Efficient Object Recognition using Convolution Neural Networks Theorem

Aarushi Thakral
VIT University
Vellore
Tamil Nadu

Shaurya Shekhar
VIT University
Vellore
Tamil Nadu

Akila Victor
VIT University
Vellore
Tamil Nadu

ABSTRACT

Object recognition is the process of identification of an object in an image. There exist various algorithms for the same. Appearance based algorithms have demonstrated good efficiency, however, their performance gets affected adversely in the presence of clutter or when background changes are affected. We hope to overcome this issue by using Convolution Neural Network (CNN) Theorem. The approach is shape based and has been proven to work well under broad range of circumstances: varied lighting conditions, affine transformations, etc. It involves tiling, which is the phenomenon of the use of multiple layers of neurons to process small portions of the image, which are then used to obtain better representations of the image. This allows CNN to be translation-tolerant. The neural elements learn to recognize objects about which they have no previous information, this 'learning' mechanism is affected by the fact that representations of the image are learned by the inner layers of the deep architectures of neurons. Unlike RBM and Auto-encoder, which are capable of learning only single global weight matrix layers, the CNN theorem makes use of shared weight in convolution layers, which means that the same filter (weight bank) is used for each pixel in the layer, which reduces the memory footprint and improves performance.

General Terms

Object Recognition, Convolution Neural Networks, Machine Learning, Artificial Intelligence.

Keywords

Recognition, Object, Neural, Features, Dataset, Training, Image

1. INTRODUCTION

The concept of convolution neural networks takes inspiration from the human brain. The human brain takes visual inputs from the eye and performs image recognition on the basis of stored information which it has learned over a period of time, which is how it is able to identify occluded objects as well. We therefore, try to develop comprehensive datasets in accordance with our needs to help test the effectiveness of algorithms. These algorithms can be trained with the help of these datasets to 'learn' to identify objects on the basis of edge recognition and image segmentation. These techniques help in dividing large images into smaller images, which thus enable faster processing to identify images. Neural networks can be used effectively for such cases. In Section 2, a number of papers have been referred to which bring out the advantages and disadvantages of the various object recognition techniques being used today. We have also referred to papers which propose the development of new datasets which would be helpful to train algorithms. In Section 3, papers which make use of Convolution Neural Networks have been referred to. In Section 4, we present a tabulated approach to understanding all the different approaches that we have come across while researching.

2. VARIOUS OBJECT RECOGNITION TECHNIQUES

The object recognition capabilities are expanded to recognizing complex multi-agent action [1]. The actions need to recognized from visual evidence and have three components: temporal structure descriptions which represent the temporal relationships between agent goals, belief networks for probabilistically representing and recognizing individual agent goals and the belief networks being automatically generated by the temporal structure descriptions that support recognition of the complex multi-agent action. The temporal structure descriptions contains the prototypical scenario of described action. They are basically behavior elements connected by temporal constraints, such as before, after and around. Individual agent goals are used as the basis for the descriptive structure and complex actions are viewed as a partially ordered set of goal directed behaviors on part of the interacting agents.The agent goals can be represented in a probabilistic framework using Bayesian belief networks. A belief network represents each goal or event and can be instantiated at any time during a play, these networks contain between 15 and 25 nodes with a tree-link complexity and can therefore be used by propagation algorithms to compute the probability of every node state. These networks consist of two states: unobservable and observable evidence states. These states aren't closed and can utilities the outputs of other belief networks. This helps in the creation of a dependency structure and lets us partition complex actions into smaller networks which simplifies and makes complex jobs more manageable. Multi-agent actions can be recognized by using multi-agent belief networks. At each of these times, the network integrates the likelihood values returned by temporal analysis functions at that time and returns a likelihood that the given event has occurred. Thus, a structure has been proposed to represent complex multi-agent actions as simpler low-order temporal graphs, thus breaking down complex actions into a series of simpler ones which results in easier recognition of occurred events. Visual evidence is probabilistically integrated into multi-agent belief networks.

We have proposed to exploit shape information via masking convolutional features. The proposal segments (e.g., super-pixels) are treated as masks on the convolutional feature maps[2]. The CNN features of segments are directly masked out from these maps and used to train classifiers for recognition. Further we propose a joint method to handle objects and "stuff" which can be anything from grass, sky to water, in the same framework. We design a convolutional feature masking (CFM) method to extract segment features directly from feature maps instead of raw images. With the segments given by the region proposal methods we project them to the domain of the last convolutional feature maps. The projected segments play as binary functions for masking the convolutional features. The masked features are then fed into the fully- connected layers for recognition. Because the convolutional features are computed from the unmasked image, their quality is not impacted. We have further shown that convolutional feature masking is applicable for joint object and stuff segmentation and state-of-the-art results are demonstrated on benchmarks of PASCAL VOC and new PASCAL CONTEXT, with a compelling computational speed.

