Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 93 (2016) 351 – 358

Procedia Computer Science

6th International Conference On Advances In Computing & Communications, ICACC 2016, 6-8 September 2016, Cochin, India

SIFT and Tensor Based Object Detection and Classification in Videos Using Deep Neural Networks

Najva N.[a], Edet Bijoy K.[b]

[a]*Electronics and Communication Engineering, MES College of Engineering, Kuttippuram, Malappuram, Kerala, 679573. India*
[b]*Electronics and Communication Engineering, MES College of Engineering, Kuttippuram, Malappuram, Kerala, 679573. India*

Abstract

Object classification in videos is very important for applications in automatic visual surveillance system. The process of classifying objects into predefined and semantically meaningful categories using its features is called object classification. As far as humans are concerned object classification in videos is a simple task but it is a complex and challenging task for machines due to different factors such as object size, occlusion, scaling, lightening etc. The need for analyzing video sequences resulted in the development of different object classification techniques. In this paper we propose a new model for detection and classification of objects in videos by incorporating Tensor features along with SIFT to classify the detected objects using Deep Neural Network(DNN. Deep Neural Networks are capable of handling large higher dimensional data with billions of parameters as like human brain. Simulation results obtained illustrate that the proposed classifier model produces more accurate results than the existing methods, which combines both SIFT and tensor features for feature extraction and DNN for classification.

Peer-review under responsibility of the Organizing Committee of ICACC 2016

Keywords: Video Object Classification; SIFT; Tensor features; Deep Neural Network ;

1. Introduction

Object classifications in video sequences is an area of continuous development which has a wide range of applications in different fields such as biomedical imaging, biometry, video surveillance, vehicle navigation, visual inspection, robot navigation and remote sensing. Due to the availability of high quality cameras and rapid development in video capture technology, video is becoming a cheap source of information. This resulted in an extensive interest in the analysis of video sequences and classification of objects in it. The various steps involved in object classification are preprocessing, conversion of videos into frames, object detection, feature extraction and classification based on features extracted. Object classification in videos is a complex process which requires highly robust and accurate

* Edet Bijoy K. Tel.9037325763, Najva N. Tel.9846286792
E-mail address: najvahassan@gmail.com, edetbijoyk@ieee.org

Peer-review under responsibility of the Organizing Committee of ICACC 2016
doi:10.1016/j.procs.2016.07.220

techniques. Unfortunately, scientists are unable to develop a precise method for object classification in real world applications. So far no effective methods have been found for this problem. Videos are sequences of images, called frames displayed at a faster rate to create an illusion of motion and continuity. The classification of objects in videos is highly important in different applications such as traffic management, public transport system, object retrieval from videos etc and it requires high accuracy, flexibility and cost effectiveness.

A number of techniques have been developed for object classification and a review of different classification methods is given in this section. The goal of object classification is to categorize various objects based on the features extracted. In[1] R.J Denham and Pringle proposed SVM classification of object based data for crop mapping. The segmented data was classified using SVM classifier. SVM classification obtained an accuracy of 87%. The most crucial factor that affected the accuracy of SVM classification was the temporal changes in the spectral characteristics, specifically through vegetation indices derived from multi-temporal dataset. The classification of vehicles in videos is a complex process due to motion blurs and varying image resolution. It founds an important application in the field of traffic management and Toll Plaza. In[2] Narhe developed vehicle classification using SIFT algorithm. The SIFT algorithm is invariant to scale and rotational change, illumination change. Drew Schmitt, Nicholas McCoy proposed a method in[3] for object classification by extracting the features using SURF(Speeded Up Robust Features). The keypoints extracted from the training data set are clustered into N centroids. K means learning algorithms were used here. The advantage of SURF over SIFT was its concise descriptor length.

