International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-4, April 2019

Data Efficient Approaches on Deep Action Recognition in Videos

Sathya R, Rugveda Muralidhar I, Sai Harsha Vardhan K, Sri Karan R, Arun Reddy B

Abstract: ***In this paper, we propose an efficient visual tracker, which specifically catches a bounding box containing the target object in a video by methods for successive activities got the hang of utilizing deep neural networks. The proposed deep neural network to control following activities is pre-prepared utilizing different preparing video sequences and calibrated amid genuine following for online adjustment to a difference in target and background. The pre-training is done by using deep Reinforcement learning just as directed learning. The utilization of RL empowers even mostly named data to be effectively used for semi-directed learning. Through the assessment of the item following benchmark data set, the proposed tracker is approved to accomplish an aggressive exhibition at three times the speed of present deep network-based trackers.***

Index Terms*: *CNN, DAP3D-Net, Computer Vision, Pre-training

I. INTRODUCTION

Visual object tracking is a well-known sub-field of computer vision. Its main objective is to find the locus of points that the target of interest follows in image coordinates by assigning labels in temporarily spaced video frames. This information may be of pertinent importance for further analysis such as to calculate the centre of mass, area, perimeter and motion vector of the target. Thus, target tracking may play a vital role in high level image analysis tasks, such as object recognition activity analysis and intelligent scene understanding [1]. The point of visual object tracking is to discover the bouncing boxes containing the objective object in each frame in a video , which is one of the essential issues in the computer vision [2]. In ongoing decades, there have been numerous advances in visual tracking calculations, yet there are as yet many testing issues emerging from various tracking hindrances, for example, impediment, and movement obscure and light change, background mess. Specifically, ordinary tracking techniques utilizing low-level[3]-[9] hand-created highlights experience the previously mentioned tracking deterrents due to their lacking element portrayal. Deep learning added an enormous lift to the starting at now rapidly developing field of computer vision. With deep learning, a lot of new employments of computer vision methods have been introduced and are presently getting the chance to be segments of our ordinary day by day presences[10]-[12]. These fuse face recognition and requesting, photo stylization or machine vision in self-driving vehicles. The objective of this course is to familiarize understudies with computer vision, starting from stray pieces and after that swinging to progressively introduce day deep learning models[13]. We will cover both picture and video recognition, including picture gathering and remark, object recognition and picture look for, changed object area techniques, development estimation, object tracking in video, human movement recognition,[14] finally picture stylization, editing and new picture age. In course adventure, understudies will make sense of how to manufacture face recognition and control framework to grasp the internal mechanics of this advancement, likely the most acclaim and as often as possible displayed in films and TV-demonstrates instance of computer vision and AI. A yield of object tracking in the object track. There are a few challenges in developing visual tracking techniques.

Object tracking is the process of finding and moving object or various objects after some time in the video. A yield of object tracking in the object track. It is the sequence of object areas in each frame of a video.[14] Visual object tracking thinks about an issue of tracking of a solitary object in the video. Object to follow is determined in the principal frame of the video. We know nothing around an object aside from its area in the main frame, in this way, object tracking is sans display tracking. We can assemble the locator to distinguish this object in different pictures. Since object appearance is changing after some time, we consider just transient tracking for this visual tracking issue in a short video sequences. Just past frames can be utilized for visual object tracking strategy, no future impressions[15]. There are a few challenges in developing visual tracking techniques. First is a computational burden. For each second of a video, we have to process a ton of N frames. Second is the appearance change of the object after some time. The appearances can change because of the object elements, perspective change, lighting change, and different reasons. Third is object communication in video.[16]-[18]Other object can block the object of intrigue. The object appearance can be like other object. For this situation, you need to recognize these distinctive objects. In visual object tracking challenge, the normal cover is utilized to combine accuracy and strength of tracker into one measurement. At the point when tracker cover achieve zero, it is re-introduced and tracking is preceded [19].

