improved SDALF are illustrated in Figure 4.3. Since the image feature extracted in the original SDALF depends on the quality of the utilized mask, the

newly proposed and improved version of the SDALF feature is named as Mask-improved SDALF (MSDALF). Thanks to the more exemplary person image masks, the Mask-improved SDALF can extract better image features that capture a person's characteristics.

An M -bin HSV color histogram or Mask-improved SDALF image feature is calculated from two images as I_i^A and I_j^B for the similarity calculation, respectively, where

$$x_i^A = f(I_i^A), \quad (4.1)$$

$$x_j^B = f(I_j^B). \quad (4.2)$$

The image feature similarity of two images will be calculated by the similarity function $s(I_i^A, I_j^B)$ which is defined as

$$s(I_i^A, I_j^B) = s_f(x_i^A, x_j^B), \quad (4.3)$$

where $s_f(x_i^A, x_j^B)$ is the histogram-intersection defined as

$$s_f(x_i^A, x_j^B) = \sum_{k=1}^M \min(x_{ik}^A, x_{jk}^B). \quad (4.4)$$

Using this appearance similarity, we sort the images and then apply SMA to find the matched image pairs by simultaneous image matching.

4.3.4 Simultaneous image pairing implementation for person re-identification

Person re-identification across two camera views is formulated as an instance of the marriage problem given person images I^A and I^B detected from two camera-views (Cameras A and B), respectively, as follows.

$$I^A = \{I_1^A, I_2^A, \dots, I_N^A\} \quad (4.5)$$

$$I^B = \{I_1^B, I_2^B, \dots, I_N^B\} \quad (4.6)$$

By following the SMA, we obtain a similarity ranking list for each person image. Table 4.1 showed an example with three images from Camera A and three images from Camera B . For matching these images, for each image in I^A , images in I^B are sorted by the similarity with respect to each image in I^A in descending order. The proposed algorithm, simultaneous image matching via SMA, is shown in Algorithm 1.

In SMA, matching will be performed by looking up both Tables 4.1 (a) and (b) alternately. First, I_1^A and I_3^B are matched as I_3^B is the best choice for I_1^A , and I_1^A is also the first rank for I_3^B . Next, I_2^A is matched with I_1^B , as I_2^A is the second rank in I_1^B and it still is not matched with any image yet. In the third loop, I_3^A is matched with I_3^B , but since I_3^B is already matched with I_1^A , I_3^A remains not matched in the first loop and I_3^B is removed from the rank list of I_3^A . In the fourth loop, I_3^A which is not matched yet is matched with I_2^B . Finally, we have three matched image pairs which are (I_1^A, I_3^B) , (I_2^A, I_1^B) , and (I_3^A, I_2^B) .

Here, we can see the query image is matched to all gallery images simultaneously by considering the similarity rank of each image. It brings us a new versatile approach

Table 4.1: Example of image similarity based on feature similarity

(a) From I^A to I^B

I^A Image List			
Query Image	First Rank	Second Rank	Third Rank
I_1^A	I_3^B	I_1^B	I_2^B
I_2^A	I_1^B	I_3^B	I_2^B
I_3^A	I_3^B	I_2^B	I_1^B

(b) From I^B to I^A

I^B Image List			
Query Image	First Rank	Second Rank	Third Rank
I_1^B	I_1^A	I_2^A	I_3^A
I_2^B	I_2^A	I_3^A	I_1^A
I_3^B	I_1^A	I_2^A	I_3^A

compared to the conventional person re-identification methods which just implement standard individual image matching for each person image. Applying SMA in person re-identification directs us to gain a stable image matching by avoiding redundant image matches.

4.4 Experiments

To evaluate the image pairing performance of the proposed method, experiments performed with several comparison methods are reported in the following Sections.

4.4.1 Dataset

For performing a comparative study, four public datasets are used.

Viewpoint Invariant Pedestrian Recognition (VIPeR) dataset [48] is widely used in Person Re-identification. It provides two outdoor camera views under several viewpoints and lighting conditions which consists of 632 person images. Each person has one image per camera and the images are scaled to 128 x 48 pixels. It provides the pose angle of each person as 0° (front), 45° , 90° (right), 135° , and 180° (back).

CUHK01 dataset [49] consists of 3,884 manually cropped person images for 971 persons. This dataset also provides a variety of outdoor camera views under several lighting conditions. Each of the person has one image per camera and the images are scaled to 60 x 160 pixels.

iLIDS-VID dataset [158] consists of 300 different pedestrians observed across two disjoint camera views in public open space. The dataset comprises 600 image sequences of 300 distinct persons, with one pair of image sequences from two camera views for each person with the images scaled to 64 x 128 pixels. The length of 23 to 192 frames, with an average of 73 frame.

PRID dataset [50] consists of two cameras, with 385 persons in camera view *A* and 749 persons in camera view *B*. The first 200 persons appear in both camera views, i.e., person 0001 in view *A* corresponds to person 0001 in view *B*, person 0002 in view *A* corresponds to person 0002 in view *B*, and so on. The remaining persons in each camera view (i.e., persons 0201 to 0385 in view *A* and persons 0201 to 0749 in view *B*) compose the gallery set of the corresponding view. Hence, a typical evaluation consists of searching the 200 first persons in one camera view from all persons in the other camera view.

4.4.2 Comparative methods

To gain a good comparison for pairing approaches in person re-identification, the conventional individual image pairing method is introduced as the baseline method. This individual image pairing method is a case of single-query pairing, while the proposed simultaneous image pairing method is a case of multiple-query pairing

person re-identification. The proposed method is compared with existing methods from individual image pairing and simultaneous image pairing to show the effectiveness of the multiple-query pairing compared to the single-query pairing in the person re-identification task. Though the main priority for the proposed method is the multiple-query pairing, individual image pairing with the newly proposed MSDALF feature is also compared.

Meanwhile, as to compare with the proposed pairing approaches, the state-of-the-art bipartite graph matching method, Greedy Matching (GM), Hungarian Matching (HM) [153], and Person Re-identification via Structured Matching (PRiSM) [115] are used. For these comparative methods, images are compared with HSV and MSDALF features. For the HSV feature, an $M = 16$ -bins HSV color histogram is used.

In summary, the following is a list of methods that are compared.

1. SDALF: The baseline person pairing method for single-query pairing for person re-identification.
2. MSDALF: The comparison method for single-query pairing for person re-identification. It utilizes the newly proposed masking method for feature extraction for person re-identification.
3. PRiSM-II [115]: The state-of-the-art method for multiple-query pairing which is an improved version of the PRiSM method [115]. It is only applied to the VIPER dataset.
4. HSV + GM: A method that uses HSV color as the image feature with Greedy Matching.
5. MSDALF + GM: A method that uses the newly proposed MSDALF as the image feature with Greedy Matching.
6. HSV + HM: A method that uses HSV color as the image feature with Hungarian Matching [153].

7. MSDALF + HM: A method that uses the newly proposed MSDALF as the image feature with Hungarian Matching [153].

4.4.3 Image pairing settings

To make the comparison study relevant with the existing methods, here, the experimental setting by Farenzena et al. [53] using the VIPeR dataset images is used as the general evaluation setting, except that here, matching rate in [53] is called pairing rate between all images in the dataset. In Farenzena et al.'s work, an experiment is repeated ten times with a random image selection, and the results are averaged as the final result in the proposed two evaluations, comparison on pairing accuracy and different number of images. Since SMA outputs only the paired result, only the Rank-1 image pairing score is evaluated. For evaluating the performance of the proposed method, the Rank-1 rate is calculated from the number of successful pairing over the number of persons.

Although the analysis is performed only on the VIPER dataset, here, the dataset is extended by using three additional public datasets introduced in Section 4.4.1. Concretely, 316, 485, 150, and 200 person images are randomly selected from VIPeR, CUHK01, iLIDS-VID, and PRID datasets, respectively, for the first evaluation purposes.

In the proposed second evaluation, analysis on the comparison of the different numbers of images, to confirm the reliability of the multiple query pairing approach for person re-identification, here, the experimental setting by Zhang et al. [115] is used, which focusses on a case where the numbers of images from two cameras are not the same. This setting is crucial as the actual scenario of the captured images from camera views is not often in the same number, e.g., in a real-life railway station and airport; therefore, this unequal image number setting is essential. They selected half (50.0%), one fourth (25.0%), and one eighth (12.5%) of the images in the query image set. Every query image will have only one matched image in the gallery set, but not all the gallery images are matched with images in the query set. To extend

the SMA to work in an unequal number setting, the condition of the pairing is restricted based on the suggestion from McVitie et al. [164]. The proposed algorithm in phase 2 in Algorithm 1 is simply extended to the male-optimal SMA to handle the unequal numbers of images. Although Zhang et al.'s work was only performed on the VIPeR dataset, here, the pairing accuracy comparison is performed also on CUHK01, iLIDS-VID, and PRID datasets.

The computational time during the identification process is a critical issue in real-world applications. Based on the Zhang et al.'s work [115], the storage and computational time are focussed on and analyzed. In this thesis, only the computational time is discussed. Since there are three types of computational times measured by Zhang et al., image descriptors $T1$, entity-matching similarities $T2$, and entity-level structured matching $T3$, here, only $T2$ is considered as $T1$ and $T3$ processes are not implemented in the proposed work. The pairing concept proposed in this thesis is similar to Zhang et al.'s concept, entity-matching similarities. Since implementations of Zhang et al.'s work are not completely reported; only the VIPeR dataset's matching performance results are referred from their papers. For other comparative methods' computational time, the provided codes from their papers are directly used without any control to extract only the image feature for MSDALF before the pairing process is conducted for the additional three CUHK01, iLIDS-VID, and PRID datasets. Our experiments were all run on a multi-thread CPU (Intel Corei7-7700) 3.6 GHz with 16 GB of RAM.

