Object Classification Using Machine Learning

Submitted By

Ami Shah

14MCEC23

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INSTITUTE OF TECHNOLOGY

NIRMA UNIVERSITY

AHMEDABAD-382481

May 2016

Object Classification Using Machine Learning

Major Project

Submitted in fulfillment of the requirements

for the degree of

Master of Technology in Computer Science and Engineering

Submitted By

Ami Shah

(14MCEC23)

Guided By

Dr. Priyank Thakkar

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INSTITUTE OF TECHNOLOGY

NIRMA UNIVERSITY

AHMEDABAD-382481

May 2016

Certificate

This is to certify that the major project entitled **"Object Classification Using Machine Learning"** submitted by **Ami Shah (Roll No: 14MCEC23)**, towards the fulfillment of the requirements for the award of degree of Master of Technology in Computer Science and Engineering of Nirma University, Ahmedabad, is the record of work carried out by her under my supervision and guidance. In my opinion, the submitted work has reached a level required for being accepted for examination. The results embodied in this major project part-II, to the best of my knowledge, haven't been submitted to any other university or institution for award of any degree or diploma.

Dr. Priyank Thakkar
Guide & Associate Professor,
CSE Department,
Institute of Technology,
Nirma University, Ahmedabad.

Dr. Priyanka Sharma
Professor,
Coordinator M.Tech - CSE
Institute of Technology,
Nirma University, Ahmedabad

Dr. Sanjay Garg
Professor and Head,
CSE Department,
Institute of Technology,
Nirma University, Ahmedabad.

Dr. P. N. Tekwani
I/c Director,
Institute of Technology,
Nirma University, Ahmedabad

Statement of Originality

I, **Ami Shah**, Roll. No. **14MCEC23**, give undertaking that the Major Project entitled **"Object Classification Using Machine Learning"** submitted by me, towards the fulfillment of the requirements for the degree of Master of Technology in **Computer Science & Engineering** of Institute of Technology, Nirma University, Ahmedabad, contains no material that has been awarded for any degree or diploma in any university or school in any territory to the best of my knowledge. It is the original work carried out by me and I give assurance that no attempt of plagiarism has been made.It contains no material that is previously published or written, except where reference has been made. I understand that in the event of any similarity found subsequently with any published work or any dissertation work elsewhere; it will result in severe disciplinary action.

_____

Signature of Student

Date:

Place:

Endorsed by

Guide Name

(Signature of Guide)

Acknowledgements

It gives me immense pleasure in expressing thanks and profound gratitude to **Dr. Priyank Thakkar**, Associate Professor, Computer Science Department, Institute of Technology, Nirma University, Ahmedabad for his valuable guidance and continual encouragement throughout this work. The appreciation and continual support he has imparted has been a great motivation to me in reaching a higher goal. His guidance has triggered and nourished my intellectual maturity that I will benefit from, for a long time to come.

It gives me an immense pleasure to thank **Dr. Sanjay Garg**, Hon'ble Head of Computer Science and Engineering Department, Institute of Technology, Nirma University, Ahmedabad for his kind support and healthy research environment.

A special thank you is expressed wholeheartedly to **Dr. P. N. Tekwani**, Hon'ble Director, Institute of Technology, Nirma University, Ahmedabad for the unmentionable motivation he has extended throughout course of this work.

I would also thank the Institute, Institute Of Technology, Nirma University for providing a good platform and facilities and all faculty members of Computer Science and Engineering Department, Nirma University, Ahmedabad for their special attention and suggestions towards the project work.

- Ami Shah
14MCEC23

Abstract

Computer vision has wealth of research. It spans over image restoration, scene reconstruction, and motion estimation. The classical problem in computer vision includes determining what the kind of object is in an image; this branch of computer vision is called object classification.

In object classification, the system is given an image as input. The system should identify the label of the object to which the object belongs to. However, it is well known that even the best object classification algorithms will produce poor results when given poor features to track.

Here, in literature survey made, different feature extraction techniques, many clustering algorithms, classification techniques and different approaches/methodologies are studied. Focus is restricted to these methods: Bag-of-Words (BoW) model and Convolution Neural Network.

Here, Experiments that are performed on BoW model are implemented using Microsoft Visual Studio with OpenCV libraries. The BOW model extracts the SIFT and SURF features from all the training images. These features are clustered using the k-means to create the dictionary of visual words. Next, SVMs with linear and RBF kernel are used then for the classification purpose. Task is also addressed through a well known convolution neural network - ALEXNET. The thesis also proposes variants of ALEXNET. Results are compared with state-of-the-art and proves the effectiveness of the proposed models.

Abbreviations

BoW	Bag-Of-Words.
SIFT	Scale-invariant feature transform.
SURF	Speeded Up Robust Features.
SVM	Support Vector Machine.
RBF	Radial Basis Function.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

For humans, it is very easy to see, identify and recognize any real-world objects with great accuracy and little effort through vision. As humans have good generalisation capability, they recognize any particular object, even if exact object is not known before.

In our endeavour to provide sensory capability equivalent to Humans, the techniques under development are very complicated tasks for system to recognize, understand and classify the object from all different aspects as similar object to some predefined object class. Thus, the Object Classification finds the category of a object (such as faces, bicycle, aeroplane, buildings, etc.) from the given image.

It has wider application areas e.g. automated systems, image retrieval, surveillance, security etc. Object Classification has also more challenging task to perform in computer vision and robotics.

The multi-utility of Object Classification has given great impetus to research on visual concept classification (object classification).Many researchers have been burning midnight oil for past 3 decades and substantial work is done in this field. Yet there remain many avenues to be explored & established before reaching a full-fledged/comprehensive module.

1.1 Problem Description

The basic question "What is the object in given image?" means whether it corresponds to any object category that machine knows. Generally, machine has no prior knowledge of any objects unlike humans. Thus, it is necessary to establish the systems for a machine prior to the classification tasks.

Generally, the object classification systems comprises of following modules:

- image acquisition
- pre-processing
- feature extraction
- classification

While designing model, we need the solution for the following problems.

- Image Representation and feature extraction:-

  Images must be represented in the form (of features) which makes system able to understand. The image features should be invariant to various image deformations. Selecting suitable feature extraction method that exhibit such invariance plays a great role in classification performance.

- Classification:-

  Classification requires recognising an unseen object from the image features and assigning to it the correct class. These algorithms should have good generalisation capability over all specific instances of each object class and learn enough distinctive information to separate the objects from the background.

Recognizing object classes remains as one of the most challenging problems, due to the undefined nature of object classes similarity. Sometimes, objects in different classes

can have more similar characteristics compared to others in the same class. Several researchers have addressed this problem in many ways.

1.2 Motivation

The great Human Vision sense: The human vision system provides most important sensory input to realize the world. What it does is it reduces the large amount of data quickly to few relevant information. So, it will be a great help to follow the human vision capability and model accordingly into computer system. So that it can give greater success in providing desired solution.

To imitate human visionary sense: Present day computer system have become powerful, intelligent and user friendly in daily life of an individual. It has become so much essential part of one's life that one can't think of life without computer. The current level of technology provides greater computational capability and/or physical characteristics. There are improvement areas where computers may become interactive and work for humans under self-actualization i.e. it can have intelligent working.