Due to its high coverage capabilities and limited costs as compared to manual method of taking samples from the seafloor, underwater object classification is an attractive approach using acoustic remote sensing techniques. Different ultrasonic techniques like multi beam echo sounder (MBES), a single-beam echo sounder (SBES) and side scan sonar(SSS) were done in[4] by Quen Feng Tan. Classification was achieved through the Bayesian approach, employing backscatter measurements per beam in MBES underwater object classification and in the SBES, echo shape parameters of the transmitted signal were determined. The classification of moving objects was done in[5] by Wu and Jian which combined the SOM with K-means. Self Organizing Map(SOM) can change the network parameters and structure by itself so that it can automatically find the sample properties. Kohonen weight adjustment rule were used to adjust the weight of win neurons as well as adjust the weight vectors of surrounding neurons. For security systems automatic detection and classification of objects is essential. A system based on probabilistic fusion of multiple features was proposed in[6] by Lipton, Alan to classify moving objects in different weather conditions. Object size, object velocity, location and difference of histogram of oriented gradients (DHoG) were the features extracted for classification.

The binary classification of either vehicles or humans was done in[7] by Zhang. Adaboost classifier was used as the tool for classification. Width,area, height, aspect ratio were the features extracted and a Adaboost classifier trained an ensemble of weak classifiers and produce a strong classifier but it was a complex process. A supervised object classification using Fuzzy logic was proposed in[8] by Nadeljkovic. ML classifier without null was used for classification. The synthetic 3D based object models were utilized to classify and recognize the moving objects in video[9] by Toshev and Alexander. Integration of feature tracking, motion grouping of tracks, and co-segmentation of successive frames were done to extract the silhouette from videos. The decision tree algorithm was proposed by Ragland,Kirubraj in[10] in which a tree was created consisting of attributes and symbols which formed the leaves of the tree. The attribute having highest entropy was calculated and the tree was built using this. The entropy depended on the number of occurrences of different attribute values. This method was undesirable when the values were undecided.

From the literature it is well understood that the existing classification methods reach the maximum accuracy of 87% and error rate were high and require further improvement. To improve accuracy, we propose an object classification model which combines SIFT and Tensor features. The classification using DNN using features(combination of both SIFT and Tensor) results in high accuracy with less error rate. Our contributions in this work are summarized as follows:

- We propose a classifier model that use deep neural networks for classification of objects in videos. The DNN can be used to represent more complex features so that the classifier efficiency is very much greater while compared to the currently existing techniques.
- From the existing methods our proposed new model incorporates SIFT and Tensor features together, which improves classifier accuracy much higher and reduces error rate.
- As SIFT is invariant to scaling and rotation, and tensor features are invariant under spatial transformation they can efficiently represent an object, combining the properties of both SIFT and tensor features helps the classifier to improve the accuracy. The proposed classifier performs well with accuracy above 90%.

2. Proposed Model for Classification of Objects in Videos

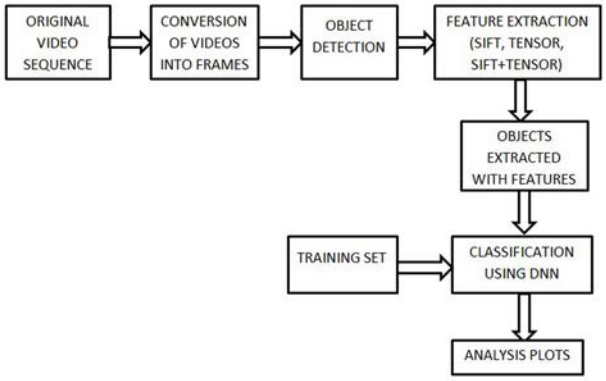

Fig. 1. Proposed Method for Classification of Objects in Videos Using Deep Neural Networks

A generalized block diagram of the object classification system in videos using deep neural networks is shown in Fig:1. The videos are the inputs which are to be converted into frames. The frames are analysed and the presence of an object is detected. Here background subtraction method is used to detect the presence of the object. The detection is followed by feature extraction. In this paper SIFT, tensor features and the combination of both SIFT and tensor features are used. The SIFT extracts the keypoint descriptors and the features are provided for training the DNN. The number of hidden layers as well as the neurons in the hidden layers are varied and the performance is evaluated. The training data should be properly selected so that the classification result has sufficient accuracy. Deep neural network is used for classifying the objects. Here we use a deep neural network having sufficient hidden layers and is trained using unsupervised learning using autoencoders. The classifier performance is analysed by its accuracy and the error rate.

