T.C.

ALTINBAS UNIVERSITY

Institute of Graduate Studies

Information Technologies

CONTENT BASED IMAGE RETRIEVAL WITH HIGH LEVEL SEMANTICS

Najm Abdullah Hussein ALMOHAMMED

Master of Science

Supervisor

Asst. Prof. Dr. Abdullahi Abdu IBRAHIM

Istanbul, 2021

CONTENT BASED IMAGE RETRIEVAL WITH HIGH LEVEL SEMANTICS

by

Najm Abdullah Hussein ALMOHAMMED

Information Technologies

Submitted to the Institute of Graduate Studies

in partial fulfillment of the requirements for the degree of

Master of Science

ALTINBAŞ UNIVERSITY

2021

The thesis titled **"**CONTENT BASED IMAGE RETRIEVAL WITH HIGH LEVEL SEMANTICSL**"** prepared and presented by Najm Abdullah Hussein ALMOHAMMED was accepted as a Master of Science Thesis in Information Technologies.

Asst. Prof. Dr. Abdullahi Abdu IBRAHIM

Supervisor

Thesis Defense Jury Members:

Asst. Prof. Dr. Abdullahi Abdu IBRAHIM	School of Engineering Architecture, Altinbas University	_____
Asst. Prof. Dr. Timur INAN	School of Engineering Architecture, Altinbas University	_____
Asst. Prof. Dr. Tareq MOHAMMED	College of information Technology, Imam Ja'afer Al-sadiq University	_____

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Science.

Approval Date of Institute of Graduate Studies: ____/____/____

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Najm Abdullah Hussein ALMOHAMMED

DEDICATION

I would like to thank Allah Almighty for the power of mind, health, strength, guidance, knowledge and skills to complete this study.

This thesis is wholeheartedly dedicated to my parents. There are no words to describe what you mean to me; there is nothing that I can repay for what you have done to me. I will continue to do my best to achieve your expectations.

Lastly, I dedicated this to my wife, brother and sisters, relatives, and friends who have been encouraging me during this study.

ACKNOWLEDGEMENTS

I would like to thank my supervisor Asst. Prof. Dr. Abdullahi Abdu IBRAHIM for all support during my study. Its great pleasure to express my deepest gratitude to my friends who have shared with me best moments during my study for the Master degree.

ABSTRACT

CONTENT BASED IMAGE RETRIEVAL WITH HIGH LEVEL SEMANTICS

ALMOHAMMED, Najm Abdullah Hussein,

M.Sc., Information Technologies, Altınbaş University,

Supervisor: Asst. Prof. Dr. Abdullahi Abdu IBRAHIM

Date: 09/2021

Pages: 55

This master's thesis focuses on Content Based Image Retrieval (CBIR) with high level semantics using image processing and deep learning techniques. The development of image edge smoothening system using Convolutional Neural Network (CNN) for CBIR optimization task is in hot pursuit, with special attention being given to the smoothening of all the edges of image. Given its high propensity to meta-size, going hand in hand with severe decreases in preservation rates, and the high inter-edge variability in image appearance, as well as a strong requirement on the training of the physician properly de-noising an image can be considered a daunting task. The purpose of this research thesis is to use a deep learning and image processing pipeline for content based image retrieval for the segmentation of edges in an image using with hybrid techniques in deep learning and imaging. The literature review of different papers was conducted with different imaging model architectures for CBIR. The CNN custom model was created for the task, and deep learning technique (CNN) was used with different levels of fine tuning of hybrid image processing techniques. Screening for high edge filter to identify edges at high accuracy has been under debate. In current discussion, the suggested smoothening procedure is bone density based including possible prescreening techniques for content based image retrieval and optimization on Corel-1000

dataset. Development of new prediction models and automated smoothening could contribute to general screening programs in the future. The custom deep learning model architectures were designed to represent different depths. The idea behind this is to analyze the effect of increasing representational capacity to the results and visualizations for all edges in an image. Additionally, deep learning CNN model was created to represent traditional automated image processing approach. Image processing has been used in some edging of images, producing good results for example in segmentation of image area structure and segmentation of image smoothening areas under consideration. The study also attempts to find solutions to practical deep learning challenges such as low training speed and lack of transparency with an accuracy of 98.47% absolutely. The imaging and deep learning pipelines are optimized in order to exploit the available parallelism using the MATLAB programming language with multiple tools under consideration.

Keywords: CBIR, Deep Learning, Segmentation, CNN, Semantics.

TABLE OF CONTENTS

Pages

LIST OF TABLES

Pages

LIST OF FIGURES

Pages

LIST OF ABBREVIATIONS

CBIR	:	CONTENT-BASED IMAGE RETRIEVAL
CNN	:	CONVOLUTIONAL NEURAL NETWORK
AIAS	:	ADVANCED IMAGING ASSISTANCE SYSTEMS
IPS	:	IMAGE PROCESSING SYSTEMS
IBC	:	IMAGE-BASED CONTROL
ISP	:	IMAGE SIGNAL PROCESSOR
DSP	:	DIGITAL SIGNAL PROCESSORS
HDR	:	HIGH DYNAMIC RANGE
IF	:	IMAGE RETRIEVAL
TM	:	TONE MAPPING
LDR	:	LOW DYNAMIC RANGES
MEF	:	MULTI-EXPOSURE FUSION
ROI	:	REGION-OF-INTEREST
RGB	:	RED GREEN BLUE
BEV	:	BIRD'S EYE VIEW
ISO	:	INTERNATIONAL ORGANIZATION FOR STANDARDIZATION
QOE	:	QUALITY OF EXPERIENCE
SVM	:	SUPPORT VECTOR MACHINE
ABC	:	ARTIFICIAL BEE COLONY ALGORITHM

RF : RECEPTIVE FIELDS

RNN : RECURRENT NEURAL NETWORK

1. INTRODUCTION

Content based image retrieval and edge smoothening is a trend that has been driven by the fact that human beings are known to be susceptible to filtering mistakes as mentioned in [1], whether it is image distractions from the direct environment. Although having a fully automated image analysis is still a challenge as described in [2], significant progress has been achieved by ongoing research in [3]; from the [4], the very first radio controlled image, in 2018, which incorporated computer vision, and to the analysis of today from most major image processing that possess numerous electronic aids such as collision avoidance, advanced imaging assistance systems (AIAS), and image processing systems (IPS). Those aids, which can be considered as a substantial step towards fully automated vehicles, use at least one camera to enhance safety. As part of a Sensing-Perception-Decision architecture [5] of autonomous driving systems, cameras are a fundamental sensing device, the data of which can be utilized in object recognition and tracking tasks that allow for proper path planning. With cameras being largely integrated into both mid and luxury-class vehicles [6], their annual growth rate is expected to be around 20% per vehicle by 2023 [7], playing a vital role in computer vision and, therefore, in image-based control (IBC).

As cameras are being increasingly adopted by the auto image industry, the urge for diverse functionality becomes highly relevant. Camera sensors need to be able to provide pleasant images for human display, as well as predictable and reliable images to be used by computer vision [7]. Those camera capabilities allow AIAS to be employed for vision-based perception tasks, such as object recognition, localization, and mapping, and path or motion planning [8], which are, in principle, IBC tasks and may enable a safer driving experience and bridge the gap towards fully autonomous driving.

As can be seen in Figure 1.1, an image colors may consist of sorted images, the image processing algorithm and the controller. A traditional image signal processor processes the RAW output of an image sensor and produces a compressed image that can be used by a vision application. Typical ISP stages include smoothening, de-noising, white balancing and color mapping, gamut mapping, tone mapping, and compression. Those stages, albeit standard, require excessive computational intensity to produce a high-quality output image, which in the case of computer vision applications is not necessarily essential [9] and can be approximated.

Approximate computing is a concept that relies on building systems with acceptable behavior from inexact hardware or software components. It allows to trade-off application accuracy in order to achieve considerable performance and energy gains [10] at design time. A key challenge in approximate computing is the identification of those sections either in hardware or software that can actually be approximated, as there is always a risk of crashing an application if a critical component is being approximated.

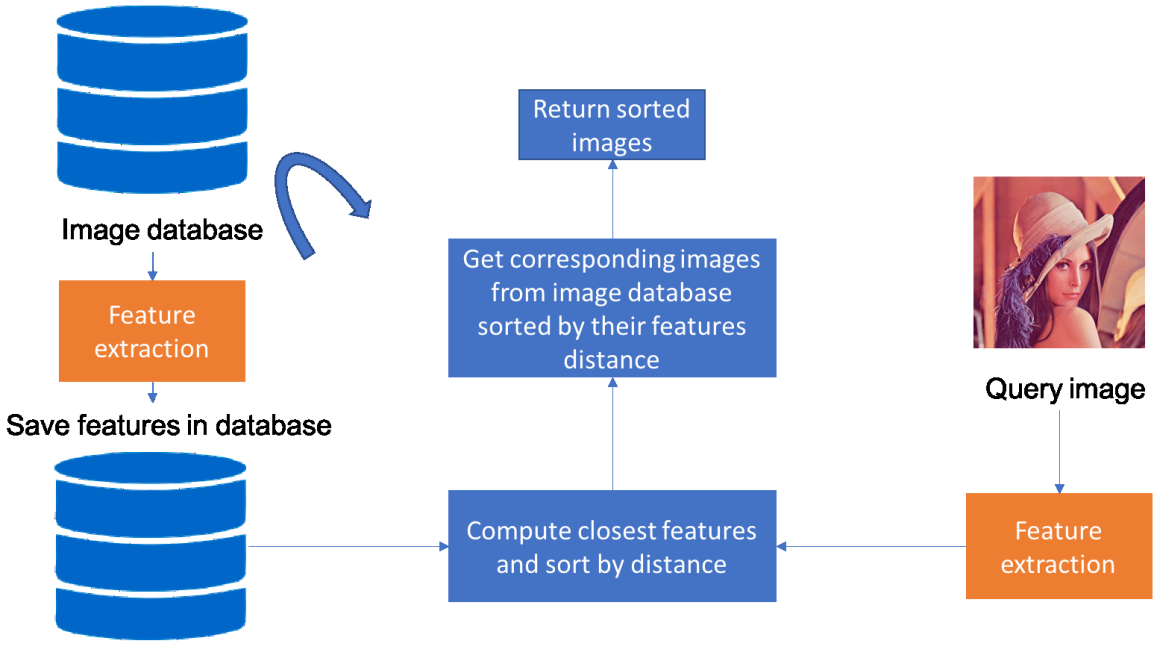

Figure 1.1: The basic flow of content based image retrieval with feature extraction [10].

1.1 MOTIVATION

A content based image retrieval uses multiple cameras for additional safety during navigation. A use case of a vision-based lateral control example using a single camera is studied in this project, where the camera output is being processed by an image processing application and the autonomous functionality is maintained by a controller that actuates based on the input from the camera. The camera produces 60 frames per second and the rest of the application needs to achieve real time performance in order to process all those frames as given in [10].

The camera is the sensor and is attached to the system. Each frame from the camera sensor goes through a series of processing stages. Initially, the frames need to be processed by an image signal processor (ISP), which converts the RAW output of the camera sensor to a format that is useful for the human and computer vision. The resulting image is, then, ready to be processed by the

image processing stage, which performs feature extraction and provides information about the image processing environment to the image-based controller as mentioned in [11]. The controller uses the visual feedback from the camera and actuates the required steering angle in order to allow the car to drive autonomously. In principle, a fundamental parameter in the control design is the sampling period. It depends on the duration the application needs in order to finish the required computations, from sensing to actuating. Typically, shorter sampling periods are necessary to achieve real time performance and maintain high quality-of-control.

The autonomous car application with the pipeline of stages can be seen in Figure 1.1. The ISP stage can be rather time consuming and is usually mapped onto specialized hardware, such as digital signal processors (DSP). Given its high computational complexity, we attempt to approximate the ISP, by coarsely skipping some of the ISP stages. Approximate computing is a technique that allows to improve the runtime performance of an error-resilient application, which is desired by the controller as mentioned in [12]. The error resilience of the application can be achieved in the image processing stage. When modified accordingly for the use case of lane detection, the image processing can handle the approximation, by successfully extracting features from the inaccurate images.

For the purpose of improving the quality-of-control of this image-based control application, seven different approximated ISP pipelines are developed. This allows to analyze the trade-off between runtime improvement and quality degradation with respect to quality-of-control for different degrees of approximation. The pipelines are optimized in order to exploit the available parallelism using the MATLAB programming language.

1.2 CONTENT BASED IMAGE RETRIEVAL

The scenes in the physical world include brusque lightening circumstances leading to highlights (over-exposed areas) or shadows (under-exposed areas) in images taken digitally. Conventional digital cameras usually fail capturing details in over as well as under exposed areas in the case when high or low exposure configurations are set in camera. To illustrate the scene as realistically as possible, some modifications are made on the hardware side, however, to reduce the costs, some are made on software side, where most of the problems were solved from this perspective. This enabled creating images from scenes in analogy to one captured via the eyes of human. Such software solutions, which are applied software-based high dynamic range (HDR) approach which

has been a major focus for various studies, in which a lot of efficient methods were created [13]. The major task of such approaches is that well-exposed images could be implemented through taking images' sequence with various levels of exposure taken via digital camera as input. Furthermore, synthesize output image which is extremely useful and perceptually attractive in comparison to input image.

