Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 6, No. 1, June : 2023

MODERN ARCHITECTURES FOR CORE COMPUTER VISION ON VIDEOS: A REVIEW

Suket Jha, Assistant Professor, Department of Information Technology, L.N. Mishra College of Business Management, Muzaffarpur, Bihar

Abstract— Technology advanced significantly as a result of machine learning's advent in the field of artificial intelligence. Practitioners can now train computers to process, analyse, categorise, and forecast various data classes thanks to today's sophisticated systems, which have the capability of being programmed to function just like the human brain. Thus, the introduction of the best network with the highest performance for the aforementioned purposes has become a hot topic for scientists and researchers in the field of machine learning. This article introduces deep neural networks, image classification implementation, and computer vision science. This article describes how models created using the idea of the human brain. There is also an introduction to the creation of a Convolutional Neural Network (CNN) and its many architectures, which have demonstrated outstanding effectiveness and evaluation in object identification, face recognition, picture categorization, and localization. Additionally, the use and applications of CNNs, such as text recognition, picture processing, voice recognition, and video processing, are carefully scrutinised. A survey of the literature is done to show the value and specifics of convolutional neural networks in different applications.

Keywords— *Modern Architectures, Core Computer, Vision on Videos, Applications of CNNs.*

INTRODUCTION

Unsupervised methods, often known as machine learning, are made up of numerous separate algorithms that are able to learn fundamental correlations and features from input data and make judgements independently. Human behaviour served as the main design inspiration for these algorithms. The goal was to create a system that can "see" like humans and then decide after analysing what it has seen. The majority of the algorithms developed in the 1990s were unable to match humans in terms of recognition task efficiency and accuracy. As a result, a number of models, including artificial neural networks, have been put into practise to address issues with picture classification, decision-making, and prediction. A subset of artificial intelligence is the field of computer vision. The primary goal is to develop an artificially intelligent system that works similarly to the human brain. In order to do tasks like image classification, motion detection, 3D modelling, and localisation, this field focuses more on how the computer can automatically attain a high level of comprehension from a system input in texts, photographs, and videos. Classifying processes is one of the most crucial tasks for machine learning techniques. The majority of classifications are for photos because this is how most systems' input data is entered. Through the use of machine learning, it is possible to classify photos into several groups based on their properties and visual traits. Various approaches and algorithms have been developed to improve accuracy and address classification issues. Deep learning is a technique that was developed to address issues with picture classification and learning capacity in machine learning. Deep learning, commonly referred to as a deep neural network, is made up of numerous hidden layers that can automatically learn from and extract features from unlabeled input. In a variety of applications, including image classification, segmentation, and object detection, it has demonstrated strong performance and convincing efficacy. Performance improvements for deep learning have been suggested utilising a variety of models and architectural approaches. The most well-known design is the Convolutional Neural Network, or CNN. A subclass of deep neural networks is convolutional neural networks. Each layer in a convolutional neural network performs a certain processing task. Large corporations like Google, Facebook, and AT&T

hired countless scientists to create and introduce the best architectures that can be used in CNNs as a result of the network's astounding efficiency. The best CNN architectures and their components are reviewed in this paper. The research produced satisfactory performance for resolving issues and avoiding human error after utilising several applications.