4.4.4 Results

4.4.4.1 Comparison on average pairing accuracy

The results for the comparison on pairing accuracy are summarized in Table 4.2 in two categories; single-query pairing and multiple-query pairing. The proposed method with a new masking, namely MSDALF, outperformed the original SDALF for the VIPeR dataset in the single query pairing comparison. This improvement

Table 4.2: Comparison of pairing methods in matching accuracy on VIPeR[48], CUHK01[49], iLIDS-VID[158], and PRID[50] datasets.

Pairing Method		Dataset			
		VIPeR [48]	CUHK01 [49]	iLIDS-VID [158]	PRID [50]
Single Query	SDALF [53]	19.87 %	—	—	—
	MSDALF	20.47 %	22.80 %	14.30 %	4.50 %
Multiple Query	PRiSM-II [115]	36.71 %	50.10 %	20.00 %	—
	HSV + GM	5.70 %	1.61 %	0.67 %	2.50 %
	MSDALF + GM	14.53 %	4.79 %	7.00 %	6.00 %
	HSV + HM	34.21 %	13.24 %	13.33 %	4.50 %
	MSDALF + HM	39.18 %	16.39 %	34.40 %	16.50 %
	Proposed : HSV + SMA	40.44 %	<u>27.03 %</u>	17.93 %	4.00 %
	Proposed : MSDALF + SMA	<u>40.32 %</u>	21.73 %	<u>34.33 %</u>	<u>13.00 %</u>

from the original image masking in Farenza’s work [53]; how to separate the foreground and the background of the images before extracting the image feature, proved that manipulating the original mask in SDALF can improve the pairing accuracy rate.

In the multiple query pairing comparison, the proposed method with HSV feature outperforms other comparative methods on the VIPeR dataset. The proposed method with MSDALF feature ranked the second. Both of these proposed methods are better than PRiSM-II by more than 3%. In the case of the CUHK01 dataset, PRISM-II showed significantly better results than the other methods. This can be because its images have many occlusions with other pedestrians compared to the other datasets. Furthermore, since the CUHK01 dataset images are captured at an outdoor scene, it consists of a large variety of lighting conditions and changes of the illumination influence in the captured images. Thus, image features may not have been accurately extracted in the proposed method. For the other two datasets, iLIDS-VID and PRID, the proposed method with MSDALF ranked the second place with 0.07% and 3.50% matching accuracy behind MSDALF + HM. The proposed method performed comparably with other comparative methods through overall comparison using all the

Table 4.3: Comparison of matching accuracy with different numbers of images in two camera views settings on the VIPeR dataset [48]

Pairing Method	VIPeR		
	158 queries	79 queries	40 queries
PRiSM-II [115]	35.90 %	36.70 %	34.50 %
HSV + GM	6.39 %	6.33 %	6.50 %
MSDALF + GM	15.00 %	15.06 %	14.75 %
HSV + HM	34.56 %	33.04 %	33.00 %
MSDALF + HM	39.49 %	40.00 %	39.75 %
Proposed : HSV + SMA	<u>40.25 %</u>	<u>40.20 %</u>	<u>39.25 %</u>
Proposed : MSDALF + SMA	41.01 %	40.38 %	39.75 %

Table 4.4: Comparison of matching accuracy with different numbers of images in two camera views settings on the CUHK01 dataset [49]

Pairing Method	CUHK01		
	243 queries	121 queries	61 queries
PRiSM-II [115]	—	—	—
HSV + GM	1.56 %	0.82 %	1.74 %
MSDALF + GM	4.44 %	4.30 %	3.77 %
HSV + HM	13.46 %	13.80 %	13.93 %
MSDALF + HM	16.50 %	16.53 %	16.07 %
Proposed : HSV + SMA	26.46 %	28.35 %	29.18 %
Proposed : MSDALF + SMA	<u>22.63 %</u>	<u>22.40 %</u>	<u>22.30 %</u>

datasets.

4.4.4.2 Comparison on different numbers of images

In this comparison study, the non-similar images for query and gallery images were set and analyzed. The analysis was divided into three subcategories based on the half, quarter, and one-eighth of the total images, which is similar to the evaluation setting utilized by Farenzena et al [53]. With the use of four datasets, VIPeR, CUHK01, iLIDS-VID, and PRID, the proposed methods are compared with the comparative methods accordingly. Comparison of the Rank-1 matching results were used and

Table 4.5: Comparison of matching accuracy with different numbers of images in two camera views settings on the iLIDS-VID dataset [158]

Pairing Method	iLIDS-VID		
	75 queries	38 queries	19 queries
PRiSM-II [115]	—	—	—
HSV + GM	0.80 %	0.26 %	0.00 %
MSDALF + GM	7.33 %	7.89 %	8.42 %
HSV + HM	13.20 %	13.95 %	14.21 %
MSDALF + HM	35.07 %	36.05 %	40.53 %
Proposed : HSV + SMA	<u>18.27 %</u>	<u>18.42 %</u>	19.47 %
Proposed : MSDALF + SMA	35.07 %	36.05 %	<u>38.42 %</u>

Table 4.6: Comparison of matching accuracy with different numbers of images in two camera views settings on the PRID dataset [50]

Pairing Method	PRID		
	100 queries	50 queries	25 queries
PRiSM-II [115]	—	—	—
HSV + GM	2.00 %	2.80 %	2.80 %
MSDALF + GM	5.80 %	7.60 %	6.80 %
HSV + HM	4.70 %	5.20 %	4.80 %
MSDALF + HM	16.00 %	17.20 %	15.60 %
Proposed : HSV + SMA	4.30 %	4.80 %	6.40 %
Proposed : MSDALF + SMA	<u>13.30 %</u>	<u>15.60 %</u>	<u>14.40 %</u>

tabulated in Tables 4.3 and 4.4 for the VIPeR and CUHK01 datasets. Here, the results show the robustness to the different numbers of query images of the proposed method compared to the other methods. Tables 4.5 and 4.6 illustrate the same comparison on the iLIDS-VID and PRID datasets. The proposed method outperforms the other comparative methods for the VIPeR, CUHK01, and iLIDS-VID datasets.

The proposed method with HSV feature (HSV + SMA) ranked the first for the CUHK01 dataset, and that with Mask-improved SDALF (MSDALF + SMA) also ranked the first for the VIPeR and the iLIDS-VID datasets. For the PRID dataset, the proposed method ranked the second after MSDALF + HM. In total, the proposed

method performs well in the case where the numbers of images are different.

4.5 Discussion

This section discusses more details about the experimental results presented in the previous Section using additional supporting information from the conducted preliminary experimental results and post-experimental results. The comparisons of results between the proposed and comparative methods with several discussions on the relative relation between the pairing approaches are conducted using dataset image characteristics with several elements. Discussing these two results from Sections 4.4.4.1 and 4.4.4.2, are essential, especially when utilizing the feature extraction for measuring the image similarity for person re-identification. The performance of the proposed method is analyzed quantitatively and qualitatively to validate the proposed method’s effectiveness.

4.5.1 Quantitative evaluation

To quantitatively evaluate the performance of the proposed method, the computational time for all datasets is evaluated and analyzed. Here, the computational time is the time dedicated to the pairing, but not the entire process for person re-identification, following the evaluation in Zhang et al.’s work [115].

Table 4.7: Comparison of computation time

Pairing Method	Dataset			
	VIPeR [48]	CUHK01 [49]	iLIDS-VID [158]	PRID [50]
PRiSM-II [115]	1.500 s	—	—	—
MSDALF + GM	0.002 s	0.004 s	0.001 s	0.001 s
MSDALF + HM	65.738 s	510.259 s	7.117 s	26.762 s
Proposed : MSDALF + SMA	0.070 s	1.095 s	0.018 s	0.066 s

The second quantitative evaluation is analyzed using the VIPeR dataset to show the importance of patch selection. An example of using the four types of image patches from one image is discussed based on the matching accuracy among all the images in the VIPeR dataset.

4.5.1.1 Comparison of computational time

The computational time during testing is one of the essential elements in the real implementation of any system other than storage and transmission rate. Here, the performance of each method is compared excluding feature extraction and distance calculation, and by only image matching. The number of images is set to be 316 in both camera views, which is the same setting as in Zhang et al.'s work [115]. In addition to the original work done by Zhang et al., the analysis is expanded by randomly selecting 485, 150, and 200 person images from the other CUHK01, iLIDS-VID, and PRID datasets respectively. The pairing methods, PRiSM-II, MSDALF + GM, MSDALF + HM and MSDALF + SMA, are analyzed with these number of images.

The results are shown in Table 4.7. The computational complexity of the proposed re-identification with SMA is considered to be $O(n^2)$, when focussing only on the pairing process. As discussed in Algorithm 1, the two loops are required in pairing the male and female based on the preferences. Zhang et al. performed the computation time study in three-phase processes: image description, entity-matching similarity, and entity-level structured matching. Here, only the entity-matching similarity is evaluated as the other two are out of the research scope of this thesis.

