An Application of pre-Trained CNN for Image Classification

Abdullah
The Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh

Mohammad S. Hasan
School of Computing and Digital Technologies
Staffordshire University
Stoke-on-Trent, UK

***Abstract*—Image Classification is a branch of computer vision where images are classified into categories. This is a very important topic in today's context as large databases of images are becoming very common. Images can be classified as supervised or unsupervised techniques. This paper investigates supervised classification and evaluates performances of two classifiers as well as two feature extraction techniques. The classifiers used are Linear Support Vector Machine (SVM) and Quadratic SVM. The classifiers are trained and tested with features extracted using Bag of Words and pre-trained Convolution Neural Network (CNN), namely AlexNet. It has been observed that the classifiers are able to classify images with very high accuracy when trained with features from CNN. The image categories consisted of Binocular, Motorbikes, Watches, Airplanes, and Faces, which are taken from Caltech 265 image archive.**

Keywords—image classification, Support Vector Machine (SVM), Bag of Words (BoW), Linear SVM, Quadratic SVM, Convolution Neural Network (CNN), AlexNet, feature extraction etc.

I. Introduction

In today's world of digital technology, mass communications take place in the form of digital pictures. From social media to typical web pages, there is a large assortment of images that need to be classified by an intelligent and efficient algorithm or techniques. Image classification techniques are already being employed in quality control, Optical Character Recognition (OCR) systems, remote sensing etc. Human eyes can recognise a familiar person from thousands of unfamiliar people, despite changes in their appearance, pose or viewpoint. Human eyes and brain classify an image using the elements of visual interpretation, computers can use machine learning approaches to classify an image. Image classification, whether biological or artificial, is the ability to perceive an image by extracting features such as shape, colour, texture etc.

Support Vector Machine (SVM) is considered a good classifier because it is very efficient and unique [1]. The research work has incorporated supervised image classification method with training data obtained by feature extraction techniques. SVM classifiers are used to classify the images.

The bag-of-words is a feature selection technique or model applied in various areas for instance: image classification, document classification etc. This model (BoW model) is also extensively used in NLP (natural language processing) and information retrieval where a text or a sentence is presented as a multiset of its words hence the name bag of words.

CNN has proven to perform outstandingly in many computer vision and machine learning problems. Out of the many applications, CNN has, one of them is image classification. Images are taken as objects that occupy most of the image and classification is done based on identifying which category this object falls under.

The rest of the paper is organised as follows. Section II discusses the previous work done in this field. The experimental setup and images used are mentioned in section III. The results and analysis are presented in section IV and finally, the paper concludes with section V.

II. Previous Work

Image classification techniques fall mainly into two broad categories: unsupervised and supervised image classification techniques. In an unsupervised learning model [2], the system is first presented with a large amount of 'unlabelled' images. The system then builds a probabilistic model from the unlabelled images by finding patterns in them. On the other hand, in a supervised learning model [2], the system is first trained with multiple examples of images. A learning model is generated using the training data. The model is then used to predict the features of an unknown image. Figure 1 shows the different categories of classifiers and few examples of each category.

Extracting features from an image is crucial in order to classify an image. In [3], the researchers employed an embedded approach for feature selection. The work is divided into two parts. In the first part of the work, the approach of Bradley and Mangasarian (1998) that minimises the training errors of a linear classifier is considered to construct a linear classifier which implicitly discards features. In the second part of the work, construction of a direct objective minimising feature selection method for non-linear SVM classifier is carried out. The results showed that the non-linear method is indispensable for feature selection.