Areas of development: The human vision need to be imitated into computer vision. So, that it can provide similar sense for humans service purpose. Work done from 1970s till to date has been successfully attempted to Optical Character Recognition (OCR), Pedestrian Detection, Face Detection, Vehicle Detection and Tracking, etc. It has simplified many work areas and greater interest is developed for high end intricate areas of development such as Object Recognition where in it is desired to reduce information from any image to abstract class of object, so that it can classify many such and similar variety of objects. With this ability, the system can identify different variety of single commodity/objects.

The basic methodology adopted in object classification makes transformation of an image into another simple representation such that it purely consists of features. This representation is invariant towards many changes in the visual perception such as brightness, contrast and position within the image while separating some aspects not related

to the object. However, the appearance of objects within each class varies. In this thesis, we have analysed the earlier approaches to object classification based on such features in detail against its pros and cons. At the same time, we have tried to explore for the improvement of some of the methods.

1.3 Problem Statement

An image is given to the system as an input. It should identify the class of the object within the image with good accuracy and efficiently.

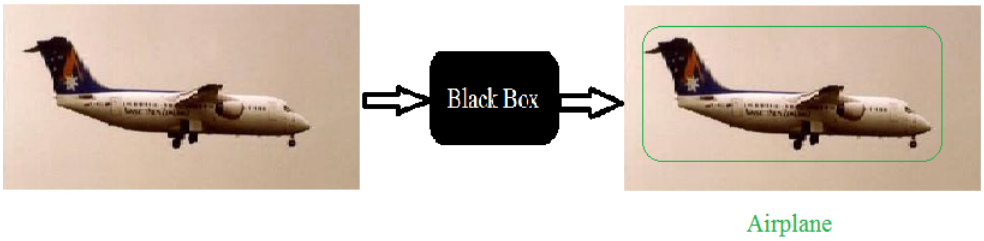

Figure 1.1: Abstract System Diagram

1.4 Outline

In chapter 2, we present analysis of the literature survey we made. It details the Challenges for the object classification. It presents brief of different classification techniques in the context of object classification. chapter 3 presents the direction of the work from analysis of the knowledge and recent trends. The chapter 4 explains tools used for the experiment and experiment results. Chapter 5 concludes the thesis.

Chapter 2

Literature Survey

2.1 List of Papers Surveyed

It was directed from dt. 13/07/2015 i.e. beginning of our 1st semester M.Tech. to carry out the survey from published papers of well-known conference and journals. In first 5 months journey of learning the earlier distinguished work, i have come across the list of papers and the brief of them is presented in the next section.

2.2 Analysis of the Survey in Brief

As it is seen in the literature survey carried, there have been many interest-point detectors and descriptors, focus on varying features, matching strategies, clustering methods and classification techniques proposed over last decade for object detection and classification. The excerpts of the study on work done is mentioned hereunder.

1. Simple Matching

   (a) D. Lowe[3] had proposed the Scale Invariant Feature Transform (SIFT) that is invariant towards various deformations. He used them to match and compare to object recognition. His study has shown good results on small datasets.

   (b) In the same period, Ke and Sukthankar[4] proposed Principle Component Analysis-SIFT (PCA-SIFT) which uses PCA to get normalized gradient patch

reducing the number of dimensions of the descriptor without compromising discrimination.

(c) In SIFT, the keypoints are determined by taking Difference of Gaussian(DoG) which takes more time to compute. In order to decrease the time, the SURF was proposed by Bay and Van[5] which uses fast-Hessian detector faster than DoG.

Many researchers worked on the simple matching and proposed the improvements in feature extraction[4][5][6] and used different matching parameters[7].

(d) Juan and Gwun[8] evaluated the performance of SIFT, PCA-SIFT and SURF to different context. They have used KNN to find the nearest neighbour match and RANSAC to reject matches which are irrelevant. The results obtained for different conditions using these three methods are shown in the table 1d.

Method	Time	Scale	Rotation	Blur	Illumination	Affine
SIFT	Common	Best	Best	Best	Common	Good
PCA-SIFT	Good	Common	Good	Common	Good	Good
SURF	Best	Good	Common	Good	Best	Good

(e) N. Y. Khan, et el. Proposed two new descriptor 64D SIFT and 96D SIFT descriptor and evaluated their performance against SIFT and SURF on benchmark dataset. They calculated the orientation histogram with different pattern from 4*4 square arrays to create 96D, 64D and 32D SIFT descriptor.

For classification, the nearest neighbour of the features of the query image are matched and recorded if they are within the threshold on the feature space. The image is classified to the label which has the maximum matches recorded.

They evaluated the performance of SURF and variations of SIFT against various conditions like scale, illumination, viewpoint, noise, blurring and rotation on the benchmarks- David Nister Dataset, Indoor Dataset, Hongwen Dataset and Caltech Dataset in 2.1:

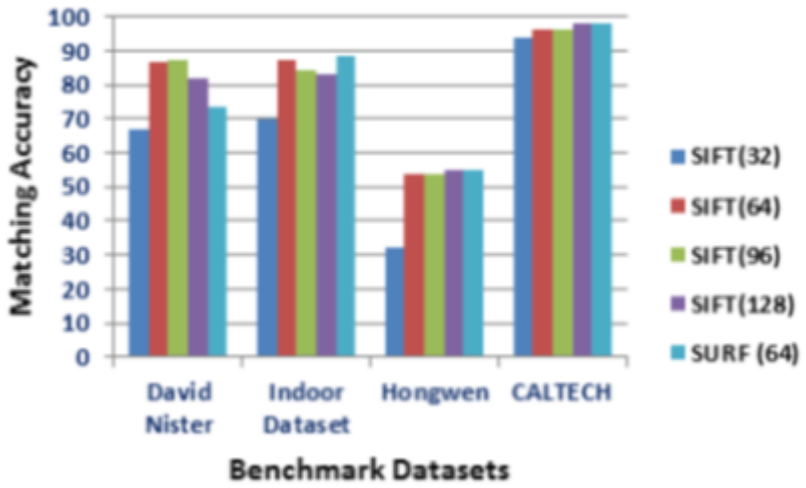

Figure 2.1: Comparison of various SIFT variants and SURF

For further advancement in object classification, the object descriptor should have to have the properties like robustness, good discriminative power and less computation intensive to find the object in real time with higher accuracy.

(f) The colour histogram is stable and easy to compute. It is invariant to occlusion, changes in scales, views and shapes [9].It has been used for texture recognition [10]. Local Binary Pattern is strongly invariant to rotation and easy to compute. To detect the object, Kwon lee, Chulhee lee, et al. [6] performed two steps process, coarse target object detection and precise target object detection. They used the colour histogram and LBP histogram which both are easy to compute.

The Sliding window histogram is very computation intensive. Then, to find the object, they used improved version of the sliding window technique which is based on histogram[11].

In the step of coarse target object detection, they only considered some windows which may be potentially contain the target object. These reduce the computation time and memory usage. The next step matches the LBP of target object with the windows generated in previous step.

The proposed algorithm[6] detects the object fast and effective compared to exhaustive search, Integral histogram and LBP matching.

