Object Detection from Video Sequences Using Deep Learning: An Overview

Dweepna Garg and Ketan Kotecha

Abstract One of the challenging topics in the field of computer vision is the detection of the stationary/non-stationary objects from a video sequence. The outcome of detection, tracking, and learning must be free from ambiguity. For effectively detecting the moving object, first the background information from the video should be subtracted. However, in the high-definition video, modeling techniques suffer from high computation and memory cost which may lead to a decrease in performance measure such as accuracy and efficiency in identifying the object accurately. It is important to identify the definite structure from a large amount of unstructured data which is a prerequisite problem to be solved. The task of finding the structure from a large amount of data is achieved using Deep Learning 'which is about learning multiple levels of representation and abstraction that help to make sense of data such as images, sound, and text'. The purpose of the paper is to survey the method with which the objects can be efficiently detected from any given video sequence along with the preferable use of the deep learning library.

Keywords Deep learning · Object detection · Video sequence · Graphics processing unit · Tensor flow

1 Introduction

'Learning'—an eight letter word means that one understands something that 'we' have understood all our life, but in a unique manner. 'Machine' when combined with 'Learning', refers to mainly three terms: task, performance, and experience. It can be framed as the performance that can be measured for a task and this performance

D. Garg (✉) · K. Kotecha
Parul University, Limda, Vadodara, India
e-mail: dweeps1989@gmail.com

K. Kotecha
e-mail: provost@paruluniversity.ac.in

R.K. Choudhary et al. (eds.), *Advanced Computing and Communication Technologies*, Advances in Intelligent Systems and Computing 562, https://doi.org/10.1007/978-981-10-4603-2_14

measure can be improved with some experience. This concept is referred to as Machine learning [1], a term coined by Arthur Samuel which can be stated as a way to make the computer intelligent. He stated in way back 1959 that machine learning is the field of study that gives the computers the ability to learn without being explicitly programmed. In 1997, Mitchell [2] defined machine learning as 'a computer program that could learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E.' Machine learning aims to bring a degree of order to the zoo of machine learning problems which is spread to a vast range of applications. The emerging area of Machine Learning research is Deep Learning which is introduced with the motive of moving the machine learning a step closer to artificial intelligence. It is becoming popular mainly because of the following three reasons: First 'drastically increased chip processing abilities', e.g., General Purpose Graphics Processing Units (GPGPUs), second due to its usage of an increased size of data used for training, and third because of the recent advances in machine learning and processing information. 'The above reasons have made the deep learning methods to compute various complex problems and to effectively make use of both unlabelled and labelled data' [3].

The paper attempts to provide an overview of detecting either non stationary or stationary objects accurately from the given video sequence using this emerging machine learning methodology. Furthermore, the purpose of the paper is limited in addressing to the research issues which can be targeted to do the above mentioned methodology. The rest of the paper is organized as follows—Sect. 2 deals with the basic structure. Section 3 targets the research issues mainly focusing on detecting the objects. A brief review of deep learning is presented in Sect. 4. Section 5 highlights the main topic of this paper giving an insight of how the work can be carried out. Section 6 describes the deep learning frameworks useful from programming point of view, Sect. 7 of the paper focuses on the applications of deep learning. Finally conclusions are drawn in Sect. 8.

2 The Basic Structure

In Fig. 1, the set of training examples are fed as an input to the learning algorithm. The choice of the learning example depends solely on the user. Hypothesis (h) is calculated taking into consideration the size of the house and the learning algorithm. Size of the house is one of the input features considered as an example. The formula for calculating the hypothesis is