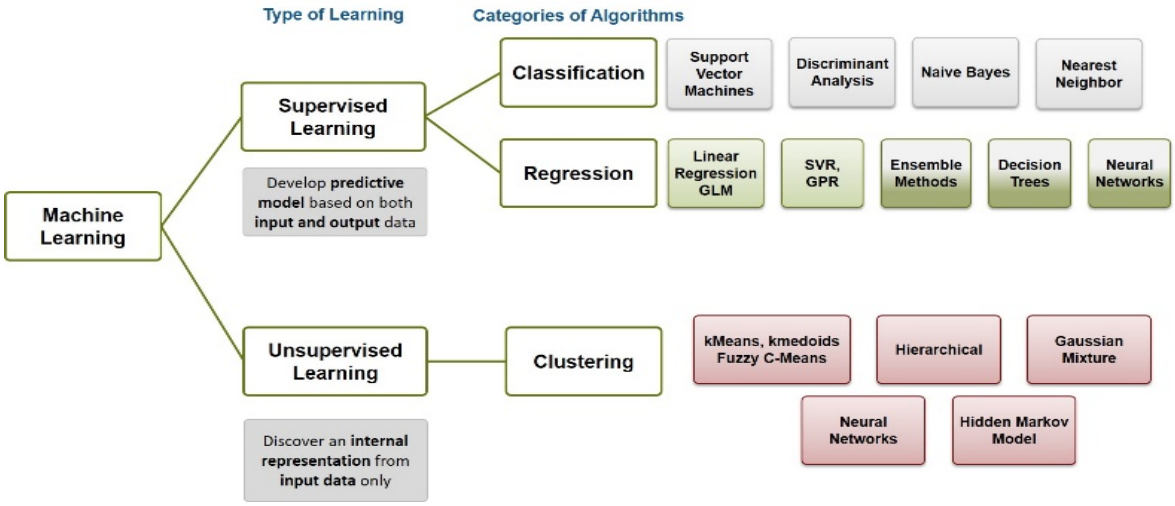

Figure 1: Supervised and unsupervised image classification [2].

In [4], the authors proposed a model which shows a quicker technique on a densely sampled grid for gaining the descriptors SIFT and SURF. A further contribution has been made to the state-of-the-art BoW pipeline which has the best performance score in the 2008 benchmarks of TRECvid and PASCAL. A model has been proposed with different techniques for descriptor extraction, projection, and classification. For example, for descriptor extraction, they have a proposed a quick and fast algorithm for sampling SIFT and SURF by comparing different variants of them. A k-means visual vocabulary with a Random Forest is compared for descriptor projection after experimenting with PCA for reducing projection time. And lastly, Support Vector Machines are used for classification part.

Another proposed model, which focuses on sentiment analysis named Delta TFIDF [5]. The mentioned approach incorporates BoW technique as feature set. The Delta TFIDF model is simple in computation and a great way to weight word scores accurately. To show improved accuracy level in sentiment analysis they have used support Vector Machines because Support vector machines along with Bag of Words feature sets ensures greater baseline accuracy. BoW technique in this way: in a bag of words (feature set) each word pair is correlated with a distinct value representing the word count in the document or sentence. By measuring the difference of TFIDF scores of the words in both the positive and negative corpora, they have assigned features values to the document. The experimental results obtained by the Delta TFIDF model are being compared with the standard bag of unigram and bigram words.

BoW is a great model and a very accurate technique in the classification task for the representation of the image. Key points are salient image patches containing details and information of an image. Moreover, an image can also be outlined as a bag of visual words or as a vector consisting the count of the visual word in that specified image based on the extraction of these key points and this theory is extensively used in scene classification. In [6], the authors have proposed a model for generating a better classification performance by providing a basis for a visual-word representation. Techniques involved applying different ways for producing numerous visual-word representations for example techniques used in the categorization of text with stop word removal, feature selection etc. The effect on the performance of the TRECVID and PASCAL collections is also investigated accordingly. As the visual-word representation is parallel to the bag-of-words representation in terms of a text document or sentence semantics, the authors have used vector quantization (VQ) technique for clustering the key point descriptors in the feature space for producing a visual-word-vocabulary outlining patterns in the images.

In [7], a method for automatic target recognition at ground level is presented by combining Convolutional Neural Network (CNN) and SVM. The results obtained by the combination of CNN and SVM have been found to be better than the results obtained from CNN alone.