Analysis:

- SIFT has overall good performance towards all kinds of deformation except for execution time as it is designed to be robust to localization error.
- PCA-SIFT having small dimension performs well and uses less storage space & time in further processing. PCA-SIFT has limitation on blur and scale performances.
- SURF is faster than SIFT and good in most situations except rotation.
- The simple matching includes two steps image representation and matching. The image can be represented by feature detection and descriptor that has good invariance towards all deformations ideally. But, no method can be invariant to all deformation. Each application requires invariance towards one or more deformation those are going to be seen more frequently. Ergo, the selection of the feature detection and extraction method depends on what kinds of invariance properties towards those deformations are required. The features should be computation effective for real time application. Hence, Its accuracy and performance also depend on the type of descriptor used.
- The Simple matching algorithm matches the object to already existing objects exactly. For example, if we have a white cup in particular shape in given image, it matches with the training image of the same cup. It cant match with other images of class cup of different colours and shapes. Ergo, it can be used to identify the object of known sizes and shapes.

2. Bag-Of-Words

   The bag-of-points approach roots from the approach 'Bag-of-Words'[12][13][14][15] widely used in texture classification. Some researchers used the small image patches as the features known which the quantization is performed to get the features, but these features are not robust[16]. Some others also performed quantization on invariant features obtained by well-known feature extraction techniques proven in literature[2].

   (a) **Csurka et al.**[2] proposed the Bag-of-keypoints approach which was extended from the approach used for text classification.

Here, the invariant features- SIFT and SURF of the training images of each class - are obtained. These features are clustered to construct the set of vocabularies i.e. the centres of each cluster. Then, the bag-of-keypoints description is obtained for the training images of each class.

These labelled descriptions are used to train the classifier- SVM and Nave Bayes.

To predict the class of given image, the bag-of-keypoints description is obtained. It is given to the classifier to predict the label[2].

The method performs well with SVM than simple Nave Bayes classifier on a seven category database. It produces good results even for background clutter[2].

With this approach, it is still difficult to identify which features to use that are robust and invariant to changes in scales, viewing angle etc. and gives better accuracy for object recognition.

Analysis:

- The Bag-Of-Words Model generalizes the variations within images of each class to capture the object content of each class[2].
- If there are more visual categories, the discriminative power of the appearance of image patches without ordering information will not sufficient. It needs to be extended to incorporate geometric information[2].

(b) B. Ganesharajah et al.[17] evaluated the performance of SIFT and e-SURF with this approach on 11 classes from PASCAL 2007. Here, the SURF descriptor of size 64 is extended to 128[17].The features computed with SIFT and e-SURF are clustered using K-means. The SVM classifier is trained and used to predict the class[17]. They have shown that e-SURF performs well than SURF and slightly better than SIFT[17].

Analysis:

- It must be able to find the objects and separate them from the background containing other objects in a given image[17].

(c) To enhance the performance of the approach, Jasper R. R. Uijlings et al. scrutinized and analysed the behaviour of descriptor extraction, visual word assignment and classification.

They improved the SIFT using fast recursive Gaussian derivative filter for diagonals and an exact derivative filter for horizontal and vertical directions and SURF using very fast approximation of the Gaussian derivative filter. These improved SIFT and SURF performs better for visual categorization.

For visual word assignment, Random Forest is used as the fast algorithm that divides the space into k dimensional vector[18]. The histogram-intersection kernel is used as the fastest classifier[18].

Analysis:

- The large amount of data and images requires computational efficiency and accuracy for concept classification[18].

(d) These features extracted by SIFT, SURF, etc. have good discrimination power for textured objects but they ignore colour information. But colour represents important information for object recognition.

Mohammad Khairul Islam, Farah Jahan, et al. use colour histogram and other features with this approach and Nave Bayes for classification[19].

They extract visual descriptor and colour histogram at each interest point from image and combine them aiming to use as single feature[19].

Results with SIFT with RGB histogram and SIFT with HSV histogram are 92% and 95.9% on dataset of images of 6 categories while SIFT and SURF gives 91% and 86.5%[19].

Analysis:

- SIFT is computationally expensive so not suitable for real-time applications. If it is combined with other global feature like RGB histogram and HSV histogram, it will be computationally cheap with good discrimination power[19].

(e) **Hakan Cevikalp, Zhal Kurt and Ahmet Okan Onarcan**[1] proposed a

Descriptor	Classification Rates for DoG Sampling
FT	83.27
SIFT	81.36
SURF	75.09
LBP	86.54
LTP	84.35
FT+LBP	91.93
FT+LTP	90.96
FT+SURF	86.15
FT+LBP+SURF	94.23
FT+LTP+SURF	93.01

Table 2.1: Classification rates for different descriptors on Caltech-4 dataset[1]

new descriptor based on weighted histograms of phase angles of Fourier transform for Bag-Of-Words Model. They compared the performance of Fourier Transform descriptor with other descriptors SIFT, SURF, Local Binary Pattern (LBP) and Local Ternary Pattern (LTP) along with the combination of FT with these descriptors on Caltech-4 and COIL-100 dataset[1].The BOW model uses the K-means clustering algorithm and nonlinear SVM to predict the class of the object[1].

Analysis:

- The results even explain that it includes further information to the other descriptor[1].

(f) **Benjamin W. Martin and Ranga R. Vatsavai** have analysed each step of the method in [20] to explore ways to enhance the performance. They analysed the effect of different feature extraction like Clustered features as in [20], SURF[5] and PCA-SIFT[4] besides the combination of Clustered features with each keypoint descriptor.

In clustering step, K-means and X-means were used to study the effects on the accuracy. First, guess-estimated value of k is used to measure the accuracy. Then, X-means is performed and the value of K estimated by it is used with K-means to again measure the accuracy. G-means takes more time to completion So, it was excluded from the comparison[21].

To improve the baseline performance of linear kernel, other kernel functions Polynomial, Radial basis and Sigmoid were tested to check the improvement of accuracy[21].

Analysis:

- As the clustered features are combined and duplicated with each keypoint descriptor, it uses more amount of memory and time[21].
- Other kernels with SVM give similar classification accuracies. Of the kernels tested, it can therefore be deduced that it is sufficient to use a linear kernel to classify the features generated[21].
- Addition of more complex image features and more complex SVM kernels is not necessary. Besides, Estimating parameters of the kernels in order to maximize the performance is a tedious task[21]. It only leads to increased execution time.
- X-means gives the optimal number of centroids for the given features. Hence, it reduces the length of feature vector by improving the quality of the clusters. It projects run-time performance boost[21].
- Use of clustering methods which automatically select the optimal number of clusters can be used to determine the optimal number of features for a set of training patches and therefore reduce classification time while generally preserving classification accuracy[21].

(g) Most of the work done is to increase discrimination power and efficiency of the feature descriptors. But, the classification method can be improved[22].

Leonardo Chang, Miriam M. Duarte, et al. proposed the Bayesian network for the classification. They have used SIFT features to represent objects because of their good invariance properties. The SIFT features extracted from all training images are clustered using agglomerative hierarchical clustering over the feature descriptors as it has been proven to perform better than K-means or EM-clustering. Next, each descriptor in each cluster is labelled with its corresponding class. For the Classification, SIFT features are extracted

from the input image. Each feature is then assigned to one of the clusters and the probability of each class given the cluster is obtained. The class of the object is the one whose sum of occurrence probabilities given each cluster is maximum than others[22].

