DEGREE PROJECT IN INFORMATION AND COMMUNICATION TECHNOLOGY,
SECOND CYCLE, 30 CREDITS
STOCKHOLM, *SWEDEN 2020*

Indoor scene verification

Evaluation of indoor scene representations for the purpose of location verification

FILIP FINFANDO

KTH ROYAL INSTITUTE OF TECHNOLOGY
SCHOOL OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

Indoor scene verification / Verifiering av inomhusbilder

Abstract

When human's visual system is looking at two pictures taken in some indoor location, it is fairly easy to tell whether they were taken in exactly the same place, even when the location has never been visited in reality. It is possible due to being able to pay attention to the multiple factors such as spatial properties (windows shape, room shape), common patterns (floor, walls) or presence of specific objects (furniture, lighting). Changes in camera pose, illumination, furniture location or digital alteration of the image (e.g. watermarks) has little influence on this ability. Traditional approaches to measuring the perceptual similarity of images struggled to reproduce this skill. This thesis defines the Indoor scene verification (ISV) problem as distinguishing whether two indoor scene images were taken in the same indoor space or not. It explores the capabilities of state-of-the-art perceptual similarity metrics by introducing two new datasets designed specifically for this problem. Perceptual hashing, ORB, FaceNet and NetVLAD are evaluated as the baseline candidates. The results show that NetVLAD provides the best results on both datasets and therefore is chosen as the baseline for the experiments aiming to improve it. Three of them are carried out testing the impact of using the different training dataset, changing deep neural network architecture and introducing new loss function. Quantitative analysis of AUC score shows that switching from VGG16 to MobileNetV2 allows for improvement over the baseline.

Keywords

computer vision, perceptual similarity, visual place recognition, indoor scene localization, deep neural networks

Sammanfattning

Med mänskliga synförmågan är det ganska lätt att bedöma om två bilder som tas i samma inomhusutrymme verkligen har tagits i exakt samma plats även om man aldrig har varit där. Det är möjligt tack vare många faktorer, sådana som rumsliga egenskaper (fönsterformer, rumsformer), gemensamma mönster (golv, väggar) eller närvaro av särskilda föremål (möbler, ljus). Ändring av kamerans placering, belysning, möblernas placering eller digitalbildens förändring (t. ex. vattenstämpel) påverkar denna förmåga minimalt. Traditionella metoder att mäta bildernas perceptuella likheter hade svårigheter att reproducera denna färdighet . Denna uppsats definierar verifiering av inomhusbilder, Indoor Scene Verification (ISV), som en ansats att ta reda på om två inomhusbilder har tagits i samma utrymme eller inte. Studien undersöker de främsta perceptuella identitetsfunktionerna genom att introducera två nya datauppsättningar designade särskilt för detta. Perceptual hash, ORB, FaceNet och NetVLAD identifierades som potentiella referenspunkter. Resultaten visar att NetVLAD levererar de bästa resultaten i båda datauppsättningarna, varpå de valdes som referenspunkter till undersökningen i syfte att förbättra det. Tre experiment undersöker påverkan av användning av olika datauppsättningar, ändring av struktur i neuronnätet och införande av en ny minskande funktion. Kvantitativ AUC-värdet analys visar att ett byte från VGG16 till MobileNetV2 tillåter förbättringar i jämförelse med de primära lösningarna.

Nyckelord

datorseende, perceptuella likheter, visuell platsigenkänning, inomhusbild lokalisering, djupa neuronnät

Acknowledgements

I would like to express my gratitude to Ying Liu and Amir Payberah for supervising this project and providing feedback. I would also like to thank the ScanNet project for sharing the dataset and SonarHome company for enabling evaluation with real-world apartment listings. Advice given by Piotr Tempczyk was invaluable during the experiment design and implementation phase. I would not be able to complete the project without computational resources shared by Polish National Institute for Machine Learning and SonarHome. I wish to acknowledge the support of EIT Digital Master School throughout the two-year education period. I would also like to thank Martyna Wojciechowska and Karolina Mroz for their help with the translation. Finally, I would like to offer my special thanks to Dorota Jedynak for proofreading the whole thesis.

Stockholm, November 2020
Filip Finfando

Contents

List of Figures

List of Tables

List of acronyms and abbreviations

AUC Area under curve

BOV bag-of-visual-words

ISD Indoor Scan Dataset

ISV Indoor scene verification

MSE Mean squared error

PR Precision and recall

PSNR Peak signal-to-noise ratio

RELD Real Estate Listings Dataset

ROC Receiver Operating Characteristic

SIFT Scale-invariant feature transform

SSIM Structural similarity index

Chapter 1

Introduction

Human's visual system is capable of recognizing similarities, despite not having seen the object of interest before. Moreover, it is able to rank entities with respect to similarity. While the task is easy for humans, learning a good perceptual similarity metric remains a challenge for computer systems. It is not clear how similarity between two objects can be measured in a quantitative way due to the complicated mechanics of human's perception [1]. *Perceptual similarity* metrics have been applied in image retrieval [2, 3], anti-piracy search [4], quality assessment [5, 6] and entity resolution [7]. Researchers in those fields use different similarity metrics depending on their use case.

1.1 Problem statement

This thesis explores the Indoor scene verification (ISV) problem. Its goal is to find a way to accurately identify whether two photographs of the indoor scenes were taken in the same indoor space or a different one.

Figure 1.1: Two images of the same indoor scene taken at a different camera pose (different perspective).

Figure 1.1 presents two images of the same room but taken from a different angle. Even though the images are significantly different (in terms of what colour or object a given pixel is presenting), it is easy for a human observer to identify that both were taken in the same room. Figure 1.2 displays two pictures of the same room taken at different daylight conditions. Again, even though the colours and tones are completely different, it is not difficult at all to notice that the photographer took the photographs in the same indoor location.

Figure 1.2: Two images of the same indoor scene taken at a different time of the day (different illumination) and moved objects.

In order for a computer to make this decision, it needs a way to measure the perceptual similarity of those images quantitatively. It has to take into account the challenges specific to pictures of the indoor scenes. To the best of our knowledge, there is currently no method designed with an aim to verify the location of indoor scene images in such a way. Authors of [8] analyze pictures of sex trafficking victims taken in hotel rooms and trying to link them to the database of photographs of hotel rooms scraped from hotel websites. They use existing and already available methods for measuring the perceptual similarity of images. It is also pointed out that there is a room for exploration and experimenting with different methods designed to analyze indoor scene similarity. Such methods, specialized in this narrow domain, should supposedly be better at solving this challenge. What are the capabilities of existing methods in solving the ISV task and how can they be improved?

1.2 Background

In order to measure the perceptual similarity of an image one first needs to extract the information about the contents of the image and be able to compare them later in a quantitative way. The technique that enables this is detecting local invariant features using Scale-invariant feature transform (SIFT) [9]. The

features extracted using this method can be later matched between the two images using a fixed threshold. The number of features matched serves as a perceptual similarity measure. Recent research in computer vision domain focuses on designing and training deep neural network architectures. Such networks are trained on a dataset of examples, ranked in terms of similarity. These models are able to learn how to convert images into lower-dimensional representations. One can later use euclidean distances between two of them to measure perceptual similarity [10, 11, 12, 13, 14].

Visual place recognition field builds on top of the methods described above and given an image, it tries to match the location where it was taken. The problem received a high interest during past years, which resulted in successful solutions applied to the outdoor scene recognition. These solutions are based on extracting image information with local invariant features [11] or deep neural network [15]. Using large databases of geo-tagged images and efficient perceptual similarity measurement, they are able to quickly retrieve matching images and recognize an accurate location. Indoor localization is also an active area of research. Using the above mentioned techniques for extracting image representations, the researchers are trying to find out the exact location of the camera within a scope of a single or several buildings [16, 17, 18].

1.3 Application and purpose

Nowadays, the Internet is a rich and valuable source of data. The number of data providers and websites where the content is generated constantly increase. This poses a challenge to entities that are willing to analyze such data. The uniqueness of the data is especially challenging. There is nothing that stops a single entity in the real world to have multiple digital counterparts. The record linkage can be achieved by using attributes of the instances such as images. Automated ISV could support detecting the same apartments across different websites. Real estate technology companies such as SonarHome [19] or short-term rental comparison engines e.g. Holidu [20] (holidu.com) may benefit from this research. Another application is to fight sex trafficking. Authors of [8] are matching images from sex services advertisements to pictures of hotel rooms.

The purpose of this thesis is to empower organizations that use indoor scene image data in applications like described above and broaden the knowledge about image clustering by exploring this field in a more narrow domain. The results of the thesis should guide anyone who needs to achieve accurate results in the ISV task.

1.4 Goals

The goal of this project is to find out what would be the best approach to solve the ISV problem as defined in section 1.1. This has been divided into the following three sub-goals:

1. create evaluation datasets that would enable assessment and comparison of the model performance in the ISV task,
2. evaluate existing solutions for measuring the perceptual similarity of images and choose the best baseline,
3. investigate and implement improvements based on existing solutions.