Present HDR imaging methods could be divided to two major categories: Content based image retrieval (IF)-based and Tone Mapping (TM)-based approaches. With regard to TM-based approaches, HDR images are initially re-constructed by using particular HDR re-construction (HDR-R) approach applied over multiple low dynamic ranges (LDR) images regarding the same scene [14]. Therefore, TM-based approaches have two important stages, which are, HDR re-construction as well as tone mapping (HDR-R+TM). A lot of efficient TM algorithms were developed recently [15].

Distinguishing from TM-based approaches require generating intermediate HDR images, IF-based approaches have the task of directly obtaining everywhere well-exposed images through a process of merging complementary information related to the input LDR image and have the ability of skipping the process of HDR image construction. Therefore, IF-based approaches are typically more effective compared to TM-based approaches [16]. Such IF-based HDR approach is referred to as the multi-exposure fusion (MEF) as well, that was developed as a major subject of study in HDR imaging and content based image retrieval. In comparison to the general content based image retrieval task [17], MEF has certain properties due to the fact that over and under exposed regions in LDR images have certain distinguishing features like extreme luminance and saturation. Even though that certain some universal fusion approaches could be used to the MEF [18], it has considerable meaning for developing extra detailed algorithms for achieving optimum performance.

For instance, in [19] suggested block-based MEF approach through selecting block that contain maximal information. At the same time, in [20] developed influential exposure fusion approach on the basis of weight maps. With regard to their approach, weight map related to each one of the source images has been estimated through combining three quality measures that are contrast, saturation and well-exposedness at pixel level, which in addition to other quality measures, were selected in the comparison conducted in this work.

MEF fill a gap between the HDR in natural scene and low LDR images captured via traditional

digital camera [21]. MEF offer inexpensive alternative to bridge a gap between LDR display and HDR imaging [22]. The main aim of exposure fusion is obtaining scene's full dynamic range through a process of blending a number of exposure images in one high-quality composite image and keeping texture information and details as much as possible [23]. Since it has been initially developed in 2017 [22], MEF was a major field of study and its importance has elevated recently. Although there are several types of approaches proposed to solve the multi-exposure content based image retrieval problems. Most of them lack the efficiency of their outcomes, due to the inaccuracy of their parameters. Finding the optimal value of parameters is a critical issue in order to provide highly accurate results. With the existence of various MEF algorithms, it is very important comparing their performance of quality measures, for the purpose of finding the optimal best in addition to directions for more development. The quality measures are important factors that influence the performance of multi-exposure content based image retrieval as given in [24]. However, the problem of this work can be defined by identifying the quality measures families and feature groups that are highly relevant to the MEF.

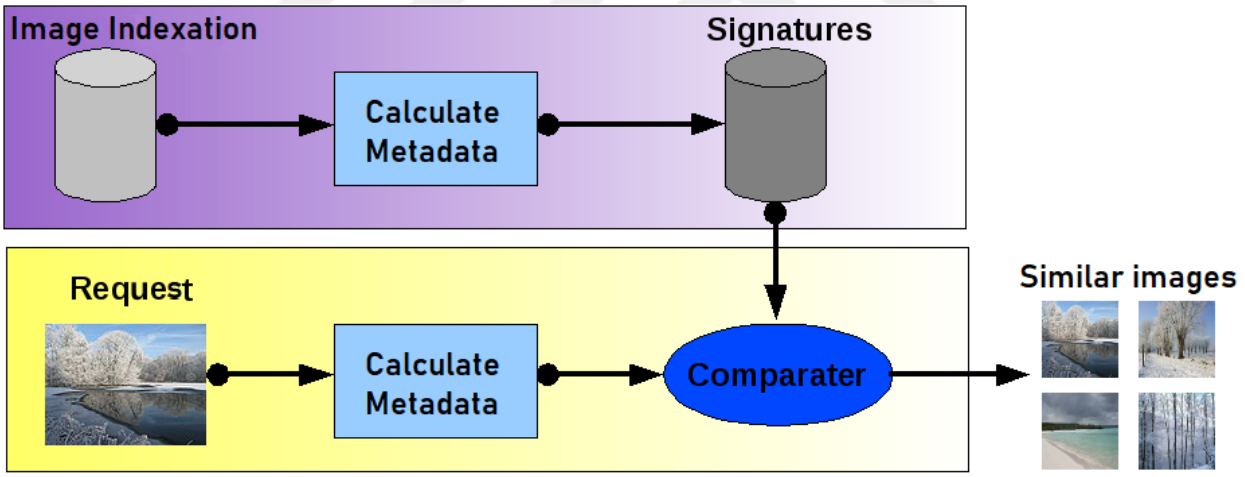

Figure 1.2: The content-based image retrieval techniques for keeping texture information and metadata details [24].

1.3 PROBLEM STATEMENT

The compute-intensive image processing stage of a content based image retrieval system result in longer sampling periods, which are considered during design time and negatively affect the quality of the underlying system. Usually the control design of data intensive feedback control applications relies on the worst case approach [25]. This includes image processing computations, the duration of which might vary, but is not taken into account. Researchers in [26] proposed to

deal with workload variations that are caused by image processing using a scenario-aware approach. Instead of designing the controller for the worst-case scenario, continuously adapt the sampling period of the controller on the actual case at run-time. In that way, the average sampling period is improved with respect to a WC design. However, this approach treats the image preprocessing pipeline as a black box and assumes a constant image preprocessing delay. The image processing step is required in order to process the output of the image sensor and extract useful information that can be used by the controller for decision making. We identified the impact that excessive computations in the image processing stage can have for real-time applications and proposed a feature-based detection algorithm, which only performs basic operations in order to identify the lane markers. In the preprocessing step, the RGB image is converted to gray-scale in order to select the region-of-interest (ROI) and extract two binary images, out of which one is used for global thresholding and the other for edge detection using the Sobel filter. Those two binaries are, then, combined using neighborhood-AND to generate the bird's eye view (BEV) of this image, and fusion identification is performed based on the maximum likelihood estimation on histogram plots. Although a feature-based lane detection approach such as the allows for increased flexibility for each one of the intermediate steps of the algorithm, it uses excessive steps, which may hinder the real-time performance. We followed a slightly different approach; using a front-mounted camera view instead of BEV for selecting the ROI in their algorithm, they attempted to minimize the steps needed by the algorithm. The preprocessing stage includes conversion from RGB to intensity image followed by an averaging filter to reduce noise. Then, the intensity image is thresholder in order to obtain the lane marks. Finally, for lane identification, the Hough transformation is used in combination with a horizontal differencing filter, which has the role of the edge detector. This approach achieves good performance, but is susceptible to the performance of the edge detection, since the Hough transformation requires accurately detected edges.

1.4 AIM OF CONTRIBUTION

The research in this works aims to contribute to the following key aspects of the use of deep learning and image processing for smoothening of edges in an image. The main contributions of this research include the following:

Integration of an image preprocessing pipeline and approximate computing in a baseline CNN algorithm for calculation of CBIR.

a. Characterization of accurate and approximate algorithms to identify the regions-of-interest for CBIR approximation.
b. Study and characterization of approximation choices of the imaging pipeline & computer vision algorithms with respect to edge smoothening with filters for CBIR.
c. Approximation of content based image retrieval without taking into account the timing impact.
d. Approximation of content based image retrieval taking into account the timing impact on sampling period.
e. Application profiling to obtain execution times for CBIR and to compute optimized sampling periods for CBIR.
f. Trade-off analysis between approximation and content based image retrieval for optimized sampling period designs.
g. Development of a tool-chain written in a high performance language (MATLAB) that allows the exploration of different approximations and their impact on quality-of-control.
h. Identify the highly performance quality measure approach.
i. Identify the optimal combination between the quality measure approaches that provide the best outcomes.

2. LITERATURE REVIEW

Approximate computing is a technique that leverages from the tolerance of applications to errors or inexact computations that reduce the quality in a controllable and acceptable manner. A large number of modern applications can tolerate those inexact computations and boost performance and energy efficiency. Software techniques such as loop perforation, memorization, precision scaling, task dropping, and data sampling or hardware techniques such as over scaling, clock over-gating, body-biasing and refreshing rate, may yield benefits of possibly up to 80% in execution time and analogous improvements in energy efficiency. In the case of content based image retrieval systems that are used in the context of content based image retrieval and smoothening, the obvious bottom line is to avoid crashing the controlled vehicle. Thus, the output of an approximate computing algorithm needs to be carefully monitored. There are different approximation strategies available, either for software-based approximation, such as precision scaling and loop perforation, or hardware-based approximation, such as using inexact hardware. A common denominator of those approaches is that they are application-specific.

2.1 CONTENT IMAGE QEURY

Load all images and read them one by one. All dimensions of each image read by the MATLAB software. In this work, eight multi-exposure sequences of images five of each one has been utilized for testing the efficiency or performance of the proposed different CNN operators of quality measure. As described in Figures below, this test set of image sequences captured with commercial cameras has been divided into many types of scene as Texture, text of books, moving object, and adding noise to image set such as Gaussian blur and salt and pepper noises. Furthermore, three most basic tools available to controlling the exposure of the scene. These tools are shutter speed, aperture, and International Organization for Standardization (ISO), also known as the "Exposure Triangle". These make a control on the brightness or darkness of the photo as mentioned in [27]. Images under various exposures in all sequence. More details concerning every one of the sequences, like name, number of source images, and spatial resolution have been stated in the caption of figures. Amongst the 8 sequences of test, two of them have been considered in the moving object scenes and the others have been considered in the static scenes. The levels of exposure are set manually according to the camera tools as given in [28]. For uncontrolled outdoors environment, moving objects such as, trees moving due to wind, make acquiring well-aligned

sequences quite a challenge. However, all the images in each dataset were registered in order to align the frames to be in the same angle and direction, so the content of each image will not vary according to the used measure. Therefore, this will be not variable in our experiments.

Figure 2.1: The content-based image query for segmentation [28].

2.2 QUALITY MEASURE OPERATORS FOR CBIR

In the presented section we will discuss a quality measures used in this comparison. Moreover, their definitions and mathematical equations will be presented.

2.2.1 Entropy

Researchers in [29] propose an algorithm that uses the line structure of the document to classify it. Their first step involves removing all characters, since their information is not used. This is done by thinning out all structures, followed by extraction and clustering of features. Finally, all clusters that belong to characters are removed, leaving an image that consists solely of lines. Next, three matrices are generated. The first contains the line intersections, where each vertical line is represented as a column, and each horizontal one as a row. When two lines intersect, the

first matrix contains a number at the appropriate location, representing the intersection type. The other two matrices encode the distances of the lines to each other, one matrix for horizontal and one for vertical lines.

2.2.2 Content Saturation

To remain scale invariant, the authors assign numbers to distances relative to the document size as mentioned in [30]. Matching a new document against a known one starts by calculating per element and matrix differences that must be above a threshold, which eliminates completely unfitting forms. Should the matrix sizes differ, a sub matrix of the larger one is used? A matching score that is derived from the similarity of the line crossing types and the distances, is calculated for the remaining candidates as mentioned in [31]. This also requires that the matrices are of the same size. Should this not be the case, the authors extract a sub matrix with the proper size that has the best match with the full version. The document type with the best matching score is the determined and returned document type.

2.2.3 Content Contrast

An approach that performs the form classification using only keywords and their location on the form. They extract all characters from the form and perform on content. Then a previously created keyword database is used to classify it. The database contains keywords and their locations, which are unique to each form type. A matching score is calculated by comparing the positions and string similarity as mentioned in [32]. Finally, the form type with the best match determines the result type. To counteract content inaccuracies, the authors match the strings by counting the number of insertions and deletions needed to transform the first string into the second. To retrieve the form contents, a list of regions is available for each form type. Those are the contents of interest which should be extracted. Each of these regions is thresholder separately to improve the content precision, and additional constrains, such as string composition suffixes, are used to extract the part of the region (excluding descriptive text) that is of interest for further processing.

2.2.4 Content Well-Exposedness

Over- and under-exposed pixels are image areas that are unwanted, due to the fact that the pixels in that area are not visible enough for showing the sufficient amount of information in the scene.

The measurement of well-exposedness eliminates this kind of pixels. This measurement is a Gauss filter, which is why, is performed the weighing of every one of the pixels according to the closeness of that pixel to 0.5. This measurement is important in the selection of a pixel via viewing its exposed value. In this approach, the map of intensity is utilized for getting a pixel's exposed value [33]. In order to get the energy, the Gaussian curve is applied for each channel separately and then the results are multiplied.