The computation time or running time is defined as how long the time it takes for a computer to perform a specific task. This computational time may be misunderstood as the typical computational complexity, but the two are different and defined as a Big O, e.g., $O(n)$, $O(N \log N)$, or $O(n^2)$. Since the existing analysis was conducted in person re-identification via computation time in Zhang et al.'s work, here, the computational time between proposed methods and other comparative methods are compared and analyzed. Here, the second computation time in Zhang et al.'s

work is suitable to be compared with the proposed method as their entity-matching similarities $T2$ could reflect the proposed pairing approach's computational time. If the computation complexity perspective is used to discuss the comparative methods' computation time, GM and HM are $O(n)$ and $O(n^3)$, respectively, HM is expected to require more computation time than GM.

Based on Table 4.2, the proposed method, HSV + SMA, perform at the first rank with the VIPeR dataset. With the other datasets, CUHK01, iLIDS-VID, and PRID, the proposed methods HSV + SMA and MSDALF + SMA are at the second rank and showed comparable results to other comparative methods. The proposed multiple-query pairing with SMA presented an additional advantage for simultaneous image pairing with various datasets. From Table 4.7, the proposed method using MSDALF + SMA showed better performance than that of other comparative methods. Although MSDALF + GM requires a shorter computation time in Table 4.7, fundamentally, in GM itself, once one of the instances is taken away during the matching, this instance can not be utilized anymore after that iteration is completed. Here, if one instance in the query is matched with one instance in the gallery, the instance in the gallery will be deleted from the preference list for the next query instance's iteration. With this deletion of this instance after one matching instance pair, MSDALF + GM is prospected to complete the pairing process earlier than the other methods, but with less pairing performance in Table 4.2.

Furthermore, the person re-identification task can be solved not only with color features but also with other features such as a geometrical and local features. Using these, MSDALF could yield better performance in matching. Although the proposed SMA-based pairing method costs $O(n^2)$, it is still comparable with PRiSM-II and HM in computation time. Here, we proved that the image pairing approach plays a vital role in achieving an excellent matching performance for person re-identification.

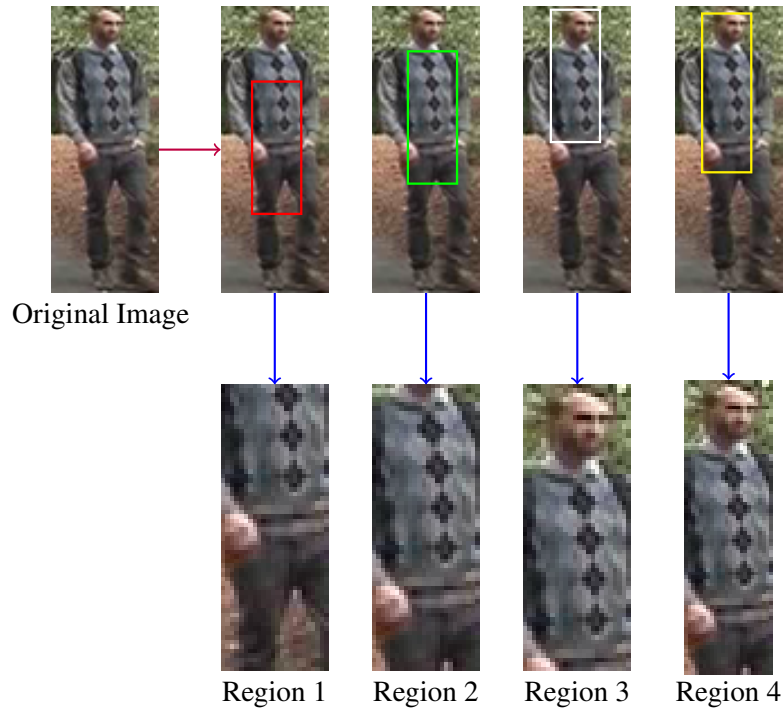


Figure 4.4: Example of image cropping from four different regions of a person.

4.5.1.2 Comparison of the selection of image patch

Next, the effect of the image patch by cropping the images is studied with the HSV color feature. Four regions from the original image are cropped as illustrated in Figure 4.4. In this preliminary experiment, HSV + SMA is used to compare all image regions. In the VIPeR dataset, images are mostly captured with noise in the background. Using one image from the VIPeR dataset as illustrated in Figure 4.4, we can see that green trees and the black background may negatively influence the image.

For this, a discussion about the most crucial region of an image to be compared in the identification context is required. The SDALF work by Farenzena et al. [53] showed that the division of an image could help the image pairing in the re-identification task. Realizing that an image has a versatile pixel characteristic in general, in this comparison, based on Figure 4.4, images are cropped as middle, middle-upper, middle-top one, and middle-top two regions. The middle region of an image delivers the most distinct characteristic of every instance since all the images in four datasets used in

Table 4.8: Comparison on Different Cropping Regions using SMA + HSV method

Cropped Region	Matching Rate VIPeR [48]
Original Image	40.44 %
Region 1	45.57 %
Region 2	46.33 %
Region 3	44.72 %
Region 4	45.52 %

this thesis were centered to a person. Nevertheless, the proposed SDALF selection could face difficulties if the image contains a vast range of illumination or scale. In Figure 4.3, since region 2 avoided this affected area and produced a good matching rate, we can conclude that by cropping images, especially with Region 2, we can achieve the best pairing rate compared to other comparative image regions in Table 4.8. This comparison concludes that the image patch or separating images into regions is one way to gain a better matching.

4.5.2 Qualitative evaluation

The performance of the proposed method is evaluated qualitatively by visualizing the image masks in example images from the four datasets. This evaluation is essential for quality checking the gained result in the previous Section's quantitative evaluation, especially during image feature extraction. The analysis on the visualization of the mask image could tell and describe the effect of the dataset images on the masked image with the newly proposed MSDALF compared to the traditional SDALF which has been discussed in Figure 4.3.

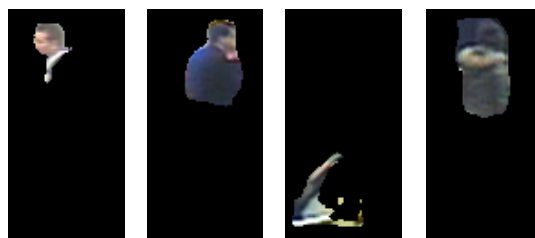
In general, SMA is quite promising in terms of the matching performance since it allows the image to be re-matched again; SMA allows the matched image pair to be



(a) VIPeR dataset [2].



(b) CUHK01 dataset [49].



(c) iLIDS-VID dataset [50].



(d) PRID dataset [51].

Figure 4.5: Examples of unsuccessfully masked images (using Mask R-CNN) for the four datasets used in the evaluation.

re-matched if the current similarity rank for one image is lower than that for an existing matched image pair. However, although the proposed method with MSDALF is expected to be the best among all the comparison methods, it ranked the second.

Figure 4.5 illustrates that Mask R-CNN could struggle in providing a good masked

4.5.3 Limitation and further consideration

Regarding the multiple query pairing approach for person re-identification, which takes the simultaneous image pairing strategy compared to the conventional individual image pairing strategy, there are some limitations and considerations for its development to be taken in the future.

- The proposed method is designed for multiple-query person re-identification task with SMA based on the Gale-Shapley (GS) algorithm [120]. The men-optimally or women-optimally SMA utilized in the proposed work could cater for the existing dataset images. However, if the dataset is extended to a larger size, it could become difficult to pair the images. Therefore, it is recommended to implement an extension of SMA, proposed by Gusfield and Irwing [165], with minimum regret and egalitarian element when considering the pairing. The three elements that could be reconsidered in bringing these ideas into person re-identification are, minimization of the sex-equality cost, minimization of the regret cost, and minimization of the egalitarian cost. However, the computation time could increase by one and a half or twice from the original SMA.
- A simple image feature was used for the similarity ranking of the multiple query pairing discussed in Section 4.3.3. However, expanding the similarity parameter, with the use of more complex image features, for analyzing the image ranking as the preference could infer a more reliable image feature extraction process to gain a higher matching rate in the person re-identification task. There is still room for further improvement of the employed image feature in the proposed method with a recent deep learning methodology such as the regression approach.
- Utilizing public dataset images considering as off-line study, is one of the limitations. The implementation of real-time or on-line person re-identification is needed to be considered as one of the future work. Comparing the effectiveness on publicly available images and real-time captured images is also needed to be compared to show the proposed method's effectiveness.

- Since person re-identification considers a person as the object for identification, a wider range of objects should be considered to ensure the versatility of the proposed method in various settings.

4.6 Summary

This chapter provided a solution for person re-identification by pairing the person images from given two viewpoints using the proposed multiple query pairing via SMA. The proposed method presents a solution to the person re-identification problem mentioned in Chapter 1. The two viewpoints scenario for obtaining the best and reliable pairs in identification discussed in this Chapter can be considered to be expanded into more than two viewpoints. Considering the similarity rank to arrange the ranking between query image to all available gallery images was conducted carefully by selecting the best pair via the SMA, especially with an exploration to the non-similar image number setting condition. The multiple-query pairing approach showed its effectiveness in the simultaneous person identification task as one of the object identification problems for this thesis. Several experiments were carried out to evaluate the quantitative and qualitative performances of the person re-identification method. As a result, the proposed method demonstrated to successfully pair the persons between query and gallery sets simultaneously compared to the traditional person re-identification.

Chapter 5

Conclusion

This chapter concludes this doctoral thesis. In Section 5.1, the discussion of proposed methods and their contributions are summarized. Section 5.2 discusses remaining challenges in this field of research and potential directions for future research. Both these discussions are towards extending the proposed methods and applications of the pairing concept in object identification. Lastly, Section 5.3 completes the thesis with closing remarks.