1.5 Research Methodology

The research process is divided into three phases. In the first phase, two evaluation datasets are to be defined and generated. The first one will be created using images from an academic dataset, called ScanNet [21]. The second one will be created using real-world indoor scene images from real estate listings collected with the support of the host company. In the second phase, several baseline candidates will be evaluated on the prepared datasets. The best one will be chosen and in the last phase, a number of experiments will be carried out aiming to improve it. Each experiment consists of designing and evaluating a different solution to the ISV task. In order to assess them in a quantitative way, metrics typical for binary classification will be used i.e. Receiver Operating Characteristic (ROC) curve, precision and recall curve and AUC score. The results will also be verified qualitatively by analyzing raw examples and using dimensionality reduction techniques.

1.6 Structure of the thesis

Chapter 2 reviews relevant background information about the methods for extracting image representation and applications in relevant fields. Chapter 3 presents the methodology and method used to answer the research question. Chapter 4 provides a detailed description of what has been done during the degree project. Chapter 5 summarizes the results and presents the most interesting insights. Chapter 6 concludes the thesis, reflects upon the work done during this project and proposes future research direction.

Chapter 2

Background

In this chapter, scientific background relevant to the ISV task is presented. In section 2.1 methods for measuring the perceptual similarity of images are discussed and evaluated one by one. Section 2.2 reviews published applications of those methods in visual place recognition and indoor localization domains. Section 2.3 is a summary of this chapter.

2.1 Image similarity measures

The following sub-sections discuss various methods for measuring the similarity between two images. Each provides a brief description of how the given technique works, it's pros and cons, as well as typical applications.

2.1.1 Mean squared error and structural similarity

The most simple approach to measure similarity between two images is calculating the average of the squared difference between corresponding pixel values also called Mean squared error (MSE). This method accurately and quickly identifies identical images, which pixel values are the same. However, any slight change in a crop, camera pose, illumination or digital alteration of an image results in significantly different pixel values and therefore high MSE distance. Peak signal-to-noise ratio (PSNR) is a metric based on MSE that is used to measure the quality of image or video compression [22]. Another metric aiming to improve PSNR is Structural similarity index (SSIM) [23]. To compute the similarity measure, it takes into account three components: luminance comparison, contrast comparison and structure comparison. It has been demonstrated it is suitable for image quality assessment.

2.1.2 Histogram methods

Another approach to measure the similarity of two images is comparing their colour histograms. The coluors of an image are grouped into many discrete bins and histogram is obtained by counting the number of times each colour appears in an image. Then the similarity metric is obtained from histogram intersection [24]. Such methods do not take into account spatial relationships between pixels and therefore are invariant to any changes in the rotation of an image. They are also robust to slight changes in scale, angle distortions or occlusion. The techniques have been successfully applied for image colour-indexing [25].

2.1.3 Local invariant features matching

It is a challenge to compare image similarity in a way that is insensitive to the changes in image crop, camera pose or illumination. Extracting distinctive invariant features using methods like SIFT [9] is a way to address these problems. Thanks to the features being distinctive, it is feasible to match them between images and compute similarity metric using a number of features matched or bag-of-visual-words (BOV) approach [26]. It has been demonstrated that matching extracted local invariant features could be used as a robust perceptual similarity metric [27].

2.1.4 Perceptual hashing

Images can be quickly and accurately compared using perceptual hashing algorithms. They are designed in a way to preserve the image features and generate hash values that are comparable between each other using hamming distance. The most basic one is the average hash (aHash), which reduces the size of an image to 8x8 (64 pixels), converts it to grayscale (64 colours) and constructs the hash by setting 64 bits to either 1 or 0 based on the colour value being below or above the mean value. Other algorithms are based on aHash and include perceptual hash (pHash) [28], difference hash (dHash) [29].

2.1.5 Deep Neural Networks

Recently it has been shown, that deep neural networks outperform methods based on hand-crafted features. Authors of [10] train a deep neural network architecture to learn a similarity metric by itself. During the training phase the model is fed triplets of images (anchor, positive, negative). It uses a triplet

loss with an objective to keep a fixed euclidean distance margin between the representation of a positive and the negative.

2.2 Related works

The research areas that are very close to the ISV task are visual place recognition and localization. This section demonstrates literature applying image similarity measures in different domains.

2.2.1 Outdoor place recognition

There were successful attempts to tackle visual place recognition problem on outdoor scenes using extracted local invariant features [11]. Authors build on top of existing solutions for image retrieval and location recognition. They manage to improve them by selecting the local features based on their distinctiveness.

Deep neural networks were also applied to the outdoor place recognition problem using NetVLAD pooling layer designed for this purpose [15]. The authors mimic vector of Locally Aggregated Descriptors (VLAD) [30] in the convolutional neural network architecture. They also apply a triplet loss similar to the one used in [10] or [12].

2.2.2 Indoor visual localization with camera pose estimation

Indoor localization problem involves predicting the camera location and sometimes also the 6 degrees of freedom camera pose based on a query image taken by this camera. It is more challenging than the outdoor place recognition problem due to several reasons: large textureless areas inside buildings (white walls), repetitive patterns, dynamic changes in illumination and frequent changes of objects position.

Authors of [16] build on top of the pre-trained NetVLAD model [15] in order to retrieve a short-list of potential candidate images. Other researchers decide on their own deep neural architecture based on InceptionV3 [18] or AlexNet [17]. They treat the problem as a classification task dividing the building space into zones and sub-zones. Other works utilize Long short-term memory networks for this purpose [31]. The solutions designed in the above works are proved to be working within the scope of up to 12 buildings.

2.2.3 Indoor scene verification

The ISV problem as described in section 1.1 is similar to the indoor visual localization task. In our case, it is not necessary to discover exact camera location and its pose, but we want to accurately decide whether two images were taken in the same indoor space. It is also desirable to discover similarities at a scale larger than a couple of buildings. This problem was tackled using pre-trained NetVLAD model to match pictures of hotel rooms and fight the problem of sex trafficking [8].

2.3 Summary

This chapter provided an overview of how the similarity between two images could be measured. Looking at all presented methods it is clear that there is no single best similarity measure for images as the concept of similarity may be different depending on the domain and application. No metric would capture every aspect of what human's visual system considers to be similar. This is why solutions designed to solve one problem in a narrow domain e.g. outdoor location recognition tend to work better than generic methods.

Indoor location recognition problem received a high interest during recent years sparked by significant improvements in computer vision tasks and increased interest in robotics. It builds on top of outdoor localization methods as the problems share common challenges however, indoor scenes seem to be more complex. The ISV problem has not achieved a high interest yet.

Chapter 3

Research method

The purpose of this chapter is to provide an overview of the research method used in this thesis. Section 3.1 describes the research process. Section 3.2 presents what kind of data will be required to answer the research question and how it will be collected. Section 3.3 describes the design of the experiments carried out in this project. Section 3.4 discusses how the reliability of and validity of the method and data will be ensured and section 3.5 describes planned data analysis.

3.1 Research Process

The ISV topic has not been addressed directly yet. Therefore there are no publicly available datasets designed to tackle this problem specifically. In the first phase of the project, the datasets will be created. Their aim is to enable the data collection and the quantitative evaluation of the solutions to the problem. In the second phase, the existing methods feasible for solving the ISV task will be chosen. Each of them is going to be evaluated on the datasets created in phase one. The most promising one will be chosen as the baseline for the last phase of the project. The improvement attempts will be considered in the last phase. Three experiments will be designed and carried out aiming to improve the result of the baseline models. The goal of each experiment is to prepare a solution to the ISV problem and apply it on the datasets created in the first phase. The results of the experiments will be compared against the results of the baseline in a quantitative way together with additional qualitative analysis if necessary.

3.2 Data Collection

The data necessary to answer the research question will be collected through a series of experiments carried out in phases two and three of the project. All experiments will be applied on the same datasets created in phase one.

The datasets should be created from a large number of images presenting indoor scenes. Some of the pictures should be taken in the same location while others in different locations. There must be a wide diversity of unique locations. The final dataset should consist of samples generated from such images. One sample is a pair of images with a binary label classifying whether two images present the same indoor scene or not. For the purpose of this project, two pictures are considered to present the same indoor scene if they were taken in the same room and present the same part of the room sharing common patterns or objects that enable identifying them as taken in the same location by a human observer.

The methods applied as baselines and experiments should be able to identify a pair of images taken in the same location or not i.e. for an input being a pair of images a normalized distance score between 0 and 1 is expected as output. The final distinction is then enabled by applying the certain threshold to the scores.

Some of the datasets will be created based on the academic datasets and it might be necessary to get access approval and preprocess the data. Other datasets will be created manually e.g. scraped from the Internet or obtained with support from the host organization. It is necessary to pay special attention to copyrights in both cases. In case of academic datasets, this will be done through sticking to rules established in agreements and licenses attached to the datasets. In the case of scraped data, it will be achieved through careful analysis of scraped website's terms of service and obeying the copyright law.

