PEOPLE'S DEMOCRATIC REPUBLIC OF ALGERIA

MINISTRY OF HIGHER EDUCATION AND RESEARCH SCIENTIST

UNVERSITY YAHIA FARES OF MEDEA



Faculty of technology

Master's degree of study memory

Faculty: Electrical engineering

Specialty : Electronics Embedded Systems

Thesis

Content based image retrieval system by merging multi convolutional neural networks

Prepared by :

BOUKHARI ABDERRAHMANE

KEBLOUTI MERZAK

Proposed and supervised by:

Dr. HANTABLI HAMZA

College year: 2020/2021

Acknowledgements:

We thank first, God for gave us the moral and physical force allowed us to complete this work, we would like to express our sincere gratitude to our supervisor; Dr. Hantabli hamza for his guidance and enthusiastic support throughout the duration of this work. He was always there to answer our questions about different aspects of the project. Many thanks go to the staff of the College of Technologie at Yahia fares University for their valuable assistance for all the time. We are grateful to our referees; chairman and members of the examination committee, for carefully reading and commenting this thesis. We will not like to miss the opportunity to acknowledge the contribution of our families it is because of them that we are able to get here and finally are presenting our thesis. We want to thank them for being there for us always.

Remerciements:

Nous remercions avant tout, Dieu de nous avoir prodiguée la force morale et physiquenous a permis d'achever ce travail, nous aimerions exprimer notre sincère gratitude à notre superviseur; Dr. Hantabli Hamza pour son orientation et son soutien enthousiaste pendant toute la durée de ce travail. Il était toujours là pour répondre à nos questions sur différents aspects du projet. Merci beaucoup au personnel du Collège de Technologie de l'université de Yahia Fares pour leur utilité précieuse pour tout le temps. Nous sommes reconnaissants à nos arbitres; Président et membres du comité d'examen, pour lire attentivement et commenter cette thèse. Nous ne voudrons pas manquer l'occasion de reconnaître la contribution de nos familles, c'est à cause d'eux que nous sommes en mesure d'arriver ici et de présenter notre thèse. Nous voulons les remercier d'être là pour nous toujours.

Abstract:

Content based image retrieval (CBIR) systems use low level image representations to measure image similarity and fetch relevant images. Color, texture and shape properties are considered as important handcrafted features in traditional CBIR systems. The success of such a CBIR system depends on the choice of the handcrafted features being used. To use relevant handcrafted features, one needs to have a good knowledge of the domain where CBIR is being applied. Usage of inappropriate handcrafted features may widen the semantic gap and can to lead to poor retrieval results.

Hence it is very important to extract features which are independent of prior domain understanding. In addition, it is beneficial if the features can be learnt automatically from an input image. Machine learning methods can be used for learning valuable representations from input image data. In machine learning area, Convolution neural networks (CNN) are able to create important expressive features from a given image data. Hence CNNs are well suited for image processing applications like classification, object recognition and clustering etc. Very large datasets, huge computing resources and processing time are required to train a deep CNN model effectively. There are many deep CNNs available which are pre-trained on massive datasets and distributed for public use. The knowledge learnt from these pre-trained deep CNN models can be applied to address image processing issues in new domains. In this thesis, we have created a framework to leverage VGG16, ResNet50 and InceptionV3 deep CNN models for extracting important features and use these features for image retrieval task. We apply this frame work for an interesting problem, to retrieve images from wang dataset. Experimental results indicate that the features which are extracted from pre-trained deep CNN model perform better than handcrafted features when used for image retrieval application.

Keywords: Content-based image retrieval, features extraction, Euclidean distance, Dataset, similarity, pre-trained, Convolutional neural network, classification, confusion matrix.

Résumé :

Les systèmes de récupération d'image basés sur le contenu (CBIR) utilisent des représentations d'image à faible niveau pour mesurer la similarité d'image et évaluer les images pertinentes. Les propriétés de couleur, de texture et de forme sont considérées comme des caractéristiques artisanales importantes dans les systèmes CBIR traditionnels. Le succès d'un tel système CBIR dépend du choix des caractéristiques artisanales utilisées. Pour utiliser des fonctionnalités artisanales pertinentes, il faut avoir une bonne connaissance du domaine où le CBIR est appliqué. L'utilisation de caractéristiques fabriquées à la main inappropriées peut élargir l'écart sémantique et peut conduire à de mauvais résultats de récupération.

Il est donc très important d'extraire des fonctionnalités indépendantes de la compréhension de domaine préalable. De plus, il est bénéfique si les fonctionnalités peuvent être apprises automatiquement d'une image d'entrée. Les méthodes d'apprentissage de la machine peuvent être utilisées pour apprendre des représentations précieuses des données d'image d'entrée. Dans la zone d'apprentissage de la machine, les réseaux de neurones de convolution (CNN) sont capables de créer des fonctionnalités expressives importantes à partir d'une donnée d'image donnée. Par conséquent, CNNS convient bien aux applications de traitement d'images telles que la classification, la reconnaissance d'objet et la clustering, etc. Très grands ensembles de données, des ressources informatiques énormes et du temps de traitement sont nécessaires pour former efficacement un modèle CNN de manière efficace. Il existe de nombreux CNN profond disponibles qui sont pré-formés à des ensembles de données massifs et distribués pour un usage public. Les connaissances acquises à partir de ces modèles de CNN pré-formés peuvent être appliqués pour traiter les problèmes de traitement d'images dans de nouveaux domaines. Dans cette thèse, nous avons créé un cadre pour tirer parti des modèles VGG16, ResNet50 et InceptionV3 Deep CNN pour extraire des fonctionnalités importantes et utiliser ces fonctionnalités pour une tâche de récupération d'image. Nous appliquons ce travail de trame pour récupérer des images d'un jeu de données d'images wang. Les résultats expérimentaux indiquent que les caractéristiques extraites du modèle CNN profond pré-formé fonctionnent mieux que des fonctionnalités fabriquées à la main lorsqu'elles sont utilisées pour une application de récupération d'image.

Mots clés : Récupération d'image basée sur le contenu, Extraction des caractéristiques, Distance euclidienne, Base de données, similarité, pré-formés, Réseau de neurones convolutionnels, classifucation, Matrice de confusion.

الملخص:

تستعمل أنظمة إسترجاع الصور على حسب المحتوى (سيبير) تمثيل الصور منخفضة المستوى من أجل قياس مدى

تشابه الصور وتقييم الصور ذات الصلة.تعتبر خصائص اللون والنسيج والشكل كخصائص حرفية مهمة في أنظمة (سيبير) التقليدية.

النجاح في أي نظام (سيبير) يعتمد على اختيار الخصائص الحرفية المستعملة،ومن أجل استعمال خصائص حرفية ذات صلة يجب أن تكون لدينا معرفة جيدة بالمجال الذي سنطبق فيه نظام (سيبير)

يمكن أن يؤدي استخدام خصائص مصنوعة يدويا غير لائقة إلى توسيع الانحراف الدلالي ويمكن أن يؤدي إلى ضعف نتائج الاسترداد، وبالتالي، من المهم للغاية استخراج الميزات التي هي مستقلة عن فهم المجال السابق، بالإضافة إلى ذلك، من المفيد إذا تم تعلم الميزات تلقائيا من صورة الإدخال.

طرق تعلم الآلة يمكن أن يمكن استخدامها لفهم تمثيلات قيمة لبيانات صورة الإدخال،في مجال تعلم الآلة الشبكات العصبية التلافيفية (سي أن أن) قادرة على خلق ميزات معبرة مهمة انطلاقا من بيانات صورة معينة وبالتالي ،(سي أن أن)مناسبة لتطبيقات معالجة الصور مثل التصنيف، التعرف على الكائن والتجميع،إلخ. هناك حاجة إلى مجموعات البيانات الكبيرة جدا، وموارد الكمبيوتر الضخمة ووقت المعالجة لتدريب نموذج (سي أن أن) بطريقة فعالة.يوجد عدد كبير من الشبكات التلافيفية العميقة متاحة ومدربة مسبقا على مجموعات ضخمة وموزعة من أجل الإستخدام العام،المعارف المكتسبة انطلاقا من من هذه النماذج المعدة مسبقا يمكن تطبيقها من أجل حل مشاكل معالجة الصور في مجالات جديدة .

في هذه المذكرة قمنا بإنشاء إطار للإستفادة من نماذج (في جي جي) و(رازنات فيفتي) و (انسابشين في ثري)العميقة من أجل استخراج ميزات مهمة واستعمالها في عملية استرجاع للصور. قمنا بتطبيق هذا النموذج على قاعدة بيانات الصور (وانغ)،فأثبتت النتائج التجريبية أن الميزات المستخرجة بإستعمال نموذج الشبكات العصبية التلافيفية العميقة المعدة مسبقا تعمل على نحو أفضل من الميزات المصطنعة يدويا عند استخدامها في تطبيقات استرجاع الصور

الكلمات المفتاحية: استرجاع الصور على حسب المحتوى،استخراج الميزات ،المسافة الإقليدية،فاعدة بيانات،تشابه،معدة مسبقا، الشبكة العصبية التلافيفية،تصنيف،مصفوفة الإرتياب.

LIST OF CONTENT

Acknowledgements 1
Abstract 2
List of abréviations 9
CHAPTER I: INTRODUCTION 10
General Introduction 10
Historical Perspective 11
Principal Components of a CBIR System 11
Deep learning 12
Project Objectives 13
Structure of the report 13
CHAPTER II : RELATED WORKS 14
Itroduction 14
Color Features 14
Texture Features 17
Shape features 20
Content-Based Image Retrieval 21
Deep Learning 21
CBIR Research Using Deep-Learning Techniques 24
Analysis in Distance Metrics in Content Based Image Retrieval 27
CHAPTER III : METHODOLOGY 29
Methodology of research 29
Neural Network 29
Convolutional Neural Network (CNN) 29
Pre-trained Neural Network Model 30
Proposed method of Feature Extraction by Deep Learning 31
Pre-trained models used 32
Inception V3 32
ResNet50 : 33
VGG16 architecture 33
Random architecture model 36

Dataset used……..36
Similarity Measure…….....38
Performances…….……38
Confusion matrix…….……..40
Tools used in this work…….41
Programmation steps..…….…...…..41
CHAPTER IV : RESULTS 42
Results and experiments :..... 43
Train set/Test set of Wang dataset: 43
Trainig random CNN model 44
Random model classification report:..... 44
Training VGG16 model……..46
VGG16 classification report.....….47
Training ResNet50 model: 47
ResNet50 classification report……49
Training InceptionV3 model : 50
InceptionV3 classification report : 51
Combined model classification report: 52
Discussion of the results.....…...53
CHAPTER V : CONCLUSION.. 54
Conclusion:..... 55
Dataset and Classes: 55
CBIR with the Visual Features Language…….55
Bibliography:..... 57
Appendix……..…62

LIST OF TABLES

Tableau 1: A summary of the performance of color features..... 17
Tableau 2: Model sequential 36
Tableau 3: Confusion matrix for binary classification. 40
Tableau 4: performances of every class 45
Tableau 5: performances of every class 47
Tableau 6: performances of every class 49
Tableau 7: performances of every class 51
Tableau 8: performances of every class 52
Tableau 9: Compare model’s performances..... 53

LIST OF FIGURES

Figure 1: Example of SVM-based classification..... 19
Figure 2: Edge operators: (a) Original image, (b) Prewitt, (c) Canny, (d) Laplacian of Gaussian, (e) Roberts 20
Figure 3: Comparison of CBIR with CNN-based features with classic SIFTS feature extraction. They have in common the creation of a Visual Features Language (VFL) and the CBIR model..... 22
Figure 4: Content-Based Image Retrieval according to the VFL model..... 24
Figure 5: Left: A two-layer Neural Network (one hidden layer of four neurons and one output layer with two neurons) with three inputs. Right: A three-layer neural network with two hidden layers of four neurons each, one output layer with a single neuron and thre..... 30
Figure 6: Every layer of a CNN converts the 3D input volume to a 3D output volume of activations. In given example, the image of the bird is the input layer, so its height and width are the dimensions of the image, and the depth would be three (Red, Green, Blue..... 30
Figure 7: CBIR image feature representation with the help of pre-trained deep learning models 31
Figure 8: A block diagram of the matching image retrieval based on the proposed system.... 32
Figure 9: The-basic-architecture-of-Inception-v3..... 33
Figure 10: ResNet50 architecture..... 33
Figure 11: VGG-16 layers..... 34
Figure 12 : The-standard-VGG-16-network-architecture 35
Figure 13: Wang dataset classes..... 36
Figure 14: The proposed framework of Automatic features extraction..... 37
Figure 15: Wang distribution classes 43
Figure 16: Training validation and loss of random model..... 44
Figure 17: Training and validation accuracy of random model..... 44
Figure 18: confusion matrix of random model..... 45
Figure 19: Training and validation loss of vgg16 model 46
Figure 20: Training and validation accuracy of vgg16 model 46
Figure 21: confusion matrix of vgg16 model..... 47
Figure 22: Training and validation loss of ResNet50 model 48
Figure 23: Training and validation accuracy of ResNet50 lodel..... 48
Figure 24: confusion matrix of ResNet50 model..... 49
Figure 25: Training and validation loss of InceptionV3 model 50
Figure 26: Training and validation accuracy of InceptionV3 model 50
Figure 27: Confusion matrix of inceptionV3 model..... 51
Figure 28: Confusion matrix of combined model..... 52
Figure 29: predicted results..…..... 54

List of abréviations:

CBIR: content-based image retrieval

CNN: convolutional neural network

VGG: Visual Geometry Group

ResNet : residual neural network

DNN : Deep neural network

VFL : visual features language

VQ : vector quantization

DRNNs : Deep Recurrent Neural Networks

DCD: dominant color descriptor

RGB : red, green, blue

HSV : hue, saturation, value

QBIR : query-based image retrieval

SVM : Support Vector Machine

CHAPTER I: INTRODUCTION

1. General Introduction:

The study growth of the Internet, the falling price of storage devices, and the increasing pool of available computing power make it necessary and possible to manipulate very large repository of digital information efficiently. Generally speaking, Content-Based Image Retrieval (CBIR) technique aims to develop techniques that support the effective searching and browsing tasks performed on large image digital libraries by using automatically derived image features [1]. The rapid increase in the size of digital image collections has motivated the research on image retrieval. The early researches on image retrieval discipline proposed annotating the images manually to facilitate their retrieval. Obviously, annotating images manually is a cumbersome and expensive task for large image databases, and is often subjective, context-sensitive and incomplete [2]. Sometimes, it is almost impossible to describe the content of an image by words, especially for the textures found in an image. As a result, it is difficult for the traditional text-based methods to support a variety of task-dependent queries [3]. To avoid manual annotation, an alternative technique was developed. It is called CBIR, by which images would be indexed according to their visual contents (such as color, texture, and shape). The basic task of a CBIR system is to find similar images (according to their visual features) within a large image database [4]. Typically, a CBIR system pre-processes an image database in order to extract information from all images in the database. This information is referred to as visual features of images. The visual features may be represented by different representations, which facilitate image's searching and indexing [5]. Mostly, the two-step approach to search the image database is still adopted. In the first step, a feature vector is extracted from each image in the database. These features vectors which characterize some image properties are stored in a dedicated features database. In the second step, for a given query image, its features vector is computed, and then compared with the features vectors contained in the features database. The most similar images to the query image are returned to the User. The similarity measure used to compare the features vectors should be effective enough to ensure correct matching between similar images, as well as being able to discriminate dissimilar ones [6]. In the next sections, the fundamental principles of CBIR are presented.

1.2 Historical Perspective:

CBIR is a technique that uses the visual contents to search the images listed in large scale image databases according to users' interests. Since 1990s, this research area has been one of most active and rapid growing disciplines [3]. Early work on image retrieval can be traced back to the late 1970s. Since 1979, the application potential of image database management techniques has attracted the attention of researchers from multiple disciplines [7]. The early techniques were not generally based on visual features but on the textual annotation of images. In other words, images were first annotated with text and then searched using a text-based approach from traditional database management systems. In the early 1990s, it was widely recognized that a more efficient and intuitive way to represent and index visual information would be based on properties that are inherent in the images them selves [8].

Researchers from the communities of computer vision, database management, human-computer interface, and information retrieval were attracted to this field. Since 1997, the number of research publications on the techniques of visual information extraction, organization, indexing, query interaction, and database management has increased enormously. Similarly, large numbers of academic and commercial retrieval systems have been developed by universities, government organizations, companies, and hospitals [3].

1.3 Principal Components of a CBIR System:

CBIR uses the visual contents of an image such as *color*, *shape*, *texture*, and *spatial layout* to represent and index the image. In a typical image retrieval system; the visual contents of the images in the database are extracted and described by multi-dimensional features vectors. The features vectors of the images in the database form a features database. Then, the system changes these examples into its internal representation of features vectors. To retrieve images, Graphical User Interface (GUI) is used to provide the retrieval engine with example images or sketched figures to search for [9].

Afterwards, the similarities /distances between the features vectors of the query example and those of the images in the database are calculated. Retrieval is performed as a last step with the aid of the built indexing structure [3]. The indexing scheme provides an efficient way to search for the image database. The image with the smallest distance to the given image is retrieved as the result. In current CBIR systems, instead of only retrieving the best image, a number of similar images are retrieved and sorted according to their distance to the given image. [5]

1.4 Deep learning:

Neural networks have a long history starting back in the 1940s, with the cornerstone paper of McCulloch and Pitts and the works of Wiener, Neumann and Hebb. From this early stage and on, there had been golden periods followed by times of stand still [10]. The research progress of the 1980s and early 1990s has stumbled upon the difficulties of training complex network structures using gradient descent methods, partially caused by the limitations of the technology at the time. It was not until the last 10 years, that new revolutionary methods combined with advances in hardware design which made massive parallelism [11] broadly available, have provided the capability to train networks with an increased number layers. A third factor that can be mentioned is the transition of the web into the social era, also known as web 2.0 that brought motivation for commercial applications of Artificial Intelligence (AI). The term Deep Neural Network (DNN) was coined to describe any network that has more than the three layers found in the reference architecture of a Multilayer Perceptron (MLP), which are input, hidden and output. Correspondingly, Deep Learning (DL) is the domain that researches the methodologies and techniques used in Deep Neural Networks, seeking solutions that will be based on them. Isolating through taxonomy of proposals, the prominent DNN groups are Deep Belief Networks (DBNs), Deep Autoencoders (DAs), Deep Recurrent Neural Networks (DRNNs), Deep Convolutional Neural Networks (DCNNs or CNNs) and some meta-groups like Network-In-Network and the recently introduced Generative Adversarial Networks.

The potential of Deep Learning for solving complex problems was shown in various image recognition tasks. In the 2012 ImageNet Large Scale Visual Recognition Challenge (ILSVRC2012), the top performing solution for the classification of images to 1000 categories was a DCNN out performing the second best by almost 10%. The AlexNet, as it is known from the name of the primary author of the corresponding paper [12], showed that Deep Neural Networks can be a powerful choice for image recognition tasks. It also shifted the focus of research to the original bio-inspired idea of receptive field mapping, incepted years ago as a solution to the T-C problem [13] and later elaborated as the first CNN networks [14]. The last years, key players in the computer industry like Microsoft and Google have demonstrated their interest, making significant advances that led to deeper and better performing networks. The results on image classification are impressive, surpassing the measured human performance on the natural image dataset that is used in ILSVRC. Models which are pre-trained on this image collection have shown that they can achieve top

performance on different image collections. Following the method of transfer learning, the learned parameters of the CNN layers are fine-tuned on new datasets which may belong to entirelydifferent domains of problems. There is an intuition that the functionality of the convolutional layers resembles that of a general purpose feature extractor, which can be used for CBIR. The visual features are extracted from images simply by recalling them through a trained deep CNN. The activation values of a convolutional layer can be considered as fine, local or global visual features. For each coordinate of the layer's output map, activation vectors are the local feature descriptors for the underlying mapped region of pixels. These can be used to create corresponding "visual words" in a language specific for the architecture and the learned state of the CNN. This intra-domain approach involves both Deep Learning and Information Retrieval and creates new research subjects:

a) Designing and training of CNNs as good feature extractors.

b) Selection of neural network layers that have the proper features for better retrieval performance.

c) Mitigation of the high dimensionality of neural layer activations for computationally effective representations and indexing of large image collections.

1.5 Project Objectives:

The basic task of this project is to find similar images (using their visual features) according to a query image with in a medium size image database. Content-Based Image Retrieval (CBIR) system aims to develop techniques that support effective searching and browsing tasks performed on a large image collection by using automatically derived image features.

1.6 Structure of the Report:

The thesis is divided into the following sections:

Chapter I: Introduction provides the necessary knowledge as a foundation for the thesis.