2.3 CONTENT ANALYSIS AND SORTING

Classification and content extraction from preprinted forms is also a closely related research topic. Instead of manual sorting and transcription of filled forms, a number of automated approaches that process scanned documents have been presented. Forms are templates that consist of text describing its purpose, captions for the blank regions, and tables or cells that must be filled in or are already preprinted. Examples include administrative forms, bank checks, payment slips, and many others. Since the use of forms is widespread, a number of different approaches have been developed over the years to automatically process them. In [34], an overview of the topic is given, including images of sample forms and descriptions of a number of popular approaches. The general layout and the small variability within the form types, where the only differences are the filled-in regions, are almost identical to our problem. In a large number of forms, the text must be entered in rectangular regions. Therefore, a number of approaches exist that exploit this fact, such as [35]. Their approach is designed to work with low quality scans that produce broken lines and noise. First, lines are detected and skew correction is performed. To correct errors introduced by broken lines and noise, the extracted cells are checked for inconsistencies under the assumption that the lines of cells and tables start and end in other lines of cells and tables. The consistency is checked using three rules in the following order: a line must end on another line. If this is not the case, the line is extended until it crosses another line. The outermost lines of tables must end in a line, otherwise they are also extended until they do. And finally, crossing points are only allowed in cell corners. The current structure is modified using these rules until no more rules are violated, where matching rules with higher precedence are preferred. The final result consists of the table and cell structure from which the values can be extracted.

2.4 ROLE OF DEEP LEARNING

With the extracted convolutional neural network (CNN) feature map, as researchers in [36] they computed the local visibility and maps of consistency for the determination of weight map for multi-exposure fusion. The CNN cannot generate the output of the MEF; rather than that, it exploits characteristics of pre-trained CNNs for the calculation of weight maps for simple and sufficient MEFs. Researchers in [37] proposes a computationally efficient algorithm to segment rectangles where corners may be missing or edges may be ragged. First, the input image is split recursively into smaller images, such that the resulting images only contain one object and little background. Next, they calculate the number of foreground and background pixels in the split images and the center of mass of the object, which is only an approximation of the real center because the objects are not perfectly rectangular. The approximate angles of the corners of the object are estimated by calculating the distance from the center to the borders. The authors note that the corners of rectangular objects appear periodically as mentioned in [38]. Therefore, their angle can be estimated from the calculated distance values and refined by applying the precondition that the objects must contain at least two parallel edges. Finally, to obtain the bounding rectangle, the authors grow a rectangle located at the already known center with the previously calculated angle as mentioned in [39]. The growing rectangle is split into four quadrants. Each of them is validated after every growing iteration with respect to the number of pixels not belonging to the object. They terminate if three quadrants are above a threshold. Researchers in [40] propose an algorithm to segment rectangular objects that may overlap and are placed on a lightly textured background with unknown color, which is the usual output of scanner preview images. The authors in [41] aims at detecting the background color, which is challenging when the image consists primarily of differently colored foreground objects, but it allows them to segment the objects more easily.

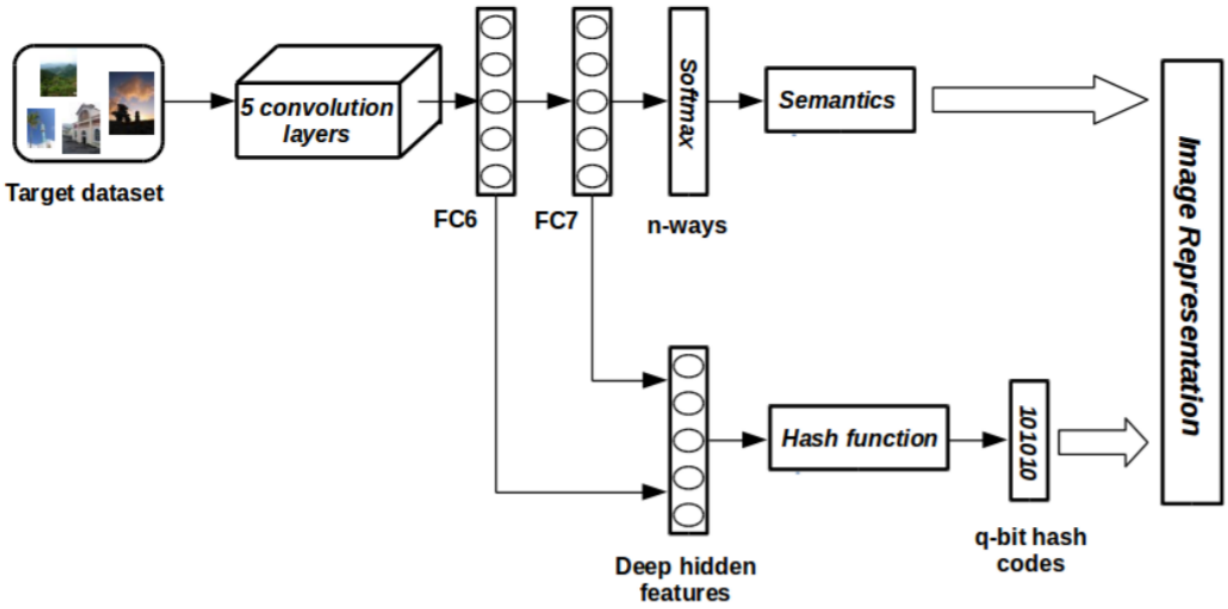

Figure 2.2: The content based image retrieval being processed through convolutional neural network [41].

To retrieve the background color, they first separate the image into line segments which have the same tint, by calculating the neighboring color differences. Under the assumption, that the background color segments are long, a voting scheme is used to extract possible background colors as mentioned in [42]. For each of the background color candidates, the edge strength at the boundaries to non-background colors are calculated. The authors observed that the gradient between background and object colors is larger than it is between foreground colors, and that the number of edge pixels correlates with the number of foreground objects. Therefore, they determine the true background color by only using line segments that exceed a certain length, are of the most frequent color, and have a large number of edge points with high gradient values as mentioned in [43].

2.5 QUALITY OF EXPERIENCE (QOE)-BASED CONTENT

Researchers in [44] have proposed an innovative approach of CNN based on two measurements of the perceptual quality, which are: the perceived local contrast and the saturation of color. For the sake of delivering maximal details of the image, they have modeled human visual system probability in the detection of local contrast based on physiological discoveries.

2.5.1 Content Perceived Local Contrast

From the content edge pixels, connected components are extracted and used to fit lines, with weights derived from the edge strength. Neighboring pixels, located in the direction of the line are added until a stopping criterion is reached. Orthogonal lines are used to calculate the corners of the objects, as well as the line segments which form the rectangle as mentioned in [45]. To eliminate wrong line segments, the median and mean color differences between foreground and background along the line segments are calculated and used to derive the parameters of a score function from a number of automatically generated images. Next, for each rectangle candidate, the same values used for line fitting and a score that relates to the number of background pixels in the rectangle are input into an SVM classifier, which determines if the rectangle candidate is a proper object or not. To find better matches, the rectangle candidates are also shifted in the local neighborhood. A candidate is accepted when at most 10% of the pixels inside the rectangle are classified as background as given in [46].

2.5.2 Content Psychometric Functions

In each iteration, the content selects a point randomly, performs the Hough voting, and removes it from the list of unprocessed points. If a cell in the accumulator exceeds the detection threshold, a line has been found. The pixels voting for the line are removed from the input, as well as the accumulator and the line is stored if it exceeds the minimum line length. The authors state that the line detection results are not satisfying, because of distortions from the camera, resulting in lines that are not perfectly straight. To detect such distorted lines, they propose to replace each pixel with a cross-shaped object. Furthermore, they define a mapping from a pixel to its corresponding cross structure and vice versa. This mapping is used to check line candidates only at their cross centers, as well as to remove the entire cross structure from the input list if it voted for an accepted line as mentioned in [47]. This is done because the surplus pixels would result in additional detections of the same line. The detected lines are grouped into rectangles by extracting parallel lines of the same length, which are then combined with orthogonal pairs to complete the rectangle.

2.5.3 Content Color Saturation

The employ of content color saturation as other measurement of quality. The definition of saturation has been incorporated in the color space of the LHS [48] for the sake of measuring how colorful a certain pixel. Researchers in [49] proposed a method for precise extraction of edge

region which is based on deep learning methodologies. To lessen noisy artifacts, the input image, after pre-processing, was applied to CNN. The CNN produced a segmentation mask which detected the area of the image edge. Using some post processing operations, the quality image of the mask is being further improved. Their input images were produced by customary cameras; henceforth, those were pre-processed in order to manage noisy artifacts. They used a filter to decrease the noise of the images as mentioned in [50]. In content analysis, all approaches are specifically designed to be fast and detect potentially imperfect rectangular objects with possible overlap or missing corners. However, they rely on uniform background color which also must be distinct from the majority of the content found in the foreground objects as mentioned in [51]. Both conditions are usually not satisfied in our application, as plates are mounted on arbitrary background, and unknown lighting conditions may cause significant brightness changes. Also, the color of the plate is not necessarily different from the background. Hence, the only way of distinguishing them from the background is by the border of the plate.

2.6 EXTRACTING THE CONTENT WITH HIGH-THROUGHPUT

An efficient rectangle detection system, intended to be used on mobile devices. First, they extract the edge map using the canny edge detector [52]. The threshold parameters are automatically selected. To improve the robustness and increase speed, all edges produced by text are removed. Consequently, a lower number of lines need to be processed. They do this by extracting connected components and calculate the bounding box for each one. The aspect ratio, height of the bounding box, and number of pixels in relation to the bounding area are used as filtering criterions. Next, the filtered image is used to extract lines by applying the Hough transform. Two lines form a line bundle if their angle does not differ more than fourteen degrees, which accounts for some amount of perspective distortion as mentioned in [53]. The authors then group line bundles to rectangle hypotheses, if all line intersections are inside the image. Finally, they compute the edge support on the dilated edge image, where the most dominant rectangle hypothesis with the highest edge support is the resulting rectangle. The edge map is dilated to increase robustness against camera distortions, causing lines that are not completely straight.

A more generic approach to extract rectangles using a windowed transform is proposed. They use a ring shaped sliding window, where the outer diameter approximately equals the size of the largest and the inner diameter the size of the smallest rectangle that can be detected. Next, the transform of the region under the sliding window is calculated, where the discretization depends

on the outer ring diameter as given in [54]. To detect the peaks in a robust manner, a modified butterfly evaluator is applied to the accumulator. To retrieve the lines, the accumulator is thresholder.

a. Using thresholding, foreground regions are identified. Thus, content based image retrieval distinguishes between pixels that belong to nuclei and background pixels.

b. Clumped nuclei are identified and segmented. For the segmentation, a seeded watershed algorithm is used. The user can choose to either apply it on the distance transform of the thresholder image or on pixel intensity values. Other traditional computer vision methods such as image smoothing are used to improve the segmentation results.

c. Nuclei whose features do not fall into user-specified ranges are discarded. For example, content based image retrieval offers to discard nuclei that are close to the image's boundary, or nuclei below a certain size as predicted in [55].

d. The segmentation quality of this approach is highly affected by the choice of parameters for the used algorithms. Thus, the user's input is required for choosing the parameters. The interface for the parameter selection of the module.

e. For parameter optimization, users usually start off with a given set of parameters and adapt these empirically based on the obtained segmentation. The user can then optimize the parameters in an iterative process as mentioned in [56]. As ground truth annotations are usually lacking, the user has no quantitative segmentation quality measure and is relying on visual inspection only.

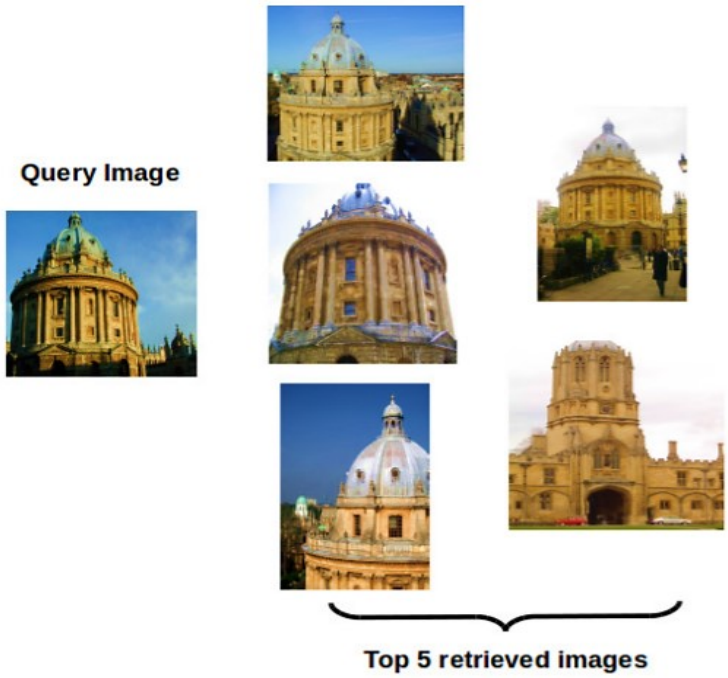

Figure 2.3: A robust retrieval of 5 images from the defined content image [56].