To distinguish individuals, fingerprint detection techniques are most reliable. In [8], the authors used four CNN models for fingerprint liveness detection. The proposed system is evaluated on a data set that comprised of 50,000 real and fake fingerprint images. One CNN model is found to be better than the rest compared in terms of correctly classified samples. In [9], CNNs are employed for medical image classification.

In [10], generic object recognition is carried out in six categories. The categories include human figures, four-legged animals, aero planes, trucks, cars, and “none of the above”. The research shows that even though CNNs are competent at learning invariant features, but they do not always produce optimal results for classification. Also, SVMs are competent at producing decision surfaces from good feature vectors, but they cannot learn complicated invariances.

III. Proposed Model

To classify images, 200 images of 5 categories, namely Binoculars, Planes, Faces of people, Watches and Motorbikes have been taken from an image archive known The Caltech 256 [11] which contains 30607 images of 256 objects. Figure 2 shows some of the photos used in the experiment.

Figure 2: Sample Images Used in the experiment from [11].

A. *Feature Extraction*

The first step is to extract features from all the images. This is done by creating an image data store object which holds all the images along with their labels.

1) *Bag of Words (BoW)*

One of the two feature extraction technique used is BoW to extract features from each set of images. BoW can be defined as the "histogram representation based on independent features" [12]. BoW represents an image in three steps: feature detection, feature description, and codebook generation [13].

After detecting features BoW needs to represent these features into a vector, which is called a feature descriptor. BoW uses Speeded-Up Robust Features (SURF) features detection algorithm to detect SURF features and to create the feature descriptor. SURF works in 3 steps: interest point detection, local neighbourhood description, and matching. SURF uses a blob detector based on the maximal determinant of the Hessian matrix around a point, to find points of interest. The local neighbourhood description creates an exclusive description of the image based on the characteristics, e.g. intensity of the pixel around the points of interests [14].

The features extracted are clustered using K-means clustering techniques. K-means clusters each observation into a cluster whose mean is nearest to the observation. This is done by alternating between two steps [15]. Assignment step where each observation is assigned to the cluster whose mean produces the least within-cluster sum of squares (WCSS) and update step that calculates the new mean.

Finally, these features are feed into Linear and Quadratic SVM classifiers. Figure 3 shows how BoW is created.

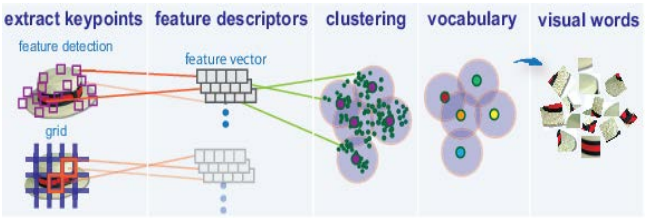

Figure 3: Algorithm workflow of Bag of Words [16].

2) *CNN Feature Extraction*

The other feature extraction technique used is CNN features extraction. CNN has two main characteristics of CNN - it uses weight sharing and local receptive field. Ideally, CNN has three main types of layers convolution layer, pooling layer, and fully-connected layer. The convolution layers main purpose is to extract features and the pooling layer is responsible for feature mapping [17]. The fully-connected layer is learning a function from the features extracted from the convolution layer.

A pre-trained CNN called AlexNet Model is used in this case. The model is trained by millions of images and can categories images into 1000 categories. But this paper uses AlexNet to extract features from the images. The last three layers of AlexNet are fully-connected layers with the last layer producing a distribution of 1000 class labels. The first convolution layer takes an input image of size 227x227. Therefore, all images are resized to the required dimensions. Each of the fully-connected layers 4096 neurons [18]. The overfitting in AlexNet is reduced in two ways: Data Augmentation and Dropout. In data augmentation, datasets are artificially enlarged using label-preserving transformations which generate transformed images with very little computation [18]. In dropout [19], the hidden neurons are set to zero with a probability of 0.5. This prevents forward pass and back-propagation and allows neurons to learn robust features as they no longer rely on other neurons [18].