Chapter II: related works introduces relevant related works that have been done in the field, and discusses some of the different approaches that have been studied.

Chapter III: methodologydescribes how the method to address the research problems have been carried out and what experiments have been done.

Chapter IV: *Results* presents the results of the experience and discussions.

Chapter V: conclusion presents some concluding remarks and future works.

CHAPTER II : RELATED WORKS

Itroduction:

In this chapter, we will look into what was done in the CBIR field over the years. Since even a simple CBIR system may encompass knowledge from several research areas, including Image Processing, Machine Learning, Information Retrieval or Computer Vision, an extensive analysis on all techniques used would be nearly impossible. Instead, we will discuss a small subset that may be seen as representatives for a significant part of the strategies followed. There is however, a reasonable consensus on the high level CBIR process.

Typically, we begin with a preprocessing stage, called Feature Extraction, to capture a representation of the image content into a set of feature vectors. Due to the potential high number of images in the database, and therefore, the high number of feature vectors, this stage if followed by an indexing stage, where we build one or more data structures to sort them. Each time a user submits a query; a feature extraction is performed for that image (query image) and compared with the features extracted from the other images. Finally, the images with the most similar features are returned to the User, usually ordered by their similarity.

2.1 Color Features :

Color is considered as one of the important low-level visual features as the human eye can differentiate between visuals on the basis of color. The images of the real-world object that are taken within the range of human visual spectrum can be distinguished on the basis of differences in color [15–18]. The color feature is steady and hardly gets affected by the image translation, scale, and rotation [19–22].) Though the use of dominant color descriptor (DCD) [15], the overall color information of the image can be replaced by a small amount of representing colors. DCD is taken as one of the MPEG-7 color descriptors and uses an effective, compact, and intuitive format to narrate the indicative color distribution and feature. Shao et al. [15] presented a novel approach for CBIR that is based on MPEG-7 descriptor. Eight dominant colors from each image are selected, features are measured by the histogram intersection algorithm, and similarity computation complexity is simplified by this.

According to Duanmu [16], classical techniques can retrieve images by using their labels and annotation which cannot meet the requirements of the customers; therefore, the researchers

focused on another way of retrieving the images that is retrieving images based on their content. The proposed method uses a small image descriptor that is changeable according to the context of the image by a twostage clustering technique. COIL-100 image library is used for the experiments. Results obtained from the experiments proved that the proposed method to be efficient [16].

Wang et al. [17] proposed a method based on color for retrieving image on the basis of image content, which is established from the consolidation of color and texture features. This provides an effective and flexible estimation of how early human can process visual content [17]. The fusion of color and texture features offers à vigorous feature set for color image retrieval approaches. Results obtained from the experiments reveal that the proposed method retrieved images more accurately than the other traditional methods. However, the feature dimensions are not higher than other approaches and require a high computational cost. A pairwise comparison for both low-level features is used to calculate similarity measure which could be a bottleneck [17].

Various research groups carried out a study on the completeness property of invariant descriptors [18]. Zernike and pseudo-Zernike polynomials which are orthogonal basis moment functions can represent the image by a set of mutually independent descriptors, and these moment functions hold orthogonality and rotation invariance [18]. PZMs proved to be more vigorous to image noise over the Zernike moments. Zhang et al. [18] presented a new approach to derive a complete set of pseudo-Zernike moment invariants. The link between pseudo-Zernike moments of the original image and the same shape but distinct orientation and scale images is formed first. An absolute set of scale and rotation invariants is obtained from this relationship. And this proposed technique proved to be better in performance in recognizing pattern over other techniques [18].

Guo et al. [19] proposed a new approach for indexing images based on the features extracted from the error diffusion block truncation coding (EDBTC). To originate image feature descriptor, two color quantizers and a bitmap image using vector quantization (VQ) are processed which are produced by EDBTC. For assessing the resemblance between the query image and the image in the database, two features Color Histogram Feature (CHF) and Bit Pattern Histogram Feature (BHF) are introduced. The CHF and BHF are calculated from the VQ-indexed color quantizer and VQ-indexed bitmap image, respectively. The distance evaluated from CHF and BHF can be used to assess the likeliness between the two images. Results obtained from the experiments show that the proposed scheme performs better than

former BTC-based image indexing and other existing image retrieval schemes. The EDBTC has good ability for image compression as well as indexing images for CBIR [19]. Liu et al. [20] proposed a novel method for region-based image learning which utilizes a decision tree named DT-ST. Image segmentation and machine learning techniques are the base of this proposed technique. DT-ST controls the feature discretization problem which frequently occurs in contemporary decision tree learning algorithms by constructing semantic templates from low-level features for annotating the regions of an image. It presents a hybrid tree which is good for handling the noise and tree fragmentation problems and reduced the chances of misclassification. In semantic-based image retrieval, the user can query image through both labels and regions of images. Results obtained from the experiments conducted to check the effectiveness of the proposed technique reveal that this technique provides higher retrieval accuracy than the traditional CBIR techniques and the semantic gap between low- and high-level features is reduced to a significant level. The proposed technique performs well than the two effectively set decision tree induction algorithms ID3 and C4.5 in image semantic learning [20]. Islam et al. [21] presented a supreme colorbased vector quantization algorithm that can automatically categorize the image components. The new algorithm efficiently holds the variable feature vector like the dominant color descriptors than the traditional vector quantization algorithm. This algorithm is accompanied by the novel splitting and stopping criterion. The number of clusters can be learned, and unnecessary overfragmentation of region clusters can be avoided by the algorithm through these criteria. Jiexian et al. [22] presented a multiscale distance coherence vector (MDCV) for CBIR. The purpose behind this is that different shapes may have the same descriptor and distance coherence vector algorithm may not completely eliminate the noise. The proposed technique first uses the Gaussian function to develop the image contour curve. The proposed technique is invariant to different operations like translation, rotation, and scaling transformation.

There are various low-level color features, and the performance of color moments is not good as it can represent all the regions of the image. Histogram-based color features require high computational cost while DCD performs better for region-based image retrieval and is computationally less expensive due to low dimensions. A detailed summary of the abovementioned color features [15–22] is represented in Table 1..

Tableau 1: A summary of the performance of color features

Author	Application	Method	Dataset	Accuracy
Duanmu [25].	Image retrieval	Color moment invariant	COIL-100	0.985
Wang et al. [26]	Content-based image retrieval	Integrated color and texture features	Corel	0.613
Zhang et al. [27]	Object recognition	Complete set of pseudo-Zernike invariants	COIL-100	_____
Guo et al. [28]	Content-based image retrieval	Error diffusion block truncation coding Features	Corel	0.797
Shao et al. [24]	Image retrieval	MPEG-7 dominant color descriptor	Corel	0.8964
Liu et al. [29]	Region-based image retrieval	High-level semantics using decision tree learning	Corel	0.768
Islam et al. [30]	Automatic categorization of image regions	Dominant color-based vector quantization	Corel	0.9767
Jiexian et al. [31]	Content-based image retrieval	Multiscale distance coherence vector Algorithm	MPEG-7 image database	0.97

2.2 Texture Features :

Papakostas et al. [23] performed their experiments on four datasets, namely, COIL, ORL, JAFFE, and TRIESCH I in order to show the discrimination power of the wavelet moments. These datasets are divided into 10, 40, 7, and 10 classes. For the evaluation of the proposed model (WMs), two different configurations of wavelets WMS-1 and WMs-2 are used where the former uses cubic B-spline and the other uses the Mexican hat mother wavelets. By keeping only effective characteristics in feature selection approach greatly improves the classification capabilities of the wavelet moments.

The performance of the proposed model is compared with Zernike, pseudo-Zernike, Fourier–Mellin, and Legendre and with two others by using 25, 50, 75, and 100 percent of the entire datasets, and each moment family behaves differently in each dataset. Classification

performance of the moment descriptors shows the better results of the proposed model (wavelet Moments and moment invariants). For the evaluation of the proposed model (MSD) for image retrieval, Liu et al. [24] perform experiments on Corel datasets as there are no specific datasets for content-based image retrieval (CBIR). Corel-5000 and Corel-10000 are used with 15000 images, and HSV, RGB, and Lab color space are used to evaluate the retrieval performance. On both datasets Corel-5000 and Corel-10000, the average retrieval and recall rates of the proposed model using different color quantization level and texture orientation quantization levels are evaluated and our proposed model performs better on HSV and Lab color space and poor on RGB color space. For getting good results between storage space, retrieval accuracy, and speed, 72 color and orientation quantization levels are used in MSD and 6 for image retrieval. The average retrieval and recall ratios of MSD are compared with other methods like Gabor MTH on Corel datasets because these algorithms are developed for image retrieval for the evaluation of MSD and the results show that our proposed model (MSD) outperforms other models.

10,000 color images [25] were collected from public resources of natural scenes such as landscapes, peoples, and textures in order to perform their experiments for image retrieval based on texture. Generally, for retrieval results, properties such as smoothness, regularity, distribution, and coarseness are considered while used additionally the color information with these properties. The precision comparison between the proposed model (color co-occurrence matrix) and the gray-level co-occurrence matrix method provides results to evaluate the proposed model. The comparison shows that the color co-occurrence matrix is better than the gray-level co-occurrence matrix because of the additionally added property (color information). For CBIR [26], Corel, COIL, and Caltech-101 datasets (those datasets are chosen that have images grouped in the form of semantic concepts) containing 10908, 7200 images, and 101 image categories for respective datasets are used. The mean precision and recall rates obtained by the proposed method (Embedded neural network with bandlet transform) on top 20 retrievals are compared with the other standard and with the state-of-the-art retrieval systems. The mean precision and recall rates obtained by the proposed method are 0.820 and 0.164 on top 20 retrievals. These results show that the research presented in [26] clearly outperformed other models in terms of mean precision value and recall rate.

With Corel image gallery containing 10900 images for categorical image retrieval, Irtaza and Jaffar [27] conducted experiments to show the effectiveness of the proposed model (SVM-based architecture; Figure 1 represents an example of binary classification while using SVM).

The Corel image gallery is divided into two sets Corel A having 1000 images that are divided into ten categories and Corel B that has 9900 images. The mean precision and recall rates obtained by the proposed method on the top 20 retrievals are compared with other standard retrieval systems. Different numbers of returned images are used to show the retrieval capacity of SVM and it shows consistent results. Thus, the results and comparison show that the proposed model has better results and is more consistent in image retrieval. Fadaei et al. [28] performed experiments on Brodatz and Vistex datasets for content-based image retrieval containing 112 grayscale and 54 color images for respective datasets. The distance between the query image and dataset image is calculated, images that have minimum distance are retrieved, and then the precision and recall rates are calculated. The results of the proposed models are compared with other prior methods. The retrieval time of Brodatz is longer than that of Vistex database because Brodatz has more images than Vistex; thus, it needs time for more feature matching and processing. The dimension of the feature vector for the proposed model is 3124 which is higher than that of other methods. Retrieving time of the proposed model is slower in feature matching and faster in feature extraction although the dimension of the feature vector is high. Comparison and results show that the proposed model (LDRP) has better performance and average precision rates and is faster in feature extraction and slower in feature matching.

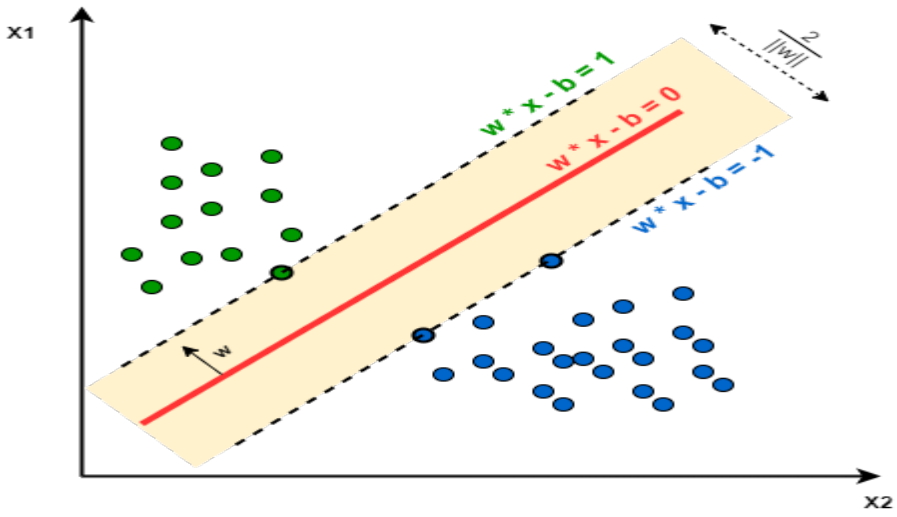


Figure 1: Example of SVM-based classification

2.3 Shape features:

Shape is one of the most important low level features for CBIR as humans tend to perceive images by decomposing them into individual objects [29]. Querying a database using shape features can allow physicians to identify malformations or tumors that otherwise might be missed [30]. In such queries, several properties, such as the shape's size or position are usually not relevant. For this reason, shape features should be invariant to translations, rotations, scales, noise and partial occlusions.To identify a shape, we must find where are its edges, that is, where a big change in the gray-level intensities occurs. In two-dimensional continuous functions, that information is calculated using the gradient. In digital images, the gradient can be approximated convolving the image with gradient edge detectors (e.g. Sobel, Prewitt), or using more sophisticated methods like zero crossing or Gaussian edge detectors [32] (see Fig. 2). The filtered image often has isolated segments, who should be removed or closed into a connected boundary [29]. There are two types of shape features that can be extracted from an image [31]: (a) global properties (e.g. perimeter) depend on the entire shape to be computed, so they are sensitive to partial occlusions; and (b) local properties (e.g. curvature) can be computed on single regions, so they are more robust, but more computational expensive. The methods used to extract those features can be classified into (a) region-based and (b) contour-based.

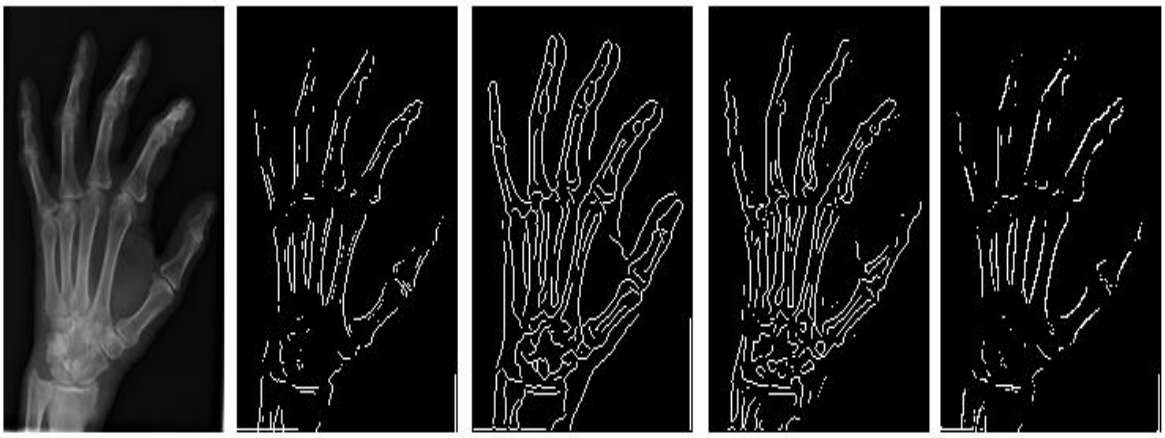


Figure 2: Edge operators: (a) Original image, (b) Prewitt, (c) Canny, (d) Laplacian of Gaussian, (e) Roberts

2.4 Content-Based Image Retrieval:

Content-based image retrieval (CBIR) for decades has been one of the most Researched fields of computer vision. CBIR aims to search for images through analyzing their visual contents, and thus image representation is the crux of CBIR. [[34], [36], [33]] In the past there has been a variety of proposed low-level feature descriptors for image representation, ranging from global features like color features,edge features, texture features, GIST and CENTRIST, and recent local feature representations, such as the bag-of-words (BoW) models using local feature descriptors (SIFT, SURF). Conventional CBIR approaches usually choose rigid distance functions on some extracted low-level features for multimedia similarity search, such as Euclidean distance or cosine similarity. However, the fixed rigid similarity/distance function may not be always optimal to the complex visual image retrieval tasks due to the grand challenge of the semantic gap between low-level visual features extracted by computers and high-level human perceptions. Therefore, in the recent past there have been a surge of active research efforts in the design of various distance/similarity measures on some low-level features by exploring machine learning techniques.Among these techniques, some works have focused on learning to hashing or compact codes. There has been a proposed mapping learning scheme for large scale multimedia applications from high dimensional data to binary codes that preserve semantic similarity.

2.5 Deep Learning:

Deep learning refers to a class of machine learning techniques, where many layers of information processing stages in hierarchical architectures are exploited for pattern classification and for feature or representation learning [35]. It lays in the intersections of several research areas, including neural networks, graphical modeling, optimization, pattern recognition, and signal processing, etc. [36] Yann LeCun adopted the deep supervised back propagation convolutional network for digit recognition. In the recent past, it has become a valuable research topic in the fields of both computer vision and machine learning where deep learning achieves state-of-the art results for a variety of tasks. The deep convolutional neural networks (CNNs) proposed by Hinton came out first in the image classification task of Imagenet classification with deep convolutional neural networks. The model was trained on more than one million images, and has achieved a winning top-5 test error rate of 15.3% over 1,000 classes. After that, some recent works got better results by improving CNN models. The top-5 test error rate decreased to 13.24% in by training the model to simultaneously classify,

locate and detect objects. Besides image classification, the object chip processing abilities (e.g., GPU units), the significantly lower cost of computing hardware, and recent advances in machine learning and signal/information processing research.Over the past several years, a rich family of deep learning techniques has been proposed and extensively studied, e.g., Deep Belief Network (DBN), Boltzmann Machines (BM), Restricted Boltzmann Machine (RBM), Deep Boltzmann Machine (DBM), Deep Neural detection task can also benefit from the CNN model, as reported in.Generally speaking, three important reasons for the popularity of deep learning today are drastically increased Networks (DNN), etc. Among various techniques, the deep convolutional neural networks, which is a discriminative deep architecture and belongs to the DNN category, has found state-of-the-art performance on various tasks and competitions in computer vision and image recognition. Specifically, the CNN model consists of several convolutional layers and pooling layers, which are stacked up with one on top of another.The convolutional Layer shares many weights, and the pooling Layer sub-samples the output of the convolutional layer and reduces the data rate from the Layer below. The weight sharing in the convolutional layer, together with appropriately chosen pooling schemes, endows the CNN with some invariance properties (e. g., translation invariance).

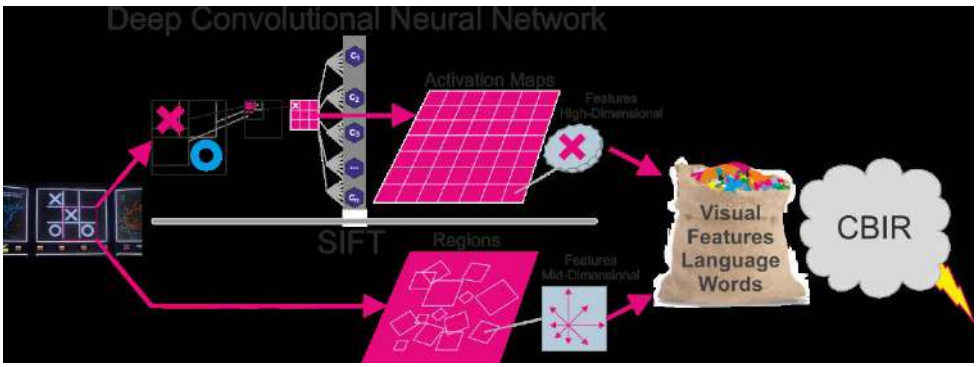


Figure 3: Comparison of CBIR with CNN-based features with classic SIFTS feature extraction. They have in common the creation of a Visual Features Language (VFL) and the CBIR model.

The research done in Deep Neural Content-based Image Retrieval (DN-CBIR) attempts to discover DNN architectures that will be used as general purposed feature extractors for CBIR. The methods of the codebook creation, indexing and retrieval can remain the same in both CNN-based and SIFT-based image retrieval. Multiple tasks are combined in a pipeline of

processes with various methods from the domains of Machine Learning, Data Mining and Information Retrieval.

These can be described by the following workflow of steps:

1. Supervised training of a DNN which is typically a CNN. This is performed on an image collection with annotations like class labels. The learned parameters of the network will incorporate the feature extraction capabilities and its layers will act as feature extractors.

2. An image is recalled through the pre-trained DNN. The output activations of a selected Layer are the feature descriptors for the image regions. Different layers correspond to different scales of the image resolution.