Image Segmentation is defined as the process of partitioning an image into smaller, yet meaningful regions with respect to a particular function. Object Recognition refers to finding a particular object in an image or video sequence. The images are divided in such a manner such that adjacent regions differ from each other on the basis of a characteristic function, whereas every point within a region demonstrates a uniformity with respect to the same function. This is done using a number of techniques: Sobel, Prewitt, Roberts, Canny, Laplacian of Gaussian (LoG), EM Algorithm, OSTU and Genetic Algorithm[3]. The two types of image segmentation are edge- based segmentation and region-based segmentation, while the former depends on abrupt changes in intensity, the latter depends on the concept of uniformity with respect to a particular function. The segmentation of the image yields in a number of edges, which leads us to edge detection, which in itself is composed of three parts: filtering, enhancement and detection. The EM algorithm presents an approach to compute the Maximum Likelihood Estimate (MLE) when missing and hidden data are encountered. Every iteration consists of an E-step, where the missing data is estimated taking into account given data and estimate of model parameters, also called conditional expectation, the M step involves maximization under the assumption that missing data is known, here, the estimations from the E step are used in place of the actual missing values. In the OTSU algorithm, threshold selection is performed by assuming every image has two basic parts: foreground and background. The OTSU algorithm involves separating the image into two clusters of pixels on the basis of the threshold, calculating the mean value of each cluster, squaring the difference between the means and multiplying it by the number of pixels in one cluster with the other. The drawback of this method is that only point intensities are considered and not the relationships between pixels. The Genetic Algorithm methodology tries to find structure in seemingly random data, either because of insufficient knowledge or complexity. It consists of three parts: selection, which involves selecting only the fittest data, and picking some less fit ones in accordance with a probability function, crossover, which involves combining two individual pixels to create new-better ones and lastly the mutation operator, which introduces changes in small number of chromosome units, its main purpose is to keep the population diversified.

We talk about how to recognize multicoloured objects invariant to a substantial change in view point, illumination and object geometry[4]. A change in spectral power distribution of the illumination is considered and a new color constant color model m1m2m3 is proposed. The result of the experimental show that highest object recognition accuracy is achieved by l1l2l3 and hue H which then is followed by c1c2c3, normalized color RGB and m1m2m3 under the constraint of white illumination and it is demonstrated that accuracy of recognition degrades substantially for all color features other than m1m2m3 with a change in illumination color. Color provides powerful information for object recognition hence the choice of which color models to use depend on their robustness against varying illumination across the scene, against changes in surface orientation of the object, and object occlusion and cluttering. Furthermore, the color models should be concise, discriminatory and robust to noise.

RGB is most appropriate for multicoloured object recognition when all imaging conditions are controlled but the discriminative power of RGB has the worst performance due to its sensitivity to varying imaging conditions. Without the presence of highlights and under the constraint of white illumination, c1c2c3 and normalized color RGB are most appropriate. When images are also contaminated by highlights, l1l2l3 or H should be taken for the job at hand. When no constraints are imposed on the SPD of the illumination, m1m2m3 is most appropriate.

Table 1. Overview of the various color models and their invariance to various imaging conditions. + denotes invariance & - denotes sensitivity of the color model to the imaging condition

	RGB	**$c_1c_2c_3$ rgb**	**$I_1I_2I_3$ H**	**$c_1c_2c_3$ rgb**	**$m_1m_2m_3$**	**$I_1I_2I_3$ H**	**$c_1c_2c_3$ rgb**	**$m_1m_2m_3$**	**$I_1I_2I_3$ H**	**$m_1m_2m_3$**	**$m_1m_2m_3$**	**$m_1m_2m_3$**	**$m_1m_2m_3$**	**$m_1m_2m_3$**	**$I_1I_2I_3$ H**	**$m_1m_2m_3$**
Viewpoint & Geometry	+	-	+	+	+	-	-	-	+	+	+	+	-	-	-	-
Highlights	+	+	-	+	+	-	+	+	-	-	+	-	+	-	-	-
Illumination Intensity	+	+	+	-	+	+	-	+	-	+	-	-	-	+	-	-
Illumination Color	+	+	+	+	-	+	+	-	+	-	-	-	-	-	+	-

Table 2. Overview of which color model to use under which imaging condition. + denotes controlled and – denotes uncontrolled imaging condition

	Viewing Direction	**Surface Orientation**	**Highlights**	**Illumination Direction**	**Illumination Intensity**	**Illumination Color**	**Inter Reflection**
I	-	-	-	-	-	-	-
RGB	-	-	-	-	-	-	-

rgb	+	+	-	+	+	-	-
S	+	+	-	+	+	-	-
c1c2c3	+	+	-	+	+	-	-
H	+	+	+	+	+	-	-
I1I2I3	+	+	+	+	+	-	-
m1m2m3	+	+	-	+	+	+	+

The problem of spatial object recognition in geographical information systems is tackled by means of reviewing methods in which it can be made easier as well as more efficient. They are done in the following steps: by improving the quality of object recognition, by creating general rules to reduce the possibility of erratical classification of objects into classes and by establishing rules defining the correlation of spatial objects of different classes[5]. Object recognition first takes place on the topological basis, wherein, objects in an image are classified into categories on the basis of image segmentation, and the various measures which can then be extracted by such measures. The concept of development of neural networks also finds mention which can be utilized after due training for determination of various factors from the images obtained by the geographical information system. The neural networks benefit from continued use as they 'learn' from every image that they encounter, eventually all the various possibilities that they encounter can be used to effectively and accurately perform object recognition. The last step involves merging of a large number of object recognition techniques, which involve methods such as binarization of the image, application of Canny method and sifting of the contours. These merged methods help in analysis of images of moving objects, development of methods of unified processing procedure of images and recognition of objects in space pictures.