Alexnet has many layers as shown in Figure 4 but not all the layers are suitable for extracting features. Like the first layer extracts features like blobs and edges. Hence, using a deeper layer will give better distinct features. Therefore, the layer before classification layer is used, ‘fc7’ to extract features [20]. These features are then used to train and test Linear and Quadratic SVM classifiers.

B. *Classifiers*

Classifiers are statistical models that can categorise a new observation to a correct category. In machine learning classification is considered to be supervised learning, where the model learns about the correct category before categorising an unknown object or observation.

The image sets used in this experiment are divided into training set and test set. Among 200 images, features from 150 images are used to train the classifiers and 50 images are used to test the classifiers. Along with the above setup, 10-time-10-fold cross-validation is performed with the 200 images. Figure 5 shows how classifiers predict images.

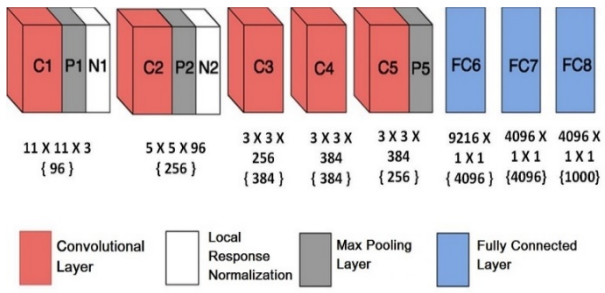

Figure 4: AlexNet Layers [21].

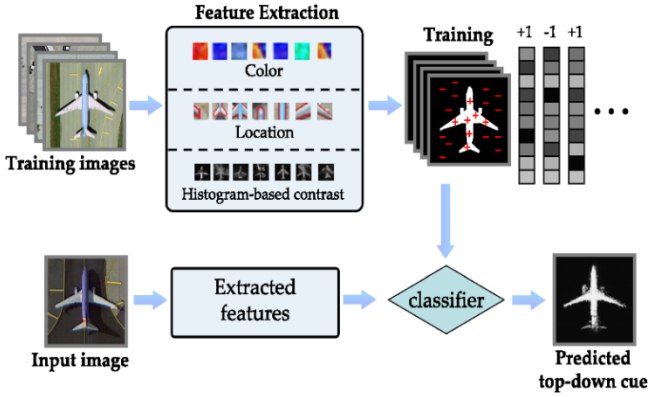

Figure 5: Overview of how classifier is used [22].

1) Linear SVM

SVM uses discriminant Hyperplane which maximises margin i.e. the distance from the nearest training point [24], as demonstrated in Figure 6.

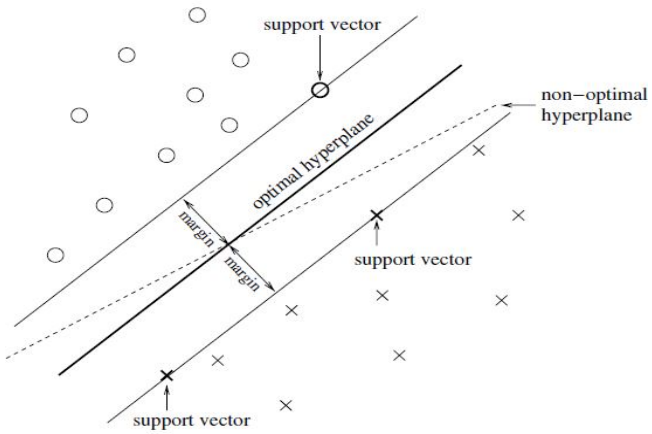

Figure 6: Hyperplane in Linear SVM [23].

2) Quadratic SVM

However, by applying Kernel Trick [25] that uses nonlinear kernel function, which transforms the hyperplane into a feature space, as shown in Figure 7.