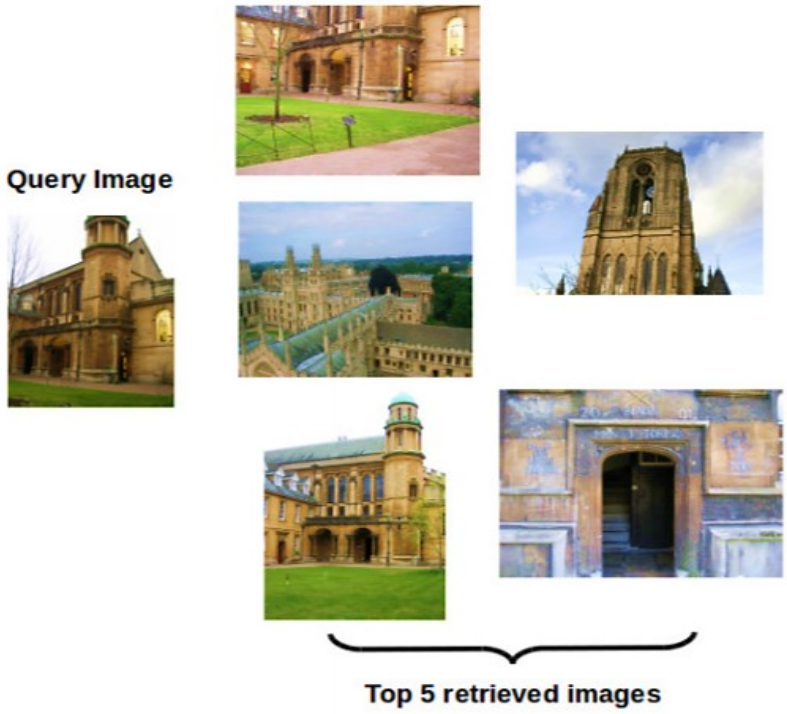

Figure 2.4: A robust retrieval of 5 images from the defined content image [56].

Under our assumption that the image consists mostly of the plate, this window size ensures that the individual pixel thresholds are adjusted to possible brightness changes, resulting from the illumination conditions with removing the noise. Yet, the size is large enough that regions in the image that contain no or very little structure are not larger than the averaging window. In such regions, bright pixels caused by noise or dirt would produce a very noisy thresholding result. The next step consists of morphological opening with a square 3 × 3 structuring element, which ensures that the plate is one connected object. The size of the structuring element is chosen to bridge small gaps caused by dirt or noise. Testing has shown, that a larger size would merge the plate with the background.

3. METHODOLOGY

It is notable that the image retrieval has successively used deeper networks but advanced deep learning techniques will provide great efficiency due to its architectural flexibility with artificial bee colony algorithm. Having more data models possess greater representational power making it possible to approximate more complex functions. Also, fast development in the field continuously introduces new techniques and architectures expanding the toolbox for design. However, science has always embraced simplicity, and it is reasonable to argue that the network should be kept as simple as possible. A study has been using the CNN technique even questioned the common practice of using pooling and sophisticated activation functions in machine learning. Instead, they suggest using just recurrent layers with properly selected stride and filter size parameters

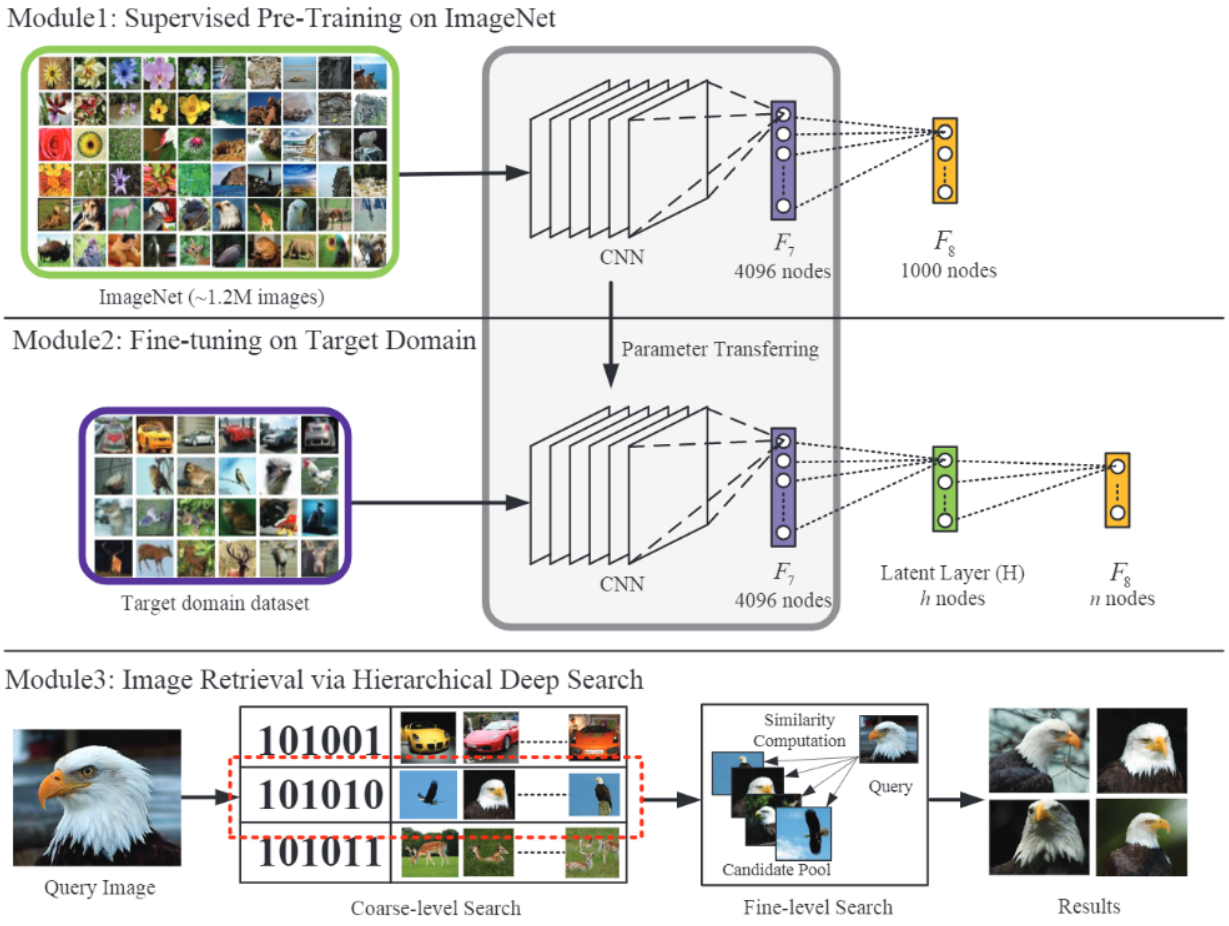

Figure 3.1: Proposed model depicting the whole content based image retrieval process being utilized in this research.

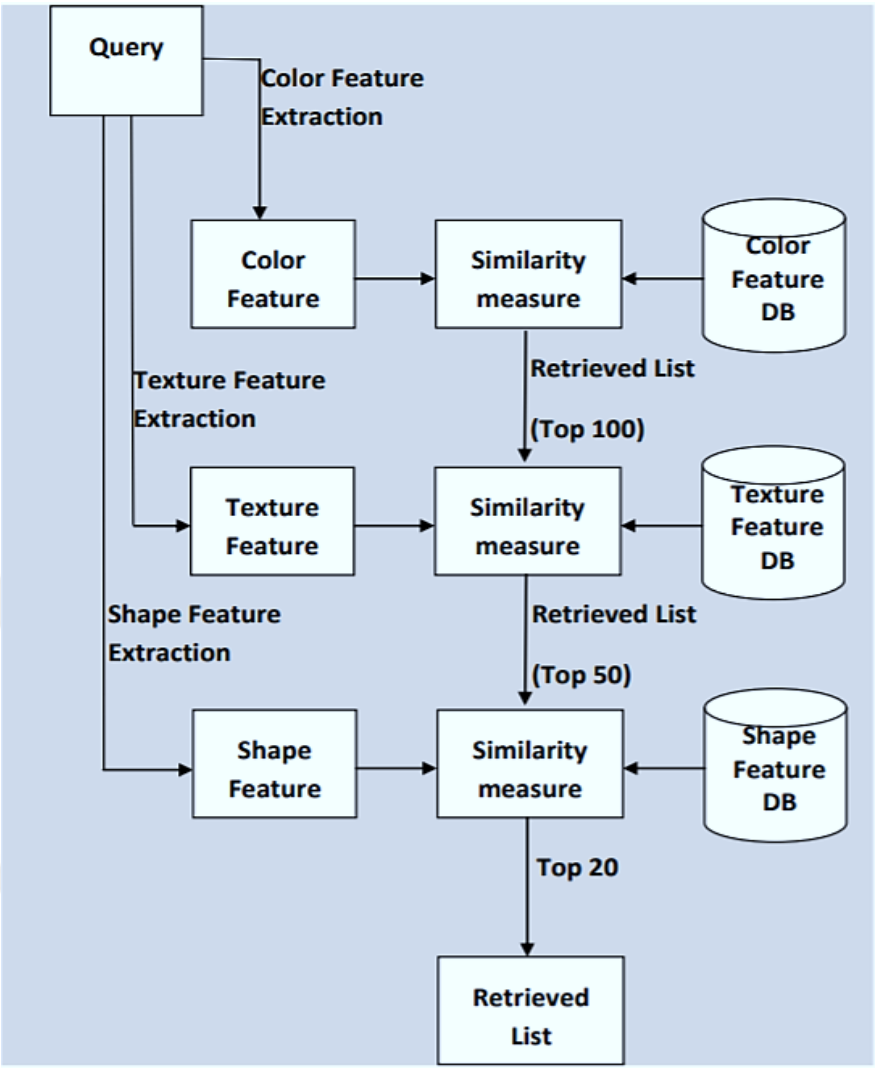

Figure 3.2: Proposed system architecture depicted by CBIR for this research.

3.1 DATASET DESCRIPTION

The dataset has been acquired from an open-source repository known as the Corel-1000 image retrieval dataset. The physical appearance (shape, size, texture, color, etc.) of objects in images can vary depending on many physical and human geographical factors, both from the ground perspective as well as in satellite imagery. Possible sources of variation include the cultivated species, plant status, topography, soil properties, weather conditions, cultivation methods and other human influences. Due to such local variations, the task of accurately differentiating between several field objects can become ambiguous. This makes it difficult to create robust algorithms that recognize all kinds of field instances without overfitting to the specific scene. In addition to the variation within a single scene, satellite images can show even stronger intra-class

variations. This applies to images from different agricultural regions, from different dates or growing seasons, or under different atmospheric and illumination conditions. Also, in view of increasing public availability of high quality geo datasets, the use of additional and more heterogeneous ground truth and satellite image data (e.g. from multiple study areas in parallel) could improve the general model accuracy, robustness and transferability. Additional data augmentation (e.g. random scaling, distortions, color offset or jitter, introduction of image noise) might also help mitigate overfitting and increase the robustness to different geographic settings as well as atmospheric and lightning conditions. More details concerning every one of the sequences, like name, number of source images, and spatial resolution have been stated in the caption of figures. Amongst the 8 sequences of test, two of them have been considered in the moving object scenes and the others have been considered in the static scenes. In addition to that, the first 5 sequences. The levels of exposure are set manually according to the camera tools. For uncontrolled outdoors environment, moving objects such as, trees moving due to wind, make acquiring well-aligned sequences quite a challenge. However, all the images in each dataset were registered in order to align the frames to be in the same angle and direction, so the content of each image will not vary according to the used measure. The dataset has been downloaded from Kaggle, link is also given by: https://www.kaggle.com/elkamel/corel-images

3.2 CONVOLUTIONAL NEURAL NETWORK

In this architecture, the CNN are responsible for feature extraction, where the features to which they will respond are determined through a learning process of the variable input connections. Simple features, such as specific orientations of lines and edges, will be extracted by lower-level CNN models, while higher-level CNN models extract more global features (e.g. parts of learned patterns). The CNN model receive their inputs through fixed connections from CNN models in the previous layers and are responsible for a robust pattern recognition (i.e. decreasing sensitivity to a deformation or location shift of a pattern). Each CNN model receives inputs from a number of CNN models (’the input window’) in one cell plane, i.e. from CNN models that extract the same feature, but at different spatial locations. The models will fire, if it receives an input from at least one of the models. Consequently, the feature will be detected even under a shift in the location of the input features, rendering the system less sensitive to the exact feature locations. The behavior of the models can however also be interpreted from an alternative point of view. The input windows for the different models strongly overlap thus models can be considered as performing a

spatial blur on the excitatory signals they receive from the models. This spatial pooling is obtained by averaging these signals from the input window, which are models that perceive the same feature at slightly different locations. Furthermore, the excitatory cell input window is often framed by a small inhibitory region. This enables the models to distinguish a continuous line from the end-point of the line and allows the network to still recognize features after blurring. A consequence of the fixed input windows of the models is a reduction in size of the CNN-planes, thus allowing the network to compress data as it progresses to higher stages.