A method for object recognition with full boundary detection by combining a region merging algorithm and affine scale invariant feature transform (ASIFT). ASIFT[6] is a fully affine invariant algorithm that means features are invariant to 6 affine parameter which are zoom, 2 translation parameters, rotation and 2 camera axis orientations. These features are very reliable and give us strong key points which we can use them for finding objects in other images. After that a robust region merging algorithm is used to recognize and detect the object with full boundary in the other images based on ASIFT key points and a similarity measure for merging regions in the image. In some other algorithm the users must indicate some of regions and locations of the background and object to run the algorithm, but the presented algorithm does not have the stated problem in region merging algorithm i.e. the algorithm does not need to indicate regions by user. We use the best key points of object that has been obtained from ASIFT results and apply them into the image. Therefore, the method will be an automatic algorithm and will not need marks by users and the achieved key points from ASIFT have been replaced with them.

A key problem in image and vision computing is pedestrian detection[7], it has several applications including robotics, surveillance and automotive safety. The pedestrian detection algorithms are based on the availability of challenging datasets to help test presently used algorithms, they include videos and low-resolution images recorded from moving vehicles, thus, the Caltech Pedestrian Dataset has been introduced to provide a better benchmark and to help identify scenarios in which current detection methods fail. For every pedestrian seen, the pedestrian was marked in the frame using a bounding box. The various statistics used to segregate the images are: scale (on the basis of height of bounding box), occlusion (this refers to the phenomena of pedestrians being covered by other pedestrians of other objects) and position (the position of most bounding boxs' with respect to the other view frame of the camera). This dataset is deemed to be better than most other datasets as it allows us to check if given algorithms can identify occluded pedestrians or not and is also being used to test all the other algorithms instead of being used to try to prove the effectiveness of any one algorithm, which is why it offers a relatively level playing field. Thus, the Caltech Pedestrian Dataset includes O(10^5) pedestrian BBs labeled in O(10^5) frames and two orders of magnitude more than any other dataset. It includes color video sequences and contains pedestrians with a large range of scales and more pose variability than typical pedestrian datasets. There also exist temporal correspondence between building boxes and detailed occlusion labels. Among the various algorithms tested, HOG, MultiFtr and FtrMine are the best algorithms, with HOG performing best on near & unoccluded pedestrians & MultiFtr trying to and outperforming HOG on more difficult cases (smaller and occluded pedestrians), which results in MultiFtr achieving a slightly better overall performance, while VJ and Shapeless perform poorly. LatSvm suffers because it has been trained on the Pascal dataset while HikSym cannot detect small people. However, the absolute performance of each of these algorithms is still poor.

We demonstrate that a properly designed integral channel feature[8] which outperforms other features including histograms of oriented gradient (HOG), and also

(1) naturally integrates heterogenous sources of information

(2) has few parameters and is insensitive to exact parameter settings,

(3) allows for more accurate spatial localization during detection, and

(4) results in fast detectors when coupled with cascade classifiers.

Integral channel features combine the diversity of information from use of image channels with the computational efficiency of the Violaand Jones detection framework. The evaluation is done by combinations of three types of channels: gradient histograms, color (including gray scale, RGB, HSV and LUV), and gradient magnitude. According to the experiment the most informative channels are the gradient histogram channels (Hist) achieving 87.2% detection performance at the reference point of 10^{-4} fppw and 89% achieved by the HOG classifier with optimized parameter setting. Combining Hist with gradient magnitude (Grad) and grayscale information (Gray) increases performance to89.4% and LUV achieves 55.8% on its own and 91.9% when combined with Grad and Hist. The combination of diverse, informative channels along with the integral image trick for fast feature computation can help us in long run.

One of the hallmarks of object recognition is scene recognition. The availability of large datasets and the rise of Convolution Neural Networks has resulted in the capability to learn high-level features and has thus made object recognition smarter and more efficient. A new scene-centric database called Places[9] is introduced which has over 7 million scenes from 476 place categories and is 60 times larger than the SUN database. It is the first scene centric database competitive enough to be able to train algorithms that require huge amounts of data such as CNNs. The design of Convolution Neural Networks is based on that of the brain and its hierarchal organization into layers of increasing complexity. Such increasingly complex architectures coupled with comprehensive databases help in extremely efficient object classification tasks. Adding adjectives to the queries allows us to download a larger number of images and thus, effectively increase the diversity of visual appearances, after which duplicate images are removed, resulting in a database with only unique URLs, duplicates are matched among the same categories from the Places and SUN database, which results in the two of them containing different images and allowing us to combine the two datasets. The pictures are then sent to Amazon Mechanical Turk for two rounds of image annotation by humans. Comparison of scene-centric databases is carried out on the basis of density (having a high degree of data concentration) and diversity (high variability of appearances and viewpoints). An ideal dataset should have both, high density as well as high diversity. The CNN theorem is then applied on this enhanced dataset, which exposes the CNN to a more varied dataset, thus allowing it to be used for more efficient object recognition as it has been exposed to a more 'denser' and a more 'diverse' set of images in comparison to those available in other datasets. This helps in greater learning of the CNN, and greater applicability. The convolution neural networks are designed in a manner to benefit and learn from massive amounts of data.