2.1. Object Detection

After converting the videos into frames the object in the video is detected using Background Subtraction technique. It is the method to detect the moving objects from the difference between the current frame and a reference frame, often called background image. The variations between current video frames to that of the reference frame in terms of pixels signify existence of moving objects. It is a simple algorithm and is highly sensitive to identify object. The success of object detection depends on the object structure, speed, frame rate and global threshold. Background is estimated to be the previous frame. Equation for Background subtraction is given by $B(x, y, t) = I(x, y, t - 1)$ and $I(x, y, t - 1) - I(x, y, t) > Th$. If the value of the difference is greater, then the presence of an object is detected. Here we compare each of the video frames with reference model to determine the presence of an object.

2.2. Feature Extraction

A group of features in the form of a feature vector are used to represent an object. Object recognition and classification is done based on this feature vectors. Every classification system comprises the process of feature extraction. Feature extraction is the process of mapping the image pixels into feature space. For the classification of objects in videos the attributes which characterize them are calculated and are used for classification. The features are used to find the similarity between objects and it is represented as a vector. In image analysis feature selection is a critical issue. Here SIFT and Tensor features are preferred because SIFT features are invariant to scaling and rotation of the image and the tensor features are invariant to spatial transformations.

2.2.1. Scale Invariant Feature Transform (SIFT)

In SIFT the image is transformed into a collection of local feature vectors. The feature vectors obtained are invariant to any scaling, rotation or translation of image. Using SIFT algorithm invariant features are obtained from the images and are used to match between different views of an object. Main advantage of SIFT are they are robust against distortion, addition of noise and change in illumination. The keypoints are evaluated and are provided as input for the classification system. The major steps are Scale Space Extrema detection, Keypoint Localization, Orientation assignment and Keypoint descriptor. In Scale Space Extrema Detection extraction of the keypoints takes place. A function, $L(x, y, \sigma)$, is the scale-space of an image which is obtained from the convolution of an input image, $I(x, y)$, with a variable scale-space Gaussian function, $G(x, y, \sigma)$.

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y) \tag{1}$$

For the detection of stable keypoints the scale space extrema is described in terms of Difference of Gaussian function (DoG). The difference of Gaussian (DoG) function is given by, $D(x, y, \sigma) = L(x, y, k\sigma) - L(x, y, \sigma)$. Accurate localization of keypoint is done by interpolation with nearby samples and those keypoints that are unstable and sensitive to noise are eliminated. In order to obtain invariance to image rotation the keypoint descriptor is represented relative to the orientation. Histogram of local gradient directions at selected scale is computed. To obtain the directions of local gradient the highest peak in the histogram is detected. Each keypoint which are invariant are described using location, scale and orientation. The above operation gives location, scale, and orientation to each keypoint, which provides invariance to these parameters [11]. For an image sample , $L(x, y)$ at this scale, the gradient magnitude $m(x, y)$, and orientation $\theta(x, y)$, is calculated using equation given below

$$m(x, y) = [L((x + 1, y) - L(x - 1, y))^2 + L((x, y + 1) - L(x, y - 1))^2]^{\frac{1}{2}} \tag{2}$$

$$\theta(x, y) = \frac{\tan^{-} 1((x, y + 1)) - L(x, y - 1)}{L(x + 1, y) - L(x - 1, y)} \tag{3}$$

In SIFT the scale space structure of an object is explored by extracting the invariant features. Here only those features that are stable over transformation and having sufficient contrast are kept after filtering.

2.2.2. Tensor Based Feature Extraction

A number of existing features which are based on the structure tensor are extracted and are provided as the input to the deep neural networks. The tensor basis ensures that vectors pointing in opposite direction reinforce each other. The analysis of shape of tensor gave the orientation and a gradient norm estimate. For an image f the structure tensor is given by

$$G = \begin{pmatrix} \overline{f_x^2} & \overline{f_x f_y} \\ \overline{f_x f_y} & \overline{f_y^2} \end{pmatrix}$$

where the subscripts indicates spatial derivatives and the bar indicates the convolution with a Gaussian filter. For color images f = (R; G; B) the color structure tensor is given by

$$G = \begin{pmatrix} \overline{R_x^2 + G_x^2 + B_x^2} & \overline{R_x R_y + G_x G_y + B_x B_y} \\ \overline{R_x R_y + G_x G_y + B_x B_y} & \overline{R_y^2 + G_y^2 + B_y^2} \end{pmatrix}$$