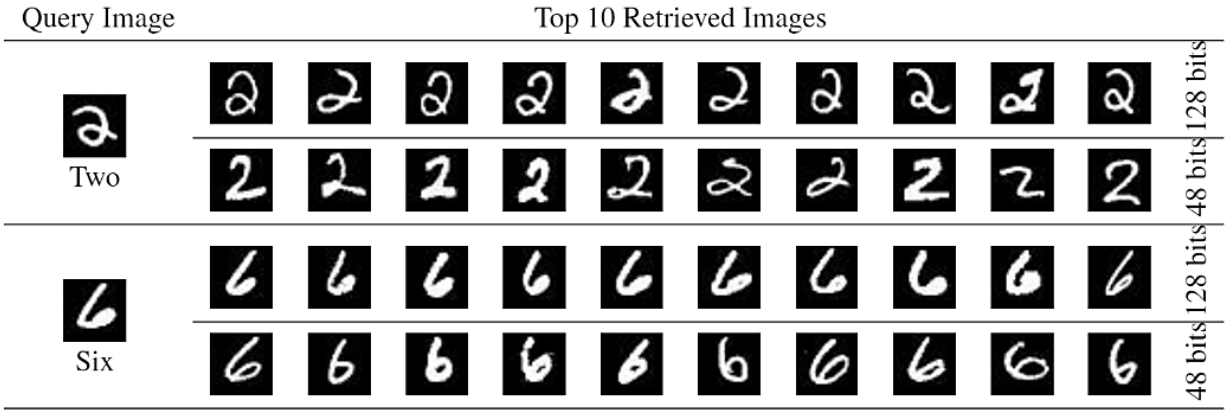

Figure 3.3: Illustration of a convolutional neural network from querying an image to retrieving an image.

CNN stands for convolutional neural networks originated from this work whose network has a similar architecture cognition. It translates the behavior of the models in mathematical formulations, respectively as convolutions and subsampling. The network primarily distinguishes itself however from the cognition in its training process. Whereas the neo-cognition relies on a layer-wise, unsupervised training process of the cell layers and supervised training of the output layer, CNN is trained to find a global minimum over all the parameters through back-propagation. The key notions that mimic biology and render CNN's so fusion for image smoothening.

Due to the high density of research in image illumination estimation, multi-exposure images often appears clumped. Thus, the main challenge for this research is to find the patch-wise solution for illumination estimation with human interference with CNN technique by separating different instances of multi-exposure fusion images. Traditional segmentation methods such as the seeded watershed algorithm, which is implemented in fusion images, are slow and sometimes make errors. In addition, the user needs to be experienced in order to use these algorithms as

parameter tuning is required. The goal of this research is to overcome these dominant problems in the patch-wise illumination estimation for multi-exposure image retrieval. We will use a

convolutional neural network for multi-exposure image retrieval images. We will train the neural network on hand-annotated images and evaluate its performance with object-based metrics.

3.3 POWERFUL CNN'S FOR IMAGE EDGE SMOOTHENING

3.3.1 Key concepts of CNN's

Convolutional layer as images contain highly correlated 2D information (generally in three RGB-channels), the idea of local connections has been exploited by creating units in a convolutional layer which receive weighted input from local patches in the previous layer, referred to as their receptive fields (RF's). By relying on local RF's, this introduces sparse connections, rather than the whole-scale image statistics passed on to units and causing high-dimensional problems in fully-connected feed-forward networks. The weight vector of a unit is also known as filter bank and through training the weights will be updated to filter for specific features, giving the units the behavior of filter detectors. The weighted inputs and additive bias are then passed through the activation function of said unit, where ReLU's have portrayed much faster learning in deep networks compared to other activation functions.

Furthermore, images tend to portray similar features at different spatial locations. Accordingly, units with receptive fields at different locations will have the same weight vectors to extract similar features, which is known as parameter sharing and further reduces the dimensionality of the problem.

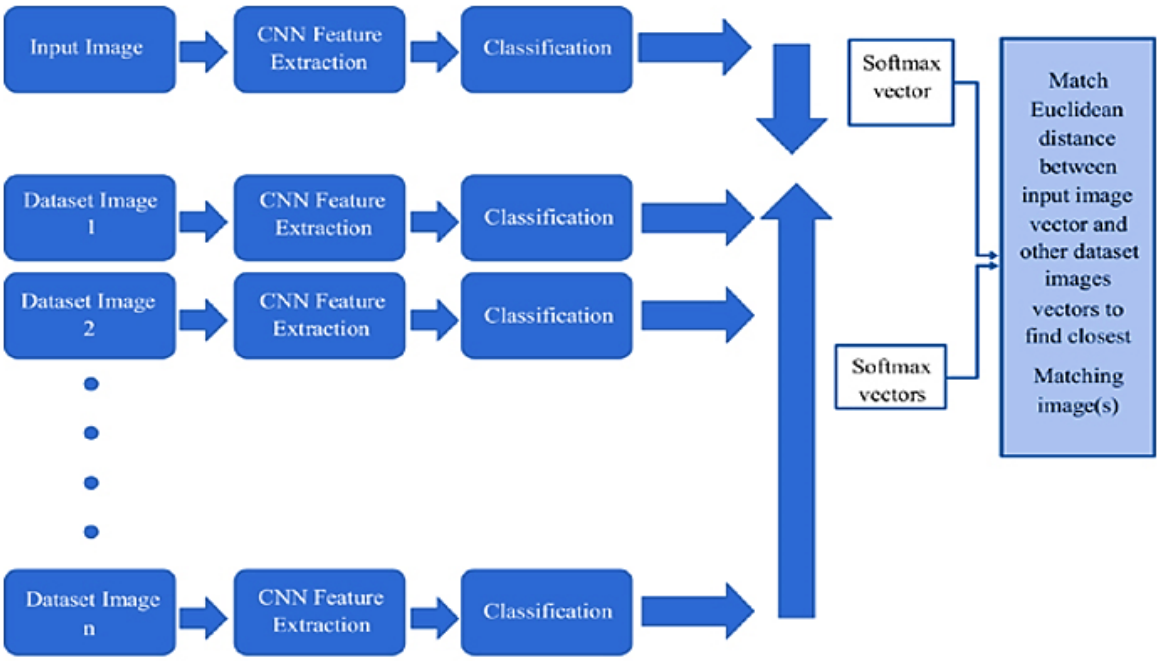

Figure 3.4: Basic building blocks of CBIR system using convolutional neural network.

On a structural level, the outputs of units with the same weight vector are organized in feature maps. Multiple feature maps (with different weight vectors) in each convolutional layer enable scanning for different features in each local image patch. The optimal number of feature maps needed to capture the image dynamics tends to increase with the complexity of the input images.

3.3.2 Pooling layer

The feature maps in the convolutional layer serve as input for the next stage, being a pooling layer. Given that the relative location of features with respect to each other contains significantly more information than their precise pixel-location, further stressed by the notion that small shifts in the precise feature locations are to be expected in different images, the idea of local pooling of features to reduce spatial resolution is implemented at this stage. This results in a network invariant to small shifts in the images and the stages of convolutional/pooling layers equip the network to handle variable size inputs.

Pooling units will scan the feature maps with (non-)overlapping p1 x p2 receptive fields and output a single value. Typical choices for pooling units are subsampling, where the output of a unit in the j^{th} layer is the average of units in the receptive field, supplemented with additive bias b and passed through an activation function and max pooling.

3.3.3 Deep hierarchies

Finally, the use of many layers represents the different stages in the simplified model of the ventral stream. Features at higher-level stages results from combinations of lower-level features, thus enabling deep structures to capture the compositional hierarchy of natural images.

3.4 TRAINING AND VALIDATION

An implementation is based on hybrid models of CNN algorithm from various approaches. The visualization method proved to be helpful in interpreting the predictions, although it produced different results than expected. The expectation was that the CNN model would learn to identify small signs of skin disease risk. Instead, with custom CNN models they learned wider patterns that possibly predict resistance to fractures. This can of course be seen as the other side of the same thing. The experiments show that deep learning techniques can be used with imaging techniques to produce predictive ability different from the standard clinical methods currently in use. However, due to limited amount of data and wide range of architectural possibilities, their true potential in this context remains to be unfolded. It is also likely that the data analysis alone cannot provide the best fracture risk prediction, regardless of the model. It could contribute to more comprehensive prediction method that utilizes a wider set of input data to cover different aspects of the image edge smoothening. Using hybrid deep learning techniques produced the best image edge smoothening prediction results in the experiments. The results support the advertised benefits of hybrid deep learning, such as faster training speed and suitability for larger datasets.

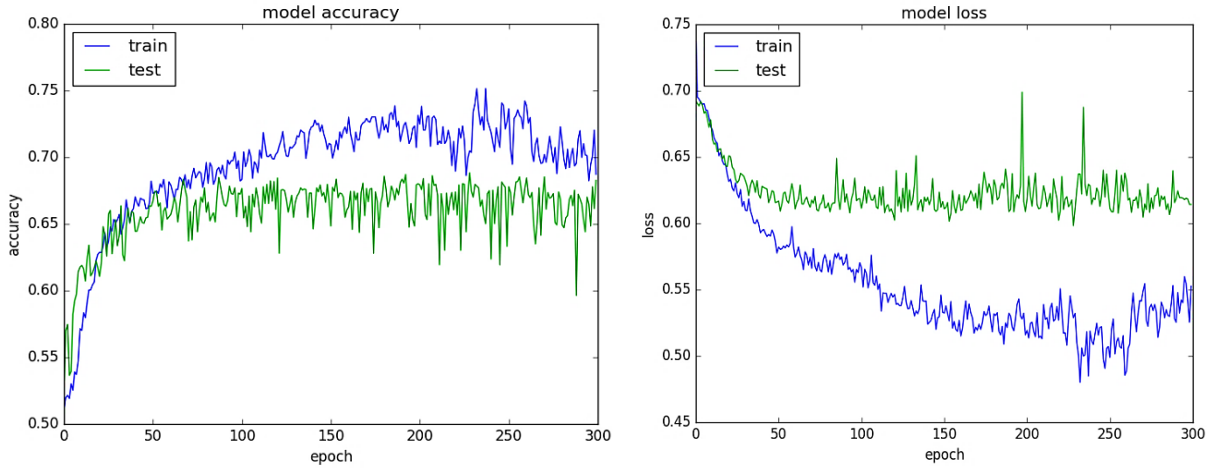

Figure 3.5: The ABC+CNN Model Accuracy (accuracy vs epochs) Model Loss (loss vs epochs).

Given the current hype of deep learning, optimizations of CNN architectures and designs is in hot pursuit. This section will therefore be devoted to some of the powerful CNN models that have been developed in the past few years and will be used throughout this work. Specific, revolutionary design-elements that have assisted these improvements are marked throughout this discourse.

3.4.1 Cross Validation

The cross validation is used to validate the data in artificial intelligence based systems, interpretable machine may even teach humans about how to make better decisions. To address the transparency problem, visualization techniques can be used to identify the image regions most important for the model's prediction. It is also possible to highlight the shapes and textures the network sees by visualizing pixel-wise impact on prediction score. Another visualization technique is called guided backpropagation. It produces an illustration of the detected features that contributed to the image edge smoothening prediction. This is done by taking account only the neurons that had a positive effect on the output, while others are set to zero. Combining different visualization techniques can produce effective visual explanations for the CNN model's predictions. To gain better coverage of the problem space, the number of training samples can be artificially expanded. This method called cross validation means modifying samples with random transformations such as rotation and flipping, varying properties like brightness and contrast or by adding random noise.

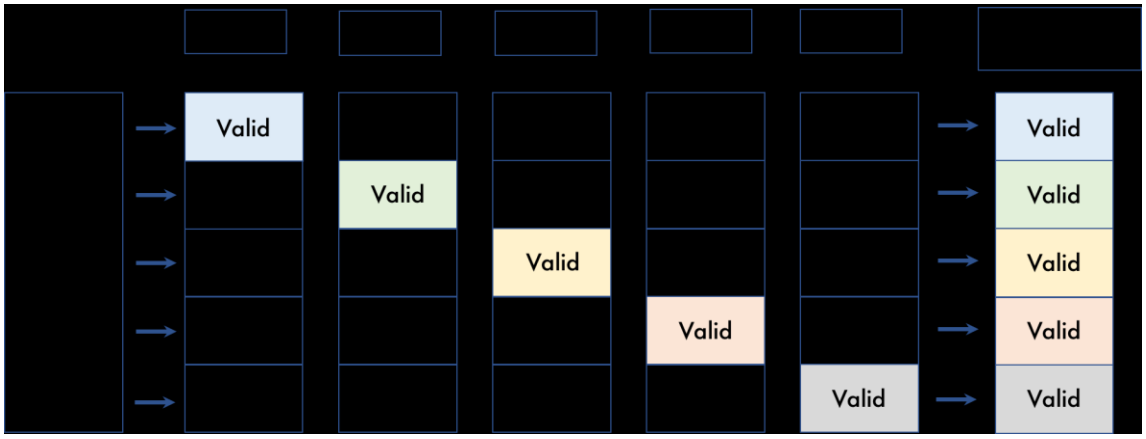

Figure 3.6: An example of dividing dataset for cross validation.