3. Dimensionality reduction is incorporated in the DNN structure or runs as an additional process using the neural activation vectors as data points.

4. The feature descriptors are quantized to create a VFL vocabulary that is specific for
a) The trained state of the DNN architecture,
b) The image dataset used for training and
c) The selected neural network layer.

5. The image regions are described by visual words of the VFL vocabulary. Each word is a representation of a region's features. All the visual words of the image compose an image representation that can be considered a visual document.

6. The ordering of visual words in the document can be disregarded as in the BoVW or exploited by the information retrieval method.

7. Visual documents can be indexed by a search engine and retrieved by queries, using standard text IR methods.

The tasks of image classification object detection and object localization can be ported as image retrieval tasks on VFL representations. Retrieving an image of the same class in the top rank can be equivalent to classification. The image areas can be segmented and indexed as different fields in the CBIR system. A fielded search with an image of an object can detect it and localize it inside the returned set of relevant images.

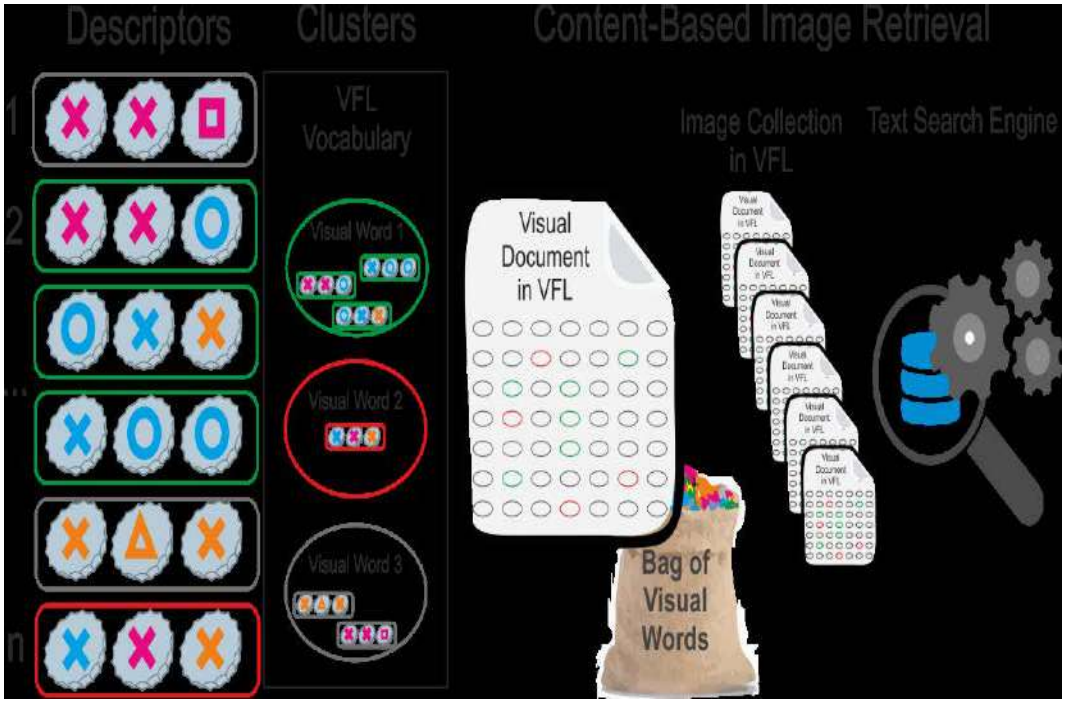


Figure 4: Content-Based Image Retrieval according to the VFL model.

VFL vocabulary is a question to be answered. An evaluation of the retrieval performance with different magnitudes of codebook sizes, could decide the best between thousands and millions of unique visual words.

Research focused in the IR domain could evaluate different weighting schemes and models for all the above combinations. Understanding the extreme complexity of the problems that the DN-CBIR research has to solve is the starting point. This thesis aims to unveil the combined aspects of the problem and contribute to the related research with the presentation of novel proposals.

2.7 CBIR Research Using Deep-Learning Techniques

Searching for digital images from lager storage or databasesis often required, so content-based image retrieval (CBIR) also known as query-based image retrieval (QBIR) is used for image retrieval. Many approaches are used to resolve this issue such as scale-invariant transform and vector of locally aggregated descriptor. Due to most prominent results and with a great performance of the deep convolutional neural network (CNN), a novel term frequency-inverse documentfrequency (TF-IDF) using as description vector the weighted convolutional word frequencies based on CNN is proposed for CBIR. For this purpose, the learned filers of convolutional layers of convolution neuron model were used as a detector of the visual words, in which the degree of the visual pattern is provided by the activation of each filter as tf

part.then three approaches of computing the idf part are proposed [37]. By providing powerful image retrieval techniques with a better outcome, these approaches concatenate the TF-IDF with CNN analysis for visual content. To prove the proposed model, the authors conduct experiment on four image retrieval datasets and the outcomes of the experiments show the existence of the truth of the model. In order to handle the large scale, Shi et al. [38] proposed a hashing algorithm that extracts features from images and learns their binary representations. The authors model the pairwise matrix and an objective function with deeplearning framework that learns the binary representations of images. Experiments are conducted on thousands of histopathology images (on 5356 skeletal muscle and 2176 lung cancer images with 4 types of diseases) to indicate the trustworthiness of the proposed algorithm. The efficiency of the proposed algorithms is achieved with 97.94% classification accuracy.

Zhu et al. [39] proposed unsupervised visual hashing approach known as the semantics assisted visual hashing (SAVH). This system uses two components that are offline learning and online learning. In offline learning firstly, the image pixel is transformed intmathematical vector representation by extracting the visual and texture feature. Then, text enhancing the visual graph is extracted with the assistance of topic hypergraph, and the semantics information is extracted from the text information and then the hash code of image is learned which preserves the correlation of image between the semantics and images, and then at the last, the hash function code is generated within the linear aggressive model. These desirable properties match the requirement of real application scenarios of CBIR [39].

In computer vision applications, the use of CNN has shown a remarkable performance, especially in CBIR models. Most of the CNN models get the features in the last layer using a single CNN with order less quantization approach and its drawback is they limit the utilization of intermediate convolutional layer for identifying local image pattern. So, in this paper, a new technique is identified as bilinear CNN-based architecture. This method used two parallel CNN models to extract the feature without the prior knowledge of the semantics of image content. The feature isdirectly extracted from the activation of the convolutional layer rather than reducing very low-dimensional feature. The experiment on this approach gives a very important conclusion:

This model reduces the image representation to the compact length as it used different quantized levels to extract the feature, so it is remarkable to boost the retrieval performance and the search time and storage cost. Secondly, the bilinear CRB-CNN is very effective in learning a very complex image having different semantics. Ten milliseconds is needed to extract the feature from the image and search from the database and very small disk size is needed to represent and store the image. And at the end, end-to-end tanning is applied without any other metadata, annotations, tags which conformed the capability of CRB-CNN to extract the feature from only the visual information in CBIR task. This technique also applies on the large-scale database image to retrieve the image and showed a high retrieval performance [40]. For efficient image search, hashing function gains efficient attention in CBIR [41]. Hashing function gives a similar binary code to the similar content of the image which maps the high-dimensional visual data into low-dimensional binary space. This approach is basically depending upon the CNN. It is to be assumed that the semantic labels are represented by the several latent layer attributes (binary code) and classification also depends upon these attributes. Based on this approach, the supervised deep hashing technique constructs a hash function from a latent layer in the deep neurons network and the binary code is learned from the objective functions that explained about the classification error and other desirable properties in the binary code. The main feature of the SSDH is that it unifies retrieval and classification in a single model. SSDH is scalable to large-scale search, and by slight modification in the existing deep network for classification, SSDH is simple and easily realizable [41].

Effective image analysis and classification of the visual information using discriminative information is considered as an open research problem [42]. Many research models are proposed using different approaches either by combining views by graph-based approach or by using transfer learning. It is difficult from the existing methods to compute the discriminative information at the image borders and to find similarity consistency constraint. The authors [42] proposed a multiview label sharing method (MVLS) for this open research problem and tried to maintain and retain the similarity. For visual classification and representation, optimization over the transformation and classification parameters is combined for transformation matrix learning and classifier training. Results on MVLS with different six views (no intra-view and no inter-view plus no intra-view) and nine views (combination of intra-view and inter-view) are conducted. Experimental results are compared with several state-of-the-art research and results shows the effectiveness of the proposed

MVLS approach [42]. For the understanding of images and object categorization, methods like CNN and local feature have shown good performance in many application domains.

The use of CNN models is still challenging for precise categorization of object and in the case with limited training information and labels. To handle the semantic gap, the smooth constraints can be used, but the performance of the CNN model degrades due to the smaller size of the training set. The authors [43] proposed a multiview algorithm with few labels and view consistency (MVFL-VC). Both labeled and unlabeled images are used together for the image view consistency with multiview information. The discriminative power of the learned parameter is also enhanced by unlabeled training images. To evaluate the proposed algorithm, experiments are conducted on different datasets. The proposed MVFL-VC algorithm can be used with other image classification and representation techniques. The algorithm is tested on unlabeled and unseen datasets. The results of experiments and analysis reveal the effectiveness of the proposed method [43]. The extraction of domain space knowledge can be beneficial to reduce the semantic gap [44]. The authors proposed multiview semantics representation (MVSR), which is a semantics representation for visual recognition.

The proposed algorithm divides the images on the basis of semantic and visual similarities [44]. Two visual similarities for training samples provide a stable and homogenous perception that can handle different partition techniques and different views. The proposed research based on MVSR is more discriminative than other semantics approaches as the semantic information is computed for future use from each view and from separate collection of images and different views. Different publicly available image benchmarks are used to evaluate this research, and the experimental results show the effectiveness of MVSR.

The result demonstrated that MVSR improved classification performance in terms of precision for image sets with more visual variations.

2.8 Analysis in Distance Metrics in Content Based Image Retrieval

There are several methods to retrieve images from a database based on its content. In the retrieving process to measure the similarity between images a distance metric can be used. Researchers are still trying to identify the best distance metric to retrieve images in image retrieval system. Euclidian distance, Manhattan distance, and Vector Cosine Angle Distance are a few of the distance measures that can be used to develop Image retrieval system [45].

Euclidian distance :

If u = (x_1, y_1) and v = (x_2, y_2) are two points, then the Euclidean distance between u and v is given by,

$$\mathrm{EU}(u,v) = \sqrt{(x1-x2)^2 + (y1-y2)^2}$$

Manhattan distance :

If u = (x1, y1) and v = (x2, y2) are two points, then the Manhattan Distance between u and v is given by :

$$MH\ (u,\ v) = |x1 - x2| + |y1 - y2|$$

Vector cosine angle distance :

If we consider two vectors X and Y where X ≡ (x_1, x2,…. x_n) and Y ≡ (y_1, y_2,…. y_n), then cosθ may be considered as the Cosine of the vector angle between X and Y in n dimensions. Formally, we define VCAD as follows :

$$VCAD(X,Y) = \frac{\sum_i x_i y_i}{\sqrt{\sum_i x^2 \sum_i y^2}}$$

Normalized cross correlation :

Cross Correlation is simply the measure of similarity between two signals. The sum of pairwise multiplications of corresponding pixel values of the images can be taken as the cross-correlation. Image correlation can be defined as the numeric measure of image similarity [46].

$$Cross - Correlation(Image1, Image2) = \sum_{x,y} Image1(x,y) \times Image2(x,y)$$

There is also an enhanced version of cross-correlation method. It is known as the Normalized cross-correlation method [46].

$$NCC(Image1, Image2) = \frac{1}{N\sigma_1\sigma_2} \sum_{xy} \left(Image1(x,y) - \overline{Image1}\right) \times \left(Image2(x,y) - \overline{Image}\right)$$

Histogram intersection distance:

Histogram Intersection Distance can be written as:

$$S(A,B) = \frac{\sum_{i=1}^{n} min\left(Hi(A), Hi(B)\right)}{\sum_{i=1}^{n} Hi(A)}$$

Where A and B represent two histograms [47].

CHAPTER III : METHODOLOGY

3.1 Methodology of research:

In a Content-Based Image Retrieval system, images are represented by a set of low level or/and high-level features. This is called feature encoding where an image from RGB or HSV space is encoded to a n-dimensional feature vector. In this thesis we propose to derive feature vectors of an image with the help of some pre-trained deep learning models. For this purpose we first present a brief description of the key concepts such as neural network, pre-trained models etc.

3.2 Neural Network

Neural Networks are set up as collections of neurons that are connected in a non-cyclic graph. These models are often depicted into separate layers of neurons. Generally, the most common type is the fully-connected Neural Network layer in which neurons between two adjacent layers are fully pairwise connected, but there is no connection between neurons within a single layer. Below are two examples of fully connected Neural Network [48].

3.3 Convolutional Neural Network (CNN)

Convolutional Neural Networks take input as images and they handle the architecture in a more sensible way. The layers of a CNN have neurons which are arranged in 3 dimensions: width, height, depth. Here, the word depth refers to the third dimension of an activation volume of a layer. The neurons in a layer are connected to a small region of the preceding CNN layer, unlike to all the neurons which is a norm in a fully-connected neural network. The visualization is shown in Figure 6.

As mentioned above, a simple CNN is a sequence of layers, and every layer of a CNN transforms one volume of activation to another by passing through certain differentiable functions. There are mainly three types of layers in CNN architectures: Convolutional Layer, Pooling Layer and Fully-Connected Layer (shown in Figure 5). We will stack these layers on top of each other to form a fully operational CNN architecture [48]. Figure 2 shows the CNN model architecture of a classification problem. This network consists of convolution, pooling, fully connected layers and some activation layers (e.g. ReLU, softmax etc).

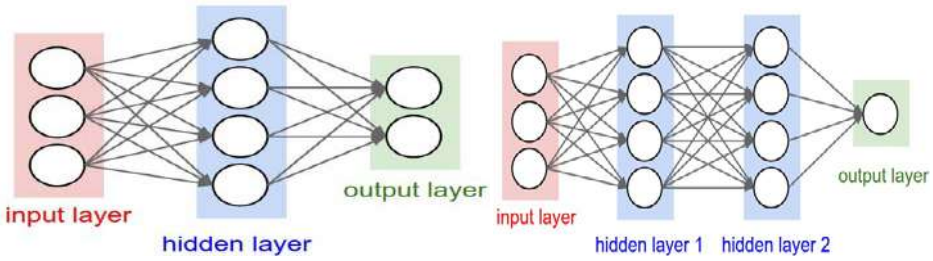


Figure 5: Left: A two-layer Neural Network (one hidden layer of four neurons and one output layer with two neurons) with three inputs. Right: A three-layer neural network with two hidden layers of four neurons each, one output layer with a single neuron and thre

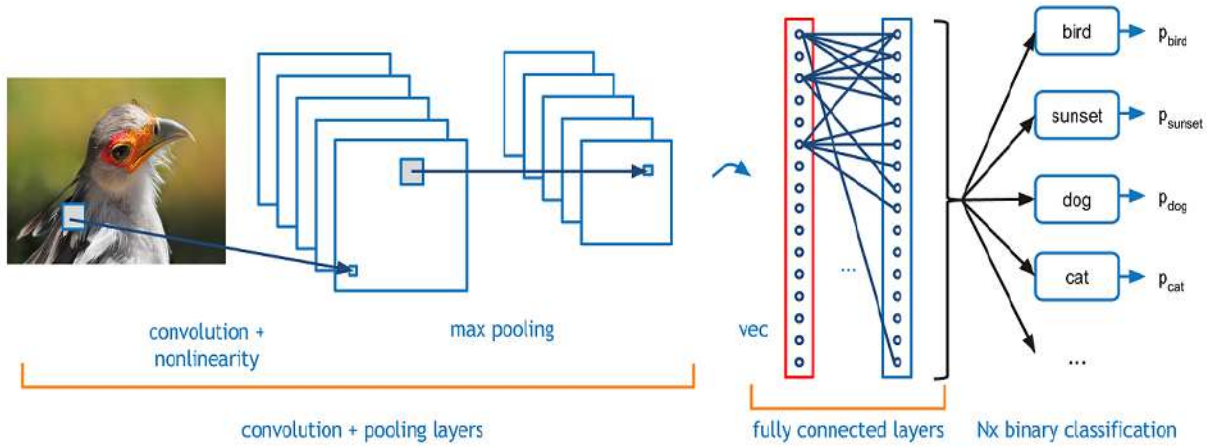


Figure 6: Every layer of a CNN converts the 3D input volume to a 3D output volume of activations. In given example, the image of the bird is the input layer, so its height and width are the dimensions of the image, and the depth would be three (Red, Green, Blue

3.4 Pre-trained Neural Network Model

Transfer learning [49] is a method where instead of starting the training process of a model from scratch, we use the learned weights of an already trained model to solve a different but similar problem. In this way we leverage previous learning through the learned weights and save time considerable amount of time. Usually much better results are also achieved compared to training from scratch. In computer vision, transfer learning is usually executed by the use of pre-trained models. A pre-trained model [50] is a model that was trained on a large benchmark dataset to solve some specific problems similar to the ones we want to solve. Accordingly, because of the high computational cost of training such deep learning models, it is a common practice to import and use models from a published architecture (e.g. VGG, ResNet, Xception etc). A comprehensive review of pre-trained models' performance on computer vision problems using data from the ImageNet [51] challenge is presented by Canziani et al. [52]. Being motivated by this, we have used a pre-trained Neural Network

model which was trained on the ImageNet Dataset. This dataset contains more than 14 million images which belong to more than 20,000 classes. It also provides bounding box annotations for around 1 million images, which can be used in Object Localization tasks. Multiple layers of convolutional layers, average pooling layers, max pooling layers etc. are stacked up with one another with different combinations to build up the model architectures. The final layer is then unwrapped to an n dimensional vector, which is called the dense (or fully connected) layer. Several fully connected layers can be stacked on this unwrapped dense layer.

Finally, a softmax activation layer of dimension: the number of class is stacked over the final dense layer to get the probabilities of each class for classification problems or a linear activation layer of single dimension can be placed to predict the regressed value for regression problems or other kind of structured activation layers are placed according to the desired problems to solve.

3.5 Proposed method of Feature Extraction by Deep Learning

From the above section it is evident that as the model architecture goes deep, it starts to learn high-level features from the low-level features. We propose to use these higher-level features for the feature representation of images in a CBIR system. So, we removed the last softmax activation layer used for calculating probabilities of each class and selected the preceding fully connected layer to be our feature vector representation for CBIR. As this vector is the deepest layé of the model, this represents the most learned high-level features. We encoded (predicted) the images of our CBIR database through our pre-trained model and got an n-dimensional feature vector for each of the images. The flowchart of this process is shown in Figure 7 for better understanding. The value of n varies with the selection of deep learning network architecture.

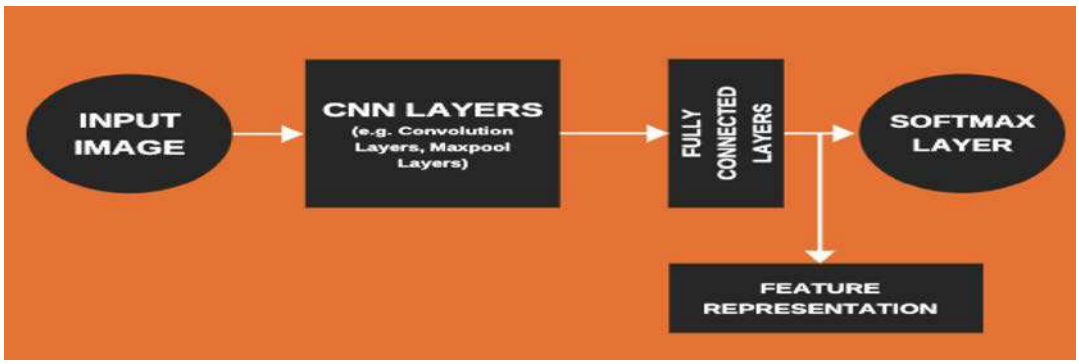


Figure 7: CBIR image feature representation with the help of pre-trained deep learning models

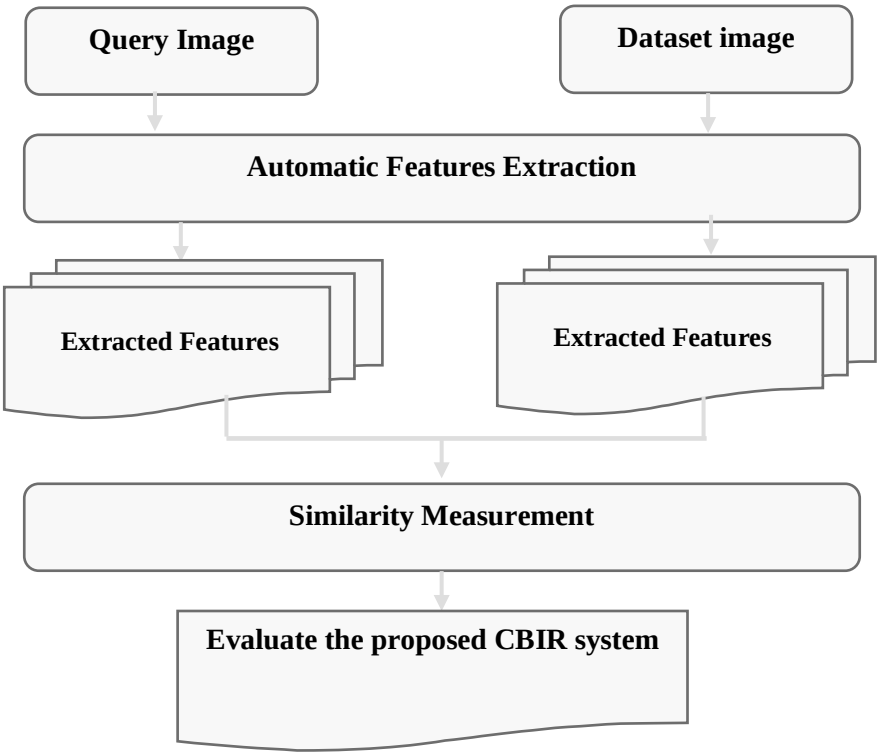


Figure 8: A block diagram of the matching image retrieval based on the proposed system.