$$h(X) = \sum_{i=1}^{n} \theta i X i \tag{1}$$

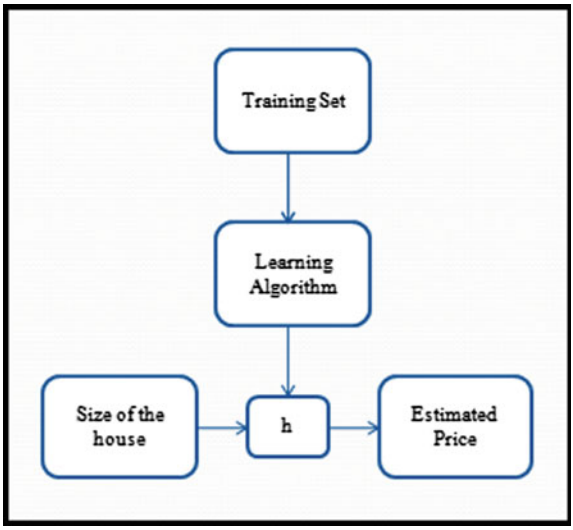

Fig. 1 Basic structure of learning algorithm

Here n denotes the number of training examples, θ is the parameter, and X is the input feature. In the above example the size of the house is one of the input features. Then the value of estimated price of the house is calculated. A model is prepared through a training process where it is required to make the predictions and is corrected when those predictions are wrong. The training process continues until the model achieves a desired level of accuracy. It is calculated using $h(X) - y$. Here $h(X)$ indicates the calculated hypothesis value and y indicates the corrected output value.

3 Research Issues

Most of the machine learning problems is arising during its applications to real-world problems. Poor performance was observed not because of the choice of learning algorithm but due to the selected training data. The problems faced by training data was—the training data selected was insufficient, training set of data was too small to learn a generalize model or the data contained noise. Due to an increased amount of data, the performance measure (for example, accuracy) was one of the main issues in machine learning. The time taken to process large amount of data was significantly high. To mimic the brain for representing information with great efficiency and robustness is a core challenge in the research of artificial intelligence. Around in 2010, the real impact of machine learning in industry began in large-scale speech recognition [3]. Artificial neural networks were involved in major of successful deep learning methods. In the context of computer vision,

object detection, identification of the image of the object in any image or in a video sequence is one of the most important fields which. It is discussed in detail later in the paper.

4 Deep Learning

It is also known as deep structured learning or hierarchical learning or deep machine learning. Geoffrey Hinton coined the term 'deep learning'. It is a large family consisting of both supervised and unsupervised learning methods. The computational models are made up of multiple processing layers and each layer learns the representation of data at different abstraction level. Artificial neural network is largely inspired by the way the human brain works. And today a new field has emerged with respect to the research of neural network and it is Deep Learning. Deep learning in collaboration with other algorithms helps in classification [4], clustering, and prediction [5]. It first reads the structure automatically and then when the deep learning algorithm trains it is able to make the guess against the training set and try to bring the guess close to the accurate answer. This is referred to as optimization [6]. It is able to learn the complicated patterns from a large amount of data by combining the computational power and special types of neural network. The word 'deep' can be used to call the network having multiple hidden layers. Adding more layers add to the difficulty in updating weights as the signals [3] propagating becomes weak. Therefore, the weight of the network becomes off the track and it becomes impossible to parameterize a 'deep' neural network with backpropagation. Here 'deep learning' comes into play which helps to train the 'deep' structures of neural network.

With deep learning one can classify, cluster, and predict about the data consisting of video, images, text, time series, and many more. It is mainly used in object detection (face detection, pedestrian detection), visual recognition of objects (ImageNet), speech recognition (conversion from speech to words), and predicting the drug activity. Deep learning [7] has already made a successful advancement in voice search on smart phones and text-to-speech conversion and speech-to-speech conversion. GPU's (graphics processing units) have greatly contributed in image segmentation, object detection, pattern recognition, and natural language processing. GPU's train the deep neural network using large training sets but in a less time. Its computing provides the efficient and parallel computation of large amount of training data. As the GPU takes less time to compute the large amount of data and works effectively well for machine learning, it is believed that the graphics processing unit can accelerate the machine learning algorithms very well. The reason for widely using GPU is its computational power and its capabilities which are growing at a faster rate as compared to CPU. The deadly combination of CPU and GPU can produce the best value in terms of computation power, performance, and price.