In computer vision a commonly applied feature detector based on structure tensor is the Harris corner detector. The color Harris operator H on an image f is be computed by using the following equation

$$Hf = \overline{f_x^T f_x} + \overline{f_y^T f_y} - \overline{f_x^T f_y}^2 - k(\overline{f_x^T f_x} + \overline{f_y^T} + f_y) \tag{4}$$

Here Harris corner detector is used for feature extraction. The features which are extracted are provided for classifying the input samples into predefined categories. Under spatial transformations the tensor based features are invariant so that they can be used as an efficient method for feature extraction.

2.3. Classification Using DNN

Traditional machine learning techniques uses shallow nets, composed of one input and one output layer, and at most one hidden layer in between. A neural network that is having more than three layers (including input and output) is called a deep neural network(DNN). A deep neural network contains an input layer and an output layer, separated by l layers of hidden units. When an input sample is provided to the input layer, the other units of the network compute their values according to the activity of the units that they are connected to in the layers below. Distinct set of features based on the previous layers output are used to train each layer in DNN. Since in DNN they aggregate and recombine features from previous layer they can be used to recognize more complex features while more advancement is provided.

2.3.1. Training of DNN

Training a DNN using back propagation algorithm produce poor result. Unsupervised learning is used to train the DNN. It was put forward by Geoffrey Hinton. Unsupervised algorithms includes Deep Belief Networks, which are based on Restricted Boltzmann Machines, and Deep Autoencoders are based on Autoencoders. An autoencoder or

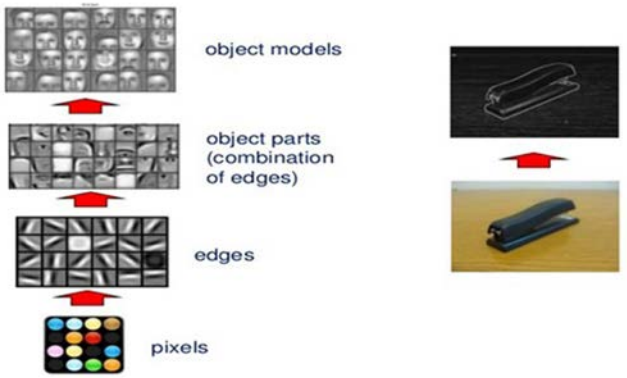

Fig. 2. Feature hierarchy in Deep Neural Networks

autoassociator is a three-layer neural network. The training of DNN is difficult in practice and training one layer at a time is an effective method to train DNN. A neural network which attempts to replicate its input at its output is called an autoencoder. Thus, the size of its input will be the same as the size of its output. The output value is set to the input itself in an autoencoder. i.e,$y(i) = x(i)$. The identity function is learned and it is given by $h(x) = x$. The hidden layer acts as the feature detectors after learning the weights. Dimensionality reduction is performed when $s_2 < s_1$ and if $s_2 > s_1$ the input is mapped to higher dimension. The sparse autoencoders extracts features from the inputs by putting a sparsity constrain on the weights in a DNN with large number of hidden units. For each input sample only particular units are activated because of this constrain. L1 regularization are used to construct sparse autoencoders. Kullback-Leibler(KL) divergence theorem is used to construct the sparse autoencoders [12]. It gives the measure of how

different two distributions are, which is given by a Bernoulli random variable with mean p and a Bernoulli random variable with mean,q

$$KL(p/q) = plog\frac{p}{q} + (1-p)log\frac{(1-p)}{(1-q)} \tag{5}$$

In the following definition, the activation of hidden unit j for input x_i is written as $a_j^2 x_i$. The average activation of unit j for a training set is defined as $p_j = \frac{1}{m}\sum_i a_j^2 x_i$. In order to compute p_j, the entire training set needs to be propagated forward first in the autoencoder. The sparsity optimization objective for unit j is $p_j = p$. RBM are two layer neural network that forms the building blocks of a deep belief network. It consists of two layers ,a visible layer and a hidden layer. The state observed or the input to the network is represented using the visible units of RBM. The feature detectors are represented by the hidden units of RBM. Activation function is one of the essential parameter in a Neural Network. Selection of an activation function for a network or it's specific nodes is an important task in order to produce an accurate result.Sigmoid function is used as the activation function for hidden layers. It is given by $p = \frac{1}{1+e^{-x}}$. Softmax function is used for activation of output layer. It is a generalization of the Sigmoid activation function. This is generalized to K outputs and can be combined ideally with the cross entropy cost function $p_j = \frac{e_j^x}{\sum_{i=1}^k e^{x^i}}$.