A category-independent shape prior for object segmentation has been introduced[10]. The main insight of this approach is that shapes are often shared between objects of different categories and to exploit this shape sharing phenomena we have developed a non- parametric prior that transfers object shapes from an exemplar database to a test image based on local shape matching. The transferred shape priors are then enforced in a graphcut formulation to produce a pool of object segment hypotheses. The approach in this paper consists of three main steps:

(1) Estimating global object shape in a test image by projecting exemplars via local shape matches

(2) Aggregating sets of similarly aligned projected shapes to form a series of hypothesized shape priors and

(3) enforcing the priors within graphcuts to generate object segment hypotheses.

Table 3. Our shape based projection and merging approach outperforms an existing merging strategy while requiring an order of magnitude fewer segments (second row). It also substantially improves the state-of-the-art bottom up segmentation (third row).

Approach	**Covering (%)**	**Num Segments**
Exemplar-based merge (Ours)	77.0	607
Neighbor merge [2]	72.2	5005
Bottom-up Segmentation [10]	62.8	1242

This approach through experiments prove that not only does shape sharing improves the quality of bottom-up segmentation, while requiring no prior knowledge of the object, it also shows that category-independent prior performs as well as a parallel category-specific one, demonstrating that shapes are truly shared across categories.

3. CONVOLUTION NEURAL NETWORKS

Successful visual object recognition methods typically rely on training datasets containing lots of richly annotated images. Annotating object bounding boxes is subjective as well as expensive. A weakly supervised convolutional neural network (CNN) [11] for object recognition that does not rely on detailed object annotation and yet returns 86.3% mAP on the Pascal VOC classification task, outperforming previous fully-supervised systems by a sizeable margin is proposed. Despite the lack of bounding box supervision, the network produces maps that clearly localize the objects in cluttered scenes. Adding fully supervised object examples to our weakly supervised setup does not increase the classification performance. Visual object recognition entails much more than determining whether the image contains instances of certain object categories. A fully supervised network architecture that consists of five convolutional and four fully connected layers and assumes as input a fixed-size image patch containing a single relatively tightly cropped object is build. To adapt this architecture to weakly supervised learning the following three modifications are introduced.

1) The fully connected layers are treated as convolutions, which allow dealing with nearly arbitrary-sized images as input.

2) The highest scoring object position in the image is searched for, by adding a single global max-pooling layer at the output.

3) A cost function that can explicitly model multiple objects present in the image is used.

An object recognition CNN trained without taking advantages of the object bounding boxes provided with the Pascal VOC training set is described. Despite this restriction, this CNN outperforms all other practical methods. The network also provides qualitatively meaningful object localization information.

We have proposed to exploit shape information via masking convolutional features.[12] The proposal segments (e.g., super- pixels) are treated as masks on the convolutional feature maps. The CNN features of segments are directly masked out from these maps and used to train classifiers for recognition. Further we propose a joint method to handle objects and "stuff" which can be anything from grass, sky to water, in the same framework. We design a convolutional feature masking (CFM) method to extract segment features directly from feature maps instead of raw images. With the segments given by the region proposal methods we project them to the domain of the last convolutional feature maps. The projected segments play as binary functions for masking the convolutional features. The masked features are then fed into the fully-connected layers for recognition. Because the convolutional features are computed from the unmasked image, their quality is not impacted. We have further shown that convolutional feature masking is applicable for joint object and stuff segmentation and state-of-the-art results are demonstrated on benchmarks of PASCAL VOC and new PASCAL CONTEXT, with a compelling computational speed. CNN has demonstrated significant results in single-label image classification tasks. [13] A flexible deep CNN infrastructure, called Hypotheses-CNN-Pooling (HCP), where an arbitrary number of object segment hypotheses are taken as the inputs, then a shared CNN is connected with each hypothesis, and then finally the CNN output results from different hypotheses are aggregated with max pooling to produce the best multi-label predictions. Some unique characteristics of this flexible deep CNN infrastructure are:

1) No ground-truth bounding box information is required for training;

2) The whole HCP infrastructure is robust to possibly noisy and/or redundant hypotheses;

3) No explicit hypothesis label is required;

4) The shared CNN may be well pre-trained with a large-scale single-label image dataset, e.g. ImageNet; and

5) It may naturally output multi-label prediction results.

The proposed HCP requires no bounding box annotation for training, and hence can easily adapt to new multi-label datasets and it is also proved that late fusion between outputs of CNN and handcrafted feature schemes can enhance the classification performance.