5.1 Summary

Overall, the research described in this thesis interprets object identification with the newly proposed concept in object identification, namely pairing, with two Research topics in the Instance-to-Instance pairing for the object identification. The pairing concept applied to object identification is defined as obtaining the best pairs for given input images for various tasks. Applying this pairing concept, mainly for Instance-to-Instance pairing, this thesis aimed to prove that the proposed concept could be considered as a new approach worth exploring. As such, an analysis of the pairing approaches would yield knowledge on considering the instance as multiple choices of objects, not limited to a person, object, viewpoint, but even more. To discuss

the proposed pairing concept, it was applied to the Research topic of object pose estimation and person re-identification, related to the sub-category of Instance-to-Instance pairing.

Chapter 1 discussed the motivation of this doctoral research and gave an overview on the background of the object identification task and their categorization: Single-query pairing and multiple-query pairing. Two Research topics were introduced to study the performance of the pairing approach applied to different applications with different target objects.

Chapter 2 reviewed existing work in the discussed two computer vision applications; object pose estimation and person re-identification, thoroughly, by giving a comprehensive analysis of the state-of-the-art on these Research topics.

Two actual computer vision tasks were discussed as Research topics 1 and 2 in detail in the following chapters:

1. In Chapter 3, Research topic 1 considered the ambiguousness of an object pose estimation. In order to disambiguate by shifting the viewpoint to a better one, the problem of selecting the next viewpoint rises. The proposed method provides a solution to this problem by selecting the best next-viewpoint via pose ambiguity minimization. Experiments with various settings showed that the proposed method produces a good estimation for the selection of two viewpoints observation in comparison to comparative methods. With the use of the proposed ambiguity scale, the possible next viewpoint from the initial viewpoint was analyzed, and accordingly, a next viewpoint that is not influenced by the ambiguity was proposed. The proposed method showed superiority to the comparative methods in Mean Squared Error (MSE) for four object categories, “Airplane”, “Car”, “Mug”, and “Toilet” and the best next-viewpoint selection method for the “Mug” object. These results showed the effectiveness of the proposed method using a versatile object.

2. In Chapter 4, Research topic 2 considered a stable pairing of images in a person re-identification task. Mispairing of image pairs could affect the overall pairing accuracy in person re-identification. A solution to this problem was proposed by simultaneous person pairing via SMA considering the person image as an instance of the identification task. By changing several experimental settings for the numbers of query and gallery dataset images; equal image number and non-similar image number, the proposed method showed its effectiveness compared to other methods; single-query pairing and multiple-query pairing approaches.

Further, the two proposed methods introduced in this thesis including the single-query pairing and the multiple-query pairing, have proved to improve the performance of the conventional single viewpoint pose estimation and the single person pairing re-identification tasks. In addition, these two proposed methods added important information and a reliable pairing approach to the yielded identification, i.e., the Instance-to-Instance pairing approach, which is helpful for the current and future improvement in existing computer vision applications such as indoor human helper robots and public surveillance. Meanwhile, both proposed methods showed their benefits and some drawbacks in the actual implementation.

For Research topic 1, at least two-viewpoints observation could obtain more information about the object characteristic compared to the single viewpoint observation. With the new idea of having an extra viewpoint, single viewpoint object pose estimation was upgraded to perform multiple viewpoints pose estimation considering the pairing approach in the identification can be easily extended to multiple rotational angles of observation instead of the fixed angular viewpoint. However, this generally applicable method has a few shortcomings in implementing in an actual situation, and hence might produce error due to the target object's surrounding or from occluding objects.

Research topic 2 targeted a higher person pairing accuracy. This method is proven to effectively and efficiently increase the accuracy of the person pairing from two

viewpoints. Therefore, the multiple-query pairing approach via SMA, utilizing the similarity rank as the image preferences, successfully sustains the performance on pairing the person in various situations even when a rigid setting is applied.

5.2 Remaining Challenges and Future Directions

Some general research directions for the improvement of the pairing approach are discussed, and some remarks on future steps for the individual research topics are given. In the end, opportunities for applications built using the proposed metrics are briefly discussed.

Pairing Approach: Applying the pairing concept to other computer vision applications than the object pose estimation and person re-identification could be the next step. Besides, the optimization issue in the pairing, related to when we select the method for pairing the instances, could be analyzed with a different strategy in each task. Not to forget about the actual implementation, investigation of the overall pairing concept could also be considered in the future research plan to ensure that the related online or real-time noise correlation with the pairing accuracy does not significantly influence the pairing performance.

Research Topic 1: In this Research topic, five categories of synthetic target objects were chosen as distinctive objects, rendered and developed from ImageNet [118] for the category-type problem setting. The objects were selected considering an indoor scene in order to help impaired people living by themselves at home. However, the selection could be extended to a broader range of objects such as outdoor objects or factory manufactured objects. Analysis on different experiment settings; sunlight, variation of sizes, and distance compared to the current settings could be of interest. With a broader range of objects, the viewpoints are expected to be from multiple

angles, not relying on one fixed z-axis angle. By expanding this z-axis angular selection in capturing the new dataset images, the optimization of the training time could be the next topic to be discussed as it could consume a long time for completing a network model. For this, an improved network than the utilized one in this thesis containing a systematic layer could be a new discussion topic in the pose estimation field.

Averaging of the pose could be considered in the future, as the number of viewpoints is increased to more than two. In that case, extension of the derivation of the probability and the likelihood of viewpoints is needed for analyzing the best next-viewpoints.

The actual implementation of the proposed single query pairing approach, in a real robot for confirming the effectiveness of the pose estimation is important for human daily life use. During observation, a robot arm or camera movement for the examples, may also introduce noise to the natural object movement. Dealing with the environmental noise in the image, for instance, is also one of the challenges.

Research Topic 2: In this Research topic, the multiple-query pairing approach was applied to the identification task. Results on the public pedestrian dataset images showed that the proposed pairing approach is effective. Interestingly, although a simple image feature was used during the analysis, it yielded good overall pairing results. Accordingly, it would be interesting to consider a new image feature or design a combination of image feature combination, similar to Farenzena et al.'s work [53].

In the past 50 years, especially for the instance pairing via the Stable Marriage Algorithm (SMA) [120], several works have improved the marriage problem. The extended works to Gale and Shapley's by McVitie and Wilson [164] provided a comprehensive conceptual coefficient option from the original SMA for the identification task, which improves the pairing approach. A new problem setting for person re-identification's evaluation is anticipated to be discussed with the utilization of a

specific or extra condition, including the male optimum stable solution, female optimum stable solution, and minimum choice stable solution. One way to consider the limitation in the original work from Gale and Shapley would be looking into the pairing condition from men to women or vice versa with a more detailed constraint.

Another direction for future research is using multiple camera views, and running simultaneously in one re-identification scenario. Similarly, the simultaneous use of various cameras could enhance the research in pairing as it reflects the actual implementation situation. At the same time, the actual pairing should be performed in real-time and embedded in a reliable and economical hardware setting.

5.3 Closing Remarks

To conclude the thesis, the following are general findings that can be conveyed based on the empirical studies that have been presented in the previous chapter. A new idea of looking at the single-query pairing approach via object pose estimation was proposed in Chapter 3 as Research topic 1. The proposed multiple viewpoint pose estimation idea solved the lack of a single viewpoint in estimating an object pose accurately. The ambiguity scale managed to suggest the best next-viewpoint for disambiguating the pose for object pose. In Chapter 4, as Research topic 2, the core assumption that the image pairing approach for available images could improve the current image pairing even with simple image features was verified via the multiple-query pairing approach. Even though most current state-of-the-art methodologies increasingly focusses on deep-learning models, the outcomes of Research topic 2 showed that a traditional approach can provide a comparative image pairing solution for an identification task.

The studies that were presented in this thesis have been introduced through the new pairing concept. Outcomes of this thesis provide solutions to different directions of identification tasks to reach the ultimate goals expressed at the beginning of this thesis. Although this doctoral research may only be the first step in fully understanding

and applying the pairing approach, hopefully, this thesis could contribute to the advancement of science and knowledge in the fields of Computer Vision and Machine Learning.

Life is short, make it sweet.

Bibliography

- [1] Emmanuel Sirimal Silva and Francesca Bonetti. Digital humans in fashion: Will consumers interact? *Journal of Retailing and Consumer Services*, 60 (102430):1–11, 2021.
- [2] Thomas Southcliffe Ashton. *The Industrial Revolution 1760–1830*. Oxford University Press, Oxford, UK, New York, NY, 1997.
- [3] Linda G. Shapiro and George C. Stockman. *Computer Vision*. Prentice Hall, Hoboken, NJ, 2001.
- [4] Victor Wiley and Thomas Lucas. Computer vision and image processing: A paper review. *Int. Journal of Artificial Intelligence Research*, 2(1):29–36, 2018.
- [5] Xin Li and Yiliang Shi. Computer vision imaging based on artificial intelligence. In *Proc. 2018 Int. Conf. on Virtual Reality and Intelligent Systems*, pages 22–25, 2018.
- [6] Matevž Kunaver and Jurij F. Tasič. Image feature extraction –An overview. In *Proc. Int. Conf. on “Computer as a Tool” 2005*, volume 1, pages 183–186, 2005.
- [7] Jun Sonoda and Tomoyuki Kimoto. Object identification from GPR images by deep learning. In *Proc. 2018 Asia-Pacific Microwave Conf.*, pages 1298–1300, 2018.