COMPUTER VISION

The most understandable and straightforward definition of computer vision is the science that endows computers or machines with the capacity for recognition. Computer vision is thought to involve more than just recording light. Memory, recognition, appraisal, and estimation are some of the practical and functional parts of the process. The basic goal is to store and extract information and features from what has been observed. The objective of computer vision is to use the processing of digital information to give machines the ability to comprehend reality, also known as picture perception. This type of machine understanding is accomplished by taking significant data and features from digital signals and executing intricate calculations. This procedure was used for typically broken down into four sub-processes: initialization, detection, estimation, and recognition. Sensors with various attributes are involved. Each of them is broken down into a plethora of various categories and processes. It should be remembered that altering system settings and using alternative data may have an effect on effectiveness and performance. Image segmentation is one of the most practical computer vision applications, and it is currently quite sophisticated. The numerous algorithms and approaches for image segmentation can be categorised into two groups: Methods: supervised and unsupervised. For specific class images or image sets with distinct classes, the supervised method classification can be utilised. It is vital to perform a mental evaluation before implementing this strategy in order to be aware of its effectiveness and the fact that the segmentation method has done its work correctly. In this method, a person must intervene and verify the procedure's accuracy by contrasting the segmented image outputs with predetermined categories. Because of these limitations, the supervised technique appears to be an issue and is not appropriate for many computer vision applications. These constraints include classification on large datasets or a large number of classes. As a result, an unsupervised approach was developed to address these problems, however it has limitations. For real-time segmentation, an unsupervised technique is very helpful and practical. This approach also allows for the adjustment of necessary parameters for subsequent algorithms in light of evaluation findings. Heavy algorithms and complicated mathematical functions have been devised to accomplish these segmentation methods, which can automatically do the needed work with high accuracy and precision. These functions can be changed depending on the varied input forms, such as image, video, and text. Using cameras and tracking algorithms, the field of computer vision has made significant advancements in overcoming localization issues during the past 20 years. By speeding up processing and calculation as well as the memory that is currently available, numerous methods are utilised to track and recognise features, solving the issue that old computer vision algorithms had. In other words, computer vision algorithms are no longer restricted to making connections between two points but can locate the desired object in real-time utilising sophisticated algorithms on a variety of cameras with microprocessors inside. Additionally, there has been a lot of attention and work from academics in the subject of context modelling recently due to a unique purpose in computer vision, which is to strengthen and better the process of analysing images and extracting information from them. Context modeling's primary function is to introduce structure, simplify it, and help users understand how to preserve data. For people who want to apply context modelling in their work today, there are several applications available that provide and explain information such as particular methodologies, applications, and approaches.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 6, No. 1, June : 2023

IMAGE CLASSIFICATION

The process of classifying images based on their content, characteristics, and other variables that may have an impact on the classification process involves several intricate computer vision algorithms. The goal of image classification is to categorise and assign each pixel in a digital image to one of several predefined classes. Colourful data is transformed into grayscale or a certain colour level before its features are recognised and retrieved. The data can then be utilised to create feature-maps in an image after being categorised. Multi-spectrum data are frequently utilised for categorization, This is a numerical basis for classification that uses the spectral pattern of the data for each pixel. An algorithm must be used in the classification of images because it was created for a different purpose. The five primary stages of the procedure are preprocessing, feature selection and extraction, training sample selection, classification processing, and accuracy assessment.

(i) Preprocessing- Before processing, the image must go through several modifications as the initial step in classification. First, a grayscale version of the RGB image is created, then it is transformed to binary data and stored. In theory, the original image is translated to new pixels based on binary equations and divisions once the adjustments are applied. The primary goal of such a transform is to enhance the image's significant details so that they are more visible during subsequent processing.

(ii) feature Selection and Extraction- information must be extracted from picture features in order to perform the classification or detection process. Additionally, the majority of unnecessary and unimportant material ought to be ignored. Additionally, calculations should be simple to extract information from, and processing should go quickly. It should be mentioned that when cameras are taking pictures prior to the process, some crucial information is automatically lost. One of the most troublesome problems that hinders feature extraction is data loss, which prevents systems from identifying actual objects in photos.

(iii) Selection of training samples- Utilising samples gathered in accordance with the goal, the algorithm preparation for the classification task is conducted. The use of a sufficient number of samples can be inferred to be crucial and to potentially result in high efficiency. Managing training samples entails removing, adding, and categorising. Keeping in mind that a chosen training sample can cause a new road map to be activated in the feature extraction stage.

(iv) Classification processing- The model is trained using samples in the step before to categorise characteristics. Electronic devices typically resulted in noisy samples that could impact the pixels that make up the classification maps in conventional image classification approaches. A procedure based on replies from distant spectrometry is introduced, and it uses a lot of filters to reduce noise. Sensors or digital cameras frequently produce noise during the saving of images, which is defined as a random variation in colour information or lightening pixels.

(v) Accuracy assessment- Two groups of datasets the training dataset and the testing dataset perform each image classification task. The testing dataset is used to assess and gauge the classification accuracy, whereas the training dataset is utilised to train the algorithm for classification. A crucial component of every categorization is the accuracy assessment. It compares the categorized image with a testing dataset that is considered as correct labeled data. This common practise creates a square error matrix, where the rows and columns show how the algorithm performed.