One of the hallmarks of object recognition is scene recognition. [14] The availability of large datasets and the rise of Convolution Neural Networks has resulted in the capability to learn high-level features and has thus made object recognition smarter and more efficient. A new scene- centric database called Places is introduced which has over 7 million scenes from 476 place categories and is 60 times larger than the SUN database. It is the first scene centric database competitive enough to be able to train algorithms that require huge amounts of data such as CNNs. The design of Convolution Neural Networks is based on that of the brain and its hierarchal organization into layers of increasing complexity. Such increasingly complex architectures coupled with comprehensive databases help in extremely efficient object classification tasks. Adding adjectives to the queries allows us to download a larger number of images and thus, effectively increase the diversity of visual appearances, after which duplicate images are removed, resulting in a database with only unique URLs, duplicates are matched among the same categories from the Places and SUN database, which results in the two of them containing different images and allowing us to combine the two datasets. The pictures are then sent to Amazon Mechanical Turk for two rounds of image annotation by humans. Comparison of scene-centric databases is carried out on the basis of density (having a high degree of data concentration) and diversity (high variability of appearances and viewpoints). An ideal dataset should have both, high density as well as high diversity. The CNN theorem is then applied on this enhanced dataset, which exposes the CNN to a more varied dataset, thus allowing it to be used for more efficient object recognition as it has been exposed to a more ‘denser’ and a more ‘diverse’ set of images in comparison to those available in other datasets. This helps in greater learning of the CNN, and greater applicability. The convolution neural networks are designed in a manner to benefit and learn from massive amounts of data.

The key challenge of face detection [15] is the large appearance variations due to the number of real-world factors such as pose changes, exaggerated expressions and extreme illuminations which can lead to the large intra-class variations and making the detection algorithm not robust enough. A locality sensitive support vector machine using kernel combination (LS-KC-SVM) algorithm to solve the above two problems is proposed. The locality-sensitive SVM (LSSVM) to construct a local model on each local region, which can handle the classification task easier due to smaller within-class variation is employed first then motivated by the idea that local features are more robust compared with global features, multiple local CNNs to jointly learn local facial features because of the powerful strength of CNN learning characteristic are used. To use this property of local features effectively, the global and local kernels to the features are applied and are introduce to the combination kernel to the LSSVM. Extensive experiments demonstrate the robustness and efficiency of the algorithm by comparing it with several popular face detection algorithms on the widely used CMU+MIT dataset and FDDB dataset. A background filter is trained in order to filter away the simple background as soon as possible. For this purpose, Adaboost is used as it is very effective and real time since Viola and Jones's work. . Setting the threshold too high may cause too many positive examples to be rejected, reducing the overall detection rate, and setting it too low can lead to too many pass-through patches that need to be classified further by the SVM, slowing down the overall detection process. So a minimum detection rate of 99.8% and a maximum false positive rate of 50% were set as the training parameters. Since local kernel is more robust than global kernel, to better exert the function of LS-KC-SVM, multiple local CNNs are adopted which apply to different face regions to jointly learn discriminate local features used in our LS-KC-SVM algorithm.

First adopting the edge box algorithm [16] to generate region proposals from edge maps for each image, and perform forward passing of all the proposals through a fine-tuned CaffeNet model. Then by getting the CNNs score for each proposal by extracting the output of softmax which is the last layer of CNN. Next, the proposals are merged for each class independently by the greedy non maximum suppression (NMS) algorithm. Then evaluating the mean average precision (mAP) for each class. The mAP of the model on PASCAL 2007 test dataset is 37.38%.

The overview of the object detection system.

1) Edge box algorithm generates proposals.

2) A fine-tuned CNN with object classes plus one background class to extract the CNN features for each proposal.

3) Employ the softmax to give the confidence (score) of all the classes for each proposal (which is typically the last layer of CNN).

4) For each class, independently use the non-maximum

suppression (NMS) to greedily merge the overlapped proposals.

The model achieves the 0.3738 mAP on VOC 2007 dataset. Changing a deeper network to increase the accuracy of classification as well as to add the ground truth bounding boxes into the training data to improve the localization accuracy.

We now apply the concept on convolution neural networks to facial expression recognition[17] which is one the most difficult to understand due to the variation in facial appearances, postures, face sizes and translation and invariance. A large number of variables allow proper use of the convolution neural network and its learning capabilities. Models have also been proposed which involve two independent CNNs, one for facial expression and the other for face identity recognition and the findings are then combined using a MLP. We have therefore, tried to refine this method itself. The concept of facial expression recognition thus depends on two types of recognitions: face recognition and expression or emotion recognition. Face recognition depends on the face model which is represented as a spatially ordered set of local features of medium complexity such as eyes, mouth, nose, eyebrow, cheek, etc. These are represented in the form of a fixed set of figural alphabets within the Convolution Neural Network. A local template is formed in the hierarchal network which helps in detection of these features. The proposed CNN is trained module-by-module and the detection result of skin colour is used as an input to the face detection module to help in narrowing down the number of closely matching faces on the basis of complexion. Training is carried out using the method of back propagation with intermediate local features such as eye corners in the FD1 and FD2 layers, while the more complex feature detectors such as mouth detectors are carried out in the FD3 or FD4 layers. As we move, from the lower layers to the higher layers, all the results begin to be collated and eventually, FD4 learns to locate faces in complex scenes. To put it simply, as they move from lower layers to higher layers they individually identify smaller features of the face and then collate all of the data to detect the face in FD4. A rule-based processing scheme has also been proposed as it was noticed that some of the lower level features extracted by the first FD layer was useful for facial expression recognition. Various changes in the distance between the various lower level features help in enhancing expression recognition and thus, enhance subject independence in facial expression recognition.