(h) **R. Muralidharan and C. Chandrasekar** proposed a combination of the hessian-laplace detector along with PCA-SIFT descriptor as local feature and the hus moment invariant as global feature based object recognition. The classifier used to identify the object from the feature vector is KNN-SVM. KNN classifier is applied first to identify the closest object from the trained features, if there is no match; SVM is performed to identify the object[23]. Their results obtained by combining SVM and KNN with local and global feature can produce better results than the traditional methods like KNN, SVM and BPN[23].

Analysis:

- Image features are generally classified into two categories. They are local and global. Local features are computed based on the interest point in the image. Global features are computed based on intensity value of the entire image. Most of the work related to object recognition is based on either the local feature or global feature, only few work were considering the local and global features for object recognition[23].
- The global features and local features are robust in finding the object even the object is partially-occluded[23].

(i) **Lazebnik** proposed Spatial Pyramid Matching (SPM) [24] as BoW model does not contain spatial information.

(j) **Drew Schmitt and Nicholas McCoy**[25] represents hierarchical pyramid scheme to localize the object. This Bag-of-words model classifies the image when the object is forefront in the image. If the object in the scene is small, it is difficult to classify it correctly. It leads to development of hierarchical

pyramid scheme in which it segments the image to find the object[25].

Analysis:

- It is still difficult to apply our framework on a real-time application, because it requires much time to process an image[26].
- The power of these features lie on their invariance against various deformation.

3. Neural Network

   Neural network models are well suited to domains where large labeled datasets are available, since their capacity can easily be increased by adding more layers or more units in each layer.

   (a) **Hinton et al.**[27] proposed a new form of regularization called Dropout. For each training example, forward propagation involves randomly deleting half the activations in each layer. The error is then back propagated only through the remaining activations.

   (b) Li Wan and Matthew Zeiler et al. propose DropConnect which generalizes Dropout by randomly dropping the weights rather than the activations. Like Dropout, the technique is suitable for fully connected layers only.

4. Convolutional Neural Network

   Current trends have pointed out that features learning performs better designed features in some tasks, since they capture the global (via multi-layers network) or inter local structures (convolutional network) of images. We argue that combining the two types of features can significantly improve visual object recognition performance.

   (a) There has been much work done in designing various pooling techniques to leverage a introduced approximate model averaging technique called dropout[27].

(b) Ian J. Goodfellow et al.[28] proposed new method called maxout as its output is the max of a set of inputs. Optimization behaves very differently in the context of dropout than in the pure SGD case.

(c) Matthew D. Zeiler and Rob Fergus[29] represents the pooling process that occurs in each convolutional layer a stochastic process. Other methods of pooling such as average and max are deterministic. Both are not much suit drawbacksable for training deep convolutional networks. In average pooling, all elements in a pooling region are considered, even if many have low magnitude. While max pooling does not suffer from these drawbacks, it easily overfits the training setin practice, making it hard to generalize well to test examples[29]. The stochastic pooling hasthe advantages of max pooling but its stochastic nature helps prevent over-fitting. It has negligible computational overhead and no hyper-parameters to tune, thus can be swapped into to any existing convolutional network architecture.

The effect of different pooling methods on the classification accuracy is represented in fig. 2.2.

(d) Deeply Supervised Network proposed by **Chen-Yu Lee, Saining Xie, et al.** provides direct supervision on the hidden layers by providing companion layers to give transparency to the hidden layers and to provide robustness to the learned features.

(e) The Convolutional Kernel Network proposed by **Julien Mairal, Piotr Koniusz et al.**[33] is combination of Convolution Neural Network and the kernels. The three types of kernels used as initial maps are Patch Map (PM), Gaussian Map (GM) and their combination (CO). The features with these kernels are extracted in unsupervised manner. These are accompanied with labels while training the SVM classifier[33].

It has opened up new dimension by combining two: kernels and CNN in the research of the good architecture with greater accuracy. As the kernel maps are learned in unsupervised manner should be turned to supervised manner to

Authors	Year	Name	Method	Acc.
Jasper Snoek, Hugo Larochelle, and Ryan Prescott Adams	2012	Practical bayesian optimization of machine learning algorithms[30]	CONV. NET + SPEARMINT	85.02
Ian J. Goodfellow, David Warde-Farley, Mehdi Mirza, Aaron Courville, and Yoshua Bengio	2013	Maxout network[28]	CONV. NET + MAXOUT	88.31
Jost Toblas Springenberg and Martin Riedmiller	2014	Improving Deep Neural Networks with Probabilistic Maxout Units[31]	CONV. NET + PROBOUT	88.65
LiWan, Matthew D. Zeiler, Sixin Zhang, Yann LeCun, and Rob Fergus	2013	Regularization of neural networks using dropconnect[32]	12 x CONV. NET + DROPCONNECT	90.68
Ian J. Goodfellow, David Warde-Farley, Mehdi Mirza, Aaron Courville, and Yoshua Bengio	2013	Maxout networks[28]	CONV. NET + MAXOUT	90.62
Jost Toblas Springenberg and Martin Riedmiller	2014	Improving Deep Neural Networks with Probabilistic Maxout Units[31]	CONV. NET + PROBOUT	90.61
Matthew D Zeiler and Rob Fergus	2013	Stochastic pooling for regularization of deep convolutional neural networks[29]	Stochastic Pooling	84.87

Table 2.2: Results for Different Pooling Methods on CIFAR-10

better approximate the features[33].

(f) Network In Network[34] proposed by **Min Lin, Qiang Chen, Shuicheng Yan** consists of mlpconv layers as convolutional layers which uses multilayer perceptrons to convolve the input and a global average pooling layer as a replacement for the fully connected layers in conventional CNN. Mlpconv layers model the local patches better, and global average pooling acts as a structural regularizer that prevents overfitting globally.

(g) there has many architectures and the improvements over time proposed and their accuracy is compared in table 2.3 and 2.4

(h) Many work has been done to improve accuracy of the architecture but it takes longer time to train. So, **Dan Ciresan, Ueli Meier and Jurgen Schmidhuber** proposed Multi-column Deep Neural Networks[40] which uses the computation power of the GPU.

5. Boltzmann Machine

There are several variations of the Boltzmann Machine proposed and their comparison is made in tables 2.5 and 2.6

2.3 Challenges Known

Object classification is the task of assignment of an image one or multiple labels corresponding to the presence of instance of object. Many applications require objects to be found in predefined pose and orientation. The system is generally trained with fixed number of training examples. If real time object may appear differently than expected, the major challenges in object classification are as follows:

1. Scale changes
2. Viewpoint changes
3. Occlusion

Authors	Year	Name	CIFAR-10	CIFAR-100
A.Coates and A.Y. Ng	NIPS, 2011	Selecting Receptive Fields in deep Networks[35]	82.0	NA
K. Sohn and H. Lee	ICML, 2012	Learning invariant representations with local transformations[36]	82.2	NA
I. J. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio	ICML, 2013	Maxout networks[28]	88.32	61.43
A. Coates, A. Y. Ng, and H. Lee	2011	An analysis of single-layer networks in unsupervised feature learning[20]	79.6	NA
L. Bo, X. Ren, and D. Fox	2013	Unsupervised feature learning for RGB-D based object recognition[37]	NA	NA
R. Gens and P. Domingos	NIPS, 2012	Discriminative learning of sum-product networks[38]	83.96	NA
M. D. Zeiler and R. Fergus	ICLR, 2013	Stochastic pooling for regularization of deep convolutional neural networks[29]	84.87	57.49