3.3 Experimental design

Each experiment's goal is to solve the ISV problem. The model tested in each experiment is supposed to take a pair of images as input and return a similarity measure of two images as output. Scores for all samples within each dataset will be generated for later evaluation. The evaluation will be performed quantitatively using ROC curve and AUC score.

The experiments will be developed and run using Python 3.7 interpreter and common scientific libraries (NumPy, SciPy, Pandas and PyTorch). PyCharm,

Visual Studio Code and Jupyter Notebook will be used as a development environment. The code will be saved either as Python script files or IPython notebook files.

Design and preparation of the experiments will require standard PC hardware and access to the internet. Some of the experiments will require GPU hardware suitable for deep learning tasks.

3.4 Assessing reliability and validity of the data collected

This section aims to provide means of ensuring reliability and validity of the chosen method and collected data.

3.4.1 Validity of method

Quantitative methodology is a typical research method used in computer vision. As long as the evaluation datasets are created according to the guidelines, the generated data is expected to return valid results. However, due to the complex nature of some models, a qualitative analysis may be required to ensure validity and better understand the pros and cons of the given solution.

3.4.2 Reliability of method

To ensure the reliability of the quantitative experiment, it needs to be ensured it is designed in the right way. First and foremost it is forbidden to use any samples from the evaluation dataset during the design and development of the experiments. The experiments are also expected to be reproducible. Therefore each experiment will be run on three different variants of the dataset and the result of each of them will be reported.

3.4.3 Data validity

The validity of the data will be checked manually by the qualitative analysis of samples randomly chosen from each dataset. Additionally, the results of each experiment will be checked by analyzing the distribution of output scores.

3.4.4 Reliability of data

To ensure the results can be relied upon, at least two evaluation datasets should be created. If the amount of data allows for such operation, the samples should be split among 2 disjoint datasets. This will ensure the results of baselines and experiments are cross-checked and reproducible on multiple datasets.

3.5 Planned Data Analysis

This section provides an overview of data analysis techniques and data analysis software used in this project.

3.5.1 Data Analysis Technique

Firstly, the data collected from the experiments will be analyzed descriptively to check for any outliers or errors. Then for each experiments evaluation metrics typical for binary classification will be computed i.e. AUC and confusion matrix. The results will be also visualized using ROC plot (recall and fall-out), as well as precision and recall plot. Qualitative analysis will consist of analysis of the false positives and false negatives as well as reducing dimensionality of a random sample and assessment of clustering capabilities.

3.5.2 Software Tools

The data analysis will be carried out in Jupyter Notebook environment using Python 3.7 and common libraries used for data preparation and visualization (NumPy, Pandas, Matplotlib, Plotly). The outcome of the analysis are image files presenting plots comparing the experiments.

Chapter 4

Indoor scene verification

This chapter explains what has been done during this degree project. The work has been divided into 3 phases as described in section 3.1. In section 4.1 the datasets created in phase one are presented. Section 4.2 describes the methods chosen as baseline candidates and compares them. Last section 4.3 introduces experiments attempting to improve the baseline. The phases of the project and their contents are illustrated in figure 4.1.

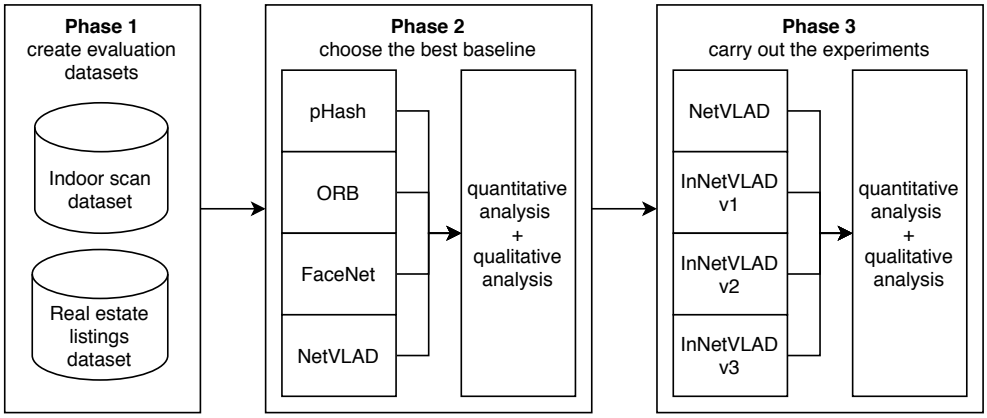

Figure 4.1: Phases of the ISV project

4.1 Evaluation datasets

In the first phase of the project, two datasets were created. Their aim is to enable performance assessment of the baseline candidates and the models prepared during the experiments. The first one called Indoor Scan Dataset

(ISD) is presented in sub-section 4.1.1 and the second one named Real Estate Listings Dataset (RELD) is described in section 4.1.2.

4.1.1 Indoor scan dataset

ScanNet [21] is an RGB-D video dataset containing 2.5 million pictures of indoor scenes across more than 1500 scans. Thanks to the generosity of ScanNet project members, I was able to download and use the data for this thesis. The scans were taken in different indoor locations using a tablet device with an RGB-D camera attached. Device owner was instructed to walk around the room while holding the camera. In the dataset, sometimes several scans were taken in the same room. The dataset is split by the authors into 1201 training and 312 validation scans.

As the dataset was created for 3D object classification and semantic segmentation, it contains much more information than it was required for this project e.g. camera pose and depth information. In the beginning, it was necessary to extract all JPG images from each scan using a Python script provided by the authors of the dataset. The script was designed to work with Python 2, so it was necessary to modify it slightly to enable seamless work in Python 3.7. For some of the scenes in the training dataset, it was not possible to extract images due to errors or missing data. In the end, final number of images was 1 477 427 across 929 scenes in the training dataset and 539 499 across 312 scenes in the validation dataset. All images within a single scan are numbered according to the sequence in which they were taken.

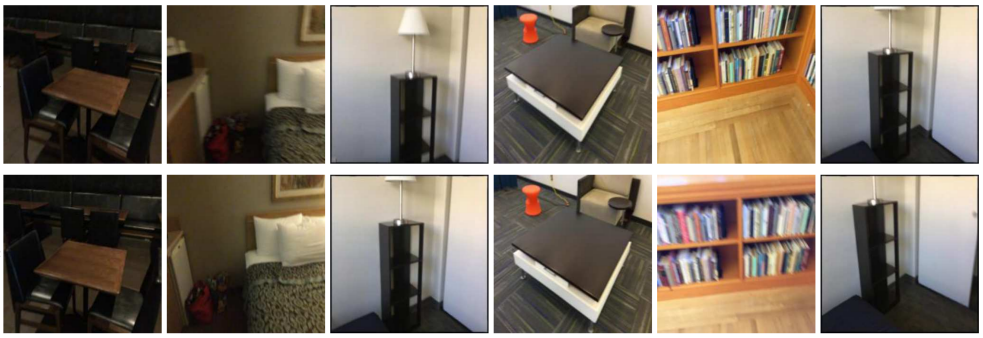

Figure 4.2: Examples of image pairs separated by 10 frames in the scan sequence

It can not be assumed that all images in a scan belong to a single class for the purpose of the ISV task. Authors of images were told to "roll" the camera

around the room and capture the whole space to be able to generate a 3D reconstruction. As a result, the images present completely different parts of the room. For someone who sees these scenes for the first time, it is impossible to tell whether two randomly chosen pictures were taken in the same indoor location or not. When two pictures taken one after the other are picked, it is clear they present the same indoor scene and therefore belong to the same class. As the camera is moved further, the images contain different objects and other parts of the room. The images gradually become less and less of the same class. One can compare image pairs separated by 10 frames (figure 4.2), 20 frames (figure 4.3) and 100 frames (figure 4.4) to get a sense of how the images become different with a growing number of frames between them.

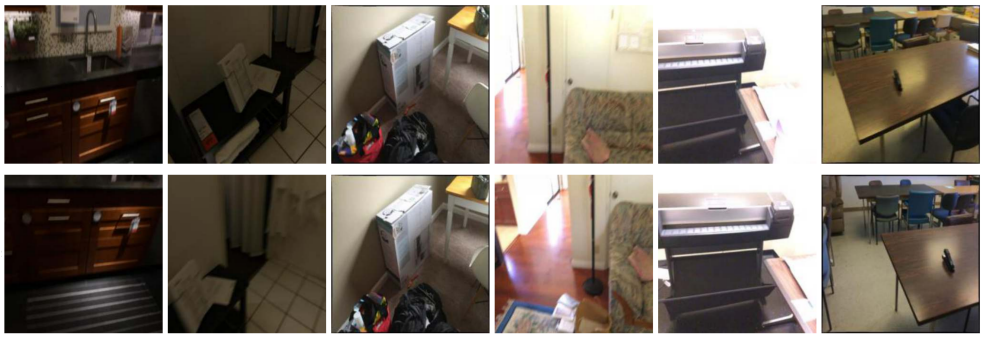

Figure 4.3: Examples of image pairs separated by 20 frames in the scan sequence

This phenomenon is also illustrated in figure 4.5, which presents mean pHash similarity distance and confidence intervals with respect to the number of frames between two images. The distance equal to 0.5 is plotted with a red dashed line as it is an expected value of pHash for two randomly taken images.