3.6 Details of the proposed algorithm and performance evaluation :

3.6.1 Pre-trained models used :

We have tried the following network architectures all pre-trained on the ImageNet dataset:

- VGG16
- InceptionV3
- ResNet50
- Random CNN architecture

3.6.2 Inception V3 :

Inception-v3 is a convolutional neural network that is 48 layers deep. You can load a pretrained version of the network trained on more than a million images from the ImageNet database. The pretrained network can classify images into 1000 object categories, such as keyboard, mouse, pencil, and many animals. As a result, the network has learned rich feature representations for a wide range of images. The network has an image input size of 299-by-299.

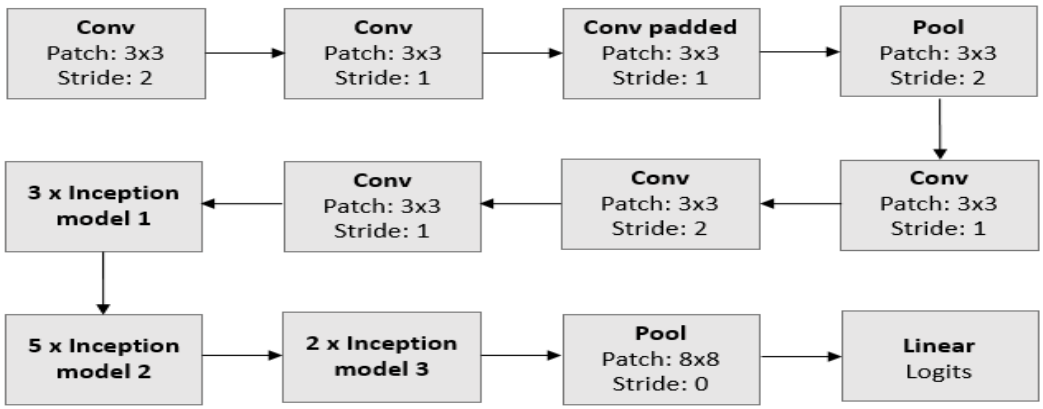


Figure 9: The-basic-architecture-of-Inception-v3

3.6.3 ResNet50：

ResNet, short for Residual Networks is a classic neural network used as a backbone for many computer vision tasks. This model was the winner of ImageNet challenge in 2015. The fundamental breakthrough with ResNet was it allowed us to train extremely deep neural networks with 150+layers successfully. Prior to ResNet training very deep neural networks was difficult due to the problem of vanishing gradients. AlexNet, the winner of ImageNet 2012 and the model that apparently kick started the focus on deep learning had only 8 convolutional layers, the VGG network had 19 and Inception or GoogleNet had 22 layers and ResNet 152 had 152 layers. In this blog we will code a ResNet-50 that is a smaller version of ResNet 152 and frequently used as a starting point for transfer learning.

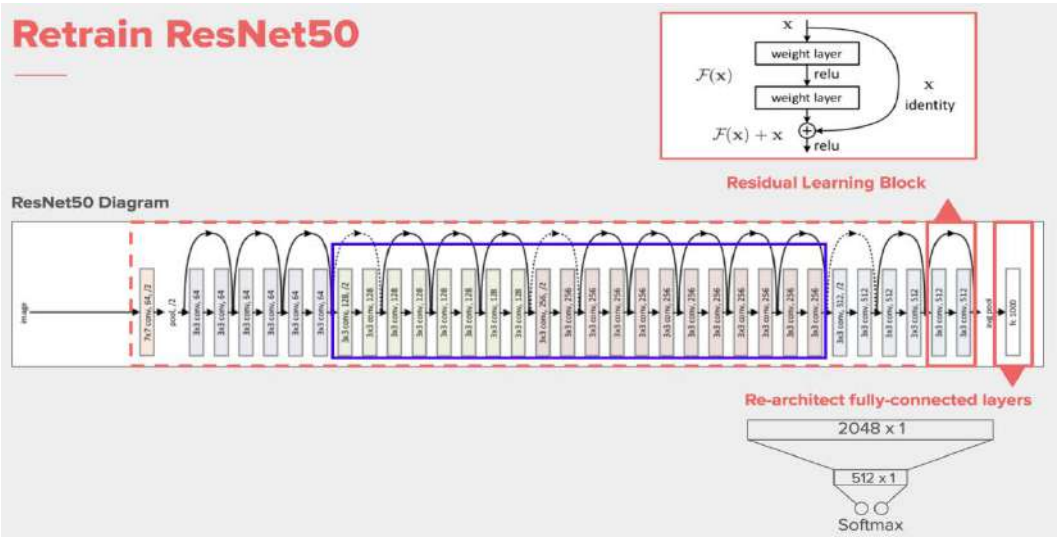


Figure 10: ResNet50 architecture

3.6.4 VGG16 architecture:

VGG16 is a convolutional neural network model proposed by K. Simonyan and A. Zisserman from the University of Oxford in the paper “Very Deep Convolutional Networks for Large-Scale Image Recognition”. The model achieves 92.7% top-5 test accuracy in ImageNet, which is a dataset of over 14 million images belonging to 1000 classes. It was one of the famous models submitted to ILSVRC-2014. It makes the improvement over AlexNet by replacing large kernel-sized filters (11 and 5 in the first and second convolutional layer, respectively) with multiple 3×3 kernel-sized filters one after another. VGG16 was trained for weeks and was using NVIDIA Titan Black GPU’s.

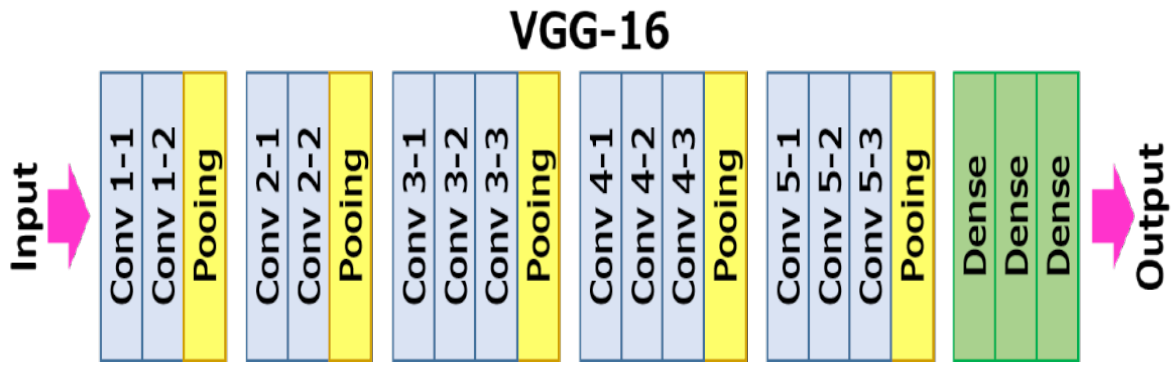


Figure 11: VGG-16 layers

ILSVRC uses a subset of ImageNet with roughly 1000 images in each of 1000 categories. At all, there are roughly 1.2 million training images, 50,000 validation images, and 150,000 testing images. ImageNet consists of variable-resolution images. Therefore, the images have been down-sampled to a fixed resolution of 256×256. Given a rectangular image, the image is rescaled and cropped out the central 256×256 patch from the resulting image.

The input to cov1 Layer is of fixed size 224 x 224 RGB image. The image is passed through a stack of convolutional (conv.) layers, where the filters were used with a very small receptive field: 3×3 (which is the smallest size to capture the notion of left/right, up/down, center). In one of the configurations, it also utilizes 1×1 convolution filters, which can be seen as a linear transformation of the input channels (followed by non-linearity). The convolution stride is fixed to 1 pixel; the spatial padding of conv. Layer input is such that the spatial resolution is preserved after convolution, i.e. the padding is 1-pixel for 3×3 conv. Layers. Spatial pooling is carried out by five max-pooling layers, which follow some of the conv. Layers (not all the

conv. layers are followed by max-pooling). Max-pooling is performed over a 2×2 pixel window, with stride 2.

Three Fully-Connected (FC) layers follow a stack of convolutional layers (which has a different depth in different architectures): the first two have 4096 channels each, the third performs 1000-way ILSVRC classification and thus contains 1000 channels (one for each class). The final layer is the soft-max layer. The configuration of the fully connected layers is the same in all networks.

All hidden layers are equipped with the rectification (ReLU) non-linearity. It is also noted that none of the networks (except for one) contain Local Response Normalisation (LRN), such normalization does not improve the performance on the ILSVRC dataset, but leads to increased memory consumption and computation time.

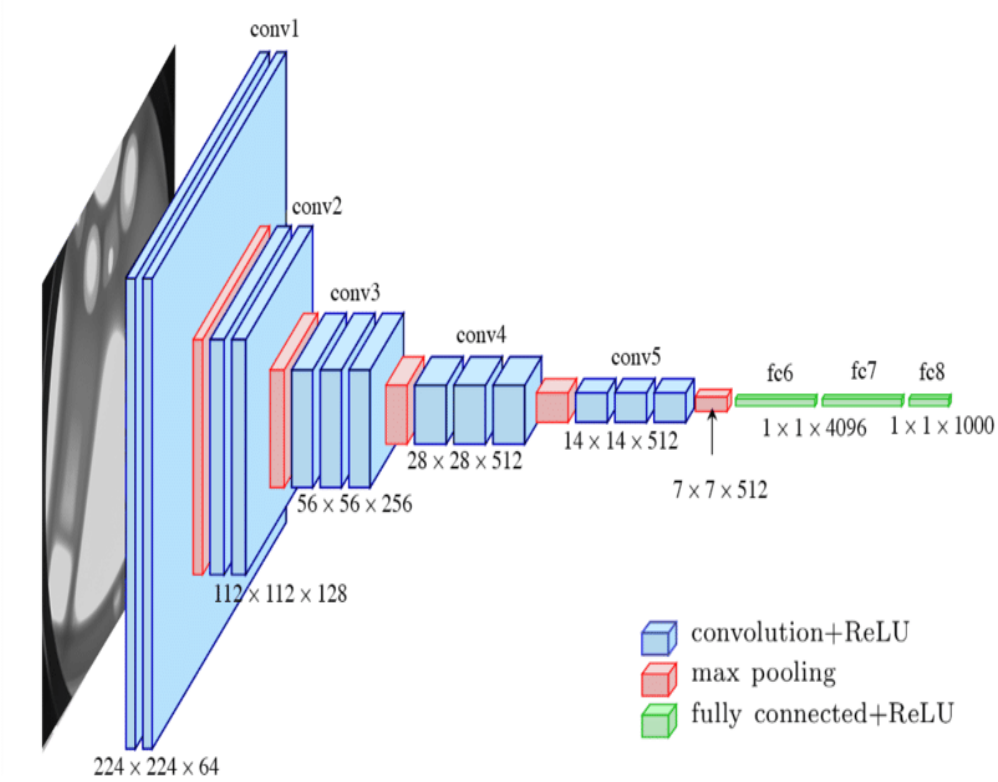


Figure 12 : The-standard-VGG-16-network-architecture

3.6.5 Random architecture model:

Tableau 2: Model sequential

Layer (type)	Output Shape	Param #
conv2d (Conv2D)	(None, 144, 144, 128)	18944
max_pooling2d (MaxPooling2D)	(None, 71, 71, 128)	0
conv2d_1 (Conv2D)	(None, 67, 67, 64)	204864
max_pooling2d_1 (MaxPooling2	(None, 32, 32, 64)	0
Flatten (Flatten)	(None, 65536)	0
Dense (Dense)	(None, 128)	8388736
dense_1 (Dense)	(None, 10)	1290

Total params: 8, 613,834
Trainable params: 8, 613,834
Non-trainable params: 0

3.6.6 Dataset used:

WANG Dataset: The dataset contains 1000 images from 10 different categories each containing 100 images.

The categories are African people, Beach, Buildings, Buses, Dinosaurs, Flowers, Elephants, Horses, Mountains, Food



Figure 13: Wang dataset classes

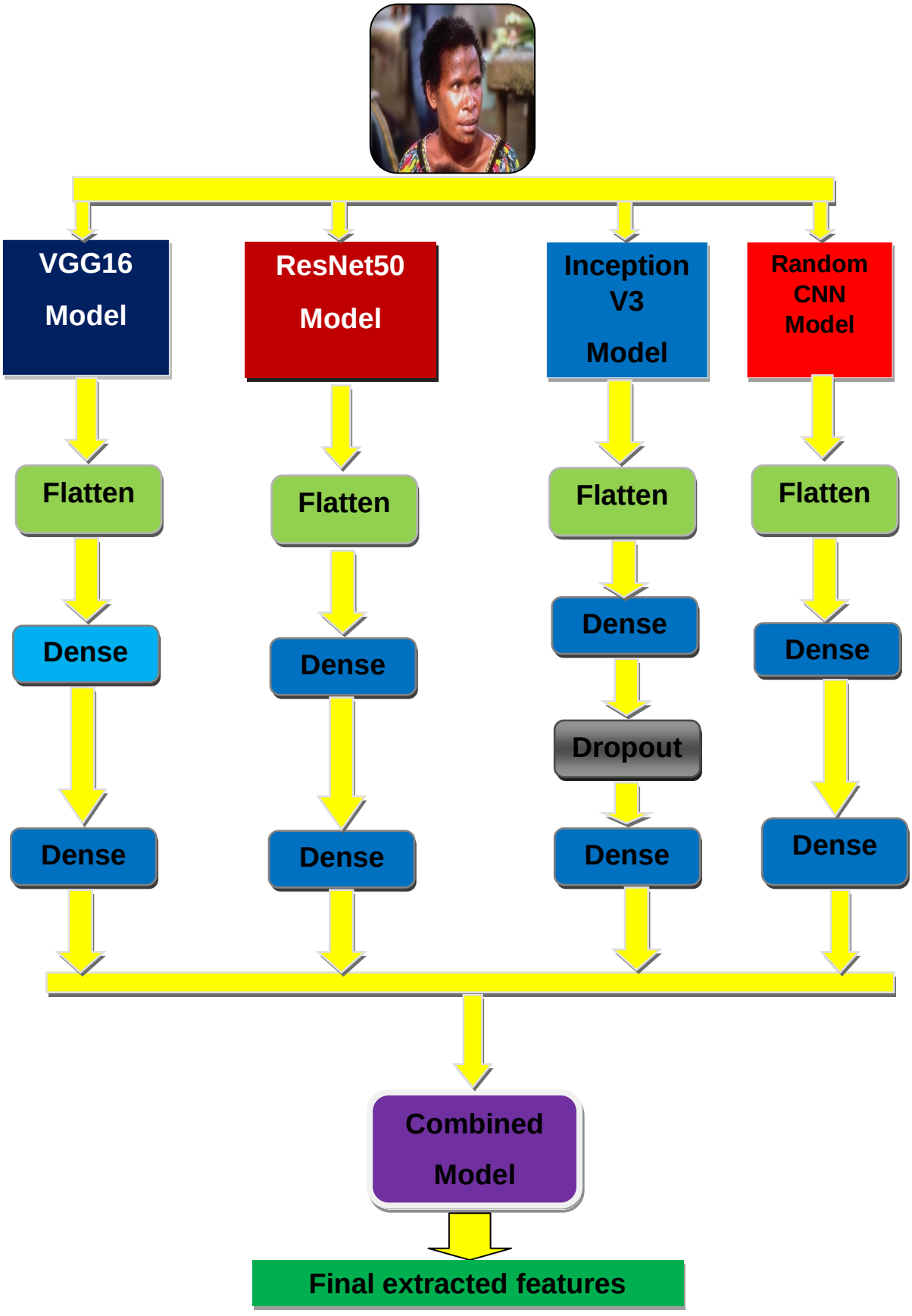


Figure 14: The proposed framework of Automatic features extraction

3.7 Similarity Measure

The similarity (or dissimilarity) between a query image (Q) given by the user and a database image (I) stored in the system is measured by some distance metric. It is assumed that this distance will accurately measure the dissimilarity (or similarity) between the images as well. A smaller calculated distance implies more similarity. The similarity between two images can be different for different user's way of perceiving the images. Broadly speaking, there are mainly two types of similarity measures, geometric and probabilistic [53]. In the first case, the similarity is based on the distance between the feature vectors. A most widely used one is Minkowski, of which L1-norm and L2-norm are most popular. In probabilistic type, a Gaussian classifier is often used to measure the relevance between the query image and the database image so that the pairs that had a high likelihood ratio were classified as relevant and the pairs having a low Likelihood ratio is considered as irrelevant. Although past research [54] shows that the probabilistic methods perform significantly better than the geometric methods, they are computationally expensive. In this work we have used L2 norm as similarity measure.

Euclidean Distance (L2 norm) between Q and I is given by :

$$D(I,Q) = \sum_{i=1}^{n} (x_i, I - x_i, Q)^2$$

Where, n is the feature dimension of the images. $x(i;\ I)$ and $x(i;Q)$ are the i-th feature value of database image and query image respectively.

3.8 Performances:

Accuracy:

Accuracy is defined as Number of Correct Predictions by Total Number of Predictions given as

$$\boldsymbol{accuracy = \frac{TP + TN}{TP + TN + FN + FP}}$$

Recall:

Recall is defined as what proportion of actual positives was identified correctly is given as

$$\boldsymbol{Recall = \frac{TP}{TP + FN}}$$

Precision:

Precision is characterized as what extent of positive identification is actually right is given as

$$\boldsymbol{Precision = \frac{TP}{TP + FP}}$$

F1-score:

The F-score, also called the F1-score, is a measure of a model's accuracy on a dataset. It is used to evaluate binary classification systems, which classify examples into 'positive' or 'negative'.The F-score is a way of combining the precision and recall of the model, and it is defined as the harmonic mean of the model's precision and recall.The F-score is commonly used for evaluating information retrieval systems such as search engines, and also for many kinds of machine learning models, in particular in natural language processing.It is possible to adjust the F-score to give more importance to precision over recall, or vice-versa. Common adjusted F-scores are the F0.5-score and the F2-score, as well as the standard F1-score.

F-score Formula :

The formula for the standard F1-score is the harmonic mean of the precision and recall. A perfect model has an F-score of 1.

$$F1 = 2 \times \frac{Precision * Recall}{Precision + Recall}$$

3.9 Confusion matrix:

Confusion matrix is a very popular measure used while solving classification problems. It can be applied to binary classification as well as for multiclass classification problems. An example of a confusion matrix for binary classification is shown in Table 2.

Tableau 3: Confusion matrix for binary classification.

		Predicted- class	
		Negative	**Positive**
Actual-class	**Negative**	TN True negatif	FP False positif
	Positive	FN False negatif	TP True positif

Confusion Matrix is used to know the performance of a Machine learning classification. It is represented in a matrix form.Confusion Matrix gives a comparison between Actual and predicted values.The confusion matrix is a N x N matrix, where N is the number of classes or outputs.

For 2 class, we get 2 x 2 confusions matrix.

For 3 class, we get 3 X 3 confusion matrix.

Confusion Matrix has 4 terms to understand True Positive(TP), False Positive(FP), True Negative(TN) and False Negative(FN).

Confusion matrices represent counts from predicted and actual values. The output “TN” stands for True Negative which shows the number of negative examples classified accurately. Similarly, “TP” stands for True Positive which indicates the number of positive examples

classified accurately. The term “FP” shows False Positive value, i.e., the number of actual negative examples classified as positive; and “FN” means a False Negative value which is the number of actual positive examples classified as negative.