Manuscript published on 30 April 2019.
***** Correspondence Author (s)
Mrs. Sathya R*, Assistant Professor, Computer Science and Engineering, SRM IST, Chennai, India.
Rugveda Muralidhar I, Computer Science and Engineering, SRM IST, Chennai, India.
Sai Harsha Vardhan K, Computer Science and Engineering, SRM IST, Chennai, India.
Sri Karan R, Computer Science and Engineering, SRM IST, Chennai, India.
Arun Reddy B, Computer Science and Engineering, SRM IST, Chennai, India.

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open access article under the CC-BY-NC-ND license http://creativecommons.org/licenses/by-nc-nd/4.0/

Along these lines, when tracking floats askew, it punishes the tracking accuracy. Along these lines, we can gauge both heartiness and accuracy in one measure. The thought is to lessen the hardware predisposition by reports in tracking speed with respect to the time required to perform channel task. As a kind of perspective, a MAX channel is utilized. MAX channel is connected in 30 by 30 window for all pixels in 600 by 600 pixel picture. The season of tracking is partitioned when of MAX channel. In 2016, very nearly 70 techniques were assessed in this challenge[20]-[22]. Different strategies depend on expound relationship channels, which likewise show great execution, yet somewhat more awful than Convolutional Neural Networks. Be that as it may, as far as speed, the best performing strategies were the slowest techniques. It is intriguing to see the most difficult and the least difficult precedent in this challenge[23]. The most difficult models incorporates tracking of an individual head in the grid, white rabbit tracking in white snow, the butterfly tracking in the flowers.

II. RELATED WORK

In visual object tracking, the principle commitment that experiences in the Gaussian fluffy calculation. There are numerous calculations for "fluffy", one of which is designated "Gaussian Blur". It utilizes a typical circulation (otherwise known as "Gaussian dissemination") for picture preparing. This article presents the "Gaussian Blur" calculation, and you will see this is a basic and straightforward calculation. Generally, it is a sort of information smoothing , which is reasonable for some events. Picture handling just gives an instinctive application model.

A. Principle of Gaussian blur

The supposed "fluffy" can be comprehended as the normal of the encompassing pixels for every pixel.

1	1	1
1	2	1
1	1	1

1	1	1
1	1	1
1	1	1

The "intermediate point" takes the normal of the "around points" and progresses toward becoming 1. In numerical terms, this is a sort of "smoothing". On the chart, it is identical to creating a "fluffy" impact, and the "intermediate point" loses detail. The above are the first picture, the impact of the haze span of 3 pixels and the haze sweep of 10 pixels. The bigger the haze sweep, the more obscured the picture. From a numerical point of view, the smoother the esteem. The following inquiry is, since each point needs to take the normal of the encompassing pixels, in what capacity would it be a good idea for us to dole out loads?

On the off chance that you utilize a basic normal, it is clearly not entirely sensible, in light of the fact that the pictures are consistent, the closer the point is, the closer the relationship is, and the more remote away the point is, the more removed the relationship is. In this way, the weighted normal is increasingly sensible, the closer the separation is, the more noteworthy the weight, and the more distant the separation is, the littler the weight is.

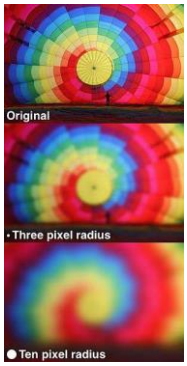


Fig-1: Depends on rate of pixel

B. Weight of the Normal distribution

A normal distribution is clearly an attractive weight distribution model. On the graph, the normal distribution is a ringer molded curve. The closer to the center, the bigger the esteem, the more distant far from the center, the littler the esteem. While computing the normal esteem, we just need to utilize the "center point" as the root. Different points can be allocated a weight as indicated by their situation on the normal curve, and a weighted normal can be acquired.