3.4.2 Training and Validation

As deep learning techniques take numerical data as input, image analysis might seem like a simple case of feeding the network with raw image data. After all, images are nothing but grids of numbers

representing pixel intensities. Actually, this approach can work well for simple recognition tasks with small, normalized images. The problem is poor scalability. When the input size and number of layers' increase, the amount of required computation explodes.

Table 3.1: The number of epochs per training instances.

Number of Epochs	Loss	Accuracy	Validation Loss	Validation Accuracy
25	0.3301	0.8492	0.6911	0.5360
50	0.2418	0.9047	0.6942	0.5360
75	0.1891	0.9350	0.7985	0.5360
100	0.1702	0.9426	0.9698	0.5379
125	0.1384	0.9606	0.4709	0.7803
150	0.1246	0.9625	0.3569	0.8220
175	0.1020	0.9768	0.3543	0.8295
200	0.0809	0.9829	0.3598	0.8371
250	0.0747	0.9853	0.3662	0.8295
300	0.0678	0.9867	0.3709	0.8314

3.5 IMAGE PRE-PROCESSING IN MATLAB

Possible sources of variation include the smoothening techniques, filter status, topography, filter properties, smoothening conditions, segmentation methods and other classification influences. Due to such local variations, the task of accurately differentiating between several field image filter can become ambiguous. This makes it difficult to create robust image processing algorithms that recognize all kinds of field instances without overfitting to the specific scene. In addition to the variation within a single scene, satellite images can show even stronger intra-class variations. This applies to images from different filter regions, from different dates or growing seasons, or under different atmospheric and illumination conditions. Also, in view of increasing public availability of high quality fusion image datasets, the use of additional and more heterogeneous ground truth and fusion image data (e.g. from multiple study areas in parallel) could improve the general model accuracy, robustness and transferability. Additional data augmentation (e.g. random scaling, distortions, color offset or jitter, introduction of image noise) might also help mitigate overfitting and increase the robustness to different fusion settings as well as image lightning

conditions. In the proceeding discussion, the image retrieval will serve as a guideline to discuss the most poignant results obtained with CNN's with indication of some remarkable new elements. To give an indication of the performance of these models, the top-5 error rate on the ImageNet dataset of the aforementioned nets is presented. It has to be noted that the submitted nets, training and testing schemes offer a myriad of other hyper-parameters, which were tuned to achieve these results (e.g. a 7-network ensemble of CNN-architectures).

The research focuses on compiling the delineate details and classifies image retrieval with edge smoothing using artificial bee colony algorithm and convolutional neural network while distinguishing between image instances of the same class. Instance segmentation lies at the intersection of edge detection (predicting the bounding box and class of a variable number of images, but not segmenting them) and semantic segmentation (labeling each image pixel with a semantic category label, but not distinguishing between edges of the same category). With semantic segmentation it would be possible to find single instances of CNN algorithm on fields if they were completely surrounded by areas of different image classes. These datasets usually image retrieval from manual tracing of field boundaries. However, manual tracing of image edge smoothening is heavily dependent on the used imagery and supplementary data. It also is a highly subjective task, inevitably leading to inaccuracies and ambiguities depending on the priorities of the operator. The ground truth dataset can be complemented by existing property parcel information that could be indistinguishable based on the imagery alone. Furthermore, a single field image edge smoothening can potentially show several subfields due to unreported land use changes within the growing season.

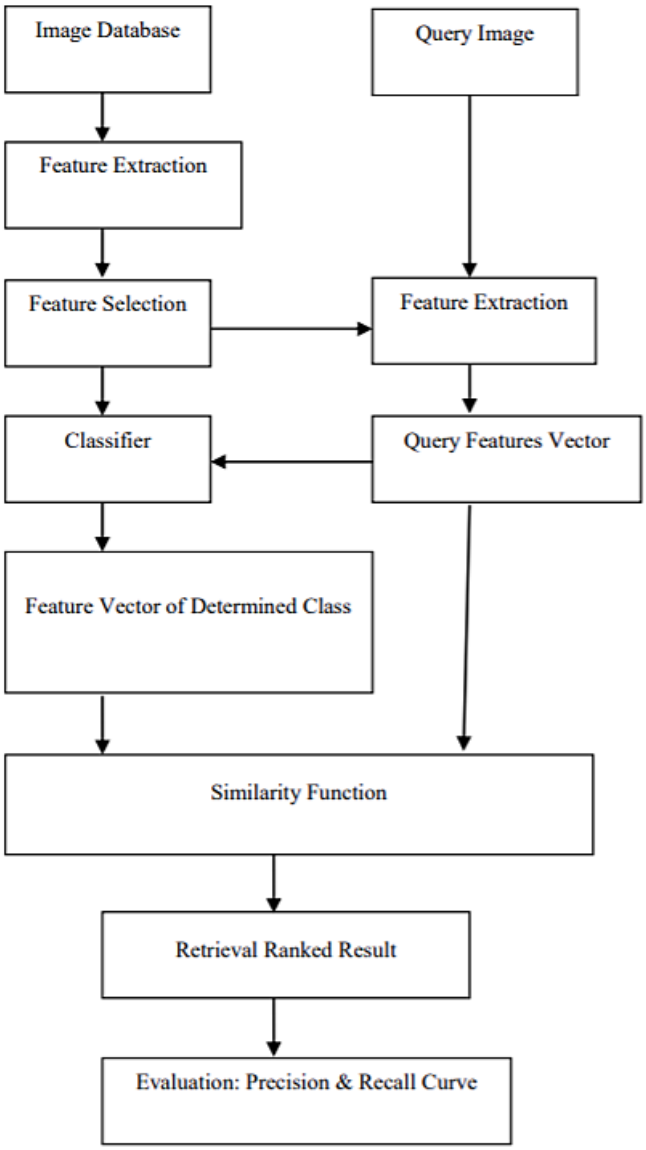

Figure 3.7: The flowchart approach for smoothened image by series of steps being followed.

Given these constraints, even a very powerful automated image processing algorithm can only become as good as the ground truth data but will never be in perfect alignment with reality. Although the model can generalize within certain limits (local generalization), its transferability is mainly limited by the variability of the training data. However, filter field are often directly adjacent to each other and have touching boundaries. In this case, edge segmentation would yield the image edge smoothening. Same filter instance segmentation is able to distinguish between these connected or overlapping instances of the same class. Deep learning has not reached mainstream adoption in the remote sensing community yet.

4. RESULTS

The goal of each image retrieval patch-wise illumination estimation experiment is to obtain a set of features for each image that describes its sequence. This set of features is called the image profile. Image profiles can be analyzed and compared to each other. In a patch-wise illumination estimation experiment for example, illumination estimation of images that were treated can be compared against profiles of images in the control set to quantify important matrix changes using CNN. For example, image patches can reveal an image sequence state or can be used for classification in image states such as phases of the image cycle or hematopoietic differentiation. Patch-wise illumination estimation using CNN can be of very different kinds. Examples include expression profiles that quantify the transcription of genes and morphological profiles that quantify the shape of the image and its compartments. Patch-wise illumination estimation is an important tool in morphological profiling and captures the images used to obtain morphological image profiles. In this experiment, the original datasets were used, without adding any further effects on them. Moreover, in order to produce the proposed image, a combination of all of the used quality measure were implemented together.

Figure 4.1: Sample of images from different semantic categories of the Corel-1000.

Furthermore, the training of the aforementioned deep CNNs generally required multiple GPU's and was a time-consuming process (up to 3 weeks). Thus, it is clear that training the same, high-performance CNN architectures on real-life Corel-1000 data is in most cases feasible. The CNNs trained to recognize the ImageNet dataset have nevertheless a strong capability to extract very distinctive features from natural images, rendering them useful for datasets of other images. This is supported by the idea that the lower levels of these architectures extract general features from the images and become increasingly more task-specific deeper in the network. From this the idea

of transfer learning sprung forth, which aims to kick start the process of the random weight initialization by using the pre-trained weights of deep CNN's to facilitate the training process. The CNN algorithm learning entails training a source network on a source dataset for a source task, transferring the learned knowledge (e.g. features or weights) to a new network for a different learning task and/or on a different target domain. The CNN learning is used with the objective of increasing performance on the target task by exploiting the trained knowledge of the source task. It can either serve as a 'Kick-starter' for the process improving initial performance, increase final performance or speed up the training of the target task.

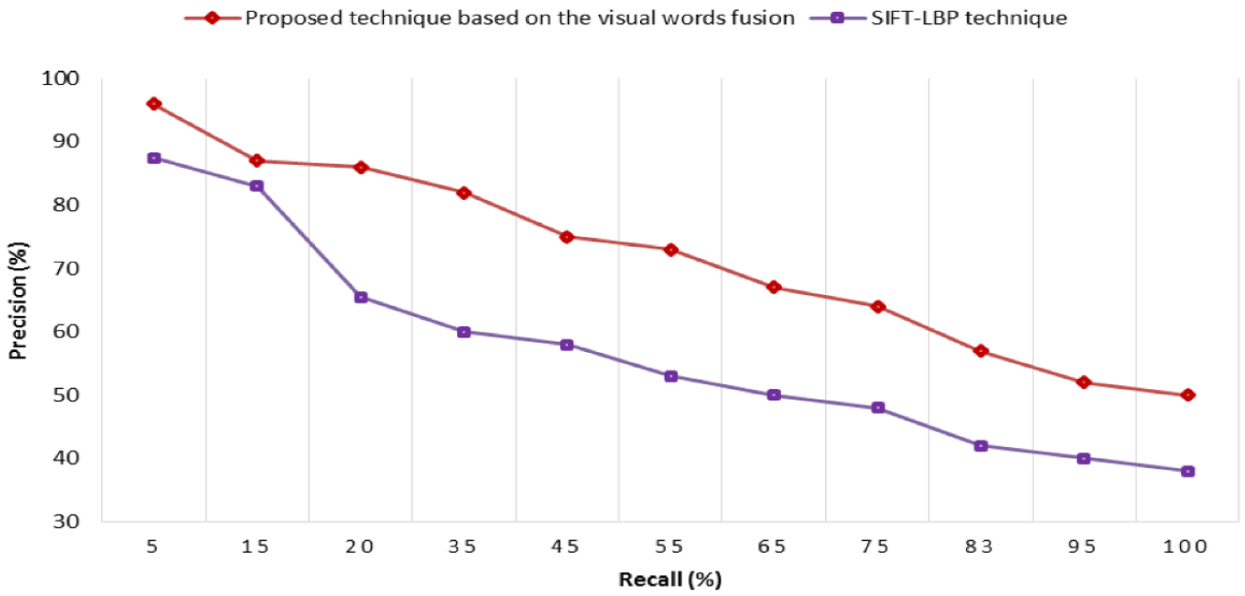

Figure 4.2: The determined precision and recall on corel-1000 dataset with proposed technique.

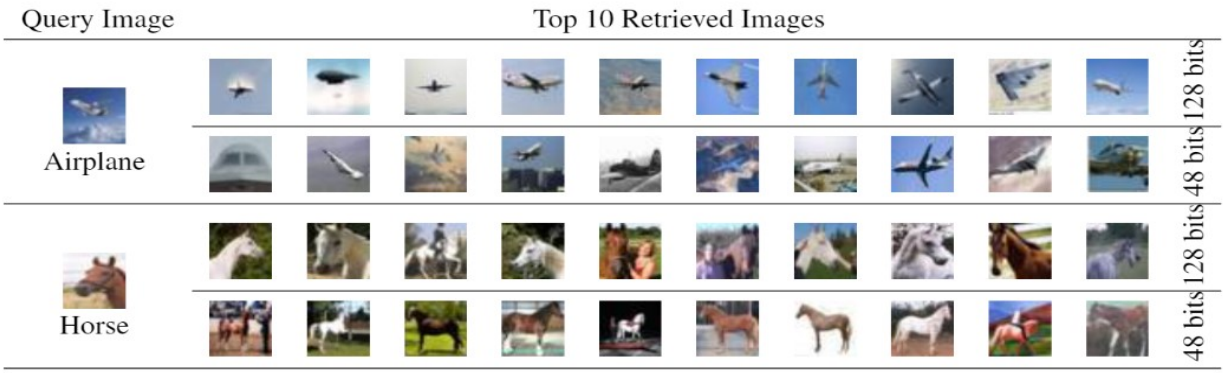

Figure 4.3: Sample of images from corel-1000 dataset being classified with respect to defined classes in CBIR.

Figure 4.4: Content based image retrieval result shows on the semantic category "Flowers" with CNN based patch-wise illumination estimation from Corel-1000 dataset.