- [8] Rita Cucchiara, Costantino Grana, Andrea Prati, and Roberto Vezzani. Computer vision system for in-house video surveillance. *IEE Procs. - Vision, Image and Signal Processing*, 152(2):242–249, 2005.
- [9] J. Harikrishnan, Arya Sudarsan, Aravind Sadashiv, and Remya A. S. Ajai. Vision-face recognition attendance monitoring system for surveillance using deep learning technology and computer vision. In *Proc. 2019 Int. Conf. on Vision Towards Emerging Trends in Communication and Networking*, pages 222–226, 2019.
- [10] Markus Weber. Frontal face dataset. *California Inst. of Technology*, <http://www.vision.caltech.edu/html-files/archive.html>, 1999. Accessed on December 2, 2021.
- [11] JoonOh Seo, SangUk Han, SangHyun Lee, and Hyoungkwan Kim. Computer vision techniques for construction safety and health monitoring. *Advanced Engineering Informatics*, 29(2):239–251, 2015.
- [12] Mark S. Nixon and Alberto S. Aguado. *Feature Extraction and Image Processing for Computer Vision*. Academic Press, Cambridge, MA, 2019.
- [13] Theo Gevers, Arjan Gijsenij, Joost Van de Weijer, and Jan-Mark Geusebroek. *Color in Computer Vision: Fundamentals and Applications*. John Wiley & Sons, Hoboken, NJ, 2012.
- [14] Andrew Senior, Sharath Pankanti, Arun Hampapur, Lisa Brown, Ying-Li Tian, Ahmet Ekin, Jonathan Connell, Chiao Fe Shu, and Max Lu. Enabling video privacy through computer vision. *IEEE Security and Privacy*, 3(3):50–57, 2005.
- [15] Liang Zheng, Yi Yang, and Alexander G. Hauptmann. Person re-identification: Past, present and future. *Computer Research Repository arXiv Preprint*, arXiv:1610.02984, 2016.
- [16] International Federation of Robotics. Service robots record: Sales worldwide up 32%, 2020. URL <https://ifr.org/ifr-press-releases/news/>

- service-robots-record-sales-worldwide-up-32. Accessed on January 26, 2022.
- [17] Feng Lu and Evangelos Milios. Robot pose estimation in unknown environments by matching 2D range scans. *Journal of Intelligent and Robotic Systems*, 18(3):249–275, 1997.
- [18] Robert Sim and Gregory Dudek. Learning and evaluating visual features for pose estimation. In *Proc. 7th IEEE Int. Conf. on Computer Vision*, volume 2, pages 1217–1222, 1999.
- [19] Michael Warren, David McKinnon, Hu He, and Ben Uprocft. Unaided stereo vision based pose estimation. In *Proc. 2010 Australasian Conf. on Robotics and Automation*, pages 354–361, 2010.
- [20] Alvaro Collet and Siddhartha S. Srinivasa. Efficient multi-view object recognition and full pose estimation. In *Proc. 2010 IEEE Int. Conf. on Robotics and Automation*, pages 2050–2055, 2010.
- [21] Takeshi Shakunaga. An object pose estimation system using a single camera. In *Proc. 1992 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, pages 1053–1060, 1992.
- [22] Vittorio Murino and Gian Luca Foresti. 2D into 3D Hough-space mapping for planar object pose estimation. *Image and Vision Computing*, 15(6):435–444, 1997.
- [23] Hiroki Tatemichi, Yasutomo Kawanishi, Daisuke Deguchi, Ichiro Ide, Ayako Amma, and Hiroshi Murase. Median-shape representation learning for category-level object pose estimation in cluttered environments. In *Proc. 25th Int. Conf. on Pattern Recognition*, pages 4473–4480, 2021.
- [24] Stephen Gould, Paul Baumstarck, Morgan Quigley, Andrew Y. Ng, and Daphne Koller. Integrating visual and range data for robotic object detection. In *Proc. Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications*, pages 1–12, 2008.

- [25] Feng Wang, Di Guo, Huaping Liu, Junfeng Zhou, and Fuchun Sun. Sound-indicated visual object detection for robotic exploration. In *Proc. 2019 Int. Conf. on Robotics and Automation*, pages 8070–8076, 2019.
- [26] Dirk Holz, Angeliki Topalidou-Kyniazopoulou, Jörg Stückler, and Sven Behnke. Real-time object detection, localization and verification for fast robotic depalletizing. In *Proc. 2015 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, pages 1459–1466, 2015.
- [27] Deepali Javale, Mohd Mohsin, Shreerang Nandanwar, and Mayur Shingate. Home automation and security system using Android ADK. *Int. Journal of Electronics Communication and Computer Technology*, 3(2):382–385, 2013.
- [28] Al-Akhir Nayan, Joyeta Saha, Jannatul Ferdaous, and Muhammad Golam Kibria. IoT based smart kitchen security system. *Applied Intelligence for Industry 4.0*, pages 354–367, 2019.
- [29] Natarajan Sundaram, Cherian Thomas, and Loganathan Agilandeewari. A review: Customers online security on usage of banking technologies in smartphones and computers. *Pertanika Journal of Science and Technology*, 27(1): 1–31, 2019.
- [30] Esteban Vázquez-Fernández and Daniel González-Jiménez. Face recognition for authentication on mobile devices. *Image and Vision Computing*, 55:31–33, 2016.
- [31] Andrew K. Hrechak and James A. McHugh. Automated fingerprint recognition using structural matching. *Pattern Recognition*, 23(8):893–904, 1990.
- [32] Marius Tico, Eero Immonen, Pauli Ramo, Pauli Kuosmanen, and Jukka Saari-nen. Fingerprint recognition using wavelet features. In *Proc. 2001 IEEE Int. Symposium on Circuits and Systems*, volume 2, pages 21–24, 2001.
- [33] Haiyun Xu, Raymond N.J. Veldhuis, Tom A.M. Kevenaar, and Ton A.H.M. Akkermans. A fast minutiae-based fingerprint recognition system. *IEEE Systems Journal*, 3(4):418–427, 2009.

- [34] Joshi Ravi, K.B. Raja, and K.R. Venugopal. Fingerprint recognition using minutiae score matching. *Computing Research Repository arXiv Preprint*, arXiv:1001.4186, 2010.
- [35] Dongjae Lee, Kyoungtaek Choi, Heeseung Choi, and Jaihie Kim. Recognizable-image selection for fingerprint recognition with a mobile-device camera. *IEEE Trans. on Systems, Man, and Cybernetics*, 38(1):233–243, 2008.
- [36] Claudia Nickel, Holger Brandt, and Christoph Busch. Classification of acceleration data for biometric gait recognition on mobile devices. In *Proc. 10th Int. Conf. of the Biometrics Special Interest Group*, pages 57–66, 2011.
- [37] Fanfeng Zeng, Shengda Hu, and Ke Xiao. Research on partial fingerprint recognition algorithm based on deep learning. *Neural Computing and Applications*, 31(9):4789–4798, 2019.
- [38] Pierre Baldi and Yves Chauvin. Neural networks for fingerprint recognition. *Neural Computation*, 5(3):402–418, 1993.
- [39] Wenyi Zhao, Rama Chellappa, P. Jonathon Phillips, and Azriel Rosenfeld. Face recognition: A literature survey. *ACM Computing Surveys*, 35(4):399–458, 2003.
- [40] Lie Gu, Stan Z. Li, and Hong-Jiang Zhang. Learning probabilistic distribution model for multi-view face detection. In *Proc. 2001 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, volume 2, pages 116–122, 2001.
- [41] Bernd Heiselet, Thomas Serre, Massimiliano Pontil, and Tomaso Poggio. Component-based face detection. In *Proc. 2001 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, volume 1, pages 657–662, 2001.
- [42] Henry Schneiderman and Takeo Kanade. Probabilistic modeling of local appearance and spatial relationships for object recognition. In *Proc. 1998 IEEE*

- Computer Society Conf. on Computer Vision and Pattern Recognition*, pages 45–51, 1998.
- [43] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *Proc. 2001 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, volume 1, pages I–511–I–518, 2001.
- [44] Kwontaeg Choi, Kar-Ann Toh, and Hyeran Byun. Realtime training on mobile devices for face recognition applications. *Pattern Recognition*, 44(2):386–400, 2011.
- [45] Emir Kremic and Abdulhamit Subasi. Performance of random forest and svm in face recognition. *Int. Arab Journal of Information Technology*, 13(2):287–293, 2016.
- [46] Weihong Wang, Jie Yang, Jianwei Xiao, Sheng Li, and Dixin Zhou. Face recognition based on deep learning. In *Proc. 1st Int. Conf. on Human Centered Computing*, pages 812–820, 2014.
- [47] Yi Sun, Ding Liang, Xiaogang Wang, and Xiaoou Tang. DeepID3: Face recognition with very deep neural networks. *Computing Research Repository arXiv Preprint*, arXiv:1502.00873, 2015.
- [48] Douglas Gray, Shane Brennan, and Hai Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *Proc. 10th IEEE Workshop on Performance Evaluation for Tracking and Surveillance*, volume 3, pages 41–48, 2007.
- [49] Wei Li, Rui Zhao, and Xiaogang Wang. Human reidentification with transferred metric learning. In *Proc. 11th Asian Conf. on Computer Vision*, volume 1, pages 31–44, 2012.
- [50] Martin Hirzer, Csaba Beleznai, Peter M. Roth, and Horst Bischof. Person reidentification by descriptive and discriminative classification. In *Proc. 17th Scandinavian Conf. on Image Analysis*, pages 91–102, 2011.