Methods Strengths Weaknesses	Methods Strengths Weaknesses	Methods Strengths Weaknesses
Support Vector Machine	Elude from Overfitting Provide Unique Explanation More Efficiency	Slow Implementation Very Intricate Algorithm
Artificial	Applicable for Big Data	Learn Slowly

Neural Networ	Practicable On Noisy Samples	Requires Powerful Systems
Decision Tree	Simple Understanding Not Much Knowledge Needed	High Loss Value Confusing Splits

Table 1: Comparison of strengths and weaknesses of SVM, ANN

DEEP LEARNING

A deep learning algorithm uses numerous layers to separate out higher-level information from unprocessed input stages. In order to extract the most evident features from an image input into a deep neural network, for example, the edges are first detected in the first layers, and then deeper features are found in the higher layers. Deep learning, one of the most significant and useful machine learning algorithms now available, has achieved great success in a variety of applications, including image analysis, speech recognition, and text recognition By extracting features from input photos using two strategies—supervised and unsupervised with a distinctive architectural characteristic—this technique teaches how to recognise and categorise items. The origins of deep learning may be traced back to 1943, when Walter Pitz and Warren McCulloch created a computer model based on neural networks seen in the human brain. For many reasons, including the unavailability of cutting-edge hardware and software, this approach was not regarded to be useful until recently. Instead, they employed a combination of algorithms and mathematics known as "threshold logic" to simulate the human thought process. Deep learning made a comeback in the industry in 2006 and has since become a popular subject for study. DL has shown a high-level efficiency in various fields such as Image Classification, Object Detection, Video Processing, Natural Language Processing, Speech Recognition. Given that each application has its algorithm and processing method, various models and algorithms of deep learning have been introduced over the past few years. The most important introduced models that have improved performance and reduced the problems of features extracting are:

(i) Stacked Autoencoder (SAE)- A stacked autoencoder is the simplest deep learning model, being a subset of the unsupervised method. SAEs are usually designed from multiple scattered layers, each composed of several auto-encoders. In this architecture, each layer input is the output of the previous layer. SAEs solve classification problems by placing several automatic encoders consisting of two main steps: Encryption, Decryption. Auto-encoders typically use backpropagation to change and reduce the size of weightlift inputs, which somehow prepares the input values for better feature extraction. In figure 1, an overview of SAE architecture can be observed.

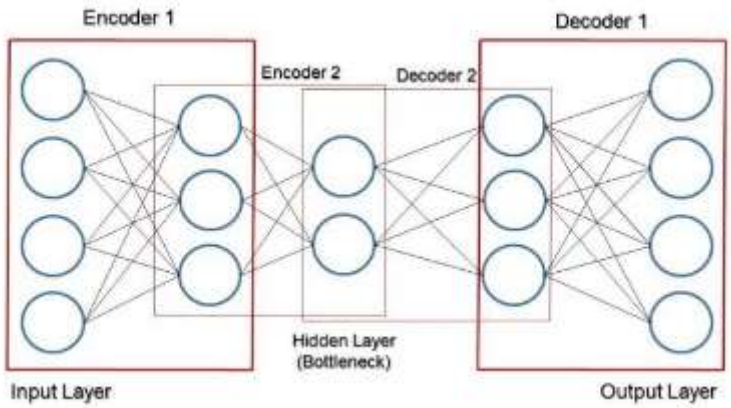

Figure 1: A simple view of stacked auto-encounter architecture

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 6, No. 1, June : 2023

(ii) Deep Belief Network (DBN)- The first deep learning model is the deep belief network, which could get fully trained. DBN and SAE‟s difference structure is that DBN comprises several restricted Boltzmann machines that include two visible (V) and hidden (H) layers. The restricted Boltzmann machine uses Gibbs sampling to train its parameters. Restricted Boltzmann (RBM) uses conditional probability P (h | v) to calculate the value of each unit in the hidden layer and then the conditional probability p (h | v) to calculate the value of each unit in the visible layer. This process is repeated until the model is fully trained. Figure 2 illustrates a deep belief network architecture. As can be implied, this model has a more complex architecture and calculations that slow down the process.