Table 2.3: Accuracy measured for different CNN Architectures

Authors	**Year**	**Name**	**CIFAR-10**	**CIFAR-100**
JulienMairal, PiotrKoniusz, ZaidHarchaoui, and CordeliaSchmid	2014	Convolutional Kernel Networks: CKN-GM[33]	74.84	NA
JulienMairal, PiotrKoniusz, ZaidHarchaoui, and CordeliaSchmid	2014	Convolutional Kernel Networks: CKN-PM[33]	78.30	NA
JulienMairal, PiotrKoniusz, ZaidHarchaoui, and CordeliaSchmid	2014	Convolutional Kernel Networks: CKN-CO[33]	82.18	NA
Chen-Yu Lee, Saining Xie, Patrick W. Gallagher, Zhengyou Zhang, Zhuowen Tu	2014	Deeply-Supervised Nets[39]	90.31	65.43
Min Lin, Qiang Chen, Shuicheng Yan	2014	Network In Network[34]	89.59	64.32

Table 2.4: Accuracy measured for different CNN Architectures(Contd.)

Authors	Year	Name	Methods	Acc.
A. Krizhevsky.	2009	Learning multiple layers of features from Tiny Images[41]	Raw pixels	37.3
M. Ranzato, A. Krizhevsky, and G. E. Hinton.	2010	Factored 3-way Restricted Boltzmann Machines for Modeling Natural Images[42]	3-Way Factored RBM	65.3
M. Ranzato and G. E. Hinton	2010	Modeling Pixel Means and Covariances Using Factorized Third- Order Boltzmann Machines[43]	Mean	59.7
M. Ranzato and G. E. Hinton	2010	Modeling Pixel Means and Covariances Using Factorized Third- Order Boltzmann Machines[43]	cRBM	64.7
M. Ranzato and G. E. Hinton	2010	Modeling Pixel Means and Covariances Using Factorized Third- Order Boltzmann Machines[43]	Mean-covariance RBM	71.0
K. Yu and T. Zhang.	2010.	Improved local coordinate coding using local tangents[44]	Improved Local Coord. Coding	74.5

Table 2.5: Accuracy measured for different BM Architectures on CIFAR-10

Authors	Year	Name	Methods	Acc.
A. Krizhevsky	2010	Convolutional Deep Belief Networks on CIFAR-10[45]	Conv. Deep Belief Net	78.9
Adam Coates, Honglak Lee, Andrew Y. Ng	2011	An Analysis of Single-Layer Networks in Unsupervised Feature Learning[20]	Sparse auto-encoder	73.4
Adam Coates, Honglak Lee, Andrew Y. Ng	2011	An Analysis of Single-Layer Networks in Unsupervised Feature Learning[20]	Sparse RBM	72.4
Adam Coates, Honglak Lee, Andrew Y. Ng	2011	An Analysis of Single-Layer Networks in Unsupervised Feature Learning[20]	K-means (Hard)	68.6
Adam Coates, Honglak Lee, Andrew Y. Ng	2011	An Analysis of Single-Layer Networks in Unsupervised Feature Learning[20]	K-means (Triangle)	77.9
Adam Coates, Honglak Lee, Andrew Y. Ng	2011	An Analysis of Single-Layer Networks in Unsupervised Feature Learning[20]	K-means (Triangle, 4000 features)	79.6
Guillaume Desjardins, James Bergstra, Yoshua Bengio	2014	The Spike-and-Slab RBM and Extensions to Discrete and Sparse Data Distributions[46]	ssRBM	76.7

Table 2.6: Accuracy measured for different BM Architectures on CIFAR-10(Contd.)

4. Illumination changes
5. Complex backgrounds
6. Presence of noise
7. Intra-class colour variation
8. Intra-class shape variation
9. Flip changes
10. Blurring

Figure 2.2: Demonstration of challanges

An object often appears together with other objects or in a cluttered background or in a different illumination conditions or occluded. Object may appear in different sizes, shapes and colours. Objects may be viewed from different aspects. This additional data needs to be discarded, because, it helps to recognise the object class under consideration. Figure 2.2 illustrates an example of the difficulties of recognising object categories.

2.4 Existing Approaches

In survey mentioned above, There are three methods found. The brief of these methods are presented in the section.

2.4.1 Bag-of-Visual-Words

The Bag-of-Words approach is derived from document classification where represents each document as the histogram formed by the frequencies of their words. This word recurrence tally is utilized for retrieving or classifying documents.

Likewise, the Bag-of-Words strategy chooses small image regions from an image which then are mapped to visual words. The frequencies of visual words are then utilized as a part of subsequent classification. Fig. 2.3 demonstrates a schematic review of this procedure, with the formation of a vocabulary of visual words on the left side, and transformation of an image to a visual word recurrence histogram on the right[2].

First of all, many small image regions are selected from the images in training set. For each small image regions, a descriptor like SIFT, SURF, etc. is computed and mapped to feature space. These set of descriptors are then clustered for the creation of the visual vocabulary using some clustering algorithm like K means. The clustering algorithm generates a set of cluster centres constituting the visual vocabulary or dictionary or visual codebook by partitioning the descriptor space. Here, each cluster centre represents one visual code.

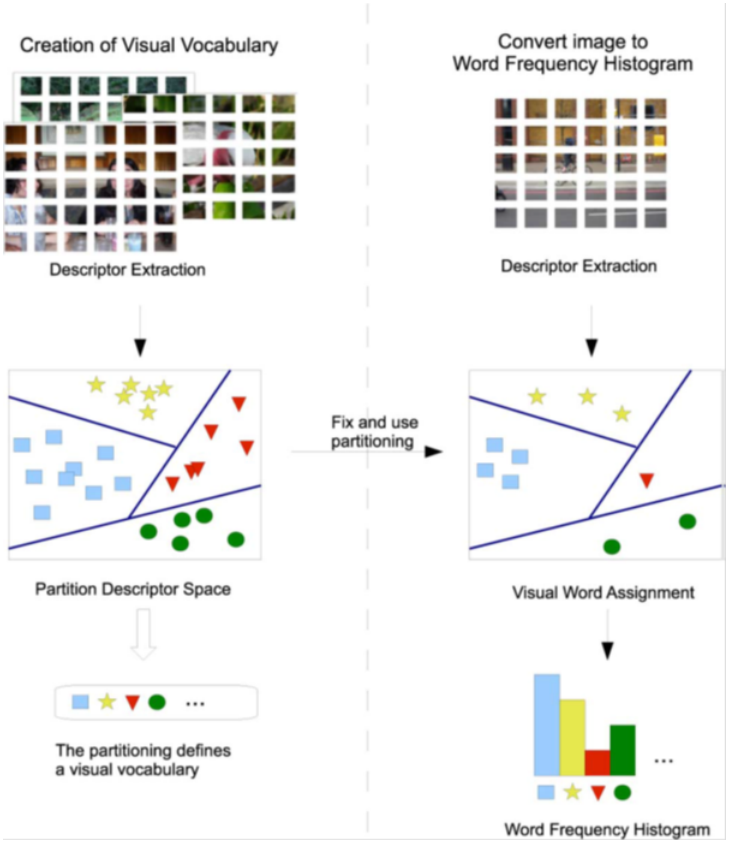

Figure 2.3: Diagram of the Bag-of-Words Model[2]

To create a bag-of-words or visual word frequency histogram, small patches are extracted from an image for which descriptors are calculated. Each descriptor is then

assigned to a visual word of the vocabulary or dictionary or cluster in the feature space by finding the nearest neighbour. Each occurrence is counted building visual word frequency histogram. Next, every histogram of the training images with its label are utilized to train the classifier.