According to the data presented in figure 4.5 image pairs separated by only up to 10 frames have almost the same contents and are easy to identify as presenting the same scene using pHash algorithm. The examples of such image pairs are presented on figure 4.2. As the camera is moved further away, the pHash algorithm becomes less and less effective. Examples of images separated by 20 frames are shown in figure 4.3. Such image pairs are already hard to identify as the same by pHash algorithm. Although they are still presenting the indoor scenes that are very similar and easy to identify by a human observer. At some point the image pairs were taken so far away from each other, they are impossible to be identified as the same indoor scene. On

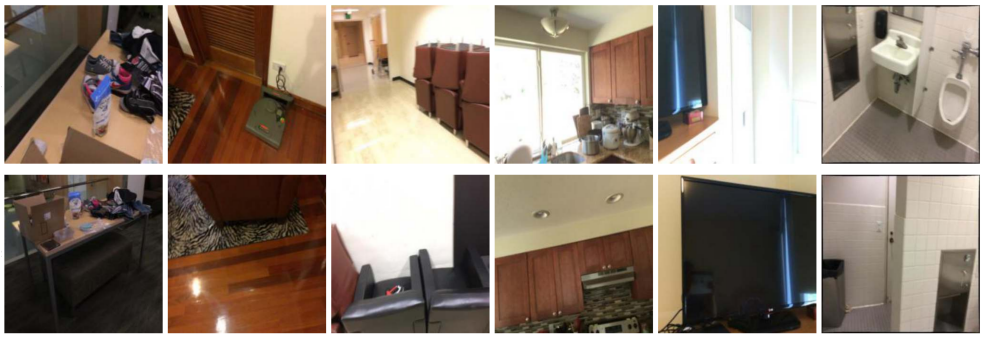

Figure 4.4: Examples of image pairs separated by 100 frames in the scan sequence

figure 4.4 examples of images separated by 100 frames are shown. These are very hard examples for human's visual system as well.

The number of frames is not a perfect measure to separate image classes. The camera could be moved across the room at a different speed. In this thesis, it is assumed that the camera was not moving at a significantly different pace between the scans and within the scan sequence. This effect may be explored further by using the camera pose data and analyzing the location and angle of taken pictures.

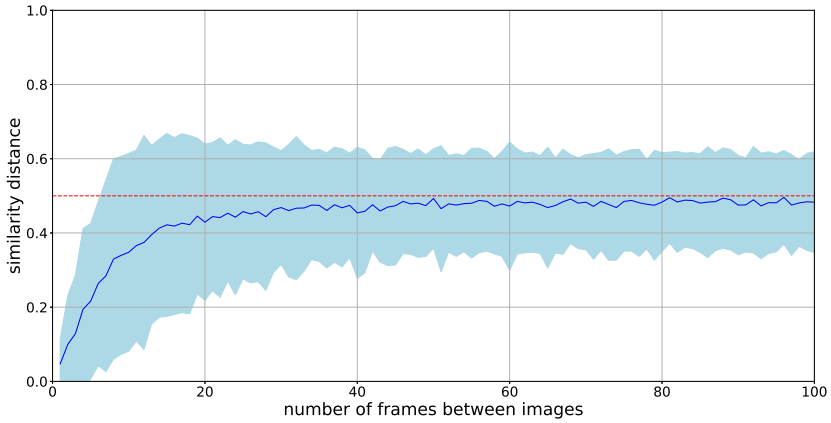

Figure 4.5: pHash similarity distance with respect to the number of frames between two images. Sample of 10 000 image pairs across 100 scene scans from validation dataset

An evaluation dataset for the ISV problem has to consist of image pairs labelled using a binary variable. It should label them as either being the same indoor scene (positive or "1") or not (negative or "0"). In order to create such dataset out of the ScanNet dataset, an assumption has to be made regarding the number of frames between images in a sequence, which separates neighbouring pictures between positive and negative. In other words: after how many frames the image does not present the same indoor scene anymore? Rather than choosing one fixed number, I propose a method that relies on sampling from truncated normal distribution to select a set of images picturing the same indoor space i.e. belonging to the same class. Firstly a frame index is sampled at random from the whole sequence of the scan. In the next step image indices are sampled from a truncated normal distribution centred around the frame index. The standard deviation parameter is constant and allows to manipulate the degree of difficulty of the dataset. In this project 3 standard deviation parameters are proposed: 10, 20 and 30. As a result, easy, medium and hard dataset variants are obtained. Figure 4.6 illustrates how image indices belonging to one class were sampled from a scan sequence consisting of 1800 images.

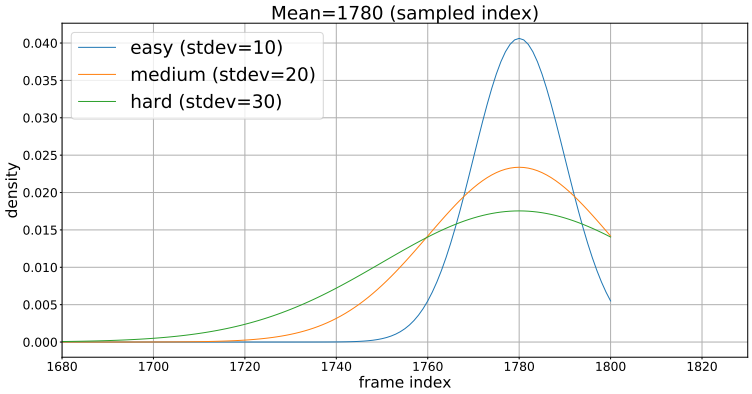

Figure 4.6: Example of truncated normal distributions used to generate positive image pairs for easy, medium and hard dataset variants

Each dataset variant will consist of 10 000 triplets. Each triplet is an anchor image, positive image pair and negative image pair. This results in 20 000 observations labeled as 0 or 1. Every triplet is sampled from the dataset of all images by taking following steps:

1. select two scenes at random without replacement,

2. sample a single image at random from all images belonging to the first scene,
3. sample a neighboring positive image pair by picking a random variate from a truncated normal distribution and rounding it to the nearest integer (the distribution is centered around image index with a standard deviation value depending on the variant of the dataset),
4. sample a random negative pair from images belonging to the second scene.

4.1.2 Real estate listings dataset

To test the baseline candidates and the experiments in a real-world scenario, a dataset that would resemble a real life application of the ISV was created. SonarHome is a company, that uses the real estate listings data from various sources to gain insights about current trends in the real estate market. One of the data sources is apartment listings scraped from the real estate portals. However, very low quality of this data requires manual processing to make it useful for advanced analysis. The company provided me with real estate listings URLs, grouped into categories based on the analysts' judgement to be related to the same apartment in the real world.

Based on this data, the images of publicly available listings were collected. They were grouped into subcategories of pictures taken in the same room. This dataset consists of 540 images of 102 scenes across 17 different apartments.

Only one variant of this dataset was created by following the steps below:

1. for each apartment room, generate unique combinations of length 2 from all images and label such image pairs as positive,
2. for each image in each apartment room, create a cartesian product with all images from other apartments. Label such image pairs as negative. Sample a number of negative image pairs equal to the number of positive image pairs.

The dataset generated according to the above guidelines consists of 3626 image pairs and an equal number of positive and negative samples.

4.2 Baseline selection

Following insights from the literature study, four techniques for image similarity measurement were selected to be evaluated as the baseline candidates. In

section 4.2.1 each of them is briefly described. In section 4.2.2 the performance of the baseline candidates is evaluated and the best one is chosen.

4.2.1 Baseline candidates

Perceptual hashing [28] was chosen as a representative of hashing approaches to measure image similarity. For every image in each dataset, 64-bit hash value was computed. In the next step, the hamming distance between two hash values of each labelled image pair was computed. ImageHash library for Python was the implementation used to test this technique.

To test the performance of methods based on matching local invariant features, the technique called **ORB** [32] was selected as the next baseline candidate. From each image, 500 descriptors were extracted. In order to compute the distance between labelled image pairs, the descriptors were paired between each other using brute force hamming distance matching. The number of matched descriptors normalized by the maximum number of descriptors served as a final similarity measure. The aforementioned tools are implemented in OpenCV2 library for Python, which was used in this project.