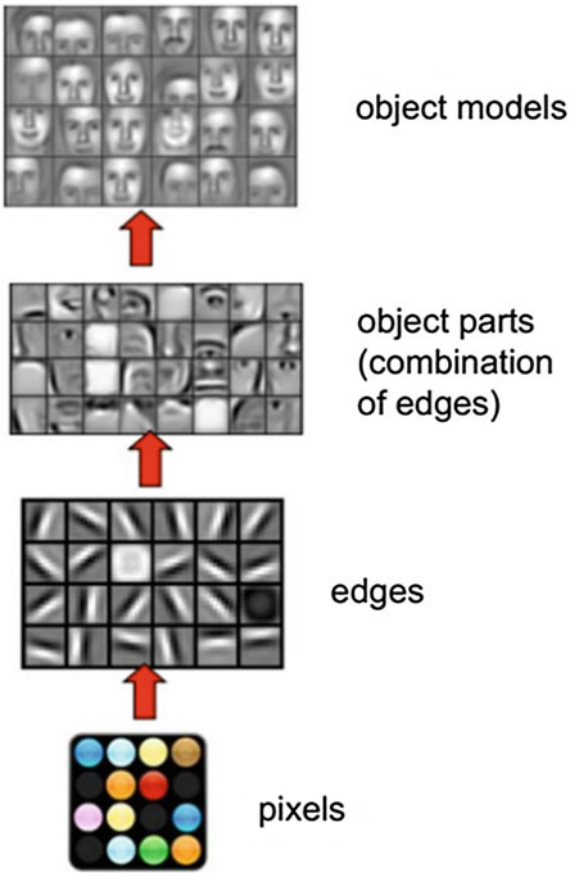

Fig. 2 Flow of backpropogation algorithm [23]

Deep Learning uses backpropagation algorithm in order to depict how the internal parameters are changed by the machine in order to compute the representation in each layer with respect to its previous layer. Initially the pixels are converted to the edges, combination of edges makes the object parts and finally the object model is prepared as depicted in Fig. 2.

Convolutional neural networks have found to succeed in a variety of tasks of computer vision especially focusing on the recognition. The concept of optical flow [8] was introduced keeping in mind the relative motion of both the viewer and the object. It highlights the important information about the spatial arrangement of the viewed object along with its displacement. It is regarded as one of the vital ingredient in solving many complex computer vision tasks. The main focus of estimating optical flow is to find per-pixel localization and then mapping the difference between the two images which are fed as inputs to the system. For this, the feature representations are to be learnt. By using various optical flow algorithms, the difference between the frames can be calculated accurately. Evaluating these

algorithms does consume time but it has shown great improvement in training the deep convolutional neural network. Computation of optical flow cannot be done for a single point in the image. Rather, the neighboring point should also be considered as each image point has two velocity fields which may change due to the change in any one of the field.

Optical flow [9] estimation has been one where CNN have found to succeed. The motion of any object is perceived when there is a change in the picture. A moving object may produce a constant pattern in brightness. In order to avoid the variations in brightness, the shading surface is considered to be flat and the illumination where the light falls is assumed to be uniform. These conditions help in determining the brightness of the image to be differentiated. Hence the motion of such patterns is determined by viewing the motion of the corresponding points [8]. Calculating the velocities of such points on the object is what known as the optical flow.

Figure 3 illustrates that the information first gets spatially compressed in a contractive part of the network and then it gets refined in the expanding part. Hence, it can be said that the 'architecture is trained from end-to-end'. The main idea behind is to exploit the capability of CNN of learning the powerful features at multiple levels of abstraction. It helps in determining the actual difference in the input images based on the features. A huge amount of training dataset is needed to predict the optical flow in training a network. Generation of optical flow from a video sequence is a difficult job and hence a popular synthetic dataset named Flying Chairs dataset [10] has been used. This dataset consists of random background images from Flicker which consists of the image segment of chairs. It is used in training the CNN. For faster processing, implementation on GPU is preferred over CPU.