3.10 Tools used in this work:

Our classification algorithm is implemented in Keras and tested on a PC with Intel(R) Core(TM) i7-4510U CPU with RAM 4 Go. Windows 10 Professionnel

We used in this work anaconda navigator for programming exactly jupyter notebook ($pyt\hbar on$ 3.8)

- Python libraries:

1) Keras & tensorflow used for image pre-processing

2) To build the convolutional neural network, the Python library Keras using the Tensorflow backend was used.

3) To normalize and calculate L2-distance between feature vectors, the Python library Numpy was utilized.

4) Other libraries ($pat\hbar\ lib, pilow$) were used for fetching and managing images.

5) Matplotlib for visualize results.

3.11 Programmation steps:

Import the importante libraries

Train set /Test set the dataset

Creat random CNN model and train it

Creat vgg16 model and train it

Creat ResNet50 model and train it

Creat InceptionV3 model and train it

Creat combine model by merging all previous models

Make predictions by testing every model

Extract features :

Creat class « features extractor » Containing the following steps **:**

1. Creat init function Which defines our extracting model

2. Creat extract function containing :

    a) Resize the images

    b) Convert the image color spase (RGB)

    c) Reformat the image

    d) Extract the image features

Importing our « features extractor » and extract wang dataset features

Saving the extracted features in a folder

Retrieve images :

Read image features

Insert the image query

Import our « features extractor » and extract query features

Calculate similarity bitween query and dataset features

Visualize retrieved images

CHAPTER IV : RESULTS

4.1Results and experiments :

Introduction

This chapter presents results and discussions of performance measures for every model. The performance evaluation, the results are given in the form of Figures and tables, We will going to visualise all categories results of the combined model in the appendix.

4.2Train set/Test set of Wang dataset:

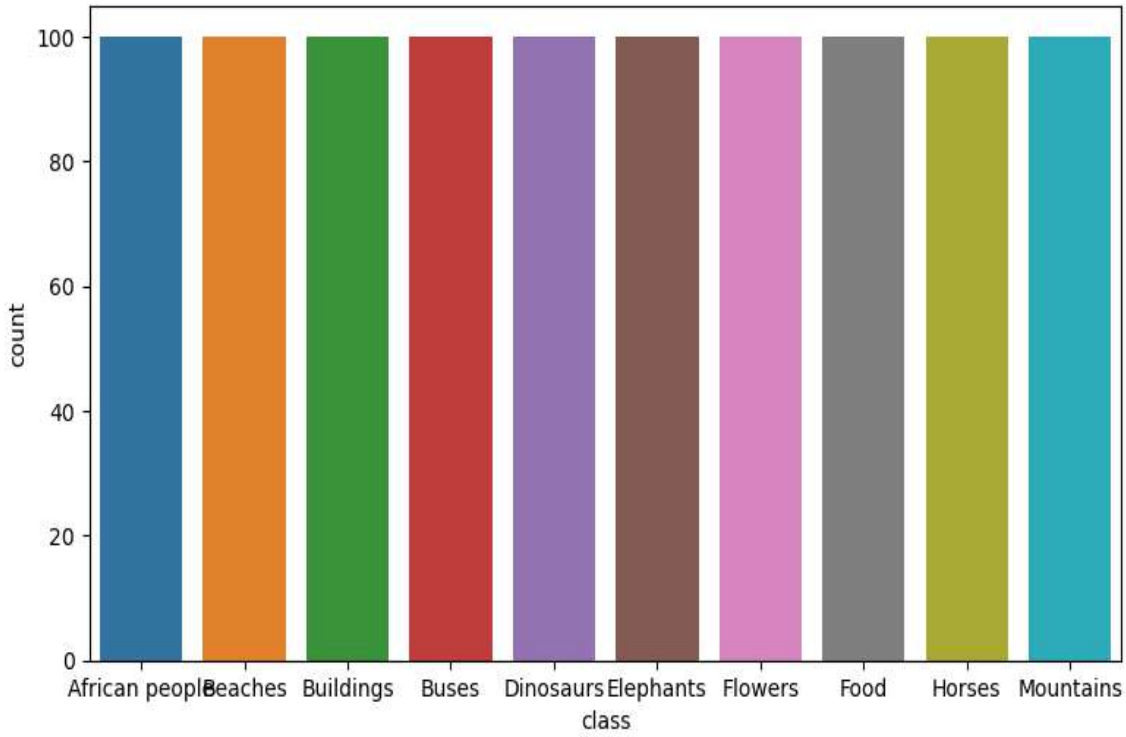


Figure 15: Wang distribution classes

4.3.1Trainig random CNN model

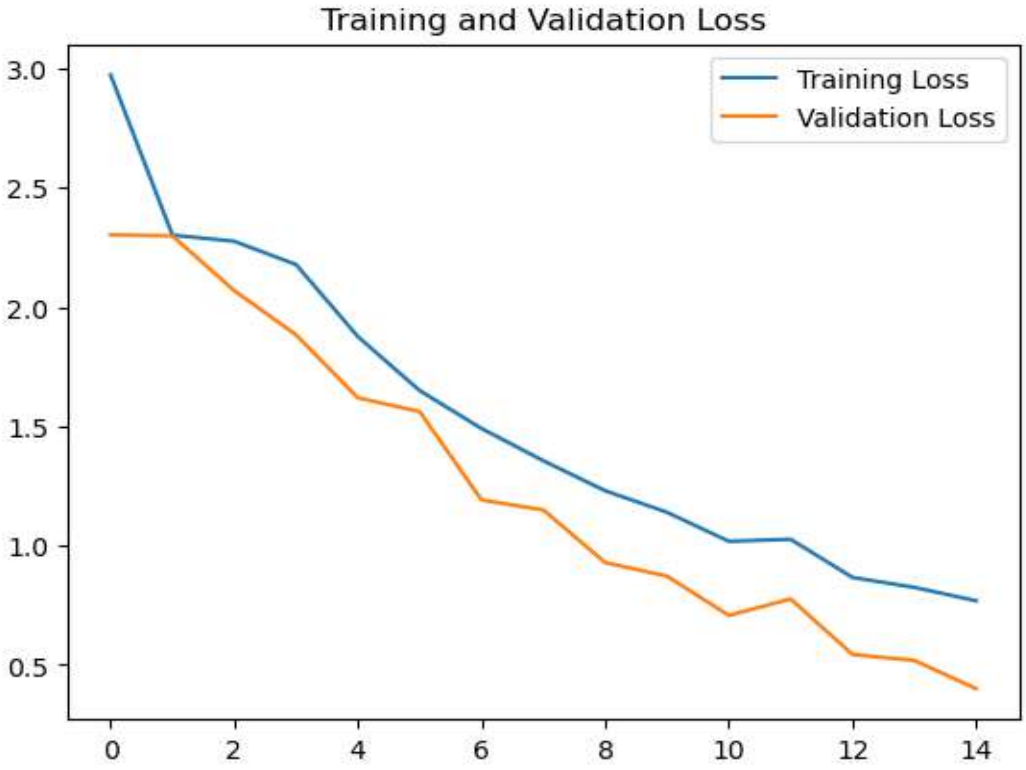


Figure 16: Training validation and loss of random model

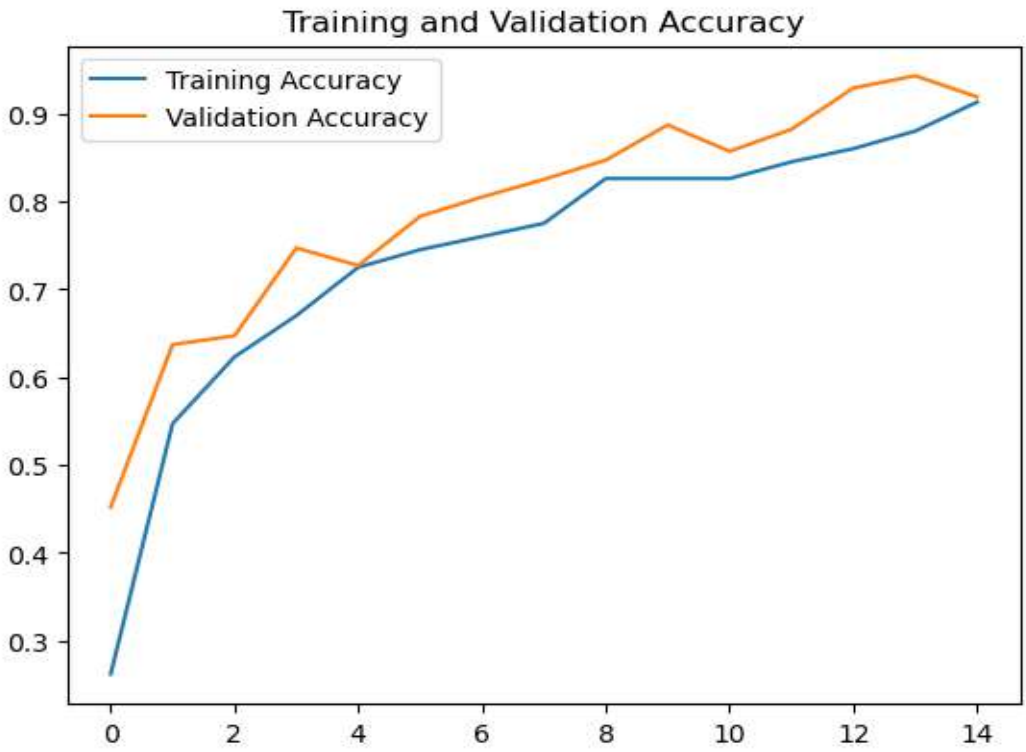


Figure 17: Training and validation accuracy of random model

4.3.2 Random model classification report:

Tableau 4: performances of every class

	Precision%	Recall%	F1-score%	Support
African people	60	72	65	100
Beaches	71	27	39	100
Buildings	46	86	60	100
Buses	89	85	87	100
Dinosaurs	96	100	98	100
Elephants	79	88	83	100
Flowers	96	51	67	100
Food	19	14	16	100
Horses	81	65	72	100
Mountains	29	39	33	100

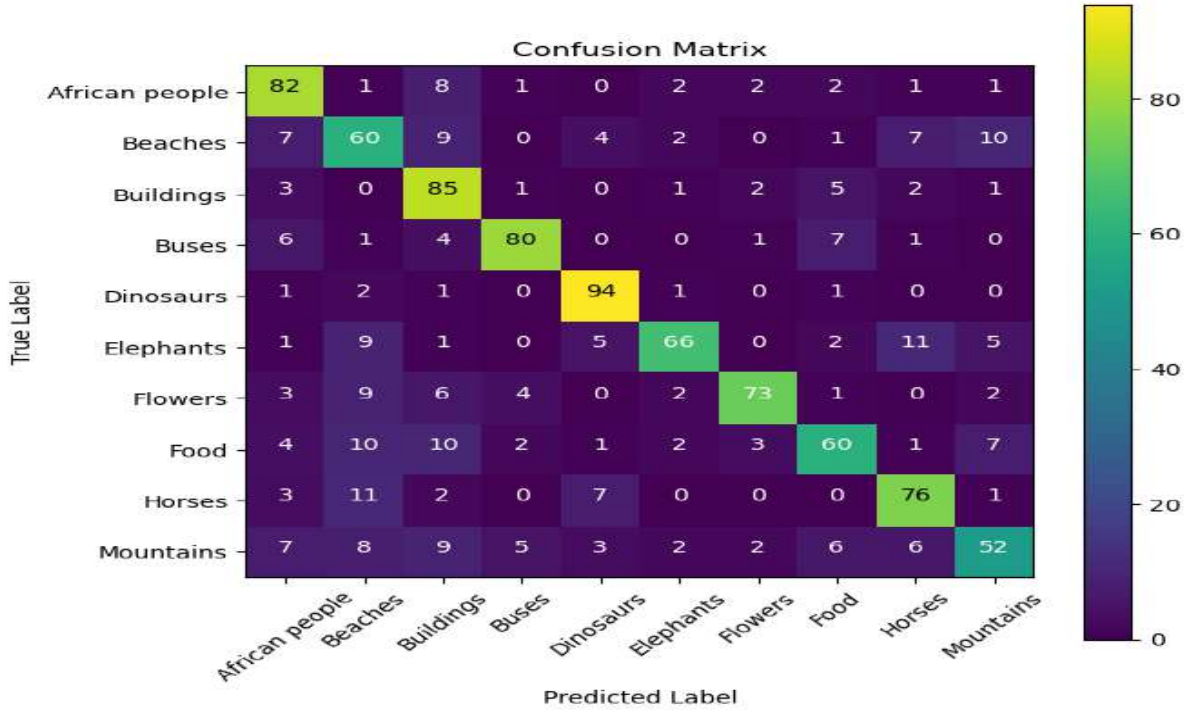


Figure 18: confusion matrix of random model

4.4.1 Training VGG16 model:

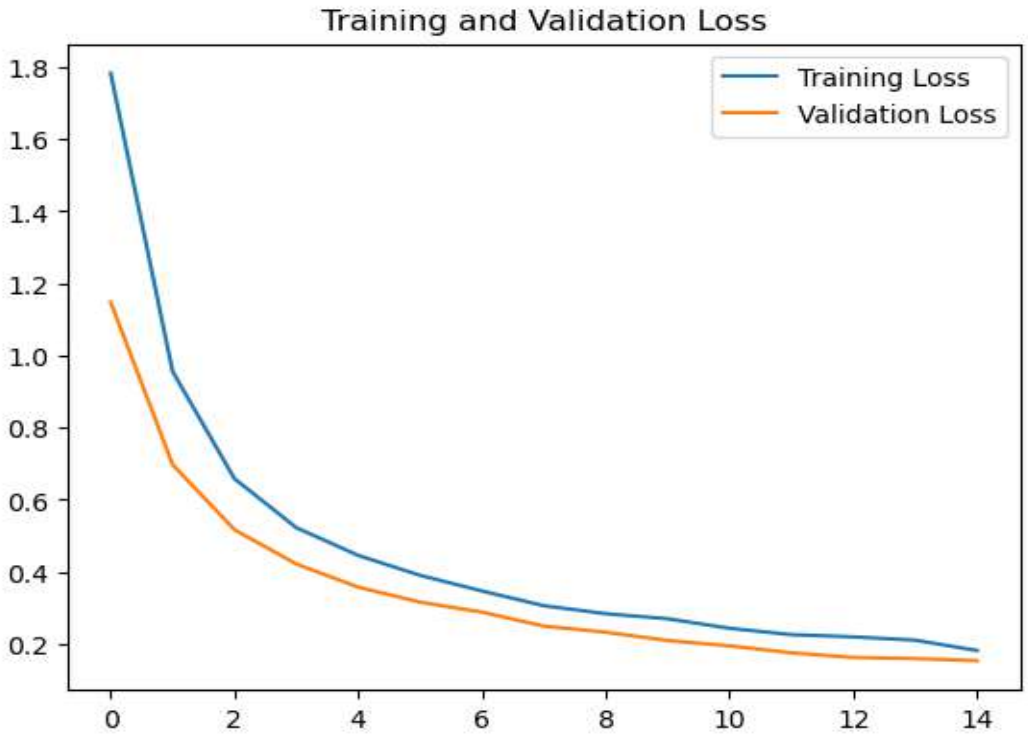


Figure 19: Training and validation loss of vgg16 model

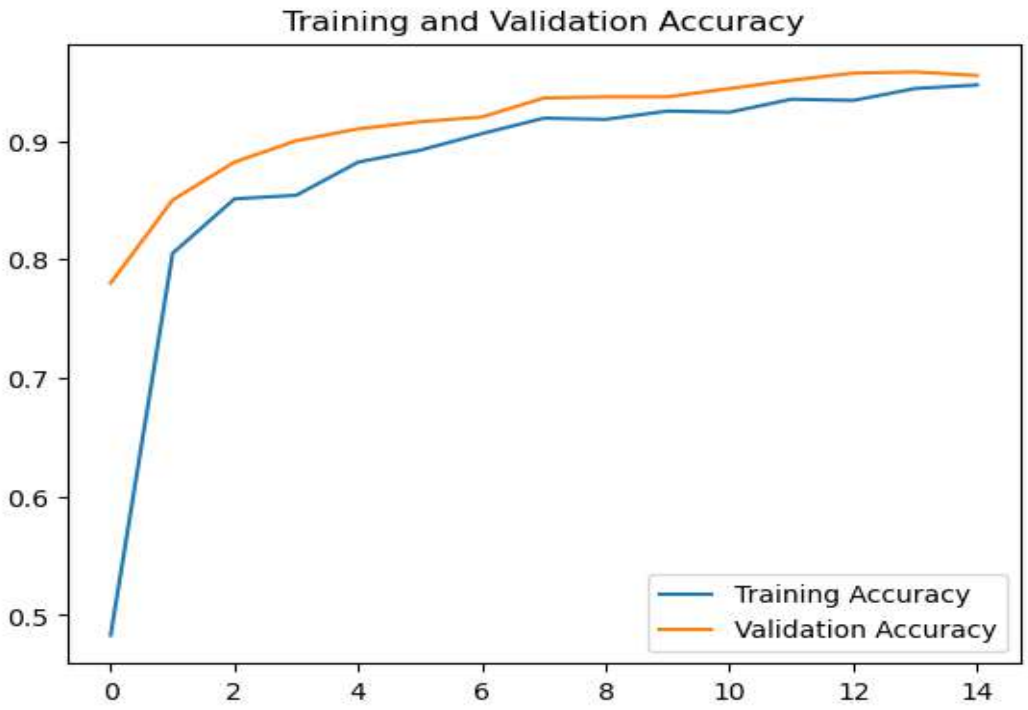


Figure 20: Training and validation accuracy of vgg16 model

4.4.2 Vgg16 model classification report:

Tableau 5: performances of every class

	Precision%	Recall%	F1-score	Support
African people	92	77	84	100
Beaches	83	75	79	100
Buildings	90	95	93	100
Buses	99	100	100	100
Dinosaurs	100	100	100	100
Elephants	87	97	92	100
Flowers	98	100	99	100
Food	94	91	92	100
Horses	96	96	96	100
Mountains	85	93	89	100