When any new unseen image comes, small image regions are extracted. For each, descriptor is formed. Now, each of them are mapped to nearest neighbouring cluster forming visual word histogram. This is fed to classifier to determine its class label.

2.4.2 Neural Network

Neural Network has one input layer, one output layer and one or many hidden layer. The image to the network is presented at input layer and class label of that image is obtained from last output layer.

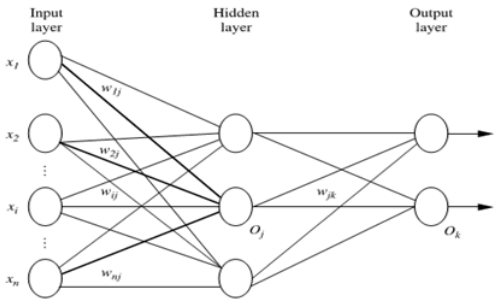

Figure 2.4: Neural Network

In this approach, the network is first trained with images with their labels using back-propagation or other technique to learn the features from them. Then, it can predict the class label of new image by this learned network, figure 2.4. Here, both stages, feature extraction and classification, are combined forming the object classifier.

2.4.3 Convolution Neural Network

Convolution Neural Network overcomes the disadvantages of neural network, over-fitting and vanishing gradient. First of all, as we keep on increasing the hidden layers, it starts mugging up the training dataset, called over-fitting. Secondly, the error signal back on

the way to the input layer, it starts vanishing. This makes the weights nearer to the output layer are trained more than distant ones. This problem is known as vanishing gradient.

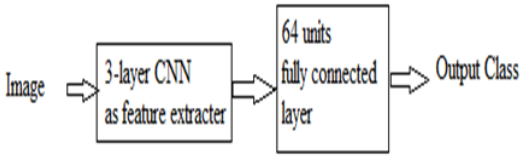

Figure 2.5: Block diagram of Convolution Neural Network

Convolution Neural network has many convolution and sub-sampling layers following small fully connected network. For object classification, the convolution and sub-sampling layer learns features from the training images. In training phase, all images with labels are used for learning features of objects of different class using supervised learning methods. After completion of training phase, an image is presented to it. Then, it learns features within and computes its label using fully connected layer.

2.4.4 Analysis of the Approaches

The bag-of-Words Model has good improvement over the straight forward matching of the interest points/features. This model works well for textured based discrimination. The performance of this model improvised using both global and local features till certain level. This model uses only designed features which have to be designed for every other application.

The Neural Network and Convolution Neural Network learn the features from the datasets by itself. These features can be called learned features. For any application, the training is required to learn the features for the same purpose.

2.5 Datasets

There are many datasets available for object classification. They have fixed number of classes of objects. These images can be divided into two sets, training set and testing

set. So, the model is constructed using Training set images and evaluated on testing set images. In our experiments, we have used caltech-101.

Name	**Size(in MB)**	**#Class**	**#Training**	**#Testing**	**Size Of Image**
CIFAR-10	175	10	5000/class	1000/class	32*32 RGB
CIFAR-100	175	100	500/class	100/class	32*32 RGB
STL-10	2.7 GB	10	500/class	800/class	96*96
Caltech-101	131	101	40-800/class		300*200

Table 2.7: Comparison of Datasets

Here, the comparison of datasets is given in Table 2.7.

Chapter 3

Proposed Approach

In this chapter, we have presented the original approach and the improvements in the approach. We have presented three architectures to improve the performance on datasets caltech-101 and caltech-256.

3.1 Original Architecture

Alex Krizhevsky, Ilya Sutskever and Geoffrey E. Hinton proposed the architecture[47] shown in fig. 3.1 in ImageNet ILSVRC-2010 contest to classify the 1.2 million high-resolution image into the 1000 different classes.

The Convolutional neural network consists of five convolutional layers and three fully-connected layers with a final 1000-way softmax. Layer 1, 2 and 5 are followed by max-pooling layers. Each layer is followed by RelU layer.

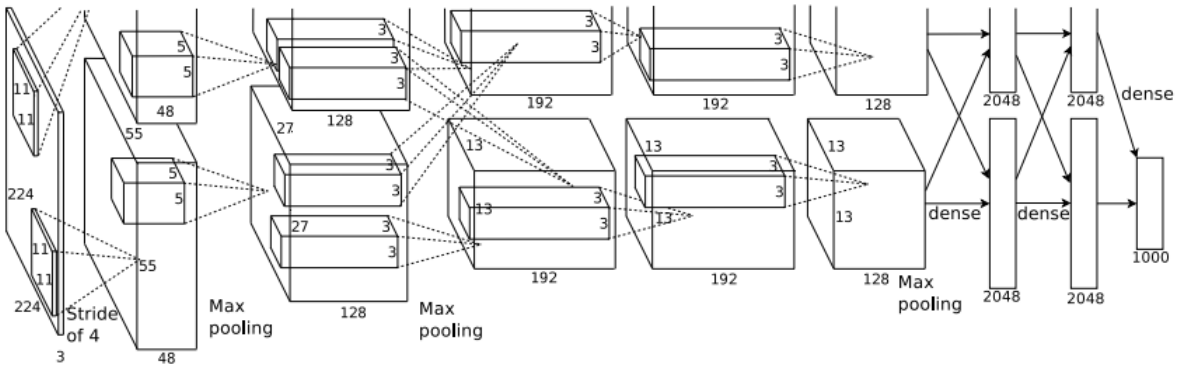

Figure 3.1: AlexNet Archtecture

The architecture gives 71.82% and 45.82% accuracy on caltech-101 and caltech-256

datasets.

3.2 Thought Process

Here, in the first layer, they have used filter of size 11x11 with stride 4. It may left out minute details from the image. if we reduce the filter size 11x11 and the stride size 4, the classification process may improve on the dataset caltech-101 and caltech-256.

3.3 Proposed Architectures

3.3.1 Architecture 1

The architecture includes one more layer between layer 3 and layer 4.

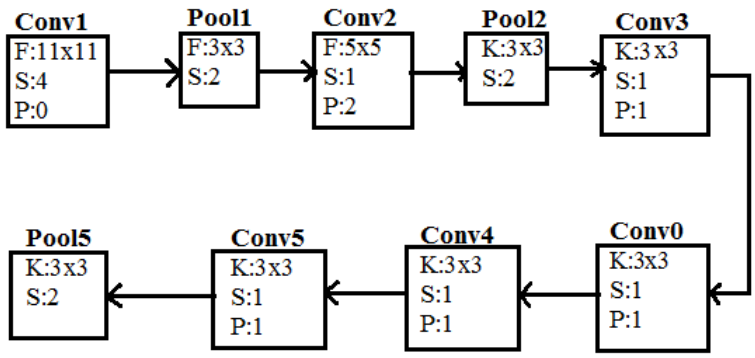

Figure 3.2: Architecture 1

3.3.2 Architecture 2

The first layer has reduced the filter size 11x11 to 9x9 and the stride size 4 to 2. In the second layer, the the size of the stride is increased to 2.