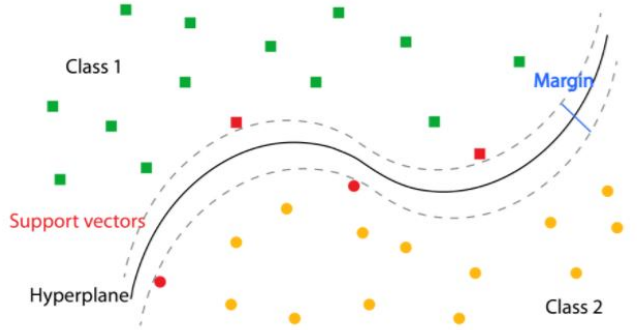

Figure 7: Hyperplane in Quadratic SVM [26].

IV. Results and Analysis

Two sets of results are obtained in four categories as illustrated in Table 1. The first set of results is obtained by keeping fixed set training and fixed testing images. The second set of results comes from 10-time-10-fold cross-validation method.

TABLE 1: RESULT CATEGORIES

Feature Extraction Technique	Classifiers
BoW	Linear SVM
BoW	Quadratic SVM
AlexNet	Linear SVM
AlexNet	Quadratic SVM

A. Fixed training and testing images

Figure 8 illustrates the accuracy of Linear SVM and Quadratic SVM. It is observed that the features obtained from AlexNet have improved the accuracy of both the classifiers. Not only the accuracy of the test images has increased but also the model accuracy has increased significantly. The cause for such an increase can be that AlexNet extracts a variety of features from blobs, edges to more distinct features which yield a better result.

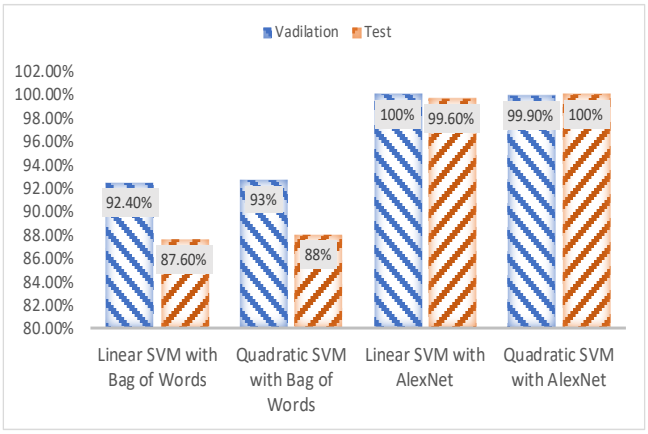

Figure 8: Validation and Test Accuracy for the fixed set of training and test images.

Next, the confusion matrix for individual classifiers and feature extract techniques are shown below.

1) Linear SVM with BoW

As shown in Figure 9 Motorbikes and Airplanes are classified with an accuracy of above 98%. Binocular is classified with an accuracy of 92% and faces having an accuracy of 84%. The least accurate classification done by the classifier is Watches, 64%.

Binocular	46	1	1	1	1
Motorbikes	0	49	1	0	0
Watches	10	3	32	3	2
Airplanes	0	0	0	50	0
Faces	2	0	6	0	42
	Binocular	Motorbikes	Watches	Airplanes	Faces

Figure 9: Confusion Matrix for Linear SVM with BoW.

2) Quadratic SVM with BoW

Figure 10 displays similar accuracy as Linear SVM with BoW for Binocular, Motorbikes, and Airplanes. The accuracy for watches has increased to 68% and a decrease of 82% for faces.

Binocular	46	1	1	1	1
Motorbikes	0	49	1	0	0
Watches	6	3	34	5	2
Airplanes	0	0	0	50	0
Faces	2	0	7	0	41
	Binocular	Motorbikes	Watches	Airplanes	Faces

Figure 10: Confusion Matrix for Quadratic SVM with Bag of Words.