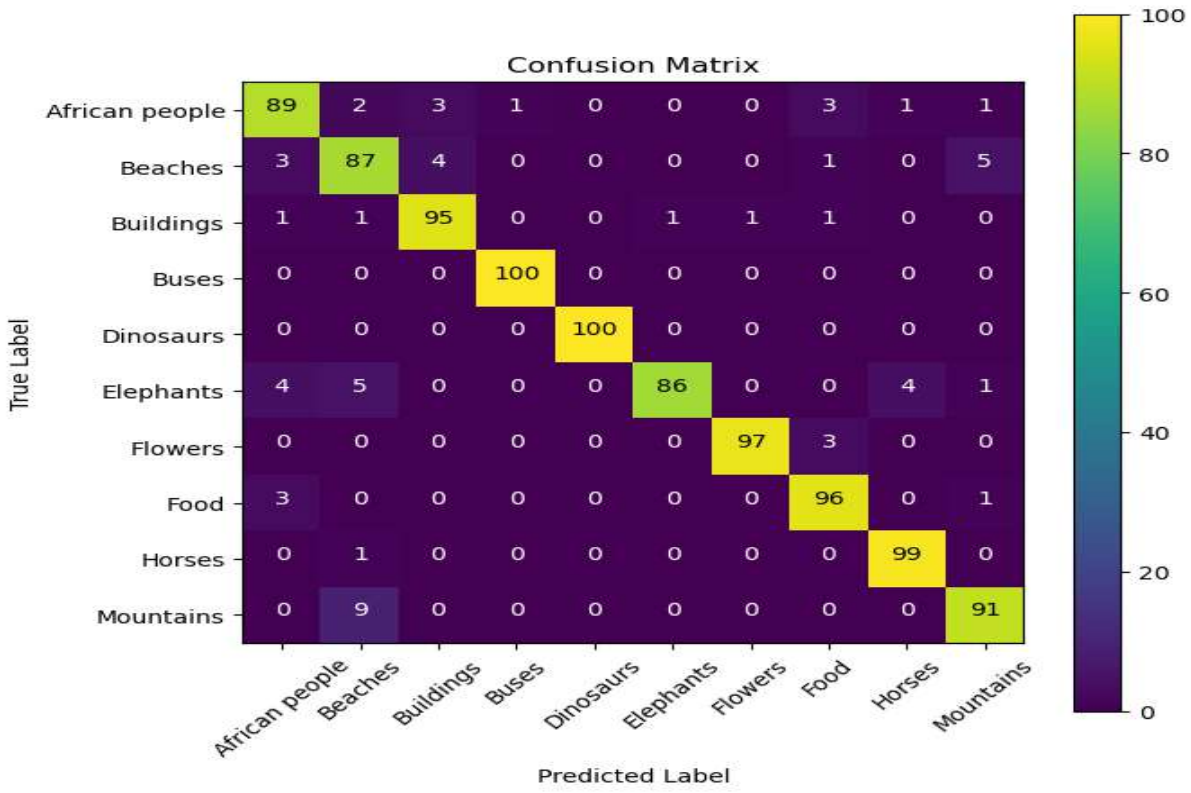


Figure 21: confusion matrix of vgg16 model

4.5.1 Training ResNet50 model:

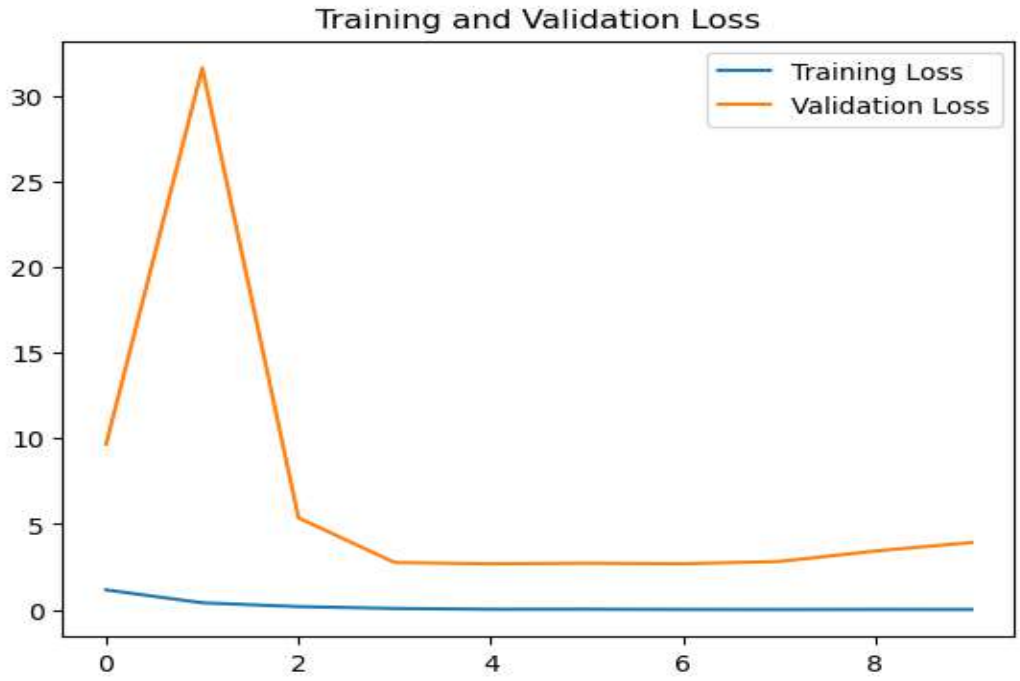


Figure 22: Training and validation loss of ResNet50 model

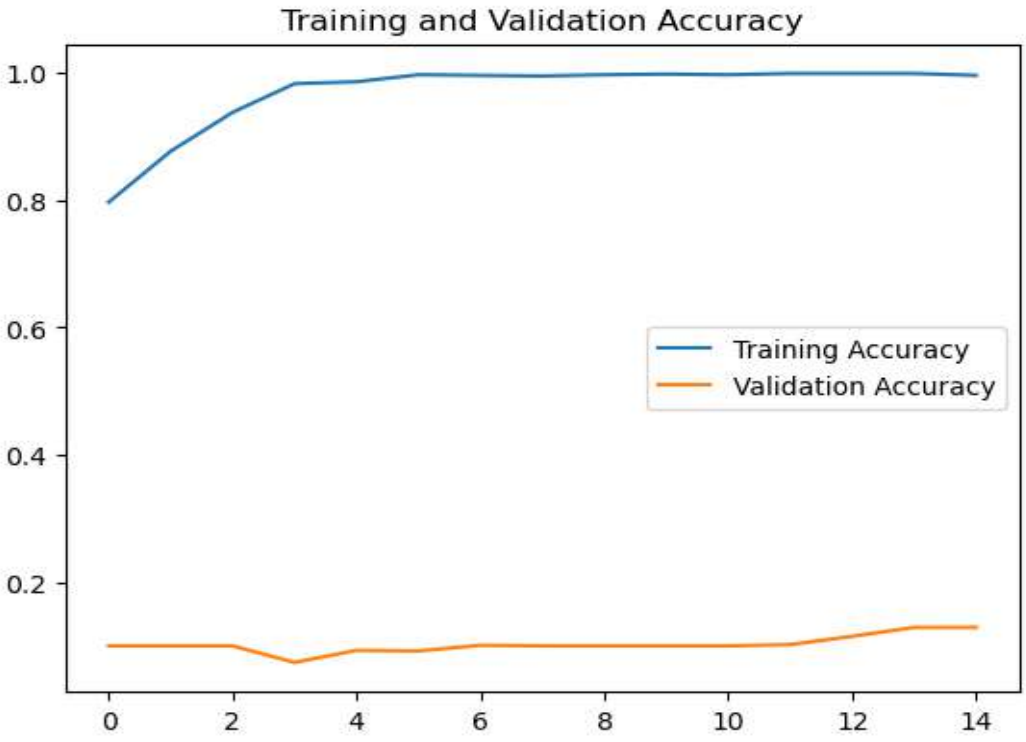


Figure 23: Training and validation accuracy of ResNet50 lodel

4.5.2 Resnet50 classification report:

Tableau 6: performances of every class

	Precision%	Recall%	F1-score%	Support
African people	0	0	0	100
Beaches	0	0	0	100
Buildings	10	100	19	100
Buses	0	0	0	100
Dinosaurs	86	30	44	100
Elephants	0	0	0	100
Flowers	0	0	0	100
Food	0	0	0	100
Horses	0	0	0	100
Mountains	0	0	0	100

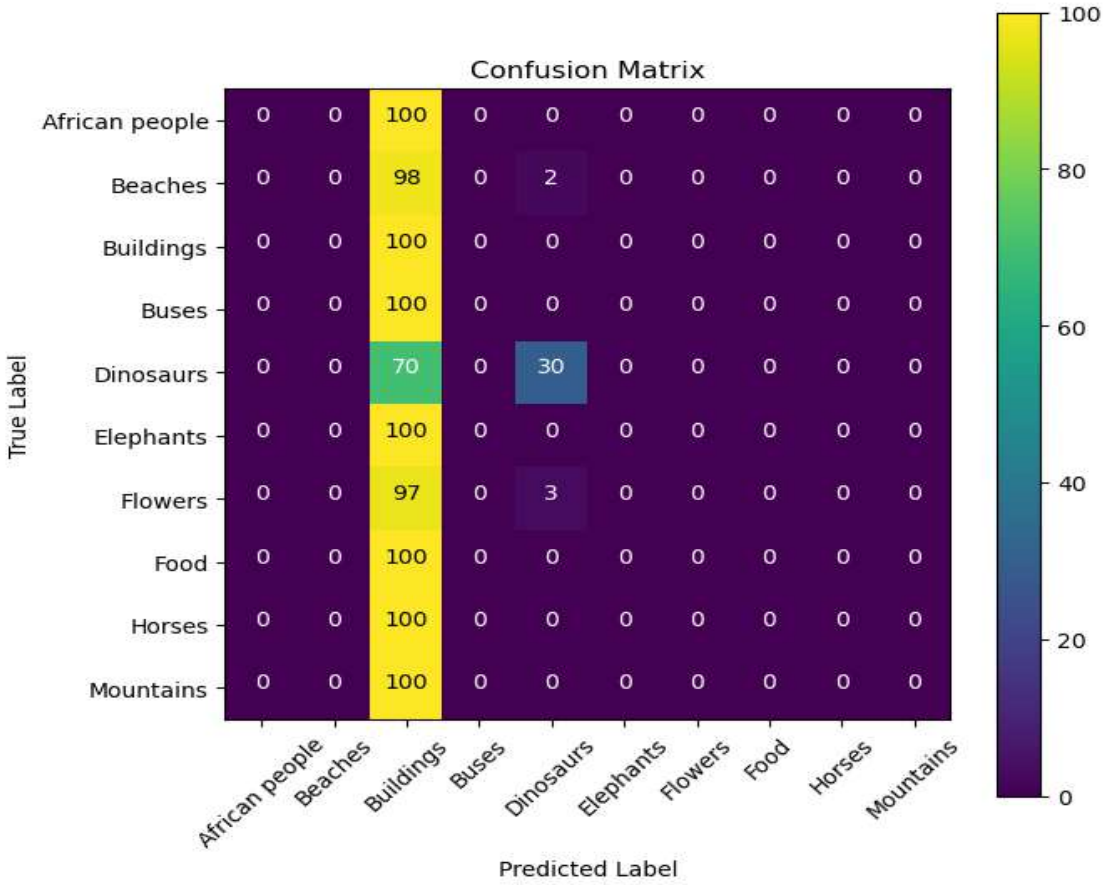


Figure 24: confusion matrix of ResNet50 model

4.6.1 Training InceptionV3 model :



Figure 25: Training and validation liss of InceptionV3 model

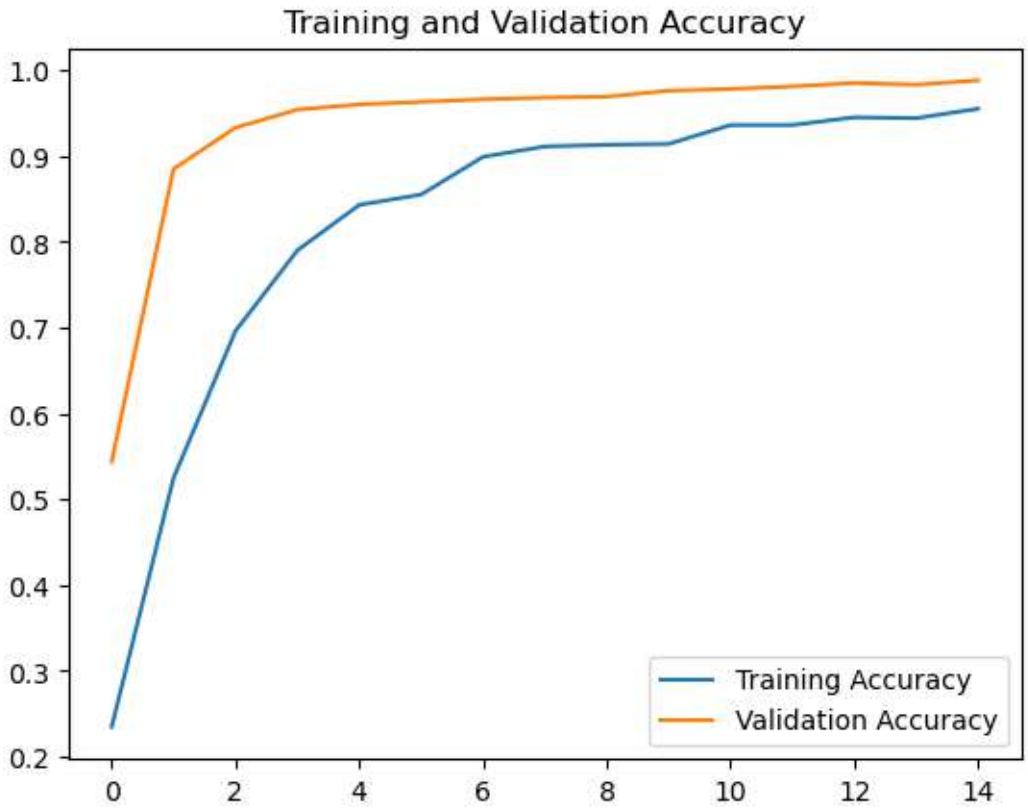


Figure 26: Training and validation accuracy of InceptionV3 model

4.6.2 InceptionV3 classification report :

Tableau 7: performances of every class

	Precision%	Recall%	F1-score%	Support
African people	78	91	84	100
Beaches	90	86	88	100
Building	86	86	86	100
Buses	100	96	98	100
Dinosaurs	100	98	99	100
Elephants	95	94	94	100
Flowers	94	99	97	100
Food	93	87	90	100
Horses	98	88	93	100
Mountains	88	92	90	100

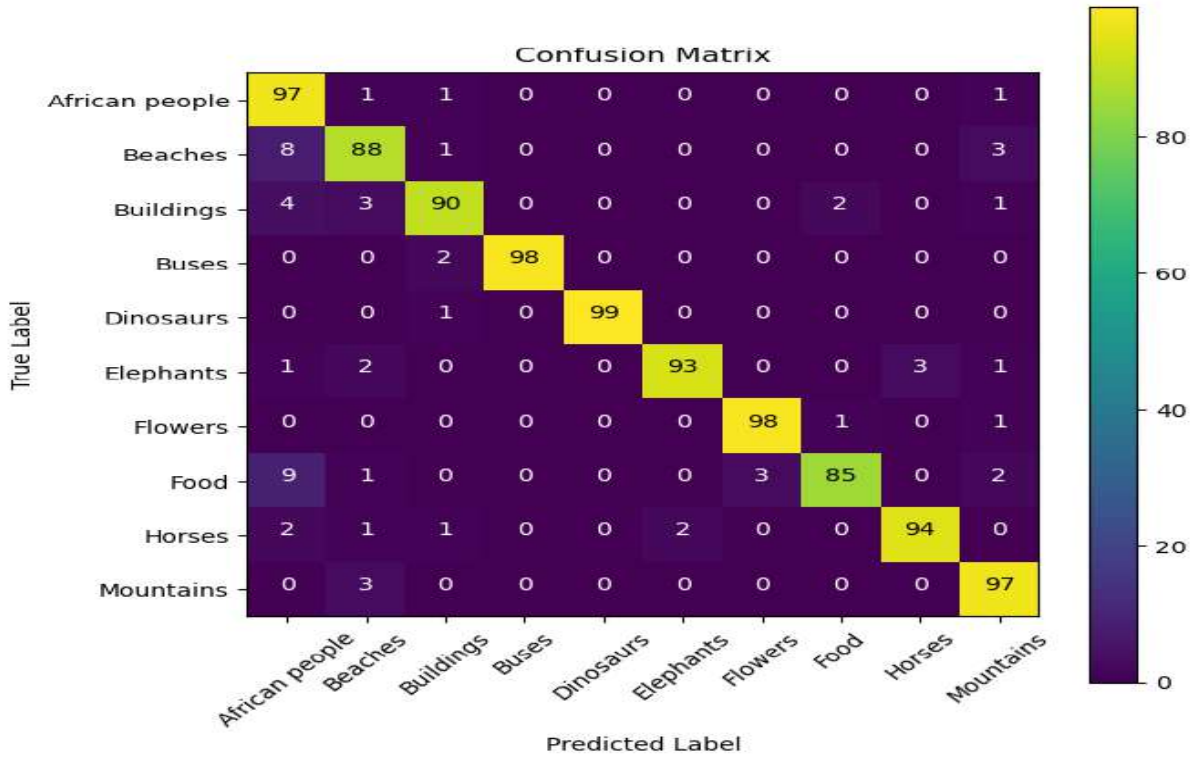


Figure 27: Confusion matrix of inceptionV3 model

4.7 Combined model classification report:

Tableau 8: performances of every class

	Precision%	Recall%	F1-score	Support
African people	92	77	84	100
Beaches	80	91	85	100
Buildings	92	92	92	100
Buses	99	100	100	100
Dinosaurs	100	100	100	100
Elephants	92	97	94	100
Flowers	98	100	99	100
Food	94	91	92	100
Horses	97	96	96	100
Mountains	93	90	91	100

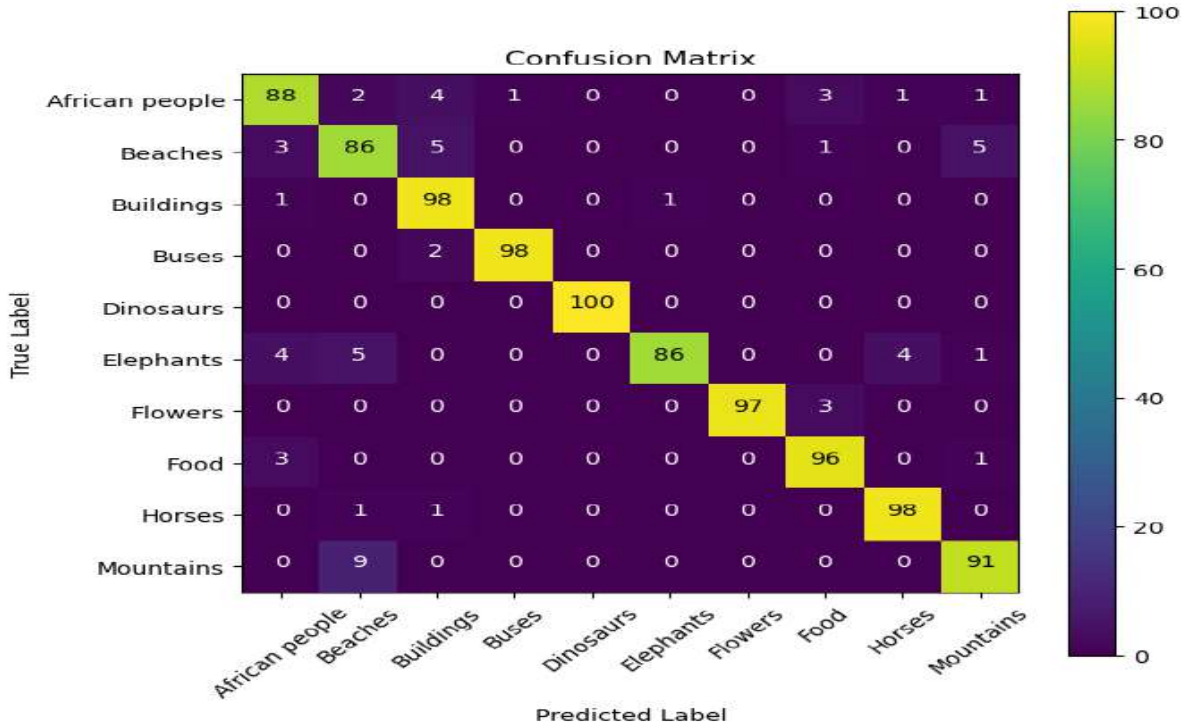


Figure 28: Confusion matrix of combined model

4.8 Discussion of the results :

Tableau 9: Compare model's performances

model / performonces	VGG16	ResNet50	InceptionV3	Random model	Combined model
Accuracy%	92.4	13	91.7	62.7	93.4
Precision %	92	10	92	67	94
Recall %	92	13	92	63	93
F1-score%	92	6	92	62	93

In this work, we demonstrated that features from pre-trained deep CNN model are suitable for image retrieval task. We created a CBIR framework using features from pre-trained VGG16 ResNet50 and InceptionV3 deep CNN models by adding random architecture of CNN. We tested our proposed CBIR framework on a multi-class dataset of Wang images.

VGG16 and InceptionV3 was good for extracting and predicting classes, but the ResNet50 model was absolutely unsatisfied perhaps it was need big large-scale dataset biggest that our sed dataset and the big raison that we haven't a multi-core processor for training it, the random architecture that we propose was satisfied.

We achieved average retrieval precision of 94% for the proposed CBIR framework, this value of precision could be higher than that. The improvement in average precision is seen in all classes of the dataset and number of images retrieved (fetch size). From experimental results we conclude that features from pre-trained deep CNN model can be successfully used for image retrieval task and perform better than handcrafted features used in traditional CBIR methods. Also, the combined model features are better suited to perform image retrieval in wang dataset.