- [51] Timothy Huang and Stuart Russell. Object identification in a Bayesian context. In *Proc. 15th Int. Joint Conf. on Artificial Intelligence*, pages 1276–1282, 1997.
- [52] Wojciech Zajdel, Zoran Zivkovic, and Ben J.A. Krose. Keeping track of humans: Have I seen this person before? In *Proc. 2005 IEEE Int. Conf. on Robotics and Automation*, pages 2081–2086, 2005.
- [53] Michela Farenzena, Loris Bazzani, Alessandro Perina, Vittorio Murino, and Marco Cristani. Person re-identification by symmetry-driven accumulation of local features. In *Proc. 2010 IEEE Conf. on Computer Vision and Pattern Recognition*, pages 2360–2367, 2010.
- [54] Lisa Gye. Picture this: The impact of mobile camera phones on personal photographic practices. *Continuum*, 21(2):279–288, 2007.
- [55] Tom Yeh, Kristen Grauman, Konrad Tollmar, and Trevor Darrell. A picture is worth a thousand keywords: image-based object search on a mobile platform. In *CHI'05 Extended Abstracts on Human Factors in Computing Systems*, pages 2025–2028, 2005.
- [56] Takuya Minagawa and Hideo Saito. Image based search system using hierarchical object category recognition technique. In *Proc. IAPR Conf. on Machine Vision and Applications 2009*, pages 219–222, 2009.
- [57] Hameed Abdul-Rashid, Juefei Yuan, Bo Li, Yijuan Lu, Song Bai, Xiang Bai, Ngoc Minh Bui, Minh N. Do, Trong Le Do, Anh Duc Duong, et al. SHREC'18 track: 2D image-based 3D scene retrieval. In *Proc 11th Eurographics Workshop on 3D Object Retrieval*, pages 37–44, 2018.
- [58] Ju-Young Kim, In-Seon Kim, Dai-Yeol Yun, Tae-Won Jung, Soon-Chul Kwon, and Kye-Dong Jung. Visual positioning system based on 6D object pose estimation using mobile Web. *Electronics*, 11(6):865, 2022.
- [59] Clark F. Olson. Probabilistic self-localization for mobile robots. *IEEE Trans. on Robotics and Automation*, 16(1):55–66, 2000.

- [60] Clemens Arth, Manfred Klopschitz, Gerhard Reitmayr, and Dieter Schmalstieg. Real-time self-localization from panoramic images on mobile devices. In *Proc. 2011 IEEE Int. Symposium on Mixed and Augmented Reality*, pages 37–46, 2011.
- [61] Douglas Gray and Hai Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *Proc. 10th European Conf. on Computer Vision*, volume 1, pages 262–275, 2008.
- [62] Omar Javed, Zeeshan Rasheed, Khurram Shafique, and Mubarak Shah. Tracking across multiple cameras with disjoint views. In *Proc. 9th IEEE Conf. on Computer Vision*, volume 2, pages 952–952, 2003.
- [63] Xiaogang Wang, Gianfranco Doretto, Thomas Sebastian, Jens Rittscher, and Peter Tu. Shape and appearance context modeling. In *Proc. 2007 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, pages 1–8, 2007.
- [64] Ziming Zhang, Yuting Chen, and Venkatesh Saligrama. A novel visual word co-occurrence model for person re-identification. In *Proc. 13th European Conf. on Computer Vision Workshop*, volume 3, pages 122–133, 2014.
- [65] Slawomir Bak, Etienne Corvee, François Bremond, and Monique Thonnat. Multiple-shot human re-identification by mean Riemannian covariance grid. In *Proc. 8th IEEE Int. Conf. on Advanced Video and Signal based Surveillance*, pages 179–184, 2011.
- [66] Martin Bäuml and Rainer Stiefelhagen. Evaluation of local features for person re-identification in image sequences. In *Proc. 8th IEEE Int. Conf. on Advanced Video and Signal-based Surveillance*, pages 291–296, 2011.
- [67] Chunxiao Liu, Shaogang Gong, Chen Change Loy, and Xinggang Lin. Person re-identification: What features are important? In *Proc. 12th European Conf. on Computer Vision Workshop and Demonstrations*, volume 1, pages 391–401, 2012.

- [68] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Unsupervised salience learning for person re-identification. In *Proc. 2013 IEEE Conf. on Computer Vision and Pattern Recognition*, pages 3586–3593, 2013.
- [69] Yinghao Cai and Matti Pietikäinen. Person re-identification based on global color context. In *Proc. 10th Asian Conf. on Computer Vision Workshop*, volume 1, pages 205–215, 2010.
- [70] M. Fatih Demirci, Ali Shokoufandeh, Yakov Keselman, Lars Bretzner, and Sven Dickinson. Object recognition as many-to-many feature matching. *Int. Journal of Computer Vision*, 69(2):203–222, 2006.
- [71] Michiel Hazewinkel. Greedy algorithm. *Encyclopedia of Mathematics*, 2001. URL http://www.encyclopediaofmath.org/index.php?title=Greedy_algorithm&oldid=34629. Accessed on January 22, 2022.
- [72] Yuan Fang, Kingsley Kuan, Jie Lin, Cheston Tan, and Vijay Chandrasekhar. Object detection meets knowledge graphs. In *Proc. 26th Int. Joint Conf. on Artificial Intelligence*, pages 1661–1667, 2017.
- [73] Yi Li, Gu Wang, Xiangyang Ji, Yu Xiang, and Dieter Fox. DeepIM: Deep iterative matching for 6D pose estimation. *Int. Journal on Computer Vision*, 128(657):683–698, 2018.
- [74] Donatello Conte, Pasquale Foggia, Carlo Sansone, and Mario Vento. Graph matching applications in pattern recognition and image processing. In *Proc. 2003 Int. Conf. on Image Processing*, volume 2, pages 21–24, 2003.
- [75] Donatello Conte, Pasquale Foggia, Carlo Sansone, and Mario Vento. Thirty years of graph matching in pattern recognition. *Int. Journal of Pattern Recognition and Artificial Intelligence*, 18(03):265–298, 2004.
- [76] Steven L. Tanimoto, Alon Itai, and Michael Rodeh. Some matching problems for bipartite graphs. *Journal of the ACM*, 25(4):517–525, 1978.

- [77] David G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. Journal of Computer Vision*, 60(2):91–110, 2004.
- [78] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. ORB: An efficient alternative to SIFT or SURF. In *Proc. 13th Int. Conf. on Computer Vision*, pages 2564–2571, 2011.
- [79] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. SURF: Speeded up robust features. In *Proc. 9th European Conf. on Computer Vision*, volume 1, pages 404–417, 2006.
- [80] Stefan Leutenegger, Margarita Chli, and Roland Y. Siegwart. BRISK: Binary robust invariant scalable keypoints. In *Proc. 13th Int. Conf. on Computer Vision*, pages 2548–2555, 2011.
- [81] Shiv Ram Dubey and Anand Singh Jalal. Robust approach for fruit and vegetable classification. *Procedia Engineering*, 38(2012):3449–3453, 2012.
- [82] Pegah Sadeghi Vasafi and Bernd Hitzmann. Comparison of various classification techniques for supervision of milk processing. *Engineering in Life Sciences*, 22(3-4):279–287, 2022.
- [83] Raymond E. Wright. *Logistic Regression*. American Psychological Association, Washington, DC, 1995.
- [84] Jiashi Feng, Huan Xu, Shie Mannor, and Shuicheng Yan. Robust logistic regression and classification. *Advances in Neural Information Processing Systems*, 27, 9 pages, 2014.
- [85] Robert Harry Riffenburgh. *Linear discriminant analysis*. PhD thesis, Virginia Polytechnic Institute, 1957.
- [86] Alan Julian Izenman. Linear discriminant analysis. In *Modern Multivariate Statistical Techniques*, pages 237–280. Springer, New York, NY, 2013.
- [87] Evgeny Byvatov and Gisbert Schneider. Support vector machine applications in bioinformatics. *Applied Bioinformatics*, 2(2):67–77, 2003.

- [88] Ashis Pradhan. Support vector machine –A survey. *Int. Journal of Emerging Technology and Advanced Engineering*, 2(8):82–85, 2012.
- [89] Soudamini Hota and Sudhir Pathak. KNN classifier based approach for multi-class sentiment analysis of Twitter data. *Int. Journal of Engineering and Technology*, 7(3):1372–1375, 2018.
- [90] Guo Haixiang, Li Yijing, Li Yanan, Liu Xiao, and Li Jinling. BPSO-Adaboost-KNN ensemble learning algorithm for multi-class imbalanced data classification. *Engineering Applications of Artificial Intelligence*, 49:176–193, 2016.
- [91] Jingnian Chen, Houkuan Huang, Shengfeng Tian, and Youli Qu. Feature selection for text classification with Naïve Bayes. *Expert Systems with Applications*, 36(3):5432–5435, 2009.
- [92] Levent Koc, Thomas A. Mazzuchi, and Shahram Sarkani. A network intrusion detection system based on a Hidden Naïve Bayes multiclass classifier. *Expert Systems with Applications*, 39(18):13492–13500, 2012.
- [93] Archana Chaudhary, Savita Kolhe, and Raj Kamal. An improved random forest classifier for multi-class classification. *Information Processing in Agriculture*, 3(4):215–222, 2016.
- [94] Juergen Gall, Nima Razavi, and Luc Van Gool. An introduction to random forests for multi-class object detection. In *Proc. 15th Int. Workshop on Theoretical Foundation of Computer Vision*, pages 243–263, 2012.
- [95] Mohammad Noor Injadat, Abdallah Moubayed, Ali Bou Nassif, and Abdallah Shami. Multi-split optimized bagging ensemble model selection for multi-class educational data mining. *Applied Intelligence*, 50(12):4506–4528, 2020.
- [96] Zillur Rahman, Md Sabir Hossain, Md Rabiul Islam, Md Mynul Hasan, and Rubaiyat Alim Hridhee. An approach for multiclass skin lesion classification based on ensemble learning. *Informatics in Medicine Unlocked*, 25(100659): 1–9, 2021.