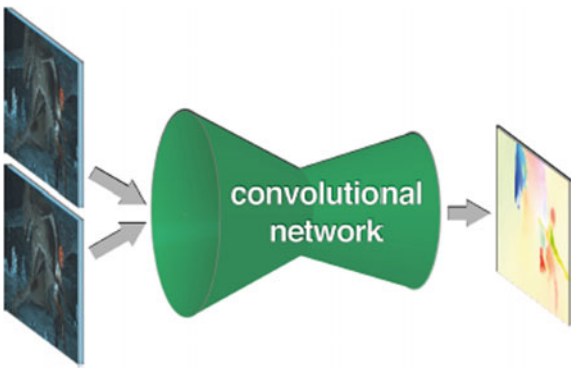

Fig. 3 End-to-end supervised learning of optical flow [24]

5 Object Detection with Deep Learning

It is easier for the humans to detect and recognize the object in the image irrespective of its different viewpoint. Vision begins with the eyes and truly takes place in the brain. So, such cameras are required which can see and understand what humans can see and understand. Cameras take pictures by converting light in 2D pictures. But these are lifeless numbers. They do not carry meaning in themselves. To take pictures does not mean to see. And to see means to understand. For this, the first thing is make the computer understand the object. If the machine is able to detect the objects same as the humans do with great accuracy, then it can lead to wonders in the real world. Another important issue related to object detection is to identify the moving object. Moving object is mainly composed of feature representation [11] and statistical modeling of the scene around. The feature representation takes into consideration the predefined features such as the color code, the texture features, the shape of the feature, and the background against which the object is positioned. The represented features are first learnt using the technique of feature learning where a transformation is carried out from such raw information to a successful representation of an object. It is necessary to discover the required features from the raw input data. Feature learning is also known as the representation learning and is mainly divided in supervised and unsupervised learning [12]. Supervised learning helps in learning the features with the labeled input data. Examples: multi-layer perceptron, neural network. Whereas unsupervised learning the features are learned with the unlabeled input data. Examples include clustering, autoencoders. A broader family of machine learning method named as deep learning is based on learning the representations of the data. *Research can be carried out to effectively represent the feature from a large amount of unstructured data.*

Statistical modeling technique mainly focuses on detecting and tracking [13] the foreground objects (such as car, humans, etc.) from a video sequence by subtracting the background image. *In order to model the foreground image—accuracy in detecting the object, noise and automatic choice of the relevant parts of the object are the key areas of research.* Background subtraction [14] is also known as foreground detection. It is widely used technique for detecting the moving objects [15] from the static cameras. Feature extraction [11], handling deformation and occlusion handling, and classification are four major components in object detection.

After preprocessing the image, the localization of object is required which makes use of this technique. It computes the difference between the current frame and the reference frame, referred to as the background image. It is carried out if the required image is in the video stream. Background subtraction is used in the areas of surveillance and learning of animals, tracking and learning of a specific car on road, human pose estimation, etc. The flow of background subtraction in Fig. 4.