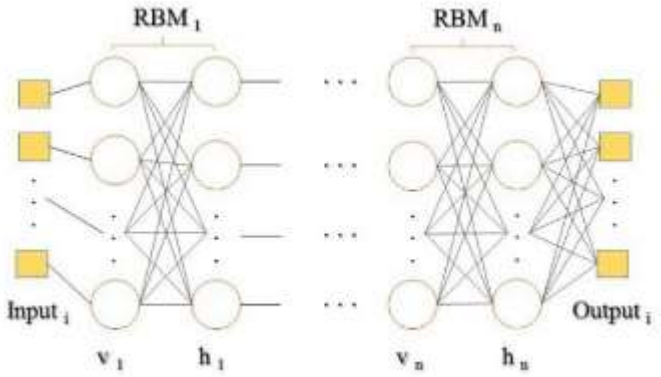

Figure 2: A deep belief network architecture

(iii) Convolutional Neural Network (CNN)- A convolutional neural network (CNN) is a special type of artificial neural network and an important subset of the supervised method. CNN uses machine learning algorithms to analyze data and extract features. CNN architecture is designed to mimic the human brain neuron pattern. Its Layers are divided into a three dimensional structure in which each set of neurons in layers analyzes a specific area of the image.

(iv) Recurrent Neural Network (RNN)- This model has been introduced due to the lack of previous models‟ ability to extract serial data‟s features. The problem was they failed to learn and extract features from texture inputs. A recurrent neural network (RNN) learns the features and information from the continuous data stored in the neural network‟s internal state‟s predefined memory input. Its purpose is like predicting the continuation of a sentence or extracting its meaning and key features. As it is known, each word is related to other words in a sentence; hence, one or more previous words must be taken to account for the model to extract features.

LITERATURE REVIEW

Many hidden brain neurons that are able to recognise and categorise specific features make up a convolutional neural network. Without changing the features of the photos, this approach shrinks the input size. Neurons are capable of extracting characteristics from speech patterns to the edge of a picture. Many convolutional cores make up the CNN structure, which helps the system identify crucial picture elements by emphasising them. This is done in order to prevent wasting time processing numerous pointless pixels that are present in each image but contain no valuable information. A Convolutional neural network has numerous neural layers, and by adding additional layers, it could extract more high-level characteristics. The simplest configuration of a convolutional neural network typically consists of two convolutional layers: the Conv layer and two sub-sampling layers that carry out the sampling function of the fully connected layer. Input is first entered into the

first hidden layer, which is made up of three filters. Three maps of the features are created after processing and are weighted. The sub-sampling layer's nonlinear activation function is then used to create a new feature map. Following their transfer to three trained filters from the Conv layer, these maps are then retrieved through the subsequent sub-sampling layer, yielding three new maps. The output of the second sub-sampling layer is passed to a fully connected layer, which then transforms it into a vectorized coordinate. The output of the fully connected layer is then inserted into the neural network used in the learning process. The model is more useful because of its three-dimensional structure. Due to the fact that CNN facilitates processing by modifying and reducing sample sizes, it can also be employed with huge datasets. They are principally employed to recognise scales, variations in space, and to categorise two-dimensional visual pictures. Additionally, because the neurons in the map with related features have the same weight, the network can carry out the learning process in parallel. The particular advantages of the convolutional neural network for speech recognition and image processing come from the local weight distribution's peculiar shape. The weight distribution simplifies the network and makes it much simpler to extract and classify features.

Cycle Improvement Year Innovation	Cycle Improvement Year Innovation	Cycle Improvement Year Innovation	Cycle Improvement Year Innovation
1940 -1979	Beginning of NN	1943 1949 1962	Evaluate neural function with predictive logic. Proposal of cellular interpretation theory. Recording Cat‟s neurons electrical activity to achieve pattern functions
1980 -1998	Creation of CNN	1980 1989	Inventing a self-learning neural network that could represent basic geometrics. Utilization backpropagation CNN for the actual application
1999 -2010	Development of CNN	1999 2006	Proposal of Max-Pooling. Presentation Max-Pooling for CNN
2011 –2015	Merging GPUs with CNN	2011 2012 2013 2014 2015	Training a CNN model with GPU for the first time. Proposal of Dropout technique by Google

			researchers. Proposal of Drop-Connect for CNN. Presentation of many more useful architectures like VGG/RCNN. Releasing different open-source libraries for CNN by Google.
2016 -2020	Introduction of advanced CNN‟s Architecture	2016 2017 2018	Improvement CNN for real-time classification by Introducing Yolo/SSD. Introduction of upgraded models for getting more performance. Pre-training language models