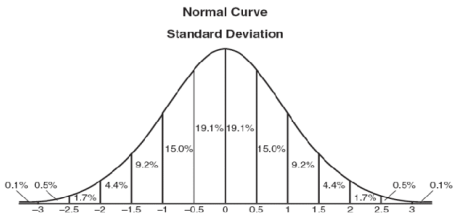


Fig-2: Normal curve standard deviation

C. Gaussian function

The density function of a normal distribution is known as a "Gaussian function" . Its one-dimensional structure is:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

Where μ is the mean of x and σ is the change of x. Since the center point is the origin while computing the normal, μ is equivalent to zero.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-x^2/2\sigma^2}$$

According to the one-dimensional Gaussian function, the two-dimensional Gaussian function can be derived as :

$$G(x,y) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2}$$

By this equation, you can calculate the weight of each point in it.

D. Weight matrix

let the coordinates of the center point taken as (0,0), the coordinates of all 8 points closest to it are as shown,

(-1,1)	(0,1)	(1,1)
(-1,0)	(0,0)	(1,0)
(-1,-1)	(0,-1)	(1,-1)

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)
© Copyright: All rights reserved.

Later points and so on. In order to find the weight matrix, it is important to set the values of σ. Let σ = 1.5, the weight matrix with a blur radius of 1 is as shown,

0.0453542	0.0566406	0.0453542
0.0566406	0.0707355	0.0566406
0.0453542	0.0566406	0.0453542

The sum of the weights of all 9 points is equal to the value 0.4787147. If only weighted the average of these all 9 points is calculated, the sum of their weights must be equal to 1, so the above 9 values are also divided by the value 0.4787147 to get the final weight matrix.

0.0947416	0.118318	0.0947416
0.118318	0.147761	0.118318
0.0947416	0.118318	0.0947416

E. Calculate Gaussian blur

By the weight matrix, you can find the value of the Gaussian blur.
Let 9 pixels, the gray value is as follows:

14	15	16
24	25	26
34	35	36

By Multiply each point by its own weight value we get,

14x0.0947416	15x0.118318	16x0.0947416
24x0.118318	25x0.147761	26x0.118318
34x0.0947416	35x0.118318	36x0.0947416

Including these nine qualities together is the estimation of the Gaussian haze of the center point. Rehashing this procedure for all points yields a Gaussian obscured picture. In the event that the original picture is a shading picture, Gaussian haze should be possible independently for the three channels of RGB.

F. Processing of boundary points

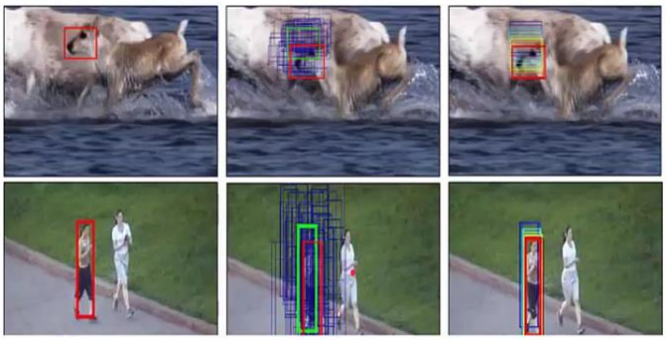
Imagine a scenario in which a point is at the outskirt and there are insufficient points around. A workaround is to duplicate the current points to the relating positions on the opposite side and recreate the complete matrix.

III. PROBLEM DESCRIPTION

A study, found that in visual article following, the methodology dependent on following by identification intends to manufacture a discriminative classifier that recognizes the objective from the encompassing foundation. Ordinarily these strategies catch the objective position by identifying the most coordinating position utilizing the classifier. the activity recognition multifaceted nature is high and even there is no help to follow 3D picture organization. We have a couple of 3D Convolutions however there are a few downsides. Article discovery, Scene Attributes Learning or Action Recognition are done as individual model. Extra layers can't be adaptable. Ordinarily these strategies catch the objective position by identifying the most coordinating position utilizing the classifier.

IV. PROPOSED MODEL

The proposed System portrays every individual activity as – Where the activity happens, what the activity is, the means by which the activity is performed.



(a)Previous Frame (b) Tracking by Detection method (c)Our Method

Fig-3: Difference between existing and proposed system

As such, it indicates precisely restrict, order and depict various activities in practical recordings. We have naturally parse the activity in recordings. 3D Images are bolstered. Item identification, Scene Attributes Learning or Action Recognition are done as brought together model. we can tweak to include extra layers.