NetVLAD [15] is a neural network architecture designed to tackle visual place recognition problem. It was trained on images of outdoor scenes using weakly supervised ranking loss. It uses a pre-trained network convolutional neural network without the last layer, which serves as a dense descriptor extractor. The output of the encoder network is transformed into a compact representation with a NetVLAD pooling layer. Such representations can be later compared between each other using euclidean distance. Authors tested two convolutional neural network architectures as dense descriptor extractors: VGG16 [33] and AlexNet [34] and achieved the best results on VGG16. In this project, I am using weights of a model based on VGG-16 and trained on Pittsburgh dataset [35].

A deep neural network architecture named **FaceNet** [12] is designed for face recognition, verification and clustering tasks. It is a successful application of online triplet loss, which accurately verify pictures of human faces. The similarity scores are indifferent to changes in illumination or camera pose. The reason for including FaceNet in the list of potential baseline candidates, is to evaluate and compare the performance of the model trained on data from a different domain.

4.2.2 Baseline evaluation and selection

To compare the performance of the models in the ISV task, Receiver Operating Characteristic (ROC) curve and Precision and recall (PR) curve were used. The distances between images were scaled to a 0-1 range for each model before the evaluation. On figure 4.7 one ROC plot is presented for each variant of the Indoor Scan Dataset (ISD).

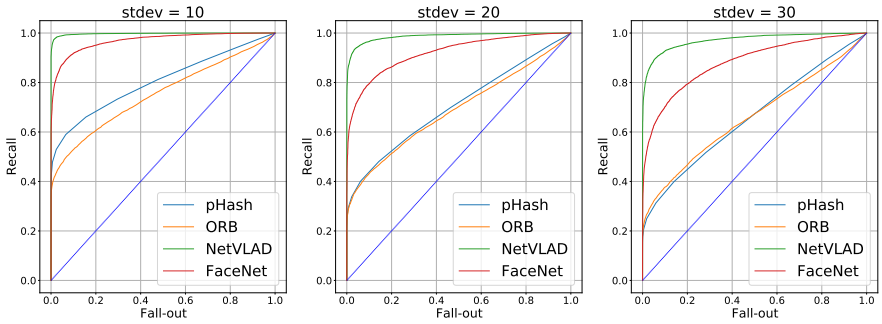

Figure 4.7: Baseline candidates - ROC curve on all variants of Indoor Scan Dataset

All methods presented in this section perform worse in terms of AUC on dataset variants with higher standard deviation parameter. This is according to the expectations as such datasets consist of image pairs that are further away from each other on average. Therefore they are harder to be correctly verified as presenting the same indoor space.

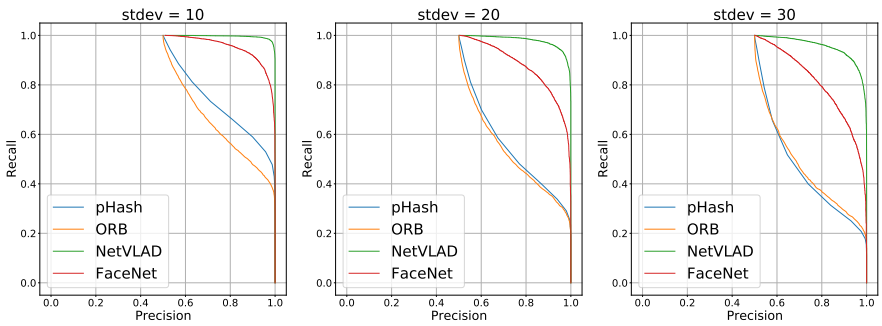

Figure 4.8: Baseline candidates - Precision and Recall curve on all variants of Indoor Scan Dataset

When comparing the AUC of the baseline candidates between each other, NetVLAD provides the best performance on all datasets and AUC is nearly 1.0 on the easiest dataset. It is followed by FaceNet, which provides satisfying performance, despite being trained on a completely different domain of images i.e. human faces. The results of pHash and ORB are similar, both being significantly worse than previously mentioned approaches based on deep neural networks. On the easiest dataset pHash is better than ORB, but the difference is not visible anymore on the medium and hard dataset variants. On the hard dataset variant, ORB is even slightly better than pHash, which exposes the limitations of the latter.

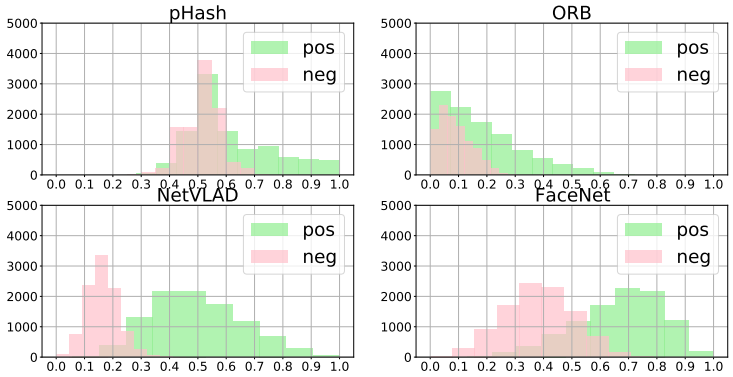

Figure 4.9: Baseline candidates - Histogram of scores on medium difficulty variant of Indoor Scan Dataset

To better understand the performance of all the methods, the PR curves are displayed in figure 4.8. NetVLAD can provide a certain threshold to detect most of the positives (more than 80% recall on all variants) while maintaining 100% precision. In comparison, pHash and ORB on the medium dataset variant are only able to detect about 25% of the true cases while maintaining 100% precision.

Histogram of scores assigned to the labelled observations allows us to gain even more insight into the behaviour of the models. Scores generated on the medium difficulty dataset variant (stdev=20) for all 4 baseline candidates were split into positive and negative samples and plotted on figure 4.9. None of the methods provides a threshold that can separate the two labels completely.

The distances between negative pairs generated from pHash form a normal distribution centred at 0.5, which is the expected value of the distance between

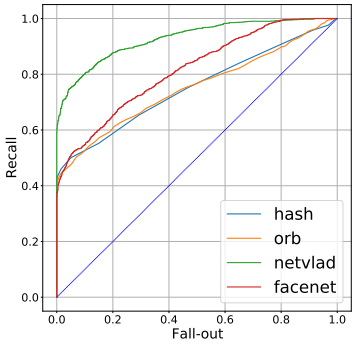

Figure 4.10: Baseline candidates - ROC on real estate dataset

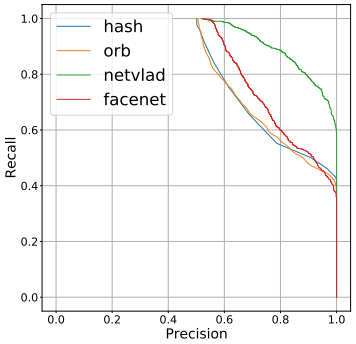

Figure 4.11: Baseline candidates - PR curve on real estate dataset

two randomly chosen images. There are hardly any negative samples above the value of 0.7 and therefore some of the positive samples can be separated using this threshold. However, most of the positive sample scores are overlapping with a distribution of negative samples.

The scores of ORB form a positively skewed distribution for both positive and negative samples. Almost 3000 out of 10000 positive pairs were in the first bin next to 0.0, which means hardly any features were matched between the images.

NetVLAD's and FaceNet's scores for negative samples form a normal distribution, while NetVLAD's distribution has a relatively lower standard deviation. Both can provide a threshold to precisely verify a large number of samples. NetVLAD is able to provide a threshold that would label most of the dataset correctly.

Baseline candidates' performance was also assessed on Real Estate Listings Dataset (RELD). The ROC curve is displayed on figure 4.10 and the PR curve is presented in figure 4.11. Real Estate Listings Dataset (RELD) contains a large number of samples, that are easy to identify using pHash algorithm. It is visible in figure 4.12, which shows the distribution of scores for each method on this dataset. These are images that only slightly different e.g. contain a watermark or are scaled. Overall, the results on this dataset confirm the results on ISD.

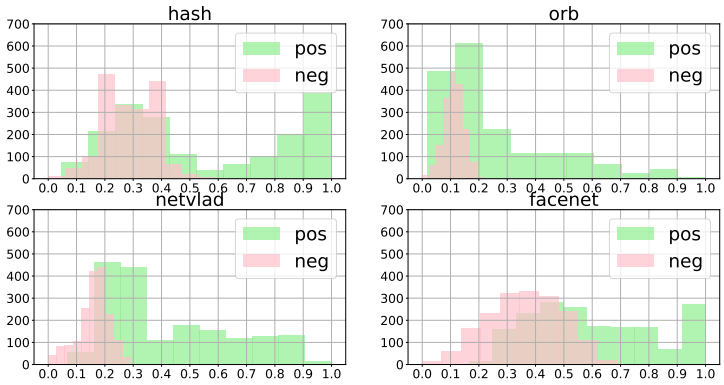

Figure 4.12: Baseline candidates - Histograms of scores on Real Estate Dataset

4.3 Experiments

The baseline candidates performance analysis in the section 4.2.2 points to NetVLAD as the most promising choice for generating comparable representations of indoor scene images and solving the ISV task. Therefore the results of the experiments in the last phase of the project will aim to gradually improve this approach and will be compared to NetVLAD as the baseline. All experiments carried out during this project use NetVLAD layer and try to gain performance improvement in the ISV task by:

1. using the new training dataset,
2. trying different dense descriptor extractor architecture,
3. changing loss function used during training.