Recognizing text images is a tedious job wherein we have to consider a two stage system. We firs break the images into smaller segments and extract the characters and then perform the image recognition on the extracted data. Segmenting and recognizing the characters go hand in hand. With a goal to amalgamate segmentation and recognition, we will follow the proposed theory of hybrid CNN-HMM [18] where the model undergoes sliding window mechanism to extract a series of frames. Feature descriptors like SIFT, HOG and LBP are widely used in vision problems. These features are not created from an optimisation process to be congruous with a particular issue. Convoluted Neural Networks is primarily used as a static model which is fixed dimensional. When we focus on the different methods, we realize that HMM is temporal modelling, whereas CNN (Convoluted Neural Networks) focuses on character appearance variations. In order to operate CNN, we have to provide initial state assignment for each frame.

The proposed operator i.e. the new agent-based operator [19] recognizes objects around one object only in one step and each point object acts as an agent-automata object. After that it senses its vicinity and identifies the objects in the surrounding. To assess this model, the operator is designed, implemented, and evaluated in a case study. Finally, the results are evaluated and presented in detail. The utility-based agent is used to for this approach. At first, the goal is defined for agent; then the agent tries to reach the defined goal optimally. The x and y coordinates of endpoints are considered as agent's observation. In this case, the agent is a point with its x, y coordinates. The objects around the agent are categorized into three types: points, lines, and polygons. The distance between the agent and point object is calculated based on Euclidian distance. In this model, four steps are defined for the process; at first, the agent observes the environment; then it gathers some data. Next, the perception of agent is obtained based on some analysis done on the observation. In the next step, agent generates its knowledge base. Finally, it selects objects based on its knowledge. The results of the implementation showed that in a complex area the agent-based buffer works correctly.

To improve the performances of current approaches, which use machine learning methods, [20] large datasets were collected and better techniques to avoid over fittings were used. A large, deep CNN was trained to classify the 1.2 million high- resolution images in the ImageNet LSVRC-2010 contest into the 1000 different classes. On the test data, top-1 and top-5 error rates of 37.5% and 17.0% which is better than the previous state-of-the-art was achieved. The neural network, which has 60 million parameters and 650,000 neurons, consists of 5 convolutional layers, some of which are followed by max- pooling layers, and three fully-connected layers with a final 1000-way softmax. Non-saturating neurons and a very efficient GPU implementation of the convolution operation was used to make training faster. To reduce overfitting in the fully- connected layers a recently-developed regularization method called "dropout" that proved to be very effective was employed. A variant of this model in the ILSVRC-2012 competition achieved a winning top-5 test error rate of 15.3%, compared to 26.2% achieved by the second-best entry. The results show that a large, deep convolutional neural network is capable of achieving record breaking results on a highly challenging dataset using supervised learning of different colors.

4. TABULATED FINDINGS

Table 4. Descriptions of Revised Papers

Sl. No.	Author	Title	Advantage	Disadvantage	Technique
1	Stephen S Intille, Aaron	A Framework for Recognizing Multi-Agent Action from	Complex multi-agent actions are being	Multi-agent belief networks cannot handle compound group of	Building of serializable simple belief networks for complex multi-

	F Bobbick	Visual Evidence	identified.	actions.	agent actions.
2	Jifeng Da, Kaiming He, Jian Sun	Convolution Feature Masking for Joint Object and Stuff Segmentation	1.Convolutional feature maps only need to be computed once. 2. The convolutional features are computed from the unmasked image, their quality is not impacted.	1. The masks on the image content can lead to artificial boundaries. These boundaries do not exhibit on the samples during the network pre-training . This issue may degrade the quality of the extracted segment feature 2. Similar to the RCNN method for object detection, these methods need to apply the network on thousands of raw image regions with/without the masks. This is very time consuming even on high-end GPUs	Convolutional feature masking (Feature map based) ,
3	Y.Ramadevi, T.Sridevi, B.Poornima, B.Kalyani	Segmentation And Object Recognition using Edge Detection Techniques	Images are divided into smaller regions (segmentation), which results in faster recognition methodologies	Only point-value intensities are taken into account, relationships between pixels are ignored	Sobel, Prewitt, Roberts, Canny, LoG, EM Algorithm, Genetic Algorithm, OTSU Algorithm
4	Theo Gevers, Arnold W. M. Smeulders	Color-based Object Segmentation	1.RGB works the best when all the imaging conditions are controlled . 2. When no constraints are imposed on the illumination, the proposed color ratio m1m2m3 works the best.	RGB has the worst performance due to its sensitivity to varying imaging conditions..	Discussing various techniques of object recognition based on color and propose m1m2m3 technique.
5	Andrianov D.E., Eremeev S.V., Kuptsov K.V.	The Review of Spatial Object Recognition Models and Algorithms	Development of Neural Networks & merging of various methodologies for enhanced object recognition	Exact replicas required to avoid erratic classification, Neural Networks require long time to 'learn' how to identify objects, false detection of green component of RGB and lack of automatic calculation of characteristics of moving objects	Neural Networks, Canny Method, Binarization of images, combination of object recognition methodologies