A. System Architecture

To do the video frame separation, now all that is left is to actualize the Camera class, which should associate with the camera hardware and download live video frames from it. The pleasant thing about embodying the hardware subordinate piece of this application in a class is that this class can have diverse executions for various individuals, however whatever is left of the application continues as before. You can think about this class as a gadget driver, which gives a uniform execution paying little mind to the genuine hardware gadget being used. The other preferred standpoint of having the Camera class isolated from whatever is left of the application is that it is anything but difficult to trick the application into speculation there is a camera when as a general rule there isn't,
since the camera class can be executed to copy a camera without genuine hardware. Indeed, while I was taking a shot at this application, the least demanding path for me to test the gushing was to do that and not need to stress over the hardware until I had everything else running.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)
© Copyright: All rights reserved.

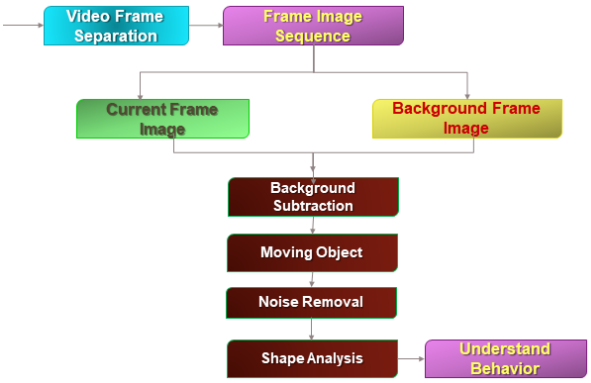


Fig-4: System Architecture

Later that the framework will isolate the casing pictures into current edge picture and foundation outline picture from that the framework will subtract the foundation outline pictures from that produced current casing pictures, A dynamic refreshing of foundation picture by edge contrast strategy and use the intensity of the foundation subtraction technique for distinguishing the moving item in all respects viably and precisely. The framework finds the moving items and evacuates the clamor in the event that it existences by utilizing the innovation RGB shading creation, Image pre-handling is the primary undertaking in moving article identification. The little changes in the pixel lead to false identification. Commotion can be added because of different reasons. Because of the clamor the pixel esteems may be changed. So picture pre-handling is exceptionally fundamental. Commotion Removing Noise is any substance which isn't of advantage to the motivation behind picture preparing. The impact of clamors on the picture flag amplitude and phase is complexity . Laterally the system makes the shape analysis of the object from the images. From the generated bounding boxes the system reads or understands the behavior of the object by storing the actions of the object from bounding boxes. From that generated actions it is very easy to track the object that are required. A few propelled foundation subtraction strategies have been proposed in the writing which is inhumane to outer ecological conditions, for example, commotion.

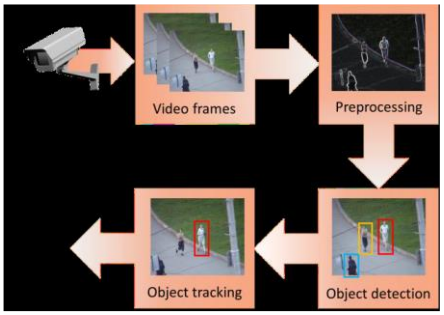


Fig-5: Process of Image detection

In the wake of distinguishing the objects of intrigue, the last advance is called object tracking which includes finding the area of the objective object in the consequent video outlines. While tracking the object, the framework administrator will take semantic activities, in view of the sort and area of the objects. For instance, if there should arise an occurrence of a traffic observing framework, the administrator can distinguish and screen the rundown of vehicles which are abusing the traffic controls and force fine naturally. To implement the project we do some requirements that are used to build the task. So to perform the functionality of the project we use language Python and we use a platform called Liclipse software and to store the data generated we use Structured Query Language, SQL database.