The training data in all experiments consists of images from ScanNet dataset. Positive and negative image pairs are picked using a custom batch sampler. Each experiment is run 3 times with 10, 20 and 30 set as standard deviation parameter in the custom batch sampler. The diagram of the training process for each model is shown in figure 4.13.

The first experiment, named **InNetVLAD-v1**, aims to use NetVLAD deep neural network architecture and train it using images from ScanNet dataset. It is using pre-trained VGG16 [33] network as feature extractor and NetVLAD layer with randomly initialized weights. The weights are updated using online triplet loss [12]. During training, the images are sampled before each update

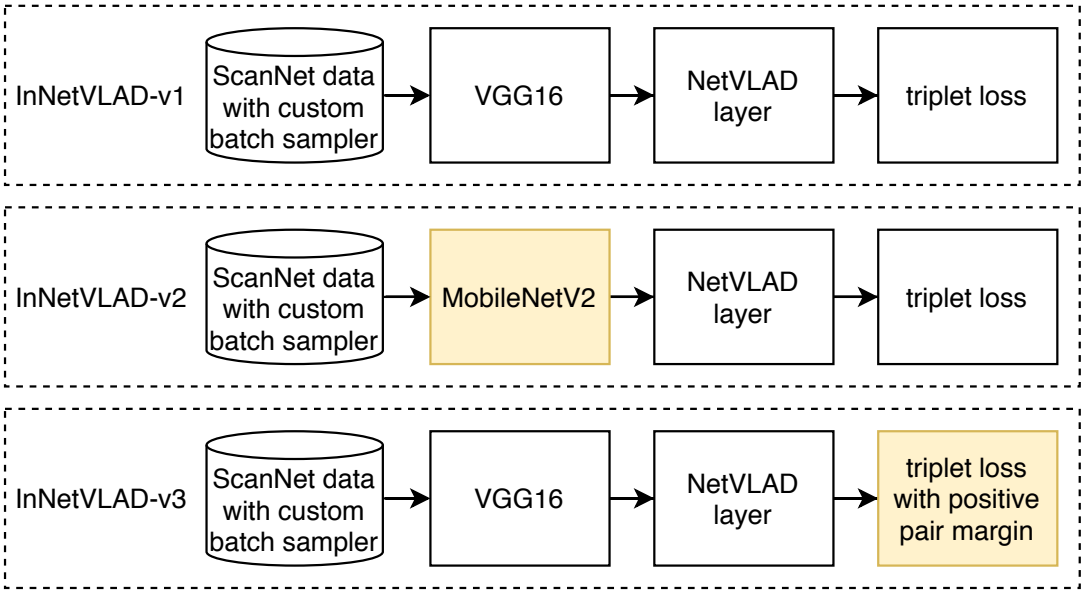

Figure 4.13: Setup of the experiments

iteration using the custom batch sampler. Before a batch of images is created, a pre-defined number of scenes is randomly sampled from the training set. For each scene, an image sequence index is selected at random. In the next step, a pre-defined number of image indices is sampled from a truncated normal distribution. The distribution is centred around the index selected initially. This is the same way of sampling as described in section 4.1.1 and shown in figure 4.6. Such images constitute one class presenting the same scene and their unique combinations are considered positive image pairs. For each combination, a semi-hard example is selected from all other classes given the current network state as described in [12]. A positive image pair combination and semi-hard negative image pair constitute a triplet used to compute the triplet loss.

The goal of the second experiment, called **InNetVLAD-v2**, is to test another deep neural network architecture as a feature extractor instead of VGG16. The choice of the architecture in this experiment is MobileNetV2 [36] as it contains a much lower number of trainable parameters than VGG16 and therefore should be easier and faster to train. Other training parameters remain the same as in InNetVLAD-v1.

InNetVLAD-v3 is the third experiment, which tests the impact of a new loss function on training. The distances between image embeddings are the

key to the success in verification task. It is important to not only ensure the different image classes are enough far away from each other, but also make the images belonging to the same class close enough to each other. In this experiment, a loss function that adds this condition to the simple triplet loss is used [37].

Chapter 5

Results and Analysis

In this chapter, section 5.1 contains the results of the experiments. Section 5.2 summarizes and reflects upon the results.

5.1 Evaluation of the experiments

In this section, the evaluation of all experiments is presented. The results on both indoor scan dataset (3 variants) and real estate dataset were generated for each model. They are analyzed and compared to the NetVLAD network, which serves as the baseline. The model designed in each experiment was trained 3 times on a different variant of the training dataset. The variants are using different standard deviation parameter to construct triplets in the custom batch sampler described in section 4.3. The standard deviation parameters used are 10, 20 and 30 and the experiments are named accordingly. The models were trained using early stopping with patience equal to 15, using AUC score on a dataset sampled from validation images. The model from the best epoch is later chosen for final evaluation. The custom batch sampler was set to generate 100 batches during each epoch. Each batch consisted of 75 images from indoor scan dataset (5 images sampled from 15 different scenes).

The setup of **InNetVLAD-v1** model is similar to the one described in [15]. The models are trained with plain triplet loss [12], Adam optimizer [38] and learning rate set to 0.000001. The models were initially trained with the higher learning rate, but it resulted in the model collapse after about 15 epochs (1500 batches). The ROC results are presented in figure 5.1. In terms of AUC all models outperform the NetVLAD baseline on indoor scan dataset. One should notice that in this experiment, models trained on datasets with higher standard deviation parameter achieved better results on all three

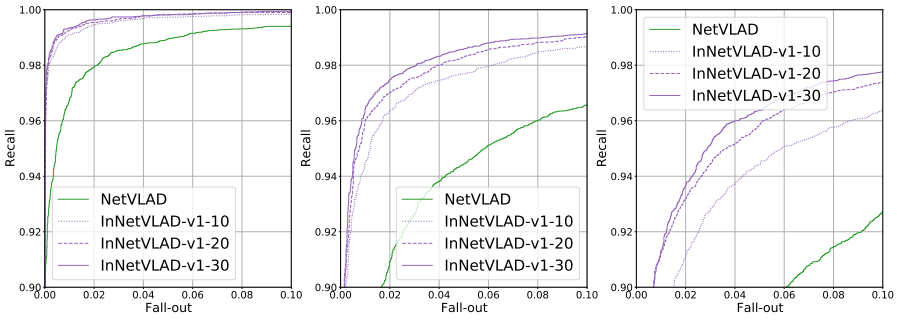

Figure 5.1: Results of InNetVLAD-v1 on indoor scan dataset

variants of the evaluation dataset. Therefore it is necessary to set the standard deviation parameter high enough so that the model is fed with both easy and hard examples. In case of the results on real estate dataset presented in figure 5.2, none of the models outperformed the baseline. Moreover, accuracy of all 3 variants was similar. This suggests that the trained models did not generalize well to other indoor scenes.

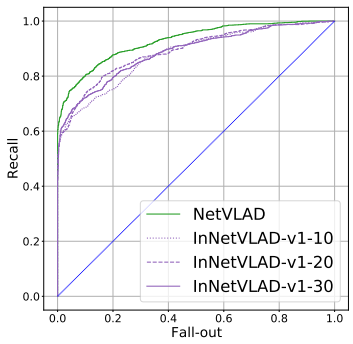

Figure 5.2: Results of InNetVLAD-v1 on real estate dataset

The experiment **InNetVLAD-v2** differs from the previous experiment only in terms of dense descriptor extractor used, which is MobileNetV2 instead of VGG16. As displayed in figure 5.3, the models outperformed the baseline in terms of AUC. However, the benefit of training with a higher standard deviation is not applicable in this case. The variant "10" achieved better score than variant "20" on easy and medium indoor scan dataset. Analyzing the

results on real estate dataset on figure 5.4 the variant "10" and "30" achieved also better AUC score than the baseline, while variant "20" achieved similar score. This shows that the architecture choice impacts the ability of the model to generalize to other data distributions. It suggests the previous model trained on VGG16 might have experienced overfitting.