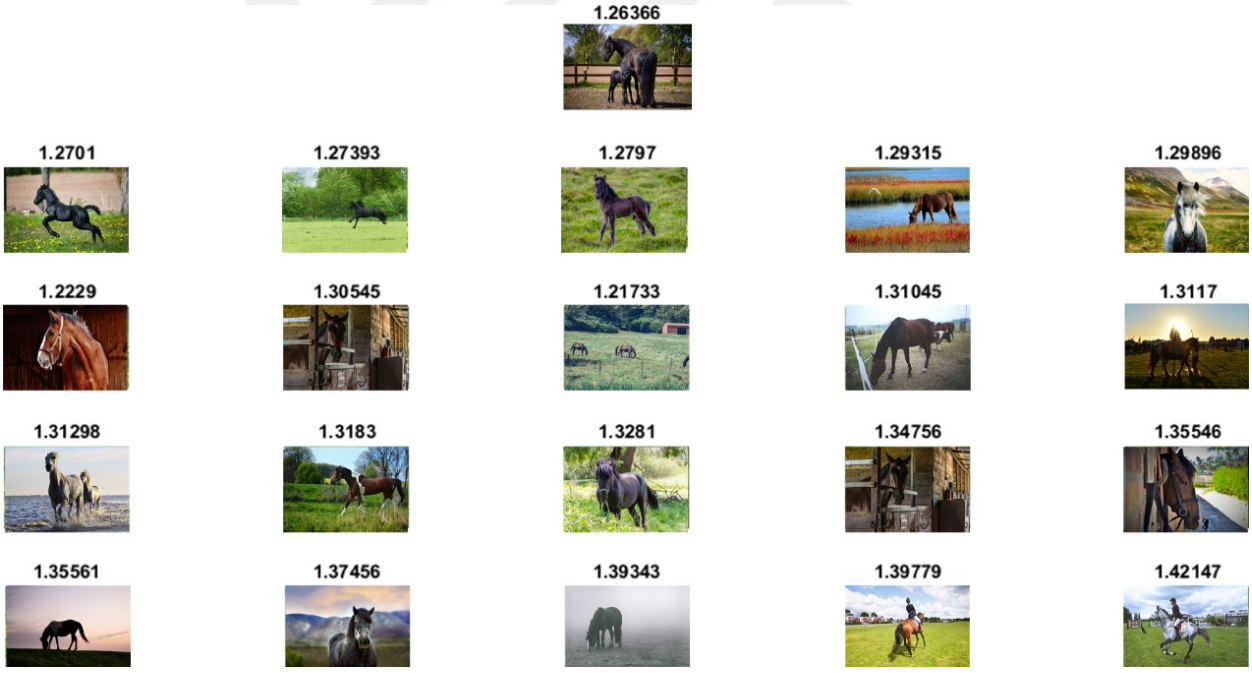

Figure 4.5: Content based image retrieval result shows on the semantic category "Horse" with CNN based patch-wise illumination estimation from Corel-1000 dataset.

The effect of gamma correction is not directly visible when compared to the output of the color transformation stage, but adds to the highlights and black tone appearance of the final result. The effect of the tone mapping stage is clearly visible, as it makes the image pleasant to the human vision, by fixing the brightness of different parts of the image.

Figure 4.6: Content based image retrieval result shows on the semantic category "Tiger" with CNN based patch-wise illumination estimation from Corel-1000 dataset.

The RAW image is basically a buffer of intensities per pixel. For this reason, the image at this stage is very dark and not suitable for a computer vision application. When the image is fusion, each pixel is mapped to a three-channel RGB representation with red, green and blue channels. The CNN algorithm determines the intensities of the other channels, based on information from neighboring pixels. This is a loss stage, as information in terms of bits per pixel is being lost. The image retrieval has a huge tone, which is due to the Bayer mosaic pattern. In the Bayer mosaic, green is the dominant color as the color arrays in the image sensor are placed in rows of green-blue and red-green. As can be seen, color transformation corrects for that hue tone. Three main types of transfer learning can be distinguished with respect to similarities in the source/target domains/tasks: (1) inductive CNN learning with the goal of inducing a generalizable, predictive model for the target task from a different source task; (2) CNN learning where a model should be derived in the target domain for the target task lacking labeled data, based on the source task containing a substantial amount of labeled data; and (3) supervised CNN learning which targets unsupervised learning in the target domain based on a different source task.

For purposes of CBIR image classification, the approach of inductive transfer learning has gained momentum in the past years. Given the large training time and requirements on labeled data, architectures are pre-trained on the ImageNet (source domain) for 1000-class classification

purposes (source task) and knowledge is transferred to a labeled dataset of choice (target domain) for e.g. classification as target task. Thus, it primarily exploits the learned feature-representations in the source domain and aims to transfer these to the target domain with the aim of improving classification performance whilst making use of deep, data-greedy CNN architectures.

The transferability of different levels of features was touched upon in this work with results confirming that the deepest layers are most task-specific and better results can be obtained by using features from levels upwards in the model. This was confirmed by the work, in which furthermore was suggested that pooling of features as well as combinations of features may improve performance. We added to the evidence by documenting higher performances for CNN's initiate with pre-trained weights and fine-tuned for the task at hand, rather than through random weight initialization and training from scratch. Furthermore, they investigated the transferability of specific feature layers and reported co-adaptation of intermediate feature layers, where splitting between frozen and trainable layers at this level might result in a drop in performance.

5. DISCUSSION

This work has shown the impact of content based image retrieval approximation to the quality-of-control for image-based control systems for smoothening. For the analysis, we used the use-case of a lateral controller that performs lane keeping for an autonomous vehicle. The application that performs the image signal processing, lane detection and control computations, was developed in the high performance MATLAB programming language, with the programmed in the domain specific language in deep learning. The simulations were run on the MATLAB simulator, using a CBIR and separate tracks for CNN algorithm, namely a straight and a curved track. The application was made error-resilient by adapting the CBIR detection algorithm to be able to process the degree of approximation that the different approximate ISP versions required. We ran this application on an Intel i7 processor, utilizing the available parallelism of the application through MATLAB on the eight available threads distributed in four cores. We, initially, benchmarked the application by conducting careful and detailed profiling on a total of 8 different pipeline versions, each of which on a dataset of 1000 images that were obtained in CNN. Then, we evaluated the degradation in quality-of-control that is caused by approximation, without taking into account the impact of approximation on runtime performance. Finally, the runtime performance gains due to approximation was taken into account and the improvement in quality-of-control was quantified for content based image retrieval. We used two separate metrics, namely the settling time and sum of squared errors, to evaluate the performance of the controller. Additionally, we used the energy, memory footprint and sensor-to-actuator delay to measure the improvement in the application performance.

Using this approach, we shed light on how real-time CNN algorithm can benefit from approximate computing. The approximated CNN pipelines achieved a maximum speedup of factor 3.5 for the sensor-to-actuator delay and a 40% improvement in settling time. The overall performance that was achieved by the multi-dimensional application was 50% better than the baseline. We showed the improvement of the application in metrics such as energy and memory footprint, which suggests that approximate computing is an efficient approach, when there are more constraints for the application, such as in an embedded environment.

We suggest a new, compact representation of patch-wise illumination estimation for multi-exposure to content based image retrieval estimation, which is independent of the camera (image projection matrix). It utilizes the projections of the 3D image sequencing to the input image as

opposed by putting away image boundaries utilizing the CNN model. This makes the indicator equipped for arranging patch-wise illumination estimation in images from input image applying the weights to the pre-processed images through feature extraction. The combination image position estimation would then be able to be reproduced when the right image projection framework and the ground plane condition.

Table 5.1: The comparison of proposed work with literature that is already available.

Article	Technique	Accuracy/Precision
[57]	Support Vector Machine (SVM)	89.36%
[58]	Recurrent Neural Network (RNN)	96.10%
Proposed	Convolutional Neural Network (CNN)	98.47%

The examination will make commitments to the patch-wise estimation remuneration and multi-exposure imaging, which comprises of testing images of light in different stances from arrangement for content based image retrieval. We will refine and finish the recursive filtering to incorporate multi-exposure fusions, all things considered, and impediments for order utilizing Convolutional Neural Network (CNN), which the classifier had the option to perceive just by taking a classifier at a solitary image. This was significant so as to do address patch-wise illumination for multi-exposure image fusion. In addition to this classification problem, we also framed the outline prediction formulation as a regression model. In this problem, we allowed the annotations to be continuous instead of categorical and thus represent a probability with which a pixel belongs to a nucleus' outline. The closer a given pixel is to the boundary, the higher we set its probability to belong to it. This boundary was created in a preprocessing step as a linear combination of the boundaries with two, four, six and eight pixels. The coefficients of the linear combinations were chosen in such a way that the probability of a pixel belonging to the outline follows a Gaussian bell in the distance of the pixel to the boundary. It is thereby assumed that the boundary is in between the two pixels of the boundary of width two. The Gaussian bell was scaled such that the two pixels adjoin to the boundary had a probability of 1 to belong to the boundary.

For this model, we keep the setup as previously defined for the classification problem. The only difference is that the mean squared error is used as a loss function.

An intuitive approach is to threshold the original image and subtract the thresholder probability map and use the remaining foreground pixels as segmented nuclei. This would separate clumped nuclei if the pixels in between them are correctly classified as boundary. However, this would also classify debris and other artifacts that are assigned as fore- ground by the thresholding method and correctly undetected as nuclei by the deep learning method.

A second method we considered is finding connected components among pixels that are not classified as boundary and filtering the largest connected component as background. This method was unsatisfactory because it failed for cases where the background was not connected, which is often the case at the image's boundaries.

We implemented a method that is very similar but uses a more involved filtering strategy. Among all connected components in the pixel's not classified as boundary, those whose mean image intensity is above a threshold are defined as nucleus. The threshold can be learned using the training data. We show the graphs for the CNN algorithm on fluorescence microscopy images from which we learned a threshold of 50 for this purpose using an 8-bit encoding that yields values ranging from 0 to 255.

In a last optimization step, we change the training procedure in such a way that for each training step, we sample a random crop from all training images which is in addition rotated by a multiple of 90◦ and flipped randomly. This approach seemed promising to us, as we achieved very high pixel-wise accuracy with the previous models. However, a common error on the test set for this models is the misclassification of boundary pixels between clumped nuclei as pixels belonging to a nucleus. On the training set, these pixels were correctly classified as boundary, which is an indicator for a model over-fit on the training data. Thus, by using data augmentation, we avoid this over-fit, and show that boundaries between nuclei get correctly classified as boundaries with this technique.

6. CONCLUSION

This master's thesis focuses on content based image retrieval using image processing and deep learning techniques. The deep learning challenge originates from the accuracy of the ground truth data sets that are used as validation and/or training data for automated image processing algorithms. These datasets usually stem from manual tracing of field boundaries. However, manual tracing of content based image retrieval is heavily dependent on the used imagery and supplementary data known as Corel-1000 dataset. It also is a highly subjective task, inevitably leading to inaccuracies and ambiguities depending on the priorities of the operator. The ground truth dataset can be complemented by existing property parcel information that could be indistinguishable based on the imagery alone. Furthermore, a single field content based image retrieval can potentially show several subfields due to unreported land use changes within the growing season. Given these constraints, even a very powerful automated image processing algorithm can only become as good as the ground truth data but will never be in perfect alignment with reality. Although the model can generalize within certain limits (local generalization), its transferability is mainly limited by the variability of the training data. Furthermore, higher spatial resolution would certainly improve the exact delineation of the field boundaries, especially for small objects. At the time of writing, presents the highest resolution of freely available satellite imagery. Instead, the methodology focuses on a simple input-output mapping. Additional pre-processing or post-processing techniques could certainly add value later on, but would distract from the current goal of establishing baseline result. To exploit the pre-trained CNN for image edge smoothening of the different datasets, a top model is to stacked on the deep learning with an accuracy of 98.17%. We explore three different top models in this thesis to maximally make use of the labeled, supervised data. A first and obvious choice is a mere fully-connected layer, applied on the flattened CNN-output with added dropout. In light of the high parametric demand this top-model places on the learning process, we opted to also investigate performance with residual blocks as top model. Not only are these residual blocks hypothesized to facilitate training by learning the residual mapping, they also decrease the number of parameters while allowing a multi-layered top model.

6.1 FUTURE RECOMMENDATIONS

In future, we want to add up the research line of deep learning with image processing, training on RNN with Apache Spark engine for combinations of different feature from different pre-trained hybrid models leads to an enhanced classification performance compared to stand-alone features for image fusion and edge smoothening.

A more advanced use of fine-tuning however proved to be adapting the pre-trained model, to the task at hand and fine-tuning up to a certain depth in the RNN model.

This scheme will give rise to the optimal results achieved in this study for edge smoothening and image filtering. Finally, these results will be compared developed using RNN framework in future for pre-training on any image dataset.

The best performance results will be achieved with high precision entered the same ballpark as the higher-scoring RNN classifiers on multiple pre-trained model framework, but could compete with the any other edge detection and smoothening work in previous studies.

We note however that in light of the trends observed in the RNN framework, a more thorough exploration (of architecture, training time and hyper-parameter settings) could potentially improve these results and open a window of opportunity to thoroughly using three labeled datasets for kick-starting the supervised training phase with limited three labeled datasets.

REFERENCES

[1] D. Ping Tian, "A review on image feature extraction and representation techniques," International Journal of Multimedia and Ubiquitous Engineering, vol. 8, no. 4, pp. 385–396, 2013.