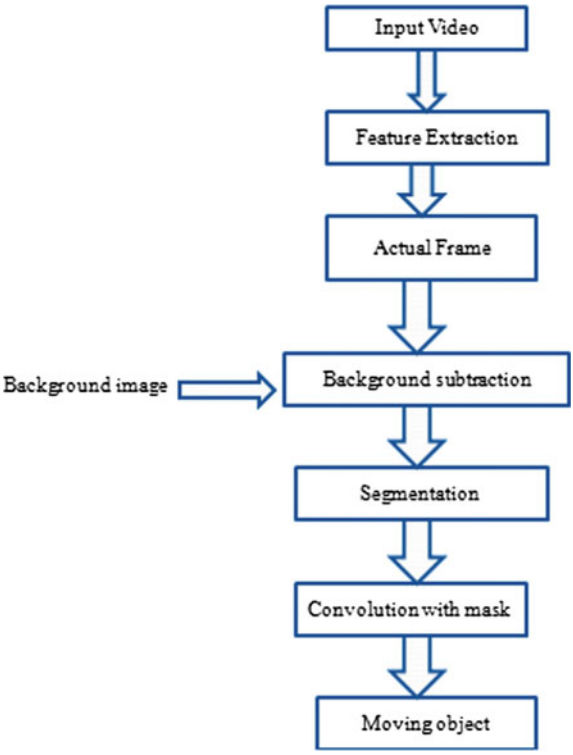

Fig. 4 Background subtraction

The parameters of study for detecting the moving object from a video sequence can be the accuracy (how accurately the object is identified), efficiency in terms of performance measure and the computational complexity of finding out the desired object.

Figure 5 illustrates the widely used example of object detection from a video sequence where in the vehicles can be monitored, detected and tracked.

The challenges in background subtraction arise with the outdoor scenes, i.e., with respect to rain, wind, different viewpoint of the object or illumination change [16]. In order to detect the object in a better way, it is mandatory to first understand the image. Understanding [17] the image is done using Deep Convolution Network. In this, the inputs are small portions of the image which is fed to the lowest layer in the hierarchical structure. In each layer where information propagates, digital filtering is carried out in order to get the most noticeable features of the observed data. The neurons, also known as the processing units process the features providing a level of invariance to shift, scale, and rotation. Nice megapixels cameras have been developed but yet there is not given vision to the blind.

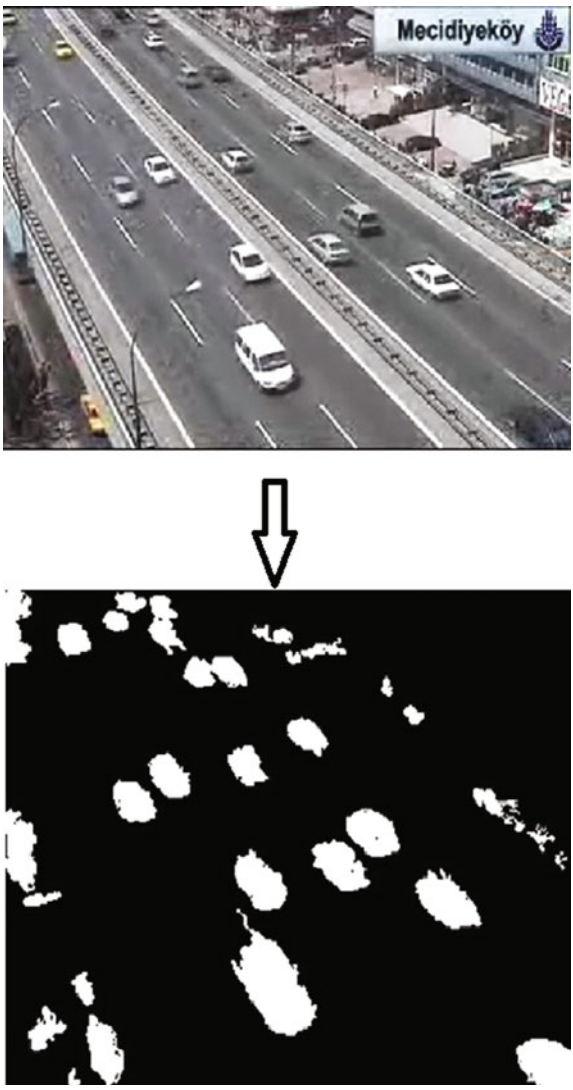

Fig. 5 Example depicting vehicle detection [25]

6 Deep Learning Frameworks

Some popular frameworks of deep learning are as follows.