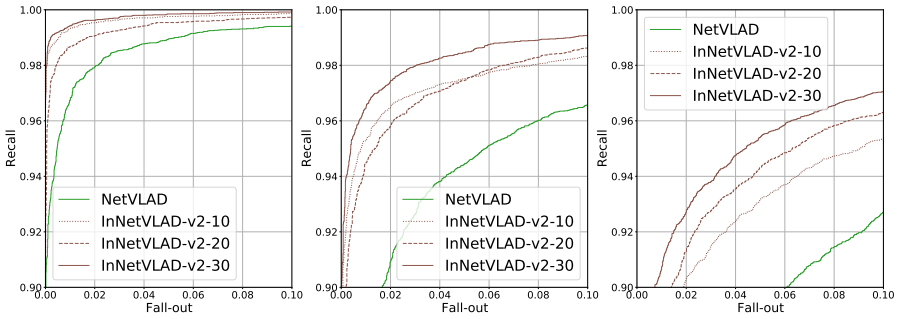

Figure 5.3: Results of InNetVLAD-v2 on indoor scan dataset

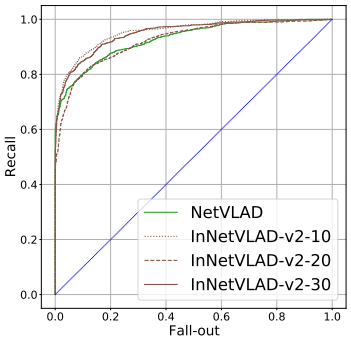

Figure 5.4: Results of InNetVLAD-v2 on real estate dataset

The variants of the third experiment, **InNetVLAD-v3** were trained in the same setup as the first experiment, except for the modified triplet loss. The chosen loss function did not stop the models from achieving better scores than the baseline on indoor scan dataset as seen in figure 5.5. However, the differences between the models trained on different variants of the dataset have vanished or changed. The best model on all 3 variants of the evaluation dataset is "20", followed by "30" and "10", but the differences between them are less

visible than in the case of InNetVLAD-v1. The results on real estate dataset displayed in figure 5.6 are not suggesting, that InNetVLAD-v3 provides any improvement over the baseline.

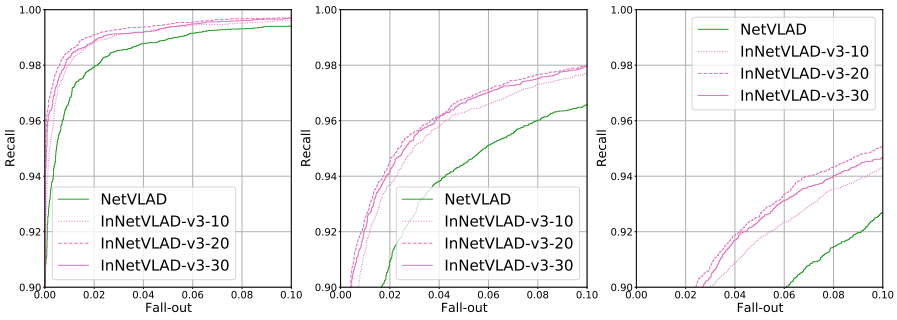

Figure 5.5: Results of InNetVLAD-v3 on indoor scan dataset

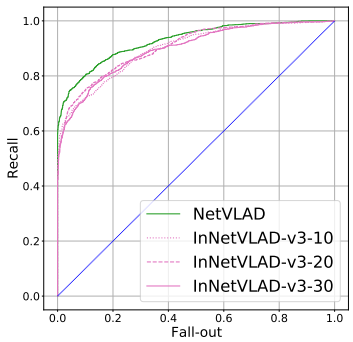

Figure 5.6: Results of InNetVLAD-v3 on real estate dataset

5.2 Discussion

In order to visually compare experiments trained on the same dataset, variant "30" was chosen for comparison on figures 5.7 and 5.8. Table 5.1 presents AUC results of all experiments and the baseline on both datasets: ISD and RELD. Two experiments demonstrate improvement over the baseline on RELD: InNetVLAD-v2-10 and InNetVLAD-v2-30. They prove it is possible

to solve the ISV task in other domains by generalizing from indoor scan images from ISD. It was anticipated that MobileNetV2 architecture used in this experiment might be easier to train than VGG16, which is thanks to the lower number of trainable parameters. These experiments demonstrate that MobileNetV2 can be used with NetVLAD layer as a dense descriptor extractor for indoor scene verification task.

The results also show that all models exceed the performance of the baseline on ISD. However in the case of InNetVLAD-v1 and InNetVLAD-v3 the results are not confirmed on RELD, therefore it is assumed the model is overfitted to the indoor scan images and can not generalize well to other domains.

Experiment name	**ISD-10**	**ISD-20**	**ISD-30**	**ISD-mean**	**RELD**
NetVLAD	0.997	0.987	0.971	0.985	0.929
InNetVLAD-v1-10	0.999	0.994	0.985	0.993	0.881
InNetVLAD-v1-20	1.000	0.996	0.989	0.995	0.892
InNetVLAD-v1-30	1.000	0.997	0.991	0.996	0.887
InNetVLAD-v2-10	0.999	0.994	0.982	0.992	0.955
InNetVLAD-v2-20	0.999	0.995	0.987	0.994	0.924
InNetVLAD-v2-30	1.000	0.996	0.988	0.995	0.950
InNetVLAD-v3-10	0.999	0.991	0.978	0.989	0.901
InNetVLAD-v3-20	0.999	0.992	0.981	0.990	0.906
InNetVLAD-v3-30	0.999	0.991	0.979	0.990	0.898

Table 5.1: AUC results of all experiments

The modified triplet loss used in InNetVLAD-v3 enforced a distance between positive image pairs lower than or equal to the half of minimum distance between negative image pairs. While this setup is only better than the baseline on ISD dataset, it provides better performance than InNetVLAD-v1 on RELD. It suggests the modified triplet loss enables learning representations that are more useful for generalization to other domains. It would be interesting to explore the combined effect of MobileNetV2 and modified triplet loss in further research.

In experiments InNetVLAD-v1 and InNetVLAD-v2, the variant "30" was better or at least as good as other variants on ISD. This shows the influence of custom batch sampler's standard deviation parameter. The higher value of the parameter allowed more difficult image pairs to be "seen" by the network during the training process. One would it expect it enables learning more

useful representations, but this effect is not confirmed on RELD. Therefore increasing the standard deviation parameter does not support learning robust representations.

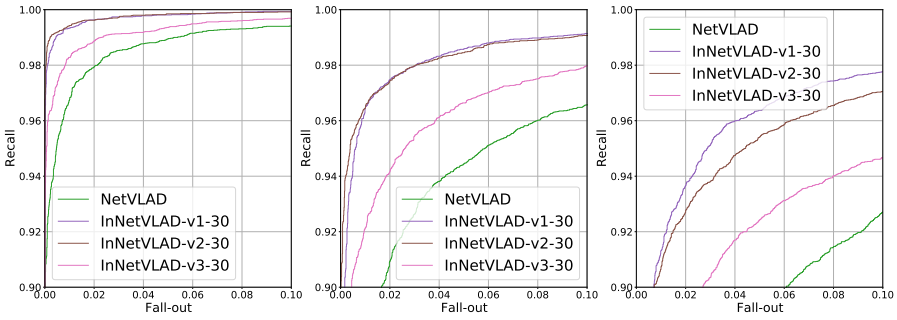

Figure 5.7: Comparison of the best experiments on Indoor Scan Dataset

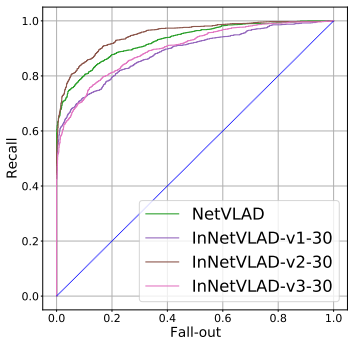

Figure 5.8: Comparison of the best experiments on Real Estate Dataset

Chapter 6

Conclusions and Future work

This chapter summarizes the degree project. The conclusions are presented, the limitations of the project and the positive and negative impacts are reflected upon.

6.1 Conclusions

The main goal of this thesis was to introduce the ISV problem and investigate ways to solve it by creating an evaluation framework and carrying out experiments. The problem was discussed theoretically by giving examples and explaining assumptions. It was supported by the literature study. Two evaluation datasets with different difficulty variants were built to ensure the results are valid in different data distributions. On both of them, four different existing methods for measuring the perceptual similarity of images were tested and the baseline was chosen. Finally, the experiments were carried out aiming to outperform the baseline. This project provides an assessment of the performance of different perceptual similarity measures on indoor scenes and proposes improvements.

6.2 Limitations

The main limitation of the project was very low availability of labelled data at large scale, which is a common problem when solving problems by training deep neural networks. ScanNet was the only dataset found that satisfied the requirements of the project. Increasing the number of images in real estate dataset by labelling the images from listing images proved to be too time-consuming. Another limitation was related to the characteristics of ScanNet

dataset used in this project. As mentioned in the section 4.1.1 the number of pictures between two images in the sequence does not imply the same similarity across scenes or even within the same scene. The camera might have been moved slower or faster by the operator and therefore affect the number of frames taken. Moreover, all images belonging to one scene were taken during a scan sequence within a couple of minutes, so recognizing changes in illumination throughout the day could not be learned using this dataset.