- [97] Guobin Ou and Yi Lu Murphey. Multi-class pattern classification using neural networks. *Pattern Recognition*, 40(1):4–18, 2007.
- [98] Arpit Bhardwaj, Aruna Tiwari, Harshit Bhardwaj, and Aditi Bhardwaj. A genetically optimized neural network model for multi-class classification. *Expert Systems with Applications*, 60:211–221, 2016.
- [99] Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010. URL <http://yann.lecun.com/exdb/mnist/>. Accessed on January 22, 2022.
- [100] Ting-Fan Wu, Chih-Jen Lin, and Ruby Weng. Probability estimates for multi-class classification by pairwise coupling. *The Journal of Machine Learning Research*, 5:975–1005, 2003.
- [101] Hansheng Lei and Venu Govindaraju. Speeding up multi-class SVM evaluation by PCA and feature selection. In *Proc. Int. Workshop on Feature Selection for Data Mining 2005*, 9 pages, 2005.
- [102] Gjorgji Madzarov, Dejan Gjorgjevikj, and Ivan Chorbev. A multi-class SVM classifier utilizing binary decision tree. *Informatica*, 33(2):233–241, 2009.
- [103] Hiroshi Murase and Shree K. Nayar. Visual learning and recognition of 3-D objects from appearance. *Int. Journal of Computer Vision*, 14(1):5–24, 1995.
- [104] Andy Zeng, Kuan-Ting Yu, Shuran Song, Daniel Suo, Ed Walker, Alberto Rodriguez, and Jianxiong Xiao. Multi-view self-supervised deep learning for 6D pose estimation in the Amazon picking challenge. In *Proc. 2017 IEEE Int. Conf. on Robotics and Automation*, pages 1386–1383, 2017.
- [105] Özgür Erkent, Dadhichi Shukla, and Justus Piater. Integration of probabilistic pose estimates from multiple views. In *Proc. 14th European Conf. on Computer Vision*, volume 7, pages 154–170, 2016.
- [106] Asako Kanezaki, Yasuyuki Matsushita, and Yoshifumi Nishida. Rotation-Net: Joint object categorization and pose estimation using multiviews from

- unsupervised viewpoints. In *Proc. 2018 IEEE Conf. on Computer Vision and Pattern Recognition*, pages 5010–5019, 2018.
- [107] Fredrik Viksten, Robert Soderberg, Klas Nordberg, and Christian Perwass. Increasing pose estimation performance using multi-cue integration. In *Proc. 2006 IEEE Int. Conf. on Robotics and Automation*, pages 3760–3767, 2006.
- [108] Ružena Bajcsy. Active perception vs. passive perception. In *Proc. 3rd Workshop on Computer Vision: Representation and Control*, pages 996–1005, 1985.
- [109] David Wilkes, Sven J. Dickinson, and John K. Tsotsos. A quantitative analysis of view degeneracy and its use for active focal length control. In *Proc. 5th IEEE Int. Conf. on Computer Vision*, pages 938–944, 1995.
- [110] Richard Pito. A sensor-based solution to the “next best view” problem. In *Proc. 13th Int. Conf. on Pattern Recognition*, volume 1, pages 941–945, 1996.
- [111] Joseph E. Banta, Laurana R. Wong, Christophe Dumont, and Mongi A. Abidi. A next-best-view system for autonomous 3-D object reconstruction. *IEEE Trans. on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 30(5):589–598, 2000.
- [112] Andreas Doumanoglou, Rigas Kouskouridas, Sotiris Malassiotis, and Tae-Kyun Kim. Recovering 6D object pose and predicting next-best-view in the crowd. In *Proc. 2016 IEEE Conf. on Computer Vision and Pattern Recognition*, pages 3583–3592, 2016.
- [113] Juil Sock, S. Hamidreza Kasaei, Luis Seacra Lopes, and Tae-Kyun Kim. Multi-view 6D object pose estimation and camera motion planning using RGBD images. In *Proc. 17th IEEE Int. Conf. on Computer Vision Workshops*, pages 2228–2235, 2017.
- [114] B.M. Monjurul Alom, Someresh Das, and Md Saiful Islam. Finding the maximum matching in a bipartite graph. *DUET Journal*, 1(1):33–36, 2010.

- [115] Ziming Zhang and Venkatesh Saligrama. PRISM: Person reidentification via structured matching. *IEEE Trans. on Circuits and Systems for Video Technology*, 27(3):499–512, 2016.
- [116] Roland T. Chin and Charles R. Dyer. Model-based recognition in robot vision. *ACM Computing Surveys*, 18(1):67–108, 1986.
- [117] Hiroshi Ninomiya, Yasutomo Kawanishi, Daisuke Deguchi, Ichiro Ide, Hiroshi Murase, Norimasa Kobori, and Yusuke Nakano. Deep manifold embedding for 3D object pose estimation. In *Proc. 12th Joint Conf. on Computer Vision, Imaging and Computer Graphics Theory and Applications*, volume 5, pages 173–178, 2017.
- [118] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.
- [119] Andy Zeng, Shuran Song, Kuan-Ting Yu, Elliott Donlon, Francois R. Hogan, Maria Bauza, Daolin Ma, Orion Taylor, Melody Liu, Eudald Romo, Nima Fazeli, Ferran Alet, Nikhil Chavan Dafle, Rachel Holladay, Isabella Morona, Prem Qu Nair, Druck Green, Ian Taylor, Weber Liu, Thomas Funkhouser, and Alberto Rodriguez. Robotic pick-and-place of novel objects in clutter with multi-affordance grasping and cross-domain image matching. *Int. Journal of Robotics Research*, 2019. DOI: [10.1177/0278364919868017](https://doi.org/10.1177/0278364919868017).
- [120] David Gale and Lloyd S. Shapley. College admissions and the stability of marriage. *The American Mathematical Monthly*, 69(1):9–15, 1962.
- [121] Kunimatsu Hashimoto, Duy-Nguyen Ta, Eric Cousineau, and Russ Tedrake. KOSnet: A unified keypoint, orientation and scale network for probabilistic 6D pose estimation. 2020. URL http://groups.csail.mit.edu/robotics-center/public_papers/Hashimoto20.pdf. Accessed on January 22, 2022.

- [122] Kumar Ashutosh, Saurabh Kumar, and Subhasis Chaudhuri. 3D-NVS: A 3D supervision approach for next view selection. *Computer Research Repository arXiv Preprint*, arXiv:2012.01743, 2020.
- [123] Renaud Detry and Justus Piater. Continuous surface-point distributions for 3D object pose estimation and recognition. In *Proc. 10th Asian Conf. on Computer Vision*, volume 3, pages 572–585, 2010.
- [124] Damien Teney and Justus Piater. Modeling pose/appearance relations for improved object localization and pose estimation in 2D images. In *Proc. 6th Iberian Conf. on Pattern Recognition and Image Analysis*, pages 59–68, 2013.
- [125] Takashi Konno, Ayako Amma, and Asako Kanazaki. Incremental multi-view object detection from a moving camera. In *Proc. 2nd ACM Int. Conf. on Multimedia in Asia*, pages 4–1–4–7, 2021.
- [126] Jianjun Ni, Tao Gong, Yafei Gu, Jinxiu Zhu, and Xinnan Fan. An improved deep residual network-based semantic simultaneous localization and mapping method for monocular vision robot. *Computational Intelligence and Neuroscience*, 2020(7490840):1–14, 2020.
- [127] David Wilkes and John K. Tsotsos. Active object recognition. In *Proc. 1992 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, pages 136–141, 1992.
- [128] Sven J. Dickinson, Henrik I. Christensen, John Tsotsos, and Göran Olofsson. Active object recognition integrating attention and viewpoint control. In *Proc. 3rd European Conf. on Computer Vision*, volume 2, pages 3–14, 1994.
- [129] Peter Gvozdjak and Ze-Nian Li. From nomad to explorer: Active object recognition on mobile robots. *Pattern Recognition*, 31(6):773–790, 1998.
- [130] Lucas Paletta and Axel Pinz. Active object recognition by view integration and reinforcement learning. *Robotics and Autonomous Systems*, 31(1–2):71–86, 2000.