3) Linear SVM with AlexNet

As illustrated in Figure 11 all categories are classified with 100% accuracy, except for watches, which has an accuracy of 98%.

Binocular	50	0	0	0	0
Motorbikes	0	50	0	0	0
Watches	1	0	49	0	0
Airplanes	0	0	0	50	0
Faces	0	0	0	0	50
	Binocular	Motorbikes	Watches	Airplanes	Faces

Figure 11: Confusion Matrix for Linear SVM with AlexNet.

4) Quadratic SVM with AlexNet

As for quadratic SVM, it has classified all images properly with 100% accuracy, as established in Figure 12.

Binocular	50	0	0	0	0
Motorbikes	0	50	0	0	0
Watches	0	0	50	0	0
Airplanes	0	0	0	50	0
Faces	0	0	0	0	50
	Binocular	Motorbikes	Watches	Airplanes	Faces

Figure 12: Confusion Matrix for Quadratic SVM with AlexNet.

It can be observed that images of Watches have suffered the highest inaccurate classification. This maybe the cause of poor images or the images contains too many variants of watches. A better accuracy for watches can be obtained if more images are used to train the classifiers. Altogether, extracting features using a pre-trained CNN and using Quadratic SVM generate better results.

B. *10-times-10-fold Cross-Validation*

Figure 13 shows that Quadratic SVM with AlexNet has performed significantly better than Quadratic SVM with BoW. Likewise, Linear SVM with AlexNet has performed significantly better than Linear SVM with BoW. Such decrease in accuracy may point out that BoW has failed to extract distinguishable features from many images. Whereas AlexNet's accuracy was not affected much as it can obtain better features from any images.

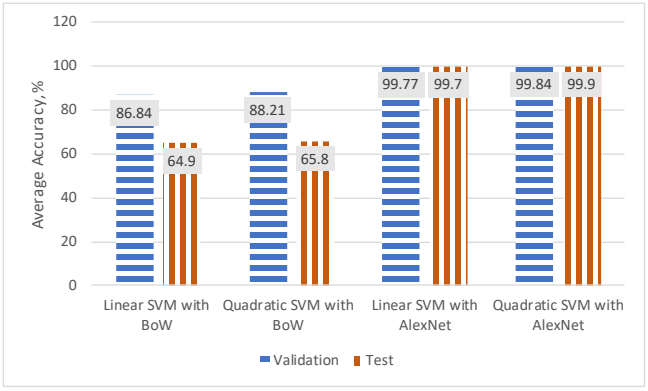

Figure 13: Validation and Test Accuracy for 10-times-10-fold cross-validation.

V. Conclusion

In this paper, the comparison between feature extraction using BoW and AlexNet CNN has been carried out. It has been noted that a pre-trained CNN can extract better distinct features from images compared to BoW. As a result, both the classifiers i.e. Linear SVM and Quadratic SVM can classify the image into correct categories, with high accuracy. It is observed that both classifiers have performed similarly but Quadratic SVM has always produced slightly better results than Linear SVM. BoW has not performed well with 10-times-10-fold cross-validation method. Its accuracy has decreased from 88% to 65.8% for Quadratic SVM and from 87.6% to 64.9 for Linear SVM. In terms 10-times-10-fold cross-validation, accuracy has also decreased for AlexNet for Quadratic SVM by 0.1%, while an increase of 0.1% for Linear SVM is observed.

Acknowledgement

The authors would like to acknowledge an Erasmus+ International Credit Mobility (ICM) fund for Bangladesh awarded to Dr Mohammad Hasan at Staffordshire University, UK in 2016.

References

[1] J. Chorowski, J. Wang, and J. M. Zurada, "Neurocomputing Review and performance comparison of SVM- and ELM-based classi fiers," vol. 128, pp. 507–516, 2014.

[2] P. Pilotte, "Analytics-driven embedded systems, part 2 – Developing analytics and prescriptive controls," 2016.