6	Reza Oji	An Automatic Algorithm For Object Recognition and Detection Based On ASIFT Keypoints	ASIFT covers two more parameters than SIFT which are longitude and latitude angle that are relevant to camera axis orientation. Which means that ASIFT is more effective and can be more robust in the changes of images.	Exceptional performance noted only in cases of affine changes.	ASIFT, City Block distance, Euclidean distance, color histogram
7	Piotr Dollar, Christian Wojek, Bernt Schiele, Pietro Perona	Pedestrian Detection: A Benchmark	A level playing field for testing of various pedestrian detection algorithms, which includes occluded and small (far) pedestrians as well.	No compatibility of dataset with detectors which utilize features computed over 2 to 4 frames.	Dataset Construction, Testing, HOG, MultiFtr, FtrMine, LatSym, HikSym
8	Piotr Dollar, Zhuowen Tu, Pietro Perona, Serge Belongie	Integral Channel Features	1. naturally integrate heterogeneous sources of information, 2. have few parameters and are insensitive to exact parameter settings, 3. allow for more accurate spatial localization during detection, and 4. result in fast detectors when coupled with cascade classifiers.	Achieving 87.2% detection performance at the reference point of 10−4 fppw. This is similar but slightly inferior to the 89% achieved by the HOG classifier with optimized parameter settings. The number of bins for the histogram has little effect once at least 6 orientations were used.	Linear transformation, histograms , gradients.
9	Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, Aude Oliva	Learning Deep Features for Scene Recognition Using Places Database	The Places Dataset exhibits high density and diversity.	Increased size may involve greater time for CNN to learn how to identify objects effectively.	Places Dataset Construction, SUN & ImageNet Dataset, Comparison of Datasets
10	Jaechul Kim, Kristen Grauman	Shape Sharing for Object Segmentation	it assumes no class specific training and thus enhances segmentation even for unfamiliar categories	Shape sharing often occurs among semantically close categories (e.g., among animals or vehicles) as well as semantically disparate classes (e.g., bottle and person	Histogram, merging, segmentation

11	Maxime Oquab, L´eon Bottou, Ivan Laptev, Josef Sivic	Weakly supervised object recognition with convolutional neural networks	1. Does not rely on detailed object annotation and yet returns 86.3% mAP on the Pascal VOC classification task. 2. The network produces maps that clearly localize the objects in cluttered scenes.	Augmenting the training set with fully labelled examples brings no benefit and instead seems to slightly decrease the performance	stochastic gradient descent training, global max pooling , zero padding,
12	Jifeng Dai, Kaiming He, Jian Sun	Convolutional Feature Masking for Joint Object and Stuff Segmentation	Convolutional feature maps only need to be computed once. The convolutional features are computed from the unmasked image, their quality is not impacted.	The masks on the image content can lead to artificial boundaries. These boundaries do not exhibit on the samples during the network pre-training . This issue may degrade the quality of the extracted segment feature.	Convolutional feature masking (Feature map based) ,
13	Yunchao Wei, Wei Xia, Junshi Huang, Bingbing Ni, Jian Dong, Yao Zhao, Senior Member, IEEE Shuicheng Yan,	CNN: Single-label to Multi-label	No ground-truth bounding box information is required for training; The whole HCP infrastructure is robust to possibly noisy and/or redundant hypotheses; No explicit hypothesis label is required; The shared CNN may be well pre-trained with a large-scale single-label image dataset, e.g. ImageNet; and 5) it may naturally output multi-label prediction results.	Since the proposed HCP is independent of the ground-truth bounding box, no object location information can be used for training. Bounding box annotation is quite costly. Therefore, approaches requiring ground-truth bounding boxes cannot be transferred to the datasets without such annotation	Mean of an image, regression

14	Yunchao Wei, Wei Xia, Junshi Huang, Bingbing Ni, Jian Dong, Yao Zhao, Senior Member, IEEE Shuicheng Yan,	CNN: Single-label to Multi-label	No ground-truth bounding box information is required for training; The whole HCP infrastructure is robust to possibly noisy and/or redundant hypotheses; No explicit hypothesis label is required; The shared CNN may be well pre-trained with a large-scale single-label image dataset, e.g. ImageNet; and 5) it may naturally output multi-label prediction results.	Since the proposed HCP is independent of the ground-truth bounding box, no object location information can be used for training. Bounding box annotation is quite costly. Therefore, approaches requiring ground-truth bounding boxes cannot be transferred to the datasets without such annotation	Mean of an image, regression
15	Qin-Qin Taoa, Shu Zhana,n, Xiao-Hong Lia, Toru Kuriharab	Robust face detection using local CNN and SVM based on kernel combination	Based on the local classifier, it imposed a global regularizer across local regions to avoid these local models from overfitting into locality-sensitive. Classification task is made easier and can measure the detailed and rough similarity comprehensively through the combination of global and local kernels used in SVM.	Setting the threshold too high may cause too many positive examples be rejected, reducing the overall detection rate, and setting it too low may lead to too many pass-through patches that need to be classified further by the SVM, slowing down the overall detection process.	Convolution, averaging , non-linear transformation, Gaussian kernel, correlation
16	Ye Yuan, Shijian Tang	Object Detection based on Convolutional Neural Network	An extreme improvement on the accuracy of image classification in ImageNet Large Scale Visual Recognition Challenge (ILSVRC). 2. Adopt the CNNs to solve the detection problem and try to improve the existing model such as rCNN.	Ignores the difficult objects. Worst case the total runtime for pass all the test images are roughly 445.68 hours.	Segmentation, CNN, non-maximum suppression algorithm,