[2] C. Zhang, J. Cheng, J. Liu, J. Pang, Q. Huang, and Q. Tian, "Beyond explicit codebook generation: visual representation using implicitly transferred codebooks," IEEE Transactions on Image Processing, vol. 24, no. 12, pp. 5777–5788, 2015.

[3] C. Zhang, C. Liang, L. Li, J. Liu, Q. Huang, and Q. Tian, "Fine-grained image classification via low-rank sparse coding with general and class-specific codebooks," IEEE Transactions on Neural Networks and Learning Systems, vol. 28, no. 7, pp. 1550–1559, 2016.

[4] C. Zhang, J. Cheng, and Q. Tian, "Structured weak semantic space construction for visual categorization," IEEE Transactions on Neural Networks and Learning Systems, vol. 29, no. 8, pp. 3442–3451, 2017.

[5] C. Zhang, J. Cheng, and Q. Tian, "Semantically modeling of object and context for categorization," IEEE Transactions on Neural Networks and Learning Systems, vol. 30, no. 4, pp. 1013–1024, 2018.

[6] C. Zhang, J. Cheng, and Q. Tian, "Unsupervised and semisupervised image classification with weak semantic consistency," IEEE Transactions on Multimedia, 2019.

[7] N. Kondylidis, M. Tzelepi, and A. Tefas, "Exploiting tf-idf in deep convolutional neural networks for content based image retrieval," Multimedia Tools and Applications, vol. 77, no. 23, pp. 30729–30748, 2018. [83] X. Shi, M. Sapkota, F. Xing, F. Liu, L. Cui, and L. Yang, "Pairwise based deep ranking hashing for histopathology image classification and retrieval," Pattern Recognition, vol. 81, pp. 14–22, 2018.

[8] Z. Yu and W. Wang, "Learning DALTS for cross-modal retrieval," CAAI Transactions

on Intelligence Technology, vol. 4, no. 1, pp. 9–16, 2019.

[9] N. Ali, D. A. Mazhar, Z. Iqbal, R. Ashraf, J. Ahmed, and F. Zeeshan, “Content-based image retrieval based on late fusion of binary and local descriptors,” International Journal of Computer Science and Information Security (IJCSIS), vol. 14, no. 11, 2016.

[10] N. Ali, Image Retrieval Using Visual Image Features and Automatic Image Annotation, University of Engineering and Technology, Taxila, Pakistan, 2016.

[11] B. Zafar, R. Ashraf, N. Ali et al., “Intelligent image classification-based on spatial weighted histograms of concentric circles,” Computer Science and Information Systems, vol. 15, no. 3, pp. 615–633, 2018.

[12] G. Qi, H. Wang, M. Haner, C. Weng, S. Chen, and Z. Zhu, “Convolutional neural network based detection and judgement of environmental obstacle in vehicle operation,” CAAI Transactions on Intelligence Technology, vol. 4, no. 2, pp. 80–91, 2019.

[13] U. Markowska-Kaczmar and H. Kwa´snicka, “Deep learning—a new era in bridging the semantic gap,” in Bridging the Semantic Gap in Image and Video Analysis, pp. 123–159, Springer, Basel, Switzerland, 2018.

[14] F. Riaz, S. Jabbar, M. Sajid, M. Ahmad, K. Naseer, and N. Ali, “A collision avoidance scheme for autonomous vehicles inspired by human social norms,” Computers & Electrical Engineering, vol. 69, pp. 690–704, 2018.

[15] H. Shao, Y. Wu, W. Cui, and J. Zhang, “Image retrieval based on MPEG-7 dominant color descriptor,” in Proceedings of the 9th International Conference for Young Computer Scientists ICYCS 2008, pp. 753–757, IEEE, Hunan, China, November 2018.

[16] X.-Y. Wang, B.-B. Zhang, and H.-Y. Yang, “Content-based image retrieval by integrating color and texture features,” Multimedia Tools and Applications, vol. 68, no. 3, pp. 545– 569, 2014.

[17] J. M. Guo, H. Prasetyo, and J. H. Chen, “Content-based image retrieval using error

diffusion block truncation coding features," IEEE Transactions on Circuits and Systems for Video Technology, vol. 25, no. 3, pp. 466–481, 2015.

[18] Y. Liu, D. Zhang, and G. Lu, "Region-based image retrieval with high-level semantics using decision tree learning," Pattern Recognition, vol. 41, no. 8, pp. 2554–2570, 2008.

[19] Z. Jiexian, L. Xiupeng, and F. Yu, "Multiscale distance coherence vector algorithm for content-based image retrieval," @e Scientific World Journal, vol. 2014, Article ID 615973, 13 pages, 2014.

[20] G. Papakostas, D. Koulouriotis, and V. Tourassis, "Feature extraction based on wavelet moments and moment invariants in machine vision systems," in Human-Centric Machine Vision, InTech, London, UK, 2012.

[21] G.-H. Liu, Z.-Y. Li, L. Zhang, and Y. Xu, "Image retrieval based on micro-structure descriptor," Pattern Recognition, vol. 44, no. 9, pp. 2123–2133, 2011.

[22] X.-Y. Wang, Z.-F. Chen, and J.-J. Yun, "An effective method for color image retrieval based on texture," Computer Standards & Interfaces, vol. 34, no. 1, pp. 31–35, 2012.

[23] R. Ashraf, K. Bashir, A. Irtaza, and M. Mahmood, "Content based image retrieval using embedded neural networks with bandletized regions," Entropy, vol. 17, no. 6, pp. 3552–3580, 2015.

[24] A. Irtaza and M. A. Jaffar, "Categorical image retrieval through genetically optimized support vector machines (GOSVM) and hybrid texture features," Signal, Image and Video Processing, vol. 9, no. 7, pp. 1503–1519, 2015.

[25] C. Zhang, J. Cheng, and Q. Tian, "Multiview semantic representation for visual recognition," IEEE Transactions on Cybernetics, pp. 1–12, 2018.

[26] N. I. Ratyal, I. A. Taj, U. I. Bajwa, and M. Sajid, "3D face recognition based on pose and expression invariant alignment," Computers & Electrical Engineering, vol. 46, pp. 241–255, 2015.

[27] N. Ratyal, I. Taj, U. Bajwa, and M. Sajid, "Pose and expression invariant alignment based multi-view 3D face recognition," KSII Transactions on Internet & Information Systems, vol. 12, no. 10, 2018.

[28] N. Ratyal, I. A. Taj, M. Sajid et al., "Deeply learned pose invariant image analysis with applications in 3D face recognition," Mathematical Problems in Engineering, vol. 2019, Article ID 3547416, 21 pages, 2019.

[29] M. Sajid and T. Shafique, "Hybrid generative–discriminative approach to age-invariant face recognition," Journal of Electronic Imaging, vol. 27, no. 2, article 023029, 2018.

[30] M. Sajid, T. Shafique, S. Manzoor et al., "Demographicassisted age-invariant face recognition and retrieval," Symmetry, vol. 10, no. 5, p. 148, 2018.

[31] M. Sajid, T. Shafique, I. Riaz et al., "Facial asymmetry-based anthropometric differences between gender and ethnicity," Symmetry, vol. 10, no. 7, p. 232, 2018.

[32] M. Sajid, T. Shafique, M. Baig, I. Riaz, S. Amin, and S. Manzoor, "Automatic grading of palsy using asymmetrical facial features: a study complemented by new solutions," Symmetry, vol. 10, no. 7, p. 242, 2018.

[33] N. Ali, K. B. Bajwa, R. Sablatnig et al., "A novel image retrieval based on visual words integration of SIFT and SURF," PLoS One, vol. 11, no. 6, Article ID e0157428, 2016.

[34] M. Naeem, R. Ashraf, N. Ali, M. Ahmad, and M. A. Habib, "Bottom up approach for better requirements elicitation," in Proceedings of the International Conference on Future Networks and Distributed Systems, p. 60, ACM, Cambridge, UK, July 2017.

[35] B. Zafar, R. Ashraf, N. Ali et al., "A novel discriminating and relative global spatial image representation with applications in CBIR," Applied Sciences, vol. 8, no. 11, p. 2242, 2018.

[36] N. Ali, B. Zafar, F. Riaz et al., "A hybrid geometric spatial image representation for scene classification," PLoS One, vol. 13, no. 9, Article ID e0203339, 2018.

[37] B. Zafar, R. Ashraf, N. Ali, M. Ahmed, S. Jabbar, and S. A. Chatzichristofis, "Image classification by addition of spatial information based on histograms of orthogonal vectors," PLoS One, vol. 13, no. 6, Article ID e0198175, 2018.

[38] N. Ali, K. B. Bajwa, R. Sablatnig, and Z. Mehmood, "Image retrieval by addition of spatial information based on histograms of triangular regions," Computers & Electrical Engineering, vol. 54, pp. 539–550, 2016.

[39] R. Khan, C. Barat, D. Muselet, and C. Ducottet, "Spatial orientations of visual word pairs to improve bag-of-visualwords model," in Proceedings of the British Machine Vision Conference, pp. 89–91, BMVA Press, Surrey, UK, September 2012.

[40] H. Anwar, S. Zambanini, and M. Kampel, "A rotation-invariant bag of visual words model for symbols based ancient coin classification," in Proceedings of the 2014 IEEE International Conference on Image Processing (ICIP), pp. 5257–5261, IEEE, Paris, France, October 2014.

[41] H. Anwar, S. Zambanini, and M. Kampel, "Efficient scaleand rotation-invariant encoding of visual words for image classification," IEEE Signal Processing Letters, vol. 22, no. 10, pp. 1762–1765, 2015.

[42] R. Khan, C. Barat, D. Muselet, and C. Ducottet, "Spatial histograms of soft pairwise similar patches to improve the bag-of-visual-words model," Computer Vision and Image Understanding, vol. 132, pp. 102–112, 2015.

[43] N. Ali, B. Zafar, M. K. Iqbal et al., "Modeling global geometric spatial information for rotation invariant classification of satellite images," PLos One, vol. 14, no. 7, Article ID e0219833, 2019.

[44] R. Ashraf, M. Ahmed, S. Jabbar et al., "Content based image retrieval by using color descriptor and discrete wavelet transform," Journal of Medical Systems, vol. 42, no. 3, p. 44, 2018.

[45] R. Ashraf, M. Ahmed, U. Ahmad, M. A. Habib, S. Jabbar, and K. Naseer, "MDCBIR-

MF: multimedia data for contentbased image retrieval by using multiple features," Multimedia Tools and Applications, pp. 1–27, 2018.

[46] Y. Mistry, D. Ingole, and M. Ingole, "Content based image retrieval using hybrid features and various distance metric," Journal of Electrical Systems and Information Technology, vol. 5, no. 3, pp. 878–888, 2017.

[47] K. T. Ahmed, M. A. Iqbal, and A. Iqbal, "Content based image retrieval using image features information fusion," Information Fusion, vol. 51, pp. 76–99, 2018.

[48] P. Liu, J.-M. Guo, K. Chamnongthai, and H. Prasetyo, "Fusion of color histogram and LBP-based features for texture image retrieval and classification," Information Sciences, vol. 390, pp. 95–111, 2017.

[49] W. Zhou, H. Li, J. Sun, and Q. Tian, "Collaborative index embedding for image retrieval," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 5, pp. 1154–1166, 2018.

[50] C. Li, Y. Huang, and L. Zhu, "Color texture image retrieval based on Gaussian copula models of Gabor wavelets," Pattern Recognition, vol. 64, pp. 118–129, 2017.

[51] H. H. Bu, N. Kim, C. J. Moon, and J. H. Kim, "Content-based image retrieval using combined color and texture features extracted by multi-resolution multi-direction filtering," Journal of Information Processing Systems, vol. 13, no. 3, pp. 464–475, 2017.

[52] A. Nazir, R. Ashraf, T. Hamdani, and N. Ali, "Content based image retrieval system by using HSV color histogram, discrete wavelet transforms and edge histogram descriptor," in Proceedings of the 2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), pp. 1–6, IEEE, Sukkur, Pakistan, March 2018.

[53] M. Srinivas, R. R. Naidu, C. S. Sastry, and C. K. Mohan, "Content based medical image retrieval using dictionary learning," Neurocomputing, vol. 168, pp. 880–895, 2015.

[54] S. Mohamadzadeh and H. Farsi, “Content-based image retrieval system via sparse representation,” IET Computer Vision, vol. 10, no. 1, pp. 95–102, 2016.

[55] Q. Li, Y. Han, and J. Dang, “Sketch4Image: a novel framework for sketch-based image retrieval based on product quantization with coding residuals,” Multimedia Tools and Applications, vol. 75, no. 5, pp. 2419–2434, 2016.

[56] H. Wu, R. Bie, J. Guo, X. Meng, and S. Wang, “Sparse coding based few learning instances for image retrieval,” Multimedia Tools and Applications, vol. 78, no. 5, pp. 6033–6047, 2018.

[57] Sindu, S. & Kousalya, R. (2020). Recurrent Neural Network for Content Based Image Retrieval Using Image Captioning Model. 10.1007/978-3-030-37218-7_112.

[58] Masroor. Ts, Shoaib & Sharma, Arvind. (2017). Content Based Image Retrieval System using SVM Technique.