3. Experimental Studies and Discussions

Here the object classification in videos is implemented using DNN. The videos are converted into 25 frames per second and each frame is analysed. Background subtraction technique is done to detect the presence of an object. The experiment is done in three categories initially consider SIFT feature only, then Tensor feature only, later then combining SIFT and Tensor features. Here DNN is used for classification since it can solve perceptual problems in a way similar to that a human brain does. It works well for speech, audio and video signals. While compared to the existing techniques the accuracy obtained using a DNN classifier system using a combination of both SIFT and tensor for feature extraction is high. An object can be efficiently represented using SIFT and tensor so that an accurate result can be obtained for the classifier system. The SIFT features are invariant to occlusion, clutter and produce highly accurate results. Highly robust and accurate estimation of edge orientations can be done in down sampled images, where the edges are not smooth by using structure tensor. The structure tensor can classify local features into several distinctive types, which is non trivial by using gradient vectors alone. Due to Gaussian filtering stage, structure tensor achieves edge orientation which is robust against noise. Combining these properties of SIFT and tensor a set of features are extracted which produces an accurate result. The dataset used for training contains 450 samples of humans using 'person' [11] dataset and 60 samples of car using 'car' [12]dataset. Here sigmoid function is used to activate the hidden layers whereas the output layer is activated using softmax function. The test input is 5 second video which has the presence of humans and cars in it.

Initially SIFT is used for feature extraction and the size of extracted feature vector is $[90 \times 1]$. The accuracy is calculated from the confusion matrix obtained using DNN tool box. In Table:1 the accuracy obtained when SIFT is used for feature extraction is given. The accuracy is 54.6% when the DNN consists of two layers and the number of neurons is 100. It increases to maximum of 70.1% when the number of neurons is raised to 400. When the hidden layer number is increased to 3 the correct classification rate increases highly to nearly 90.2% and the number of neurons in each hidden layer is 400. As we increase it to 5 layers the maximum accuracy that can be obtained is 90.2% but the accuracy was 84.3% when the number of neurons is 200 whereas in a DNN with 3 hidden layers with 200 neurons the accuracy is only 81.2%. In Fig:3(a) a plot between accuracy and hidden layers is shown when SIFT is used for feature extraction. The rate of correct classification is maximised when the DNN consisted of 3 hidden layers with 400 neurons. From Table:1 and Fig:3(a) it is clear that by using DNN for classification accurate and precise results are obtained while compared to the existing classifier models.

The accuracy is increased when the feature is extended from SIFT to tensor based features. The length of feature vector now increases to $[370 \times 1]$. In Table:2 the accuracy obtained for DNN with hidden layers 2,3 and 5 is described

Table 1. Accuracy Comparison For DNN(Using SIFT)

HiddenLayers	100Neurons	200Neurons	400Neurons
2	54.6%	62.5%	70.1%
3	65.7%	81.2%	90.2%
5	73.5%	84.3%	90.2%

Table 2. Accuracy Comparison For DNN(Using Tensor)

HiddenLayers	100Neurons	200Neurons	400Neurons
2	57.5%	68.4%	79.3%
3	75.7%	85.2%	92%
5	77.6%	85.2%	92%