Table 2: Creation and progress of convolutional neural networks

ARCHITECTURE

Since its creation, the convolutional neural network has undergone a lot of adjustments to enhance performance and address many problems, including feature extraction. Convolutional layers, Pooling Layers, and Fully Connected Layers are the three primary neural layers that make up the CNN architecture today, with all the adjustments and advancements. These layers perform a significant amount of numerical processing operations. However, regulatory methods like dropout or batch normalisation are occasionally applied to network settings, which enhance the neural network's functionality. Additionally, there are two stages used to complete the network training: forward and backward steps. When going into each layer, the forward step strives to retain parameters like input weight and value. Then, data loss is calculated using output prediction. Chain rules calculations determine the gradient of each parameter in the backward step based on the loss estimated in the first step, updating parameters for the following forward step. Up until the network is completely trained, these two processes are repeated. concept of CNN architecture is shown in Figure 3.

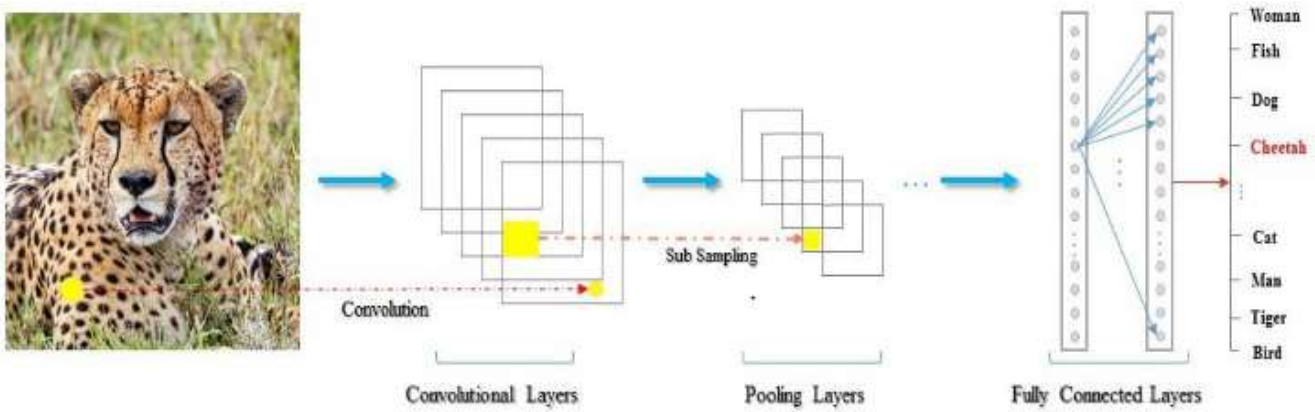

Figure 3: The concept of convolutional neural network architecture

CONCLUSION

Human behaviour serves as the foundation for the unsupervised machine learning process. Without explicit instructions or direct human involvement, the computer can independently learn fundamental relationships and features from input data. For specific tasks like image classification, decision-making, and prediction, primary ML algorithms have been substituted by a number of models, such as an artificial neural network. The failure of previous models to perform recognition tasks as accurately and efficiently as humans led to this progression. The primary goal of computer vision is to develop an artificially intelligent system that works similarly to the human brain. More emphasis is placed on how a computer can autonomously develop a high level of understanding without any assistance from humans. Deep learning techniques and artificial neural networks have been introduced to increase classification accuracy. In a variety of applications, including object identification, segmentation, and picture classification, they have demonstrated strong performance. Convolutional Neural Network is the most well-known and practical artificial neural network type. CNNs belong to the Deep Neural Network class and have an architecture with multiple layers, each of which performs a particular processing task. The CNN models have been widely used in several methodologies, such as Alzheimer's disease diagnosis. The primary focus in the future might be on improving algorithms for identifying specifics of these illnesses and perhaps forecasting how they will manifest in human brains.