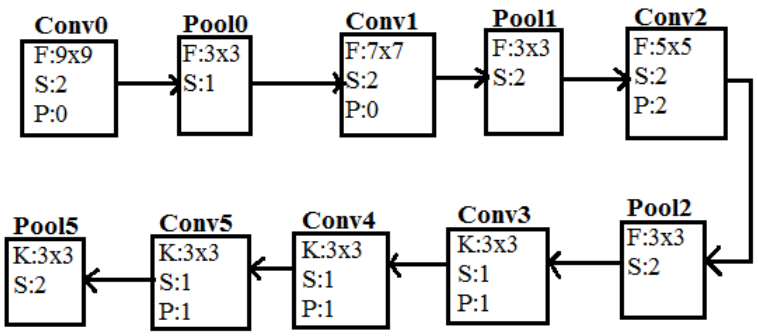

Figure 3.3: Architecture 2

3.3.3 Architecture 3

Here, the filter size of the first layer is reduced to 7x7.

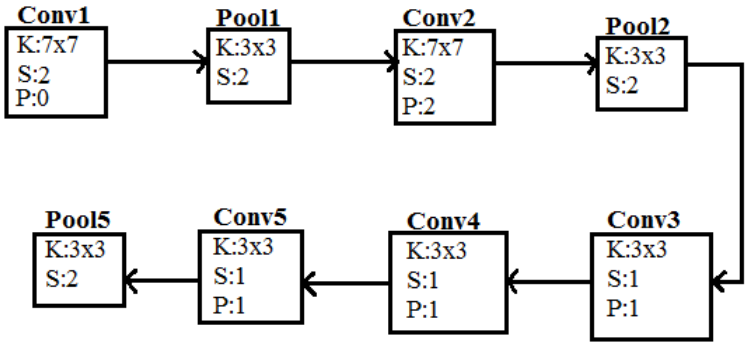

Figure 3.4: Architecture 3

Chapter 4

Experimental Setup and Results

In this chapter, The details of experiment setup is narrated. The comparison of our results with that of base results is presented.

4.1 Tools for Experimentation

Here, for our experiment, we use microsoft visual studio 2010 as IDE and opencv libraries for image processing and machine learning. OpenCV (open source Computer Vision) provides programming functions mainly aimed at computer vision. It was developed by intel research centre.The OpenCV libraries are configured on microsoft visual studio 2010.

4.2 Experimental Setup and Results

As per given in section 3.2, We have performed Experiment 5 on BOW model with SVM. The Experiment 1-4 are performed Convolutional Neural Network on caltech-101 and caltech-256. The experiment setup and results are presented in the section.

4.2.1 Experiment 1

Here, we have used the architecture proposed by Alex et al.[47]. The results we get for Caltech-4, Caltech-101 and caltech-256 are as shown in table 4.1:

AlexNet Architecture	
Dataset	Classification Rates
Caltech-4	99.03
Caltech-101	71.83
Caltech-256	45.83

Table 4.1: AlexNet results

Dataset	**Architecture 1**	**AlexNet**
Caltech-101	71.87	71.83
Caltech-256	46.16	45.83

Table 4.2: Comparison between Architecture 1 and AlexNet

4.2.2 Experiment 2

Here, we have used the architecture 1 in fig. 3.2 with same configuration in [47]. The results we get for Caltech-101 and caltech-256 as shown in table 4.2:

4.2.3 Experiment 3

Here, we have used the architecture 2 in fig. 3.3 with same configuration in [47]. The results we get for Caltech-101 and caltech-256 as shown in table 4.3:

4.2.4 Experiment 4

Here, we have used the architecture 3 in fig. 3.4 with same configuration in [47]. The results we get for Caltech-101 and caltech-256 as shown in fig. 4.4:

4.2.5 Experiment 5

Hakan Cevikalp et al.[1] have considered dataset of 1074, 526, 450, and 826 images of 4 categories of objects (i.e. airplanes, cars, faces, and motorbikes) respectively from caltech-101 dataset.

The SIFT and SURF features are extracted from the training set of images. The clustering was performed on those extracted features.The dictionary size was set to 1000.

Dataset	**Architecture 2**	**AlexNet**
Caltech-101	73.16	71.83
Caltech-256	46.88	45.83

Table 4.3: Comparison between Architecture 2 and AlexNet

Dataset	Architecture 3	AlexNet
Caltech-101	70.84	71.83
Caltech-256	43.58	45.83

Table 4.4: Comparison between Architecture 3 and AlexNet

SVM with Linear Kernel (Base)	
Descriptor	Classification Rates
SIFT	81.36
SURF	75.09

Table 4.5: Base results on Caltech-4 dataset

After the formation of dictionary, the SVM with linear kernel was trained using training images with labels. They performed 5-fold testing evaluate the performance of the feature extraction techniques, SIFT and SURF. Their experiment results are in table 4.5:

We have performed 5-fold testing for the BoW Model with same configuration. The comparison with base results are in the table 4.6.

4.2.6 Discussions

Here, it is learnt that if we reduce the size of the kernel, it gives better results. Increasing one layer like we added in architecture-1, it increases the accuracy. this change has analogy with cognition process in humans. If we combine both changes in one architecture it may produce enhanced & better results.

SVM with Linear Kernel (Base)	
Descriptor	Classification Rates
SIFT	81.36
SURF	75.09
SVM with Linear Kernel (Ours)	
Descriptor	Classification Rates
SIFT	89.29
SURF	93.70
SVM with Linear Kernel (Ours)	
Descriptor	Classification Rates
SIFT	87.74
SURF	93.64

Table 4.6: Comparison of base approach with Our results

Dataset	**Architecture 1**	**Architecture 2**	**Architecture 3**	**AlexNet**
Caltech-101	71.87	73.16	70.84	71.83
Caltech-256	46.16	46.88	43.58	45.83

Table 4.7: Comparison of AlexNet with proposed architectures

Chapter 5

Conclusion And Future Scope

5.1 Conclusion

From the Literature Survey,

Current trend is towards the designing the feature learners and classifier as an system which both uses machine learning instead of image processing techniques. The Bag-Of-Words model works well for small datasets as its discrimination power decreases with increase in amount of datasets. The Convolutional Neural Network performs better with good accuracy for larger dataset.

From simulation results,

If we increase number of layers in CNN as architecture 1, it shows improvement in classification rates by 0.04% and 0.33% on Caltech-101 and caltech-256 datasets respectively. To collect minute details, the filter size in the initial layer of CNN is decreased in architecture 2. It shows good improvements on caltech-101 and caltech-256 by 1.33% and 1.05%.

5.2 Future Scope

- The method applied has potential of getting enhanced and near to cent percent results by perfecting the method.
- It has can have wider application in professional use of object classification.

Bibliography

[1] H. Cevikalp, Z. Kurt, and A. Onarcan, "Return of the king: The fourier transform based descriptor for visual object classification," in *Signal Processing and Communications Applications Conference (SIU), 2013 21st*, pp. 1–4, IEEE, 2013.

[2] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Workshop on statistical learning in computer vision, ECCV*, vol. 1, pp. 1–2, Prague, 2004.