Caffe: The earliest known mainstream deep learning toolkit was developed by Berkeley Vision and Learning Center (BVLC) and by community contributors [5]. Caffe has worked well with innovation and application. Caffe can process over 60M images per day with a single NVIDIA K40 GPU [18]. It is one of the best choices with respect to deployment because of its cross-platform nature. It is C++ based and

has the ability to be complied on variety of devices. Caffe works well with Convolution Neural Network principles and parts. Mainly the steps to train a CNN with Caffe are: (i) The first step involves data preparation wherein the images are preprocessed and are stored in the format which can be used by Caffe, (ii) The second step moves with model definition where CNN architecture is chosen and the parameters are defined in a configuration file, (iii) Then comes the solver definition where solver holds the responsibility of optimizing the model. The parameters of the same are stored in the configuration file, (iv) The final step is model training wherein the model is trained by executing a Caffe command from the terminal. The trained model is then stored in the file.

Theano: Developed by LISA group (now MILA), run by Yoshua Bengio at the University of Montreal [19] is a framework developed by keeping in mind the deep learning algorithms. Theano compiles the program written in Python to efficiently run on GPUs or CPUs. More than a deep learning library, it is a research platform. Theano works well with deep convolutional network, deep belief network and stacked denoising autoencoders.

Torch: It was developed by Ronan Collobert, Clement Farabet, and Koray Kavu Cuoglu for research and development in deep learning algorithms. It was promoted by the CILVR lab at New York. Facebook AI lab, Twitter and Google Deepmind later used and further developed this deep learning library. Torch makes use of C/C++ libraries as well as CUDA for GPUs [20].

TensorFLow: An open source software library used for machine intelligence is **TensorFlow** [2]. The data flow graphs are used for numerical computation. The mathematical operations are represented by the nodes of the graph and the edges represent the tensors (the multi-dimensional data arrays) to communicate between the nodes. The flexible architecture of TensorFlow allows the users to deploy the computation to one or more central processing unit or the graphics processing unit in a server or desktop or any cellular device with a single API. TensorFlow was originally developed by the engineers working in Google Machine Intelligence Research organization for conducting the research on deep neural networks and machine learning. The system claims the developers is ‘general enough to be applicable in a wide variety of other domains as well’.

7 Applications

The machine learning technologies are being widely used, e.g., in marketing, finance, telecommunications, and network analysis. One of the well-known examples of machine learning is the concept of web page ranking [21], where the search engine finds the web pages relevant to the query given by the user according to the order of relevance. Another such application is collaborative filtering [5], where past purchases and viewing decisions can be used to attract the users to purchase additional goods. Amazon, as an example uses this technology to entice the users to have additional goods. Other applications are in speech

recognition [1], handwriting recognition [7], face detection [11], face recognition [22], automated driving [2], fraud detection [21], text-based sentiment analysis, email spam filtering [21], network intrusion detection, anomaly detection, classification, signal diagnosis [17], weather forecasting, anti-virus software, anti-spam software [1], genetics, image classification [14] etc. So, by using Machine Learning, the guess work can be avoided as well as time to solve a problem can be reduced thereby providing good guarantees for the solutions.

8 Conclusion

The techniques developed from deep learning research have already been 'impacting a wide range of signal and information processing work within the traditional and the new'. This emerging area of research in Machine Learning is expected to move it closer to Artificial Intelligence.