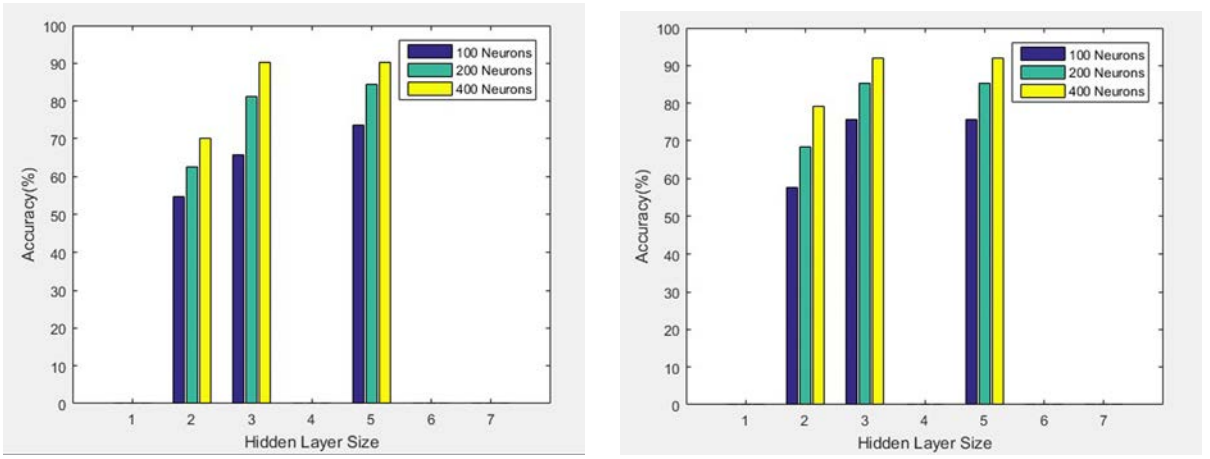

Fig. 3. (a) Accuracy vs Hidden Layer Size(Using SIFT); (b) Accuracy vs Hidden Layer Size(Using Tensor).

while varying the number of neurons from 100, 200 and 400 when tensor features are used. The maximum accuracy of 92% is obtained when feature extraction is extended to tensor based features and DNN consisting of 3 and 5 hidden layers with 400 neurons is used. When the DNN consisted of 3 hidden layers the accuracy increases from 75.7% to 85.2% by varying the number of neurons from 100 to 200. The accuracy increases when the hidden layer size extended to 5 but the maximum accuracy obtained is same as that of DNN with 3 hidden layers. Fig:3(b) is a plot between hidden layer size and accuracy(%) when tensor based features is used. The accuracy obtained for DNN with 3 and 5 is same and is very much greater than the accuracy obtained using DNN with 2 hidden layers. As we extend the features to tensor based features the accuracy of classification increased while compared to SIFT even though the vector size is increased.

Table 3. Accuracy Comparison For DNN(Using SIFT and Tensor)

HiddenLayers	100Neurons	200Neurons	400Neurons
2	67.8%	67.8%	89.3%
3	76.4%	87.2%	93.2%
5	77.6%	88.1%	93.2%

Table 4. Comparison of Performance

Features Extracted	Accuracy
SIFT	90.2%
Tensor	92%
SIFT + Tensor	93.2%

The accuracy of the classification system is maximized when the features used was the combination of both tensor and SIFT. Now the feature vector size is large [460×1]. The time taken for training the DNN also increased but we obtained an accurate classifier system for classifying the objects in videos. From Table:3 it is clearly evident that the efficiency of classifier system using DNN is highly increased when a combination of both SIFT and tensor is used. The accuracy increased from 67.8% to 89.3% when the number of neurons in DNN having 2 hidden layers is varied from 100 to 400 neurons. The maximum accuracy obtained is 93.2% using DNN of 3 hidden layers and 400 neurons. A plot between accuracy and hidden layer size is shown in Fig:4(a). When a combination of both SIFT and tensor is used the accuracy reached to a maximum of 93.2% which is greater while compared to accuracy obtained using SIFT and tensor alone.