REFERENCES

[1]. Sergey Bartunov, Adam Santoro, Blake A. Richards, Luke Marris, Geoffrey E. Hinton, and Timothy P. Lillicrap. Assessing the Scalability of Biologically-Motivated Deep Learning Algorithms and Architectures. In Proceedings of the 32nd International Conference on Neural Information Processing Systems, NIPS'18, pages 9390–9400, Red Hook, NY, USA, Dec. 2018. Curran Associates Inc. 4.

[2]. Ruha Benjamin. Race after Technology: Abolitionist Tools for the New Jim Code. Polity, Medford, MA, 2019. 2

[3]. Abeba Birhane and Vinay Uday Prabhu. Large Image Datasets: A Pyrrhic Win for Computer Vision? In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 1537–1547, 2021.

[4]. Rupert Brown, Susan Condor, Audrey Mathews, Gillian Wade, and Jennifer Williams. Explaining intergroup differentiation in an industrial organization. Journal of Occupational Psychology, 59(4):273–286, 1986. 3

[5]. my Bruckman. Studying the amateur artist: A perspective on disguising data collected in human subjects research on the Internet. Ethics and Information Technology, 4(3):217–231, Sept. 2002. 3

[6]. Joy Buolamwini and Timnit Gebru. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In Conference on Fairness, Accountability and Transparency, pages 77–91. PMLR, Jan. 2018.

[7]. T. B. Moeslund and E. Granum, "A survey of computer vision-based human motion capture," Comput. Vis. image Underst., vol. 81, no. 3, pp. 231–268, 2001.

[8]. H. Zhang, J. E. Fritts, and S. A. Goldman, "Image segmentation evaluation: A survey of unsupervised methods," Comput. Vis. image Underst., vol. 110, no. 2, pp. 260–280, 2008.

[9]. R. J. Radke, "A survey of distributed computer vision algorithms," in Handbook of Ambient Intelligence and Smart Environments, Springer, 2010, pp. 35–55.

[10]. O. Marques, E. Barenholtz, and V. Charvillat, "Context modeling in computer vision: techniques, implications, and applications," Multimed. Tools Appl., vol. 51, no. 1, pp. 303–339, 2011.
[11]. X. Yang, S. Sarraf, and N. Zhang, "Deep learning-based framework for Autism functional MRI image classification," J. Ark. Acad. Sci., vol. 72, no. 1, pp. 47–52, 2018.
[12]. J. C. S. J. Junior, S. R. Musse, and C. R. Jung, "Crowd analysis using computer vision techniques," IEEE Signal Process. Mag., vol. 27, no. 5, pp. 66–77, 2010.
[13]. D. Lu and Q. Weng, "A survey of image classification methods and techniques for improving classification performance," Int. J. Remote Sens., vol. 28, no. 5, pp. 823–870, 2007.
[14]. S. Sarraf, "Binary Image Segmentation Using Classification Methods: Support Vector Machines, Artificial Neural Networks and K th Nearest Neighbours," Int. J. Comput., vol. 24, no. 1, pp. 56–79, 2017.
[15]. P. Babaniamansour, M. Ebrahimian-Hosseinabadi, and A. Zargar-Kharazi, "Designing an optimized novel femoral stem," J. Med. Signals Sens., vol. 7, no. 3, p. 170, 2017.
[16]. S. Sarraf, "Hair color classification in face recognition using machine learning algorithms," Am. Sci. Res. J. Eng. Technol. Sci., vol. 26, no. 3, pp. 317–334, 2016.
[17]. A. Sarraf, "Binary Image Classification Through an Optimal Topology for Convolutional Neural Networks," Am. Sci. Res. J. Eng. Technol. Sci., vol. 68, no. 1, pp. 181–192, 2020.
[18]. S. Sarraf, "French Word Recognition Through a Quick Survey on Recurrent Neural Networks Using Long-Short Term Memory RNN-LSTM," Am. Sci. Res. J. Eng. Technol. Sci., vol. 39, no. 1, pp. 250–267, 2018.
[19]. D. Tuia, M. Volpi, L. Copa, M. Kanevski, and J. Munoz-Mari, "A survey of active learning algorithms for supervised remote sensing image classification," IEEE J. Sel. Top. Signal Process., vol. 5, no. 3, pp. 606–617, 2011.