A few predictions for the retrieved results :

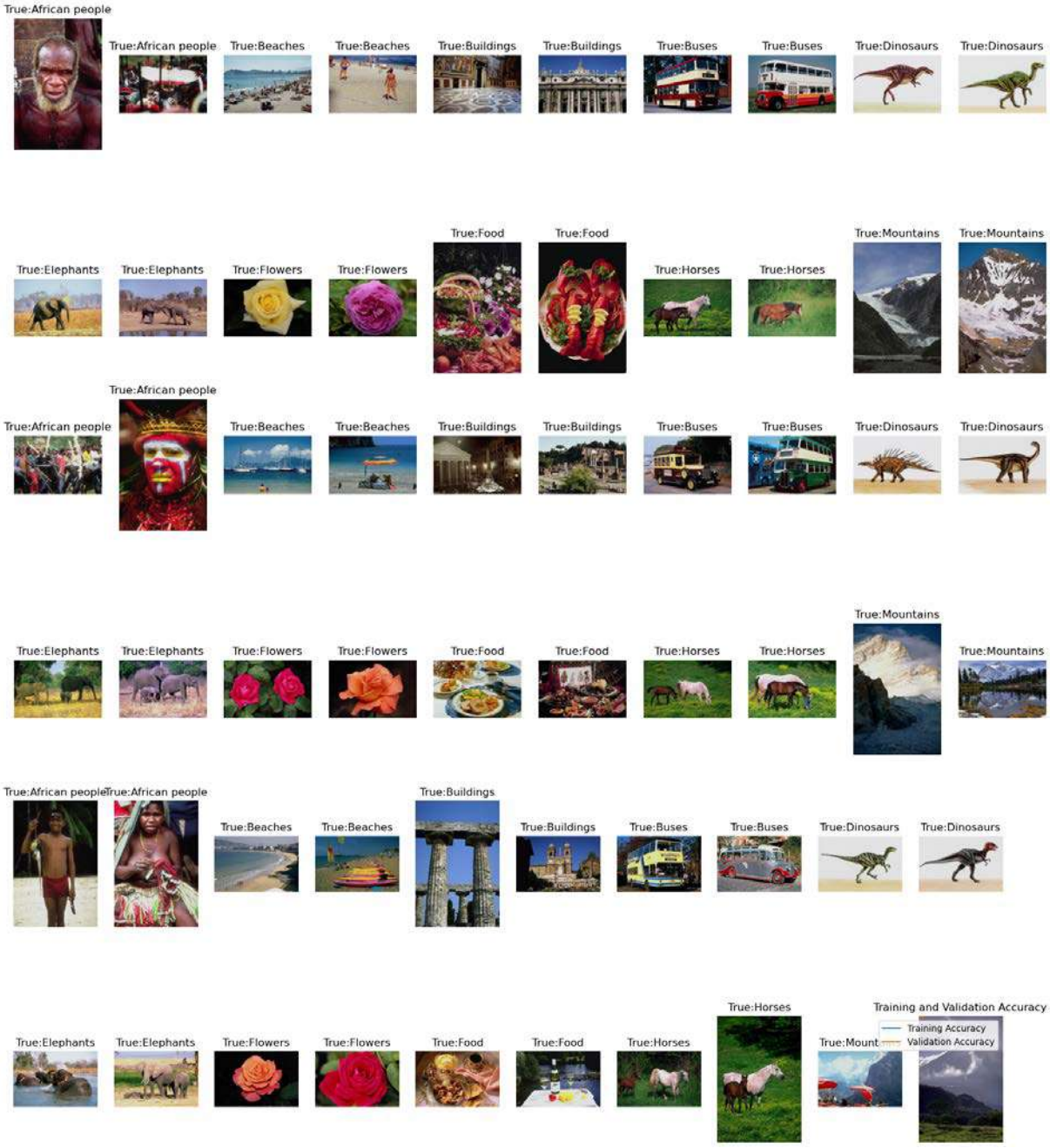


Figure 29 : predicted results

CHAPTER V : CONCLUSION

5.1 Conclusion:

Deep Neural Networks have created new prospects for solving image recognition problems with Convolutional Neural Networks as a suitable model to capture visual information in images. Nevertheless there is still a great distance to cover in order to reach the destination of using them as universally applicable visual feature extractors. The various architectures and methods of the current state-of-the-art are connected with increased computational costs. Moreover their superior accuracy is achieved in controlled conditions with a predefined number of categories and hundreds of samples for each one of them.

In order to use CNNs in real-world applications, like robot agents and medical imaging equipment, they should become scalable allowing the addition of new samples of unknown categories, without the need of fine-tuning the entire model. Also a small number of samples for each category used to train the network should be sufficient for accurate and precise prediction. New architectures and methods are needed which will be not overfit on the ground truth image set. A decreased overfitting during supervised gradient descent training can be an indicator of a more generic CNN feature extractor.

5.2 Dataset and Classes:

In this study we worked with only 1000 images divided in 10 classes. In future to make the system more generalized and efficient the dataset can be increased and more number of classes can be added. Just by considering 10 classes we showed that the system is efficient and even if more classes were added to it the performance of the system would only increas

5.3 CBIR with the Visual Features Language:

This work has not sufficiently addressed the problem of discovering well performing CNN-based features for increased CBIR performance. The experimental results showed inadequate performance of the overall model. Nevertheless the evaluation system used for experiments could be considered a systematic approach to CBIR research. Moreover the complexity is fully explored and the factors that contribute to the CNN-based CBIR performance are identified. This can be an appropriate starting point for a more targeted research that will aim in increasing the retrieval metrics. A contribution may be considered the Visual Features Language that generalizes the Bag-of-Visual-Words model and extends it with characteristics of a natural language like syllables and stopwords.

The creation of fabricated textual words may be proven beneficial, as the image retrieval problem is abstracted into the more comprehensible text retrieval problem. Existing image collections can be translated into VFL datasets and reused as text collections for IR research, without the need of the CNN and the clustering model. The interesting observation that was made through the CBIR experiments was the appearance of Zipf's law in visual features, which needs to be further studied in order to support a theory: Quanta of visual information inside images follow the Zipfian distribution. An evaluation of different clustering methods for the creation of the VFL could not fit in the time restrictions of this Master's thesis. The Rooting k-Means algorithm is fast, simple for the current needs and suitable for low computational resources. Its quality needs to be evaluated individually in different problem domains and compared with the constant branching factor of Hierarchical k-Means and the binary tree of Bisecting k-Means. Creating a content-based image retrieval task in an annual competition that will a large scale image collection could bring advances in image retrieval, like the advances in image classification which have been achieved through ILSVRC.

For improving Content-Based Image Retrieval based on CNN features, there is a vast search space that is shaped by many factors. Individual solutions may include an ensemble of high-accurate CNN classifiers trained on the same data, providing more than one feature descriptors for an image patch on the same spatial resolution. This increase in dimensionality could be handled by an Autoencoder that will embed dimensionality reduction directly in the deep architecture. The last convolutional module in this multi-functional Deep Neural Network would feed both the Classifier and the Autoencoder. Investigating multiple depths and rooting factors for the Rooting k-Means clustering alone could lead to increased image retrieval performance. Testing on a variety of image collections from different problem domains could discover a universally applicable model as a solution for Intelligent Vision.

Bibliography:

[1] Chen, C., Wang, J.Z., and Krovetz, R., "Content-Based Image Retrieval by Clustering", Department of Computer Science, University of New Orleans, New Orleans, 2003.

[2] Mojsilović, A., Kovačević, J., Hu, J., Safranek, R.J., and Ganapathy, S.K., "Retrieval of Color Patterns Based on Perceptual Dimensions of Texture and Human Similarity Rules", Bell Laboratories, Lucent Technologies, Murray Hill, New Jersey, 2002.

[3] Long, F., Zhang, H.J., and Feng, D.D., "Fundamentals of Content-Based Image Retrieval", Microsoft Research, Asia, 2003.

[4] Muller, H., Michoux, N., Bandon, D., and Geissbuhler, A., "A Review of Content-Based Image Retrieval Systems in Medical Applications: Clinical Benefits and Future Directions", Division for Medical informatics, University Hospital of Geneva, Switzerland, 2003.

[5] Zhang, Y., "On The Use of CBIR in Image Mosaic Generation", Technical Report, Dept. of Computing Science, University of Alberta, Edmonton, Alberta, Canada, July, 2002.

[6] Tian, Q., Sebe, N., Lew, M.S., Loupias, E., and Huang, T. S. , " Image Retrieval Using Wavelet-Based Salient Points", Journal of Electronic Imaging, Special Issue on Storage and Retrieval of Digital Media, Vol.10(4), pp. 835-849, October, 2001.

[7] Rui, Y., and Huang, T.S., and Chang, S., "Image Retrieval: Past, Present, and Future", Department of Electrical Engineering & New Media Technology Center, Columbia University, New York, 1998.

[8] Veltkamp, R.C., and Tanase, M., "Content-Based Image Retrieval Systems: A Survey", Department of Computing Science, Utrecht University, Netherlands, 2001.

[9] Celebi, E., and Alpkocak, A., "Clustering of Texture Features for Content Based Image Retrieval", DokuzEylul University, Department of Computer Engineering, Izmir, Turkey, 2001.

[10]N. Yadav, A. Yadav, and M. Kumar, *An introduction to neural network methods for differential equations*.Springer, 2015.

[11] S. Che, M. Boyer, J. Meng, D. Tarjan, J. W. Sheaffer, and K. Skadron, "A performance study of general-purpose applications on graphics processors using CUDA," *J. Parallel Distrib.* Comput. vol. 68, no. 10, pp. 1370–1380, 2008.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105 ,2012,.

[13]D.E. Rumelhart, G .E. Hinton, and R.J. Williams, “Learning internal representations by error propagation,” DTIC Document, 1985.

[14] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[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 2008.

[16] X. Duanmu, “Image retrieval using color moment invariant,” in *Proceedings of the 2010 Seventh International Conference on Information Technology: New Generations (ITNG)*, pp. 200–203, IEEE, Las Vegas, NV, USA, April 2010.

[17] 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.

[18] H. Zhang, Z. Dong, and H. Shu, “Object recognition by a complete set of pseudo-Zernike moment invariants,” in *Proceedings of the 2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pp. 930– 933, IEEE, Dallas, TX, USA, March 2010.

[19] 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.

[20] 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.

[21] M. M. Islam, D. Zhang, and G. Lu, “Automatic categorization of image regions using dominant color based vector quantization,” in *Proceedings of the Digital Image Computing: Techniques and Applications*, pp. 191–198, IEEE, Canberra, Australia, December 2008.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] S. Fadaei, R. Amirfattahi, and M. R. Ahmadzadeh, “Local derivative radial patterns: a new texture descriptor for content-based image retrieval,” *Signal Processing*, vol. 137, pp. 274–286, 2017.

[29] D. Zhang and G. Lu. Shape retrieval using fourierdescriptors. Conference on Multimedia and Distance Education, 2001.

[30] P. Korn, N. Sidiropoulos, C. Faloutsos, E. Siegel, and Z.Protopapas. Fast and effective retrieval of medical tumor shapes. IEEE Trans KDE, 10:889-904, 1998.

[31] Rajiv Mehrotra and James E. Gary. Similar-shape retrieval in shape data management. IEEE Computer, 28(9):57-62, September 1995.

[32] M. Sharifi, M. Fathy, and M. T. Mahmoudi. A classified and comparative study of edge detection algorithms. In Information Technology: Coding and Computing, 2002. Proceedings. International Conference on, pages 117-120, 2002.

[33] Ying Liu, Dengsheng Zhang, and Guojun Lu. Region-based image retrieval with High-level semantics using decision tree learning. Pattern Recogn. 41(8):2554–2570, August 2008.

[34] Dengsheng Zhang, MdMonirul Islam, Guojun Lu, and Jin Hou. Semantic image retrieval using region based inverted file. In Proceedings of the 2009 Digital Image Computing: Techniques and Applications, DICTA ’09, pages 242–249, Washington, DC, USA, 2009. IEEE Computer Society.

[35] JiWan, DayongWang, Steven Chu Hong Hoi, PengchengWu, Jianke Zhu, Yongdong Zhang, and Jintao Li. Deep learning for content-based image retrieval: A comprehensive study. In Proceedings of the ACM International Conference on Multimedia, pages157–166. ACM, 2014.

[36] Alex Krizhevsky, IlyaSutskever, and Geoffrey E Hinton. Imagenet classification with deep convolution neural networks. In Advances in neural information processing systems, pages 1097–1105, 2012.

[37] N. Kondylidis, M. Tzelepi, and A. Tefas, “Exploiting tf-idf in deep convolutional neural networks for content based image retrieval,” *Multimédia Tools and Applications*, vol. 77, no. 23, pp. 30729–30748, 2018.

[38] 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.

[39] L. Zhu, J. Shen, L. Xie, and Z. Cheng, “Unsupervised visual hashing with semantic assistant for content-based image retrieval,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 2, pp. 472–486, 2017.

[40] A. Alzu’bi, A. Amira, and N. Ramzan, “Content-based image retrieval with compact deep convolutional features,” *Neurocomputing*, vol. 249, pp. 95–105, 2017.

[41] H.-F. Yang, K. Lin, and C.-S. Chen, “Supervised learning of semantics-preserving hash via deep convolutional neural networks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 2, pp. 437–451, 2018.

[42] C. Zhang, J. Cheng, and Q. Tian, “Multiview label sharing for visual representations and classifications,” *IEEE Transactions on Multimedia*, vol. 20, no. 4, pp. 903–913, 2018.

[43] C. Zhang, J. Cheng, and Q. Tian, “Multiview, few-labeled object categorization by predicting labels with view consistency,” *IEEE Transactions on Cybernetics*, vol. 49, no. 11, pp. 3834–3843, 2019.

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

[45] Vadivel, Majumdar, and S. Sural, "Performance comparison of distance metrics in content-based image retrieval applications," in *International Conference on Information Technology (CIT), Bhubaneswar, India*, 2003.

[46] "http://docs.adaptive-vision.com/current/studio/machine_vision_guide/TemplateMatching.html," Adaptive Vision, 2016. [Online].

[47] M. D. Chaudhary and P. V. Pithadia, "Multi-feature histogram intersection for efficient Content based image retrieval," in *International Conference on Circuit, Power and Computing Technologies*, 2014.

[48] CS231n: Convolutional Neural Networks for Visual Recognition. URL: http://cs231n.stanford.edu/. (Accessed: 23.12.2019).

[49] Lisa Torrey and Jude Shavlik. “Transfer Learning”. In: (2009).

[50] Pedro Marcelino. Transfer learning from pre-trained models. URL: https://towardsdatascience.com/transfer-learning-from-pre-trained-models-f2393f124751. (Accessed: 23.12.2019).

[51] M. D. Chaudhary and P. V. Pithadia, "Multi-feature histogram intersection for efficient Content based image retrieval," in *International Conference on Circuit, Power and Computing Technologies,* 2014.

[52] Alfredo Canziani, Adam Paszke, and Eugenio Culurciello. “An Analysis of Deep Neural Network Models for Practical Applications”. In: arXiv e-prints, arXiv:1605.07678 (May 2016), arXiv:1605.07678. ArXiv: 1605.07678 [cs.CV].

[53] Jing Peng, Bir Bhanu, and Shan Qing. “Probabilistic Feature Relevance Learning for Content-Based Image Retrieval”. In: Computer Vision and Image Understanding 75 (1999), pp. 150–164.

[54] S. Aksoy and R. M. Haralick. “Probabilistic vs. geometric similarity measures for image retrieval”. In: Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No.PR00662). Vol. 2. June 2000, 357–362 vol.2. DOI: 10.1109/CVPR.2000.854847.

APPENDIX A

RETRIEVED RESULTS:

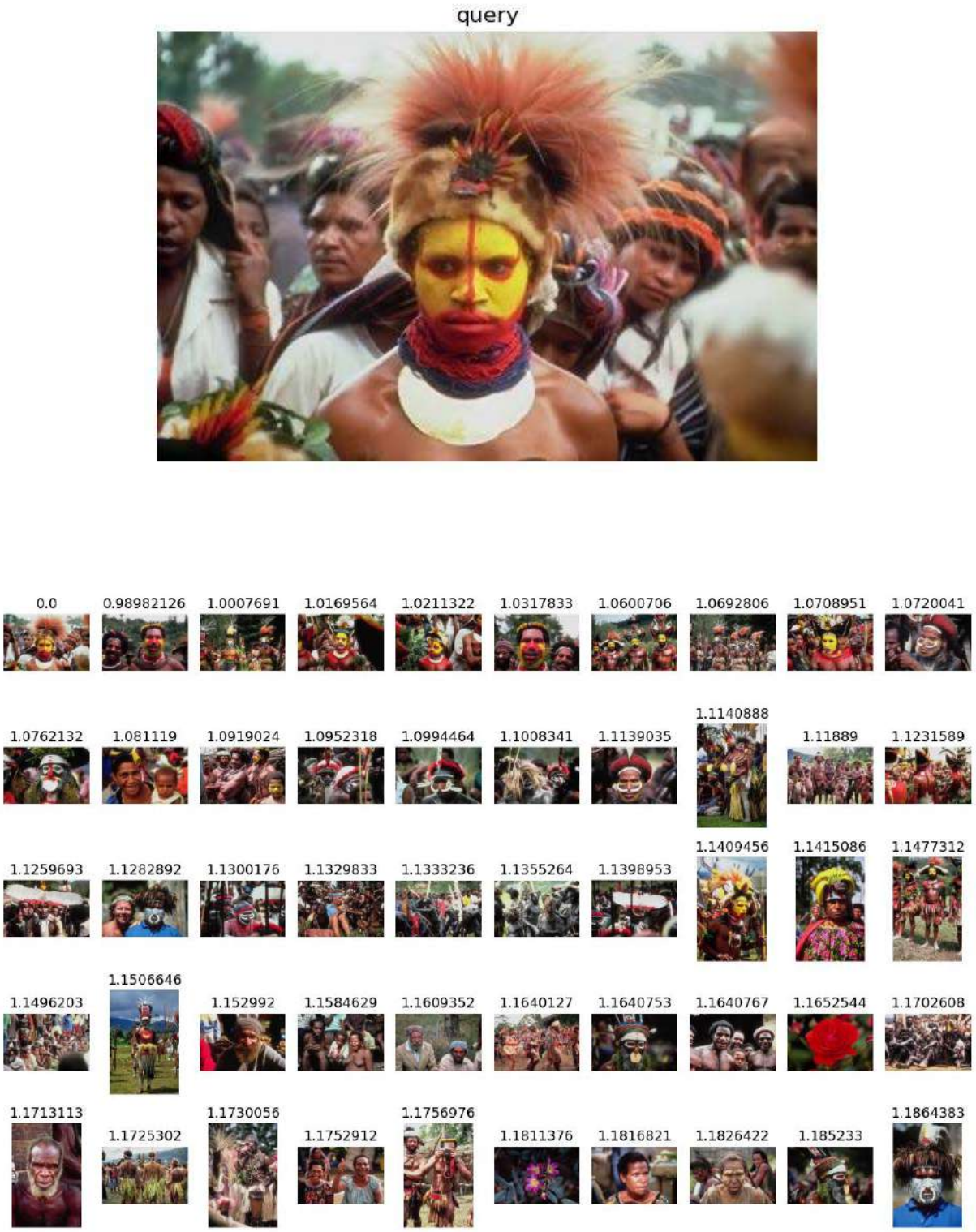


Figure A.1: African people

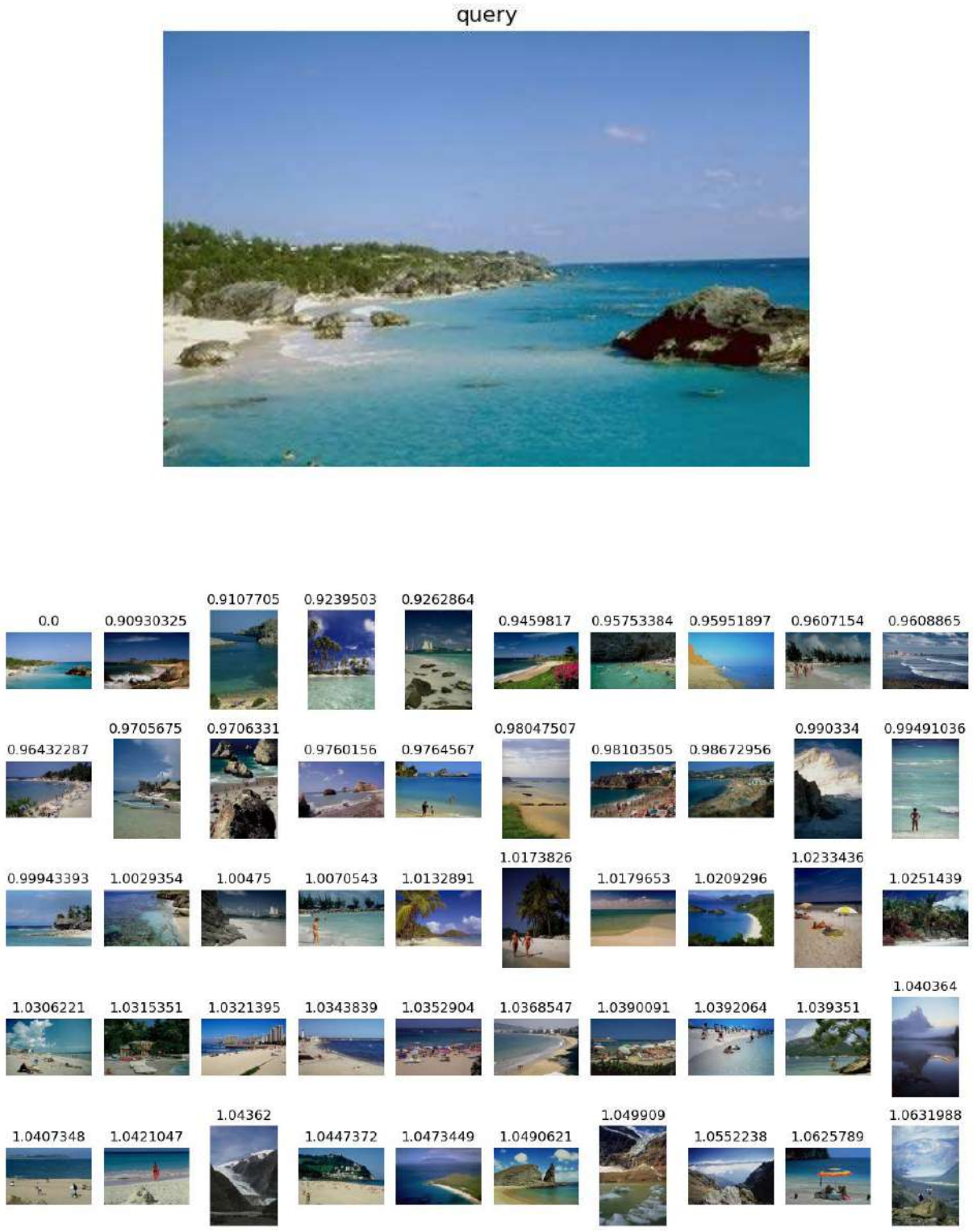


Figure A.2: Beach



Figure A.3 : Buses

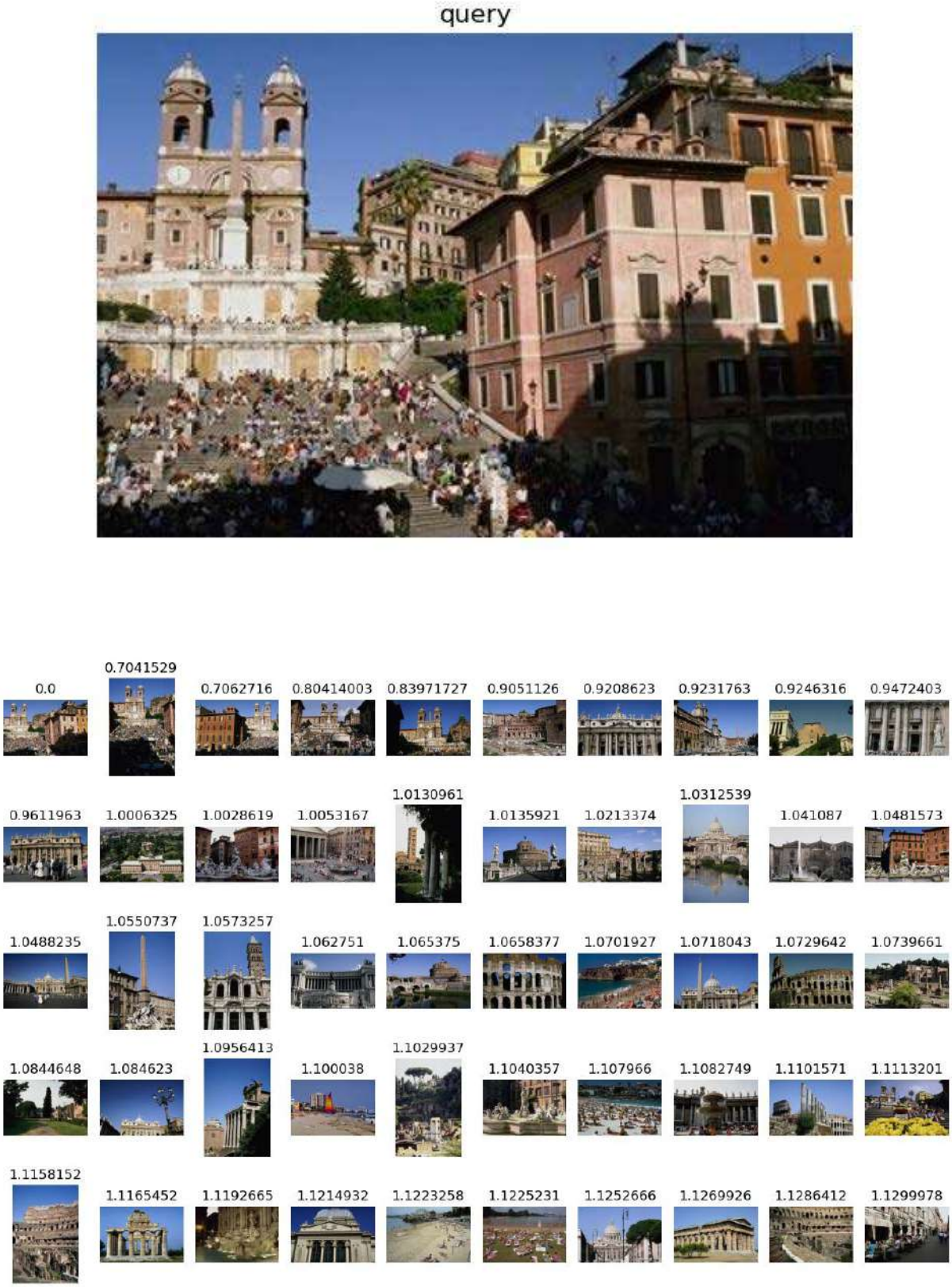


Figure A.4 : buildings



Figure A.5 : Dinosaurs

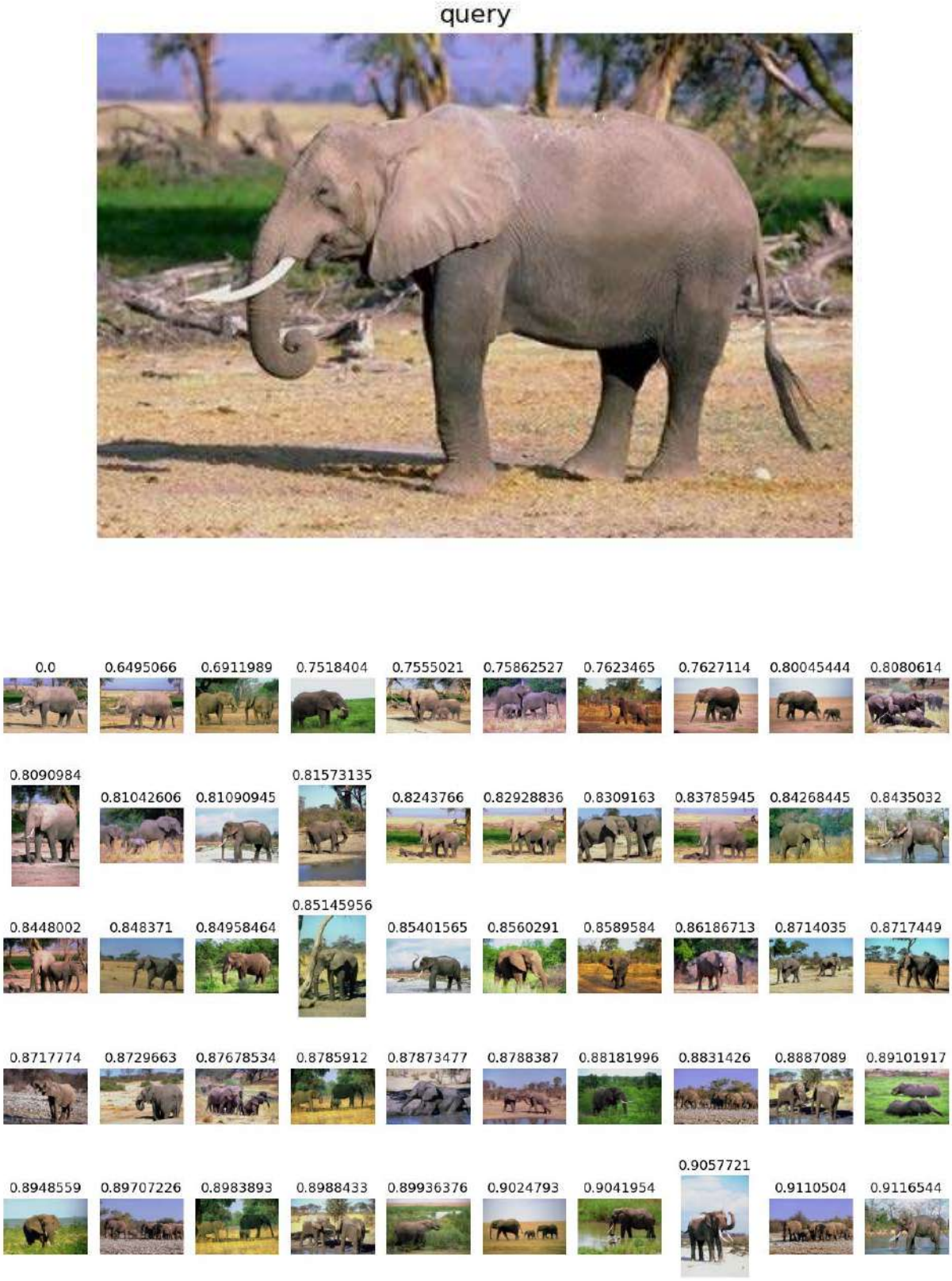


Figure A.6 : Elephants

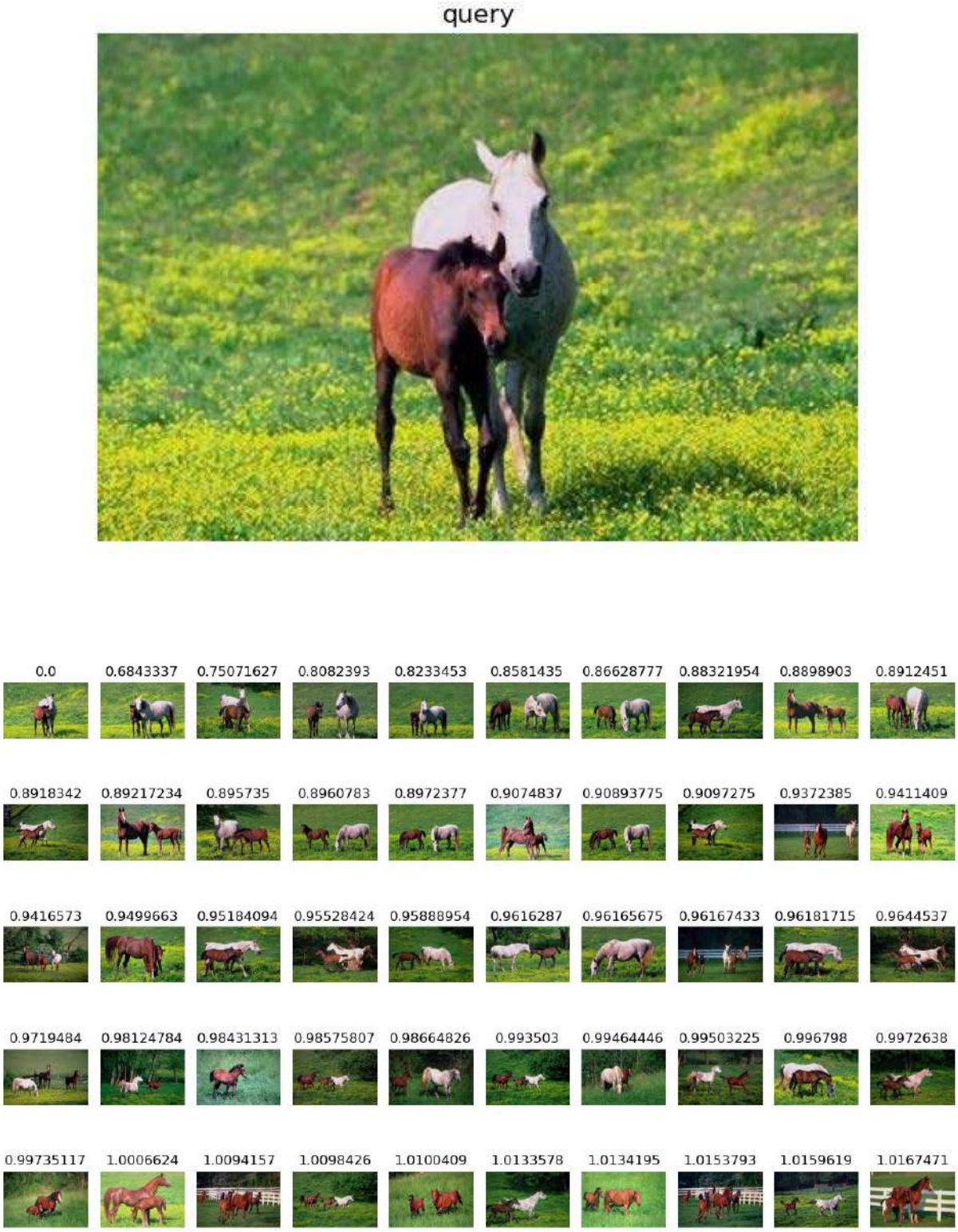


Figure A.7 : Horses



Figure A.8 : Flowers

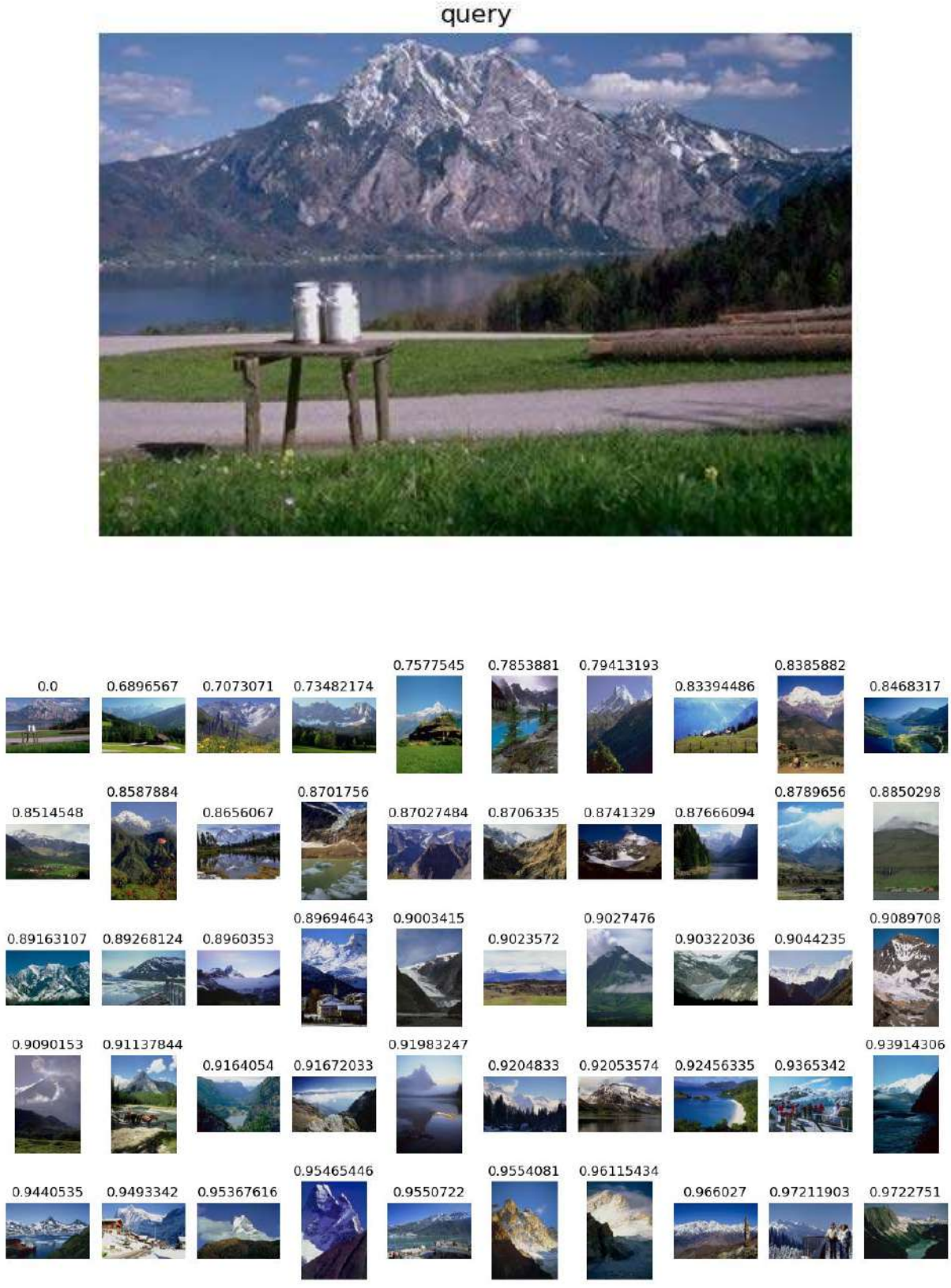
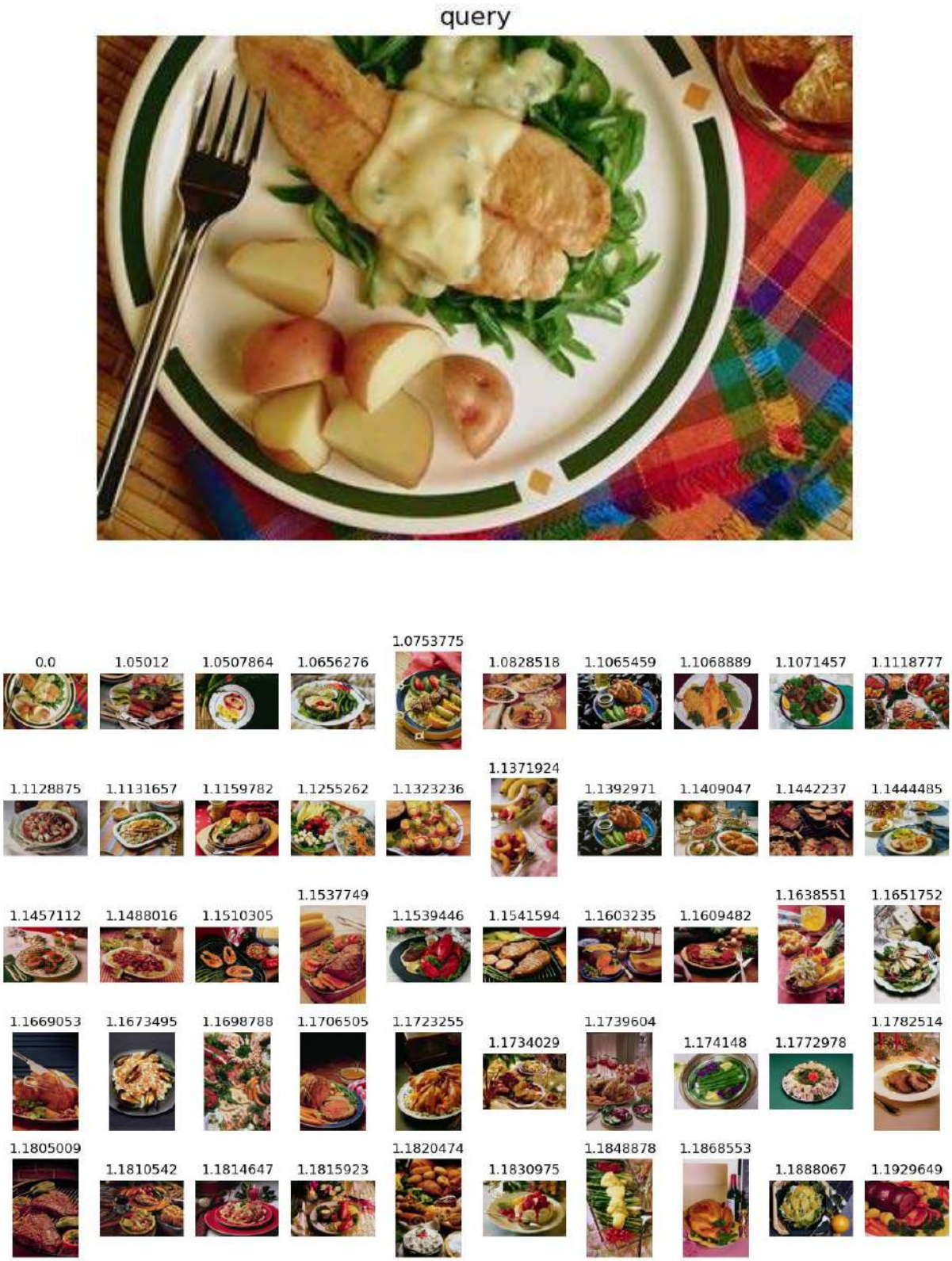


Figure A.9 : Mountains



FigureA.10 : Food