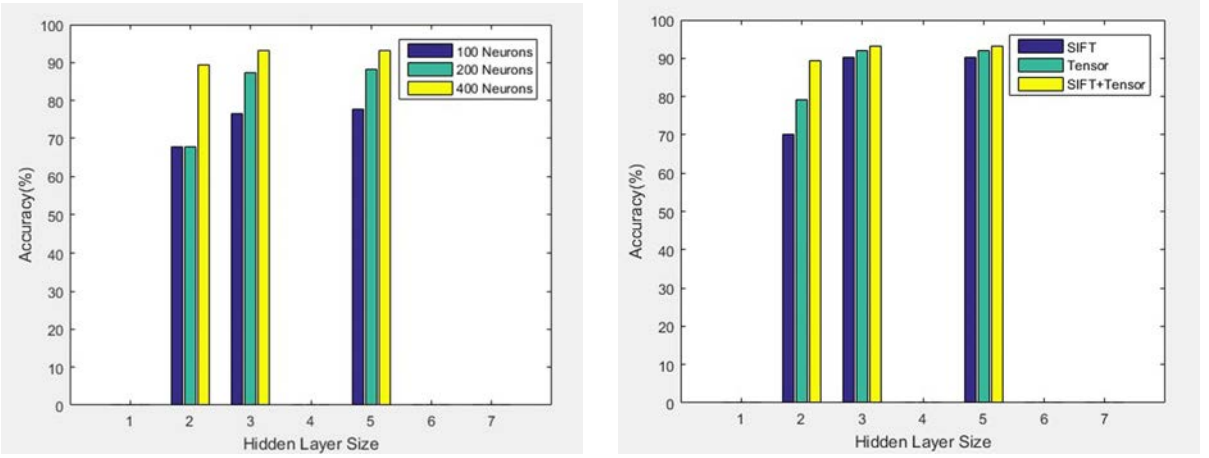

Fig. 4. (a) Accuracy vs Hidden Layer Size(Using Both SIFT and Tensor); (b)Accuracy vs Hidden Layer Size(For Various Features).

In Table:4 a comparison is made among the classifier system using SIFT, tensor and a combination of both for feature extraction. The accuracy obtained is high(93.2%) when both SIFT and tensor based features while compared to SIFT(90.2%) and tensor(92%) alone. Fig:4(b) is a graphical representation of the accuracy obtained for DNN using different features when varying the number of hidden layers in it. The classifier that uses a combination of both SIFT and tensor for feature extraction and DNN with 3 hidden layers produced the accurate result.

4. Conclusion

In this paper, we proposed a novel classification model which is used to classify the detected objects in videos using deep neural networks. Here the work done is to detect the presence of human and car from a video clip and to classify them. Experimental results are observed and analysed by varying the number of hidden layers and number of neurons in each hidden layer. DNN is trained using autoencoders and the performance of the classifier is analysed. The features used to train the network was a combination of SIFT and Tensor features. Experimental results shows that DNN with combining SIFT and Tensor features outperforms existing classifiers with accuracy reaching upto 93.2%.

References

1. Devadas, R, R. J. Denham, M. Pringle, Support vector machine classification of object-based data for crop mapping, using multi-temporal landsat imagery. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci 39 (2012): 185-190*
2. Narhe, Megha C., and M. S. Nagmode, Vehicle Classification using SIFT. *International Journal of Engineering Research and Technology. Vol. 3. No. 6 (June-2014). ESRSA Publications, 2014.*
3. Schmitt, Drew; Nicholas McCoy, Object classification and localization using SURF descriptors. *2011*
4. Kumar, Naveen, Qun Feng Tan, Shrikanth S. Narayanan, Object classification in sidescan sonar images with sparse representation techniques.*Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on. IEEE, 2012.*
5. Wu, Jian, et al, Moving object classification method based on SOM and K-means. *Journal of computers 6.8 (2011): 1654-1661.*
6. Lipton, Alan J., et al, A system for video surveillance and monitoring. *Vol. 2. Pittsburg: Carnegie Mellon University, the Robotics Institute, 2000.*
7. Zhang Lun, et al, Real-time object classification in video surveillance based on appearance learning. *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on. IEEE, 2007.*
8. Nedeljkovic, Igor, Image classification based on fuzzy logic. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 34.Part XXX (2004): 685.*
9. Toshev; Alexander, Ameesh Makadia, Kostas Daniilidis, Shape-based object recognition in videos using 3D synthetic object models. *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. IEEE, 2009.*
10. Ragland, Kirubaraj, P. Tharcis, A Survey on Object Detection, Classification and Tracking Methods. *International Journal of Engineering Research and Technology. Vol. 3. No. 11 (November-2014). ESRSA Publications, 2014.*
11. http://pascal.inrialpes.fr/data/human/
12. http://lear.inrialpes.fr/data