[3] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[4] Y. Ke and R. Sukthankar, "Pca-sift: A more distinctive representation for local image descriptors," in *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, vol. 2, pp. II–506, IEEE, 2004.

[5] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *Computer vision–ECCV 2006*, pp. 404–417, Springer, 2006.

[6] K. Lee, C. Lee, S.-A. Kim, and Y.-H. Kim, "Fast object detection based on color histograms and local binary patterns," in *TENCON 2012-2012 IEEE Region 10 Conference*, pp. 1–4, IEEE, 2012.

[7] N. Y. Khan, B. McCane, and G. Wyvill, "Sift and surf performance evaluation against various image deformations on benchmark dataset," in *Digital Image Computing Techniques and Applications (DICTA), 2011 International Conference on*, pp. 501–506, IEEE, 2011.

[8] L. Juan and O. Gwun, "A comparison of sift, pca-sift and surf," *International Journal of Image Processing (IJIP)*, vol. 3, no. 4, pp. 143–152, 2009.

[9] M. J. Swain and D. H. Ballard, "Color indexing," *International journal of computer vision*, vol. 7, no. 1, pp. 11–32, 1991.

[10] T. Leung and J. Malik, "Representing and recognizing the visual appearance of materials using three-dimensional textons," *International journal of computer vision*, vol. 43, no. 1, pp. 29–44, 2001.

[11] C. H. Lampert, M. B. Blaschko, and T. Hofmann, "Beyond sliding windows: Object localization by efficient subwindow search," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pp. 1–8, IEEE, 2008.

[12] T. Joachims, *Text categorization with support vector machines: Learning with many relevant features*. Springer, 1998.

[13] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," *The Journal of Machine Learning Research*, vol. 2, pp. 45–66, 2002.

[14] H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, and C. Watkins, "Text classification using string kernels," *The Journal of Machine Learning Research*, vol. 2, pp. 419–444, 2002.

[15] N. Cristianini, J. Shawe-Taylor, and H. Lodhi, "Latent semantic kernels," *Journal of Intelligent Information Systems*, vol. 18, no. 2-3, pp. 127–152, 2002.

[16] L. Zhu, A. B. Rao, and A. Zhang, "Theory of keyblock-based image retrieval," *ACM Transactions on Information Systems (TOIS)*, vol. 20, no. 2, pp. 224–257, 2002.

[17] B. Ganesharajah, S. Mahesan, and U. Pinidiyaarachchi, "Robust invariant descriptors for visual object recognition," in *Industrial and Information Systems (ICIIS), 2011 6th IEEE International Conference on*, pp. 158–163, IEEE, 2011.

[18] J. R. Uijlings, A. W. Smeulders, and R. J. Scha, "Real-time visual concept classification," *Multimedia, IEEE Transactions on*, vol. 12, no. 7, pp. 665–681, 2010.

[19] M. K. Islam, F. Jahan, J.-H. Min, and J.-h. Baek, "Object classification based on visual and extended features for video surveillance application," in *Control Conference (ASCC), 2011 8th Asian*, pp. 1398–1401, IEEE, 2011.

[20] A. Coates, A. Y. Ng, and H. Lee, "An analysis of single-layer networks in unsupervised feature learning," in *International conference on artificial intelligence and statistics*, pp. 215–223, 2011.

[21] B. W. Martin and R. R. Vatsavai, "Exploring improvements for simple image classification," in *Southeastcon, 2013 Proceedings of IEEE*, pp. 1–6, IEEE, 2013.

[22] L. Chang, M. M. Duarte, L. E. Sucar, and E. F. Morales, "A bayesian approach for object classification based on clusters of sift local features," *Expert Systems With Applications*, vol. 39, no. 2, pp. 1679–1686, 2012.

[23] R. Muralidharan and C. Chandrasekar, "Combining local and global feature for object recognition using svm-knn," in *Pattern Recognition, Informatics and Medical Engineering (PRIME), 2012 International Conference on*, pp. 1–7, IEEE, 2012.

[24] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 2, pp. 2169–2178, IEEE, 2006.

[25] D. Schmitt and N. McCoy, "Object classification and localization using surf descriptors," 2011.

[26] D. Doan, N.-T. Tran, D.-P. Vo, B. Le, *et al.*, "Learned and designed features for sparse coding in image classification," in *Computing and Communication Technologies, Research, Innovation, and Vision for the Future (RIVF), 2013 IEEE RIVF International Conference on*, pp. 237–241, IEEE, 2013.

[27] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[28] I. J. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio, "Maxout networks," *arXiv preprint arXiv:1302.4389*, 2013.

[29] M. D. Zeiler and R. Fergus, "Stochastic pooling for regularization of deep convolutional neural networks," *arXiv preprint arXiv:1301.3557*, 2013.

[30] J. Snoek, H. Larochelle, and R. P. Adams, "Practical bayesian optimization of machine learning algorithms," in *Advances in neural information processing systems*, pp. 2951–2959, 2012.

[31] J. T. Springenberg and M. Riedmiller, "Improving deep neural networks with probabilistic maxout units," *arXiv preprint arXiv:1312.6116*, 2013.

[32] L. Wan, M. Zeiler, S. Zhang, Y. L. Cun, and R. Fergus, "Regularization of neural networks using dropconnect," in *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pp. 1058–1066, 2013.

[33] J. Mairal, P. Koniusz, Z. Harchaoui, and C. Schmid, "Convolutional kernel networks," in *Advances in Neural Information Processing Systems*, pp. 2627–2635, 2014.

[34] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.

[35] A. Coates and A. Y. Ng, "Selecting receptive fields in deep networks," in *Advances in Neural Information Processing Systems*, pp. 2528–2536, 2011.

[36] K. Sohn and H. Lee, "Learning invariant representations with local transformations," *arXiv preprint arXiv:1206.6418*, 2012.

[37] L. Bo, X. Ren, and D. Fox, "Unsupervised feature learning for rgb-d based object recognition," in *Experimental Robotics*, pp. 387–402, Springer, 2013.

[38] R. Gens and P. Domingos, "Discriminative learning of sum-product networks," in *Advances in Neural Information Processing Systems*, pp. 3248–3256, 2012.

[39] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, "Deeply-supervised nets," *arXiv preprint arXiv:1409.5185*, 2014.

[40] D. Ciresan, U. Meier, and J. Schmidhuber, "Multi-column deep neural networks for image classification," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 3642–3649, IEEE, 2012.

[41] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," 2009.

[42] A. Krizhevsky, G. E. Hinton, *et al.*, "Factored 3-way restricted boltzmann machines for modeling natural images," in *International conference on artificial intelligence and statistics*, pp. 621–628, 2010.

[43] M. Ranzato and G. E. Hinton, "Modeling pixel means and covariances using factorized third-order boltzmann machines," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pp. 2551–2558, IEEE, 2010.

[44] K. Yu and T. Zhang, "Improved local coordinate coding using local tangents," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pp. 1215–1222, 2010.

[45] A. Krizhevsky and G. Hinton, "Convolutional deep belief networks on cifar-10," *Unpublished manuscript*, vol. 40, 2010.

[46] A. Courville, G. Desjardins, J. Bergstra, and Y. Bengio, "The spike-and-slab rbm and extensions to discrete and sparse data distributions," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 36, no. 9, pp. 1874–1887, 2014.

[47] 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.