17	Masakazu Matsugu, Katsuhiko Mori, Yusuke Mitari, Yuji Kaneda	Subject independent facial expression recognition with robust face detection using a convolutional neural network	A large number of variables allow proper use of the convolution neural network and its learning	1% false rejection rate and 6% false acceptance rate	Gradient,
18	Fenglei Wang, Jun Lei, Dan Tu, Guohui Li	Convolutional feature learning and Hybrid CNN-HMM for scene number recognition	Makes convoluted neural networks applicable to dynamic issues	It lacks context consideration	Convolution neural networks and HMMs
19	S. Behzadi, A. Ali. Alesheikh	Introducing An Agent-Based Object Recognition Operator For Proximity Analysis	The proposed method is also timesaving; it solves the problem 23% faster than the common method for the region of 792 parcels.	There is a need to find objects with the distance between D1 and D2 from the river, a specific angle must be selected for agent.	Sensing function ,
20	GeoffreyE.Hinton, Ilya Sutskever, Alex Krizhevsky	ImageNet Classification with Deep Convolutional Neural Networks	Faster, reduced overfitting in the fully-connected layers	Model needs also have lots of prior knowledge to compensate for all the data that it doesn't have. Error rate of 15.3%,	ImageNet, CNN

5. CONCLUSIONS

Humans are able to identify a large number of objects with very little effort. This is because of the human brain which works in mysterious ways due to the presence of a special type of cell called the neuron. Such neurons impart the ability to think, remember and apply learnings from past experiences into current scenarios for improved decision making. The working of the brain can be considered similar to a directed graph representation. Neural networks refers to the field of computer science engineering aimed at imitating this directed graph representation in computer science. Objects are even recognized by humans when they are barricaded from sight. This poses a great deal of challenge for computer vision systems. Countless approaches like appearance based and feature based methods to overcome such difficulties have been taken into consideration with some being discussed in the paper and have been successfully implemented. Certain genetic algorithms can work without prior knowledge of a dataset and can successfully develop recognition procedures without human intervention.

6. REFERENCES

[1]Intille, S. S., & Bobick, A. F. (1999). A framework for recognizing multi-agent action from visual evidence. AAAI/IAAI, 99, 518-525.

[2]Dai, J., He, K., & Sun, J. (2015). Convolutional feature masking for joint object and stuff segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 3992-4000).

[3]Ramadevi, Y., Sridevi, T., Poornima, B., & Kalyani, B. (2010). Segmentation and object recognition using edge detection techniques. International Journal of Computer Science & Information Technology (IJCSIT), 2(6), 153-161.

[4]Gevers, T., & Smeulders, A. W. (1999). Color-based object recognition. Pattern recognition, 32(3), 453-464.

[5]Andrianov, D. E., Eremeev, S. V., & Kuptsov, K. V. (2015). The Review of Spatial Objects Recognition Models and Algorithms. Procedia Engineering, 129, 374-379.

[6]Oji, R. (2012). An automatic algorithm for object recognition and detection based on ASIFT keypoints. arXiv preprint arXiv:1211.5829.

[7]Dollár, P., Wojek, C., Schiele, B., & Perona, P. (2009, June). Pedestrian detection: A benchmark. In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on (pp. 304-311). IEEE.

[8]Dollár, P., Tu, Z., Perona, P., & Belongie, S. (2009). Integral channel features.

[9]Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., & Oliva, A. (2014). Learning deep features for scene recognition using places database. In Advances in neural information processing systems (pp. 487-495).

[10] Kim, J., & Grauman, K. (2012, October). Shape sharing for object segmentation. In European Conference on Computer Vision (pp. 444- 458). Springer Berlin Heidelberg.

[11] Oquab, M., Bottou, L., Laptev, I., & Sivic, J. (2014). Weakly supervised object recognition with convolutional neural networks. In Proc. of NIPS.

[12] Dai,J.,He,K.,&Sun,J.(2015).Convolutionalfeaturemaskin gforjoint object and stuff segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 3992-4000).

[13] Wei, Y., Xia, W., Huang, J., Ni, B., Dong, J., Zhao, Y., & Yan, S. (2014). CNN: Single-label to multi-label. arXiv preprint arXiv:1406.5726.

[14] Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., & Oliva, A. (2014). Learning deep features for scene recognition using places database. In Advances in neural information processing systems (pp. 487-495).

[15] Tao, Q. Q., Zhan, S., Li, X. H., & Kurihara, T. (2016). Robust face detection using local CNN and SVM based on kernel combination. Neurocomputing.

[16] Tang, S., & Yuan, Y. Object Detection based on Convolutional Neural Network.

[17] Matsugu, M., Mori, K., Mitari, Y., & Kaneda, Y. (2003). Subject independent facial expression recognition with robust face detection using a convolutional neural network. Neural Networks, 16(5), 555-559

[18] Guo, Q., Wang, F., Lei, J., Tu, D., & Li, G. (2016). Convolutional feature learning and Hybrid CNN-HMM for scene number recognition.Neurocomputing, 184, 78-90.

[19] Behzadi, S., & Ali Alesheikh, A. (2013). Introducing AN Agent-Based Object Recognition Operator for Proximity Analysis. ISPRS- International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences,1(3), 91-95.

[20] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105).