- [131] Joachim Denzler and Christopher M. Brown. Information theoretic sensor data selection for active object recognition and state estimation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(2):145–157, 2002.
- [132] Björn Browatzki, Vadim Tikhanoﬀ, Giorgio Metta, Heinrich H. Bülthoﬀ, and Christian Wallraven. Active object recognition on a humanoid robot. In *Proc. 2012 IEEE Int. Conf. on Robotics and Automation*, pages 2021–2028, 2012.
- [133] Dennis Stampfer, Matthias Lutz, and Christian Schlegel. Information driven sensor placement for robust active object recognition based on multiple views. In *Proc. 2012 IEEE Int. Conf. on Technologies for Practical Robot Applications*, pages 133–138, 2012.
- [134] Juergen Gall and Victor Lempitsky. Class-specific Hough forests for object detection. In *Proc. 2009 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, pages 1022–1029, 2009.
- [135] Shangxuan Wu, Ying-Cong Chen, Xiang Li, An-Cong Wu, Jin-Jie You, and Wei-Shi Zheng. An enhanced deep feature representation for person re-identification. In *Proc. 2016 IEEE Winter Conf. on Applications of Computer Vision*, pages 1–8, 2016.
- [136] Tetsu Matsukawa and Einoshin Suzuki. Person re-identification using CNN features learned from combination of attributes. In *Proc. 23rd Int. Conf. on Pattern Recognition*, pages 2428–2433, 2016.
- [137] Tetsu Matsukawa, Takahiro Okabe, Einoshin Suzuki, and Yoichi Sato. Hierarchical Gaussian descriptor for person re-identification. In *Proc. 2016 IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1363–1372, 2016.
- [138] Tetsu Matsukawa, Takahiro Okabe, Einoshin Suzuki, and Yoichi Sato. Hierarchical Gaussian descriptors with application to person re-identification. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 42(9):2179–2194, 2019.

- [139] Srikrishna Karanam, Yang Li, and Richard J. Radke. Person re-identification with discriminatively trained viewpoint invariant dictionaries. In *Proc. 16th IEEE Int. Conf. on Computer Vision*, pages 4516–4524, 2015.
- [140] Faqiang Wang, Wangmeng Zuo, Liang Lin, David Zhang, and Lei Zhang. Joint learning of single-image and cross-image representations for person re-identification. In *Proc. 2016 IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1288–1296, 2016.
- [141] Martin Hirzer, Peter M. Roth, and Horst Bischof. Person re-identification by efficient impostor-based metric learning. In *Proc. 9th IEEE Int. Conf. on Advanced Video and Signal-based Surveillance*, pages 203–208, 2012.
- [142] Martin Koestinger, Martin Hirzer, Paul Wohlhart, Peter M. Roth, and Horst Bischof. Large scale metric learning from equivalence constraints. In *Proc. 2012 IEEE Conf. on Computer Vision and Pattern Recognition*, pages 2288–2295, 2012.
- [143] Fei Xiong, Mengran Gou, Octavia Camps, and Mario Sznajder. Person re-identification using kernel-based metric learning methods. In *Proc. 13th European Conf. on Computer Vision*, volume 7, pages 1–16, 2014.
- [144] Dapeng Tao, Lianwen Jin, Yongfei Wang, Yuan Yuan, and Xuelong Li. Person re-identification by regularized smoothing KISS metric learning. *IEEE Trans. on Circuits and Systems for Video Technology*, 23(10):1675–1685, 2013.
- [145] Sakrapee Paisitkriangkrai, Chunhua Shen, and Anton van den Hengel. Learning to rank in person re-identification with metric ensembles. In *Proc. 2015 IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1846–1855, 2015.
- [146] Wei Li, Yang Wu, and Jianqing Li. Re-identification by neighborhood structure metric learning. *Pattern Recognition*, 61:327–338, 2017.

- [147] Chong Sun, Dong Wang, and Huchuan Lu. Person re-identification via distance metric learning with latent variables. *IEEE Trans. on Image Processing*, 26(1):23–34, 2017.
- [148] Jifei Song, Yongxin Yang, Yi-Zhe Song, Tao Xiang, and Timothy M. Hospedales. Generalizable person re-identification by domain-invariant mapping network. In *Proc. 2019 IEEE Conf. on Computer Vision and Pattern Recognition*, pages 719–728, 2019.
- [149] Chaojie Mao, Yingming Li, Zhongfei Zhang, Yaqing Zhang, and Xi Li. Pyramid person matching network for person re-identification. In *Proc. 9th Asian Conf. on Machine Learning*, pages 487–497, 2017.
- [150] Ejaz Ahmed, Michael Jones, and Tim K. Marks. An improved deep learning architecture for person re-identification. In *Proc. 2015 IEEE Conf. on Computer Vision and Pattern Recognition*, pages 3908–3916, 2015.
- [151] Jiawei Liu, Zheng-Jun Zha, Qi Tian, Dong Liu, Ting Yao, Qiang Ling, and Tao Mei. Multi-scale triplet CNN for person re-identification. In *Proc. 24th ACM Int. Conf. on Multimedia*, pages 192–196, 2016.
- [152] Li Zhang, Tao Xiang, and Shaogang Gong. Learning a discriminative null space for person re-identification. In *Proc. 2016 IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1239–1248, 2016.
- [153] Harold W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1–2):83–97, 1955.
- [154] Slawomir Bak and Peter Carr. One-shot metric learning for person re-identification. In *Proc. 2017 IEEE Conf. on Computer Vision and Pattern Recognition*, pages 2990–2999, 2017.
- [155] Yang Shen, Weiyao Lin, Junchi Yan, Mingliang Xu, Jianxin Wu, and Jingdong Wang. Person re-identification with correspondence structure learning. In *Proc. 16th IEEE Int. Conf. on Computer Vision*, pages 3200–3208, 2015.

- [156] Mang Ye, Andy J. Ma, Liang Zheng, Jiawei Li, and Pong C. Yuen. Dynamic label graph matching for unsupervised video re-identification. In *Proc. 17th IEEE Int. Conf. on Computer Vision*, pages 5142–5150, 2017.
- [157] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An information-rich 3D model repository. *Computing Research Repository arXiv Preprint*, arXiv:1512.03012, 2015.
- [158] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. Person re-identification by video ranking. In *Proc. 13th European Conf. on Computer Vision*, volume 4, pages 688–703, 2014.
- [159] David F. Manlove. Hospitals residents problem. In *M.-Y., Ming-Yang Kao (Ed.), Encyclopedia of Algorithms*, pages 390–394. Springer, Boston, MA, 2008.
- [160] Shamik Sural, Gang Qian, and Sakti Pramanik. Segmentation and histogram generation using the HSV color space for image retrieval. In *Proc. 2002 Int. Conf. on Image Processing*, volume 2, pages 589–592, 2002.
- [161] Annalisa Barla, Francesca Odone, and Alessandro Verri. Histogram intersection kernel for image classification. In *Proc. 2003 IEEE Int. Conf. on Image Processing*, volume 3, pages 513–516, 2003.
- [162] Nebojsa Jojic, Alessandro Perina, Marco Cristani, Vittorio Murino, and Brendan Frey. Stel component analysis: Modeling spatial correlations in image class structure. In *Proc. 2009 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, pages 2044–2051, 2009.
- [163] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Proc. 17th IEEE Int. Conf. on Computer Vision*, pages 2961–2969, 2017.

- [164] David G. McVitie and Leslie B. Wilson. Stable marriage assignment for unequal sets. *BIT Numerical Mathematics*, 10(3):295–309, 1970.
- [165] Dan Gusfield and Robert W. Irving. *The Stable Marriage Problem: Structure and Algorithms*. MIT press, Cambridge, MA, 1989.

Publication list

Peer-reviewed Journal

- [1] Nik Mohd Zarifie Hashim, Yasutomo Kawanishi, Daisuke Deguchi, Ichiro Ide, Ayako Amma, Norimasa Kobori, and Hiroshi Murase. Best next-viewpoint recommendation by selecting minimum pose ambiguity for category-level object pose estimation. *Journal of the Japan Society for Precision Engineering*, 87(05):440–446, May 2021. DOI: <https://doi.org/10.2493/jjspe.87.440>.
- [2] Nik Mohd Zarifie Hashim, Yasutomo Kawanishi, Daisuke Deguchi, Ichiro Ide, and Hiroshi Murase. Simultaneous image matching for person re-identification via the stable marriage algorithm. *IEEJ Transactions on Electrical and Electronic Engineering*, 15(6):909–917, April 2020. DOI: <https://doi.org/10.1002/tee.23133>.

International Conference

- [1] Nik Mohd Zarifie Hashim, Yasutomo Kawanishi, Daisuke Deguchi, Ichiro Ide, Hiroshi Murase, Ayako Amma, and Norimasa Kobori. Next viewpoint recommendation by pose ambiguity minimization for accurate object pose estimation. In *Proc. 14th Int. Joint Conf. on Computer Vision, Imaging and*

Computer Graphics Theory and Applications (VISIGRAPP), volume 5, pages 60–67, Feb. 2019. DOI: 10.5220/00073667006000067.

- [2] Nik Mohd Zarifie Hashim, Yasutomo Kawanishi, Daisuke Deguchi, Ichiro Ide, and Hiroshi Murase. A preliminary study on optimizing person re-identification using stable marriage algorithm. In *Proc. 2018 Int. Workshop on Frontiers of Computer Vision*, pages 1–6, Feb. 2018.

Domestic Conference

- [1] Nik Mohd Zarifie Hashim, Yasutomo Kawanishi, Daisuke Deguchi, Ichiro Ide, and Hiroshi Murase. Viewpoint recommendation for object pose estimation via pose ambiguity minimization. In *Proc. 22nd Meeting on Image Recognition and Understanding*, number PS1-69, Aug. 2019.
- [2] Nik Mohd Zarifie Hashim, Yasutomo Kawanishi, Daisuke Deguchi, Ichiro Ide, and Hiroshi Murase. An analysis of simultaneous image matching on various datasets for person re-identification. In *2019 Tokai-Section Joint Conf. on Electrical, Electronics, Information, and Related Engineering*, number G5-2, Sept. 2019.