[3] J. Neumann, C. Schnörr, and G. Steidl, "Combined SVM-based feature selection and classification," *Mach. Learn.*, vol. 61, no. 1–3, pp. 129–150, 2005.

[4] U. J.R.R., S. A.W.M, and S. R.J.H., "Real-time Bag of Words , Approximately."

[5] J. Martineau and T. Finin, "Delta TFIDF : An Improved Feature Space for Sentiment Analysis," pp. 258–261, 2009.

[6] J. Yang and A. G. Hauptmann, "Evaluating Bag-of-Visual-Words Representations in Scene Classification," pp. 197–206, 2007.

[7] S. Wagner, "Combination of convolutional feature extraction and support vector machines for radar ATR," pp. 1–6, 2014.

[8] R. F. Nogueira, R. de Alencar Lotufo, and R. C. Machado, "Fingerprint Liveness Detection using Convolutional Networks," *Ieee Trans. Inf. Forensics Secur.*, vol. 11, no. 6, pp. 1206–1213, 2016.

[9] L. Lu, H. Shin, H. R. Roth, M. Gao, L. Lu, S. Member, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, "Deep Convolutional Neural Networks for Computer-Aided Detection : CNN Architectures , Dataset Characteristics and Transfer Learning Deep Convolutional Neural

Networks for Computer-Aided Detection : CNN Architectures , Dataset Characteristics and Transfer," *IEEE Trans. Med. Imaging*, vol. 35, no. 5, pp. 1285–1298, 2016.

[10] F. J. Huang and Y. LeCun, "Large-scale learning with SVM and convolutional nets for generic object categorization," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 1, pp. 284–291, 2006.

[11] P. Griffin, G. Holub, AD. Perona, "The Caltech 256," 2006.

[12] L. Fei-Fei; R. Fergus & A. Torralba, "Recognizing and Learning Object Categories, CVPR 2007 short course," *ICCV*, 2005.

[13] P. Fei-Fei Li; Perona, "A Bayesian Hierarchical Model for Learning Natural Scene Categories," *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2, p. 524, 2005.

[14] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF : Speeded Up Robust Features."

[15] David and MacKay, *An Example Inference Task : Clustering*. Cambridge University Press, 2003.

[16] "Image Classification with Bag of Visual Words," 2017. [Online]. Available: https://www.mathworks.com/help/vision/ug/image-classification-with-bag-of-visual-words.html. [Accessed: 02-Aug-2017].

[17] X. Du, Y. Cai, S. Wang, and L. Zhang, "Overview of Deep Learning," *31st Youth Acad. Annu. Conf. Chinese Assoc. Autom.*, pp. 159–164, 2016.

[18] A. Krizhevsky and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," pp. 1–9.

[19] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors arXiv : 1207. 0580v1 [cs . NE] 3 Jul 2012," pp. 1–18.

[20] B. Athiwaratkun, "Feature Representation In Convolutional Neural Networks," pp. 6–11.

[21] R. V. Babu, "A Taxonomy of Deep Convolutional Neural Nets for Computer Vision A Taxonomy of Deep Convolutional," no. January, 2016.

[22] B. Li, "Robust aircraft segmentation from very high-resolution images based on bottom-up and top-down cue integration."

[23] F. Lotte and M. Congedo, "TOPICAL REVIEW A review of classification algorithms for EEG-based brain–computer interfaces," vol. 1, 2007.

[24] C. J. C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," vol. 43, pp. 1–43, 1997.

[25] M. A. Aizerman, Braverman, M. Emmanuel, and L. l Rozonoer, "Theoretical foundations of the potential function method in pattern recognition learning," *Autom. Remote Control*, no. 25, pp. 821–837, 1964.

[26] "Classification parameter optimization," 2015. [Online]. Available: http://www.coxdocs.org/doku.php?id=perseus:user:activities:matrixprocessing:learning:classificationparameteroptimization. [Accessed: 04-Aug-2017].