References

1. Machine Learning.: http://www.mlplatform.nl/what-is-machine-learning. Accessed 01 Jan 2016
2. Mitchell, T.M.: Machine Learning, p. 421. McGraw-Hill Science/Engineering/Math (1997)
3. Deng, L., Yu, D.: Deep learning methods and applications. Found. Trends Sign. Process. **7**(3–4), 197–387 (2014) [Now Publishers Inc. Hanover, MA, USA]
4. Jang, H., Yang, H.-J., Jeong, D.-S.: Object classification using CNN for video traffic detection system. In: 21st Korea-Japan Joint Workshop on Frontiers of Computer Vision (FCV), pp. 1–4. IEEE (2015)
5. Collaborative Filtering.: http://benanne.github.io/2014/08/05/spotify-cnns.html. Accessed 20 Sept 2016
6. Le, Q.V., Ngiam, J., Coates, A., Lahiri, A., Prochnow, B., Ng, A.Y.: On optimization methods for deep learning. In: Proceedings of the 28th International Conference on Machine Learning (ICML 2011), Bellevue, WA, USA, pp. 265–272 (2011)
7. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**, 436–444 (2015)
8. Horn, B.K.P., Schunck, B.G.: Determining optical flow. Artif. Intell., Elsevier, North Holland **17**, 185–203 (1981) [Technical Report, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, USA (1980)]
9. Dosovitskiy, A., Fischer, P., Ilg, E., Häusser, P., Hazırbaş, C., Golkov, V., Smagt, P., Cremers, D., Brox, T.: Flownet: learning optical flow with convolutional networks. In: IEEE International Conference on Computer Vision (ICCV 2015), Santiago, Chile (2015)
10. Flying Chairs Dataset.: http://lmb.informatik.unifreiburg.de/resources/datasets/FlyingChairs.en.html. Accessed 19 Sept 2016
11. Feature Detection and Extraction.: http://in.mathworks.com/help/vision/feature-detection-andextraction.html. Accessed 14 July 2016
12. Nguyen, K., Fookes, C., Sridharan, S.: Improving deep convolutional neural networks with unsupervised feature learning. In: International Conference on Image Processing (ICIP 2015), pp. 2270–2274. IEEE (2015)

13. Chen, Y., Yang, X., Zhong, B., Pan, S., Chen, D., Zhang, H.: CNNTracker: online discriminative object tracking via deep convolutional neural network. Appl. Soft Comput. **38**, 1088–1098 (2015) [Elsevier]
14. Ouyang, W., Wang, X.: Joint deep learning for pedestrian detection, computer vision foundation. In: Proceedings of the 2013 IEEE International Conference on Computer Vision (ICCV'13), pp. 2056–2063. IEEE Computer Society, Washington, DC, USA (2013)
15. Zhang, Y., Li, X., Zhang, Z., Wu, F., Zhao, L.: Deep learning driven blockwise moving object detection with binary scene modeling. Neurocomputing **168**, 454–463 (2015) [Elsevier]. http://arxiv.org/abs/1601.07265
16. LeCun, Y., Huang, F.J., Bottou, L.: Learning methods for generic object recognition with invariance to pose and lighting. In: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'04), pp. 97–104. IEEE Computer Society, Washington, DC, USA (2004)
17. Jin, L., Gao, S.: Hand-crafted features or machine learnt features? Together they improve RGB-D object recognition. In: Proceedings of the 2014 IEEE International Symposium on Multimedia (ISM'14), pp. 311–319. IEEE Computer Society, Washington, DC, USA (2014)
18. Caffe.: http://caffe.berkeleyvision.org. Accessed 30 Mar 2016
19. Deep Learning Frameworks.: https://github.com/zer0n/deepframeworks#architecture. Accessed 28 Mar 2016
20. Deep Learning Libraries.: http://machinelearningmastery.com/popular-deep-learning-libraries. Accessed 28 Mar 2016
21. Applications.: https://www.quora.com/What-are-the-practical-applications-of-deep-learning-What-are-all-the-major-areas-fields. Accessed 18 Feb 2016
22. Cascade.: http://www.svcl.ucsd.edu/projects/Cascades. Accessed 10 June 2016
23. Convnet.: http://fastml.com/object-recognition-in-images-with-cuda-convnet. Accessed 28 July 2016
24. Optical Flow.: http://www.slideshare.net/xavigiro/deep-learning-for-computer-vision-34-video-analytics-lasalle-2016. Accessed 19 Sept 2016
25. Background Subtraction.: http://www.slideshare.net/ravi5raj_88/background-subtraction. Accessed 04 Aug 2016