6.3 Future work

Firstly, the future work on the ISV problem should be focused on the development of bigger datasets. Indoor scan dataset created during this project consisted of images taken in 958 (training dataset) or 312 (validation dataset) different scenes. A bigger dataset with more variance might be the key to performance improvement.

Secondly, the camera pose data associated with each image in ScanNet dataset might be used to estimate whether two pictures in the dataset contain the same objects. If this is possible, such a distance metric might be used instead of frame distance used in this project and lead to building a more accurate dataset.

6.4 Reflections

This project should support anyone willing to work with indoor scene images data by describing challenges regarding perceptual similarity and guiding through methods for comparing the contents of such data. It should also support researchers aiming to explore the ISV problem further.

References

[1] B. Li, E. Chang, and Y. Wu, "Discovery of a perceptual distance function for measuring image similarity," *Multimedia Systems*, vol. 8, pp. 512–522, 2003.

[2] R. Datta, D. Joshi, J. Li, and J. Wang, "Image Retrieval: Ideas, Influences, and Trends of the New Age," *ACM Comput. Surv.*, vol. 40, 2008.

[3] H. Liu, R. Wang, S. Shan, and X. Chen, "Deep supervised hashing for fast image retrieval," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2064–2072.

[4] V. Monga and B. Evans, "Perceptual Image Hashing Via Feature Points: Performance Evaluation and Tradeoffs," *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, vol. 15, pp. 3452–65, 2006. doi: 10.1109/TIP.2006.881948

[5] A. Horé and D. Ziou, "Image Quality Metrics: PSNR vs. SSIM," in *2010 20th International Conference on Pattern Recognition*, 2010, pp. 2366–2369.

[6] R. Reisenhofer, S. Bosse, G. Kutyniok, and T. Wiegand, "A Haar wavelet-based perceptual similarity index for image quality assessment," *Signal Processing: Image Communication*, vol. 61, pp. 33–43, 2018, publisher: Elsevier.

[7] M. Chen, S. Wang, and L. Tian, "A High-precision Duplicate Image Deduplication Approach." *JCP*, vol. 8, no. 11, pp. 2768–2775, 2013.

[8] A. Stylianou, "Indoor Scene Localization to Fight Sex Trafficking in Hotels," 2016. doi: https://doi.org/10.7936/K7J38QX2

[9] D. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision*, vol. 60, pp. 91–, 2004. doi: 10.1023/B:VISI.0000029664.99615.94

[10] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu, "Learning Fine-Grained Image Similarity with Deep Ranking," *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014. doi: 10.1109/cvpr.2014.180. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2014.180

[11] R. Arandjelović and A. Zisserman, "DisLocation: Scalable Descriptor Distinctiveness for Location Recognition," in *Computer Vision – ACCV 2014*, D. Cremers, I. Reid, H. Saito, and M.-H. Yang, Eds. Cham: Springer International Publishing, 2015. ISBN 978-3-319-16817-3 pp. 188–204.

[12] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. doi: 10.1109/cvpr.2015.7298682. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2015.7298682

[13] A. Hermans, L. Beyer, and B. Leibe, "In Defense of the Triplet Loss for Person Re-Identification," *ArXiv*, vol. abs/1703.07737, 2017.

[14] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric," in *CVPR*, 2018.

[15] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN Architecture for Weakly Supervised Place Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, pp. 1437–1451, Jun. 2018. doi: 10.1109/tpami.2017.2711011. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2017.2711011

[16] H. Taira, M. Okutomi, T. Sattler, M. Cimpoi, M. Pollefeys, J. Sivic, T. Pajdla, and A. Torii, "InLoc: Indoor Visual Localization with Dense Matching and View Synthesis," in *CVPR*, 2018.

[17] F. Zhang, F. Duarte, R. Ma, D. Milioris, H. Lin, and C. Ratti, "Indoor Space Recognition using Deep Convolutional Neural Network: A Case Study at MIT Campus," *CoRR*, vol. abs/1610.02414, 2016. [Online]. Available: http://arxiv.org/abs/1610.02414

[18] F. Bidoia, M. Sabatelli, A. Shantia, M. A. Wiering, and L. Schomaker, "A Deep Convolutional Neural Network for Location Recognition and Geometry based Information," in *Proceedings of the 7th International Conference on Pattern Recognition Applications and Methods, ICPRAM 2018, Funchal, Madeira - Portugal, January 16-18, 2018*, M. D. Marsico, G. S. d. Baja, and A. L. N. Fred, Eds. SciTePress, 2018. doi: 10.5220/0006542200270036 pp. 27–36. [Online]. Available: https://doi.org/10.5220/0006542200270036

[19] "Sonarhome - main website." [Online]. Available: https://sonarhome.pl

[20] Matthew Hughes, "This travel site taps AI for cheap vacation rentals," Sep. 2017. [Online]. Available: https://thenextweb.com/artificial-intelligence/2017/09/26/this-travel-site-taps-ai-for-cheap-vacation-rentals/

[21] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "ScanNet: Richly-annotated 3D Reconstructions of Indoor Scenes," in *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017.

[22] Q. Huynh-Thu and M. Ghanbari, "The accuracy of PSNR in predicting video quality for different video scenes and frame rates," *Telecommunication Systems*, vol. 49, no. 1, pp. 35–48, Jan. 2012. doi: 10.1007/s11235-010-9351-x. [Online]. Available: https://doi.org/10.1007/s11235-010-9351-x

[23] Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[24] X. Wan and C.-C. J. Kuo, "Color distribution analysis and quantization for image retrieval," in *Storage and Retrieval for Still Image and Video Databases IV*, I. K. Sethi and R. C. Jain, Eds., vol. 2670. SPIE, 1996. doi: 10.1117/12.234782 pp. 8 – 16, backup Publisher: International Society for Optics and Photonics. [Online]. Available: https://doi.org/10.1117/12.234782

[25] S. Lee, J. Xin, and S. Westland, "Evaluation of image similarity by histogram intersection," *Color Research & Application: Endorsed by Inter-Society Color Council, The Colour Group (Great Britain), Canadian Society for Color, Color Science Association of Japan, Dutch*

Society for the Study of Color, The Swedish Colour Centre Foundation, Colour Society of Australia, Centre Français de la Couleur, vol. 30, no. 4, pp. 265–274, 2005, publisher: Wiley Online Library.

[26] Sivic and Zisserman, "Video Google: a text retrieval approach to object matching in videos," in *Proceedings Ninth IEEE International Conference on Computer Vision*, 2003, pp. 1470–1477 vol.2.

[27] M. Toews and W. Wells, "SIFT-Rank: Ordinal description for invariant feature correspondence," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 172–177.

[28] C. Zauner, "Implementation and benchmarking of perceptual image hash functions," 2010, publisher: na.

[29] D. J. Oftedal, "DifferenceHash." [Online]. Available: http://01101001.net/DifferenceHash.py

[30] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010, pp. 3304–3311.

[31] F. Walch, C. Hazirbas, L. Leal-Taixé, T. Sattler, S. Hilsenbeck, and D. Cremers, "Image-based localization using LSTMs for structured feature correlation," *arXiv:1611.07890 [cs]*, Aug. 2017, arXiv: 1611.07890. [Online]. Available: http://arxiv.org/abs/1611.07890

[32] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: an efficient alternative to SIFT or SURF," in *Proceedings of the IEEE International Conference on Computer Vision*, 2011. doi: 10.1109/ICCV.2011.6126544 pp. 2564–2571.

[33] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *international conference on learning representations*, 2015.

[34] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105. [Online]. Available: http://papers.nips.cc/paper/

4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

[35] A. Torii, J. Sivic, M. Okutomi, and T. Pajdla, "Visual Place Recognition with Repetitive Structures," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 11, pp. 2346–2359, 2015. doi: 10.1109/TPAMI.2015.2409868

[36] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510–4520.

[37] K. Ho, J. Keuper (Fehr), F.-J. Pfreundt, and M. Keuper, "Learning Embeddings for Image Clustering: An Empirical Study of Triplet Loss Approaches," 2020.

[38] D. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *International Conference on Learning Representations*, 2014.

For DIVA

{
"Author1": { "name": "Filip Finfando"},
"Degree": {"Educational program": "Master's Programme, ICT Innovation, 120 credits"},
"Title": {
"Main title": "Indoor scene verification",
"Subtitle": "Evaluation of indoor scene representations for the purpose of location verification",
"Language": "eng" },
"Alternative title": {
"Main title": "Verifiering av inomhusbilder",
"Subtitle": "Bedömning av en inomhusbilder framställda i syfte att genomföra platsverifiering",
"Language": "swe"
},
"Supervisor1": { "name": "Ying Liu" },
"Examiner": {
"name": "Amir Payberah",
"organisation": {"L1": "School of Electrical Engineering and Computer Science" }
},
"Other information": {
"Year": "2020", "Number of pages": "xiii,39"}
}

TRITA TRITA-XXXX