B. Flow Chart

The general flowchart of the proposed article following calculation is delineated in Figure There are three primary strides in the proposed calculation: (1) object segmentation and detection block, (2) noise reduction and (3) object tracking using Kalman filter. In this calculation, information of three distinctive thresholding strategies is utilized to make "assuming at that point" fluffy principles. The planned fluffy based thresholding strategy joins the referenced three distinct limits so as to give the best possible and ideal edge which will be used to section the item from the foundation. At last, the fragmented casing is connected to a Kalman filter to foresee the following way when the article moves. The accompanying subsections will depict the calculation in detail.

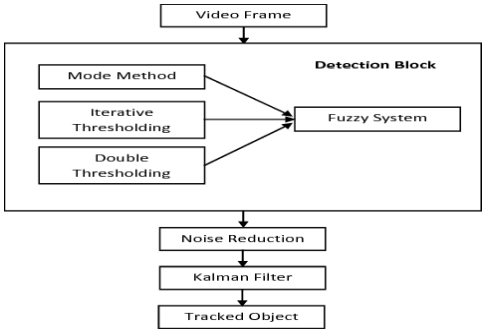


Fig 6 : Flow Chart

a. Object Detection

Object representation is the foundation of tracking. Thresholding methods are image divisions dependent on image-space locales. The basic rule of thresholding procedures depends on the attributes of the image. It picks appropriate edges to partition image pixels into a few classes and separate the objects from the foundation. For this reason, in this paper, three unique strategies, for example mode strategy, iterative thresholding and twofold thresholding are utilized. The mode strategy is reasonable for images with comparable objects and comparative light power field. The iterative thresholding is appropriate for non-uniform lighting and twofold thresholding is reasonable for non-uniform object shading. Information of these three distinctive thresholding strategies is utilized to make "assuming at that point" fluffy standards.

b. Noise Reduction

Images procured through present day sensors might be polluted by an assortment of commotion sources. In this stage, for take out clamor, the end activity is utilized that can fill little openings and holes.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)
© Copyright: All rights reserved.

This activity creates a specific measure of smoothing on an object's shape. At that point, acquired zones contrasted and the region which has 10 pixels utilizing the accompanying basic principle: y If object territory littler than 10 then the object is commotion.

c. Kalman Filter

Kalman filter has been effectively utilized in various forecast applications or state assurance of a framework, for example, object tracking. Kalman filter is a recursive estimator which has two particular stages: anticipate and update .The foresee stage utilizes the gauge from the past time venture to deliver a gauge of the present state. In the update stage, estimation data from the present time step is utilized to refine this forecast to touch base at another , progressively precise gauge.

C. Module Description

In this paper, to present the task we need some modules to implement the project. So we are using four various modules as followed below.

Fig-7: Module Description

a. Uploading video & Parsing

You can break down the problem to the various segments for the data you want to train the model with, and a predicted value that you can evaluate and then use operationally. Download the data sets and save them. The first dataset trains the machine learning model and then the second is used to find how accuracy your model is. It stacks a dataset with pictures and the comparing activities. This is utilized to produce the model, and after that train it. Pre-handling and cleaning information are imperative assignments that are utilized for a dataset to get more effectiveness for AI. Crude information is right off the bat loud and untrustworthy, and may make them miss esteems. Load the data Extract features.

b. Frame Separation

In this module, we recover the information one by one from the recordings transferred. Each recovered information called as edge will be put away in the server for further investigation. In the pictures research and applications, pictures are the main intrigued by some specific parts. These parts are regularly treated as objectives or closer view picture (as different pieces of the foundation). So as to distinguish and investigation the objective in the picture, we have to confine them from frontal area to foundation picture. The picture division is to isolate the picture and to get distinctive areas of it, each casing shows diverse qualities and to extricate the objective from that outline is the fascinating errand with regards to the procedure.

c. Background & Noise Removal

In this module, the principle task is Separating frontal area from foundation and if assumes a noteworthy job in numerous PC vision frameworks, including activity acknowledgment, video packing, remotely coordinating, movement catch, and reconnaissance following. Picture pre-preparing is one of the principle task in moving article recognition. Clamor can be added because of different reasons. Because of the clamor the pixel esteems may be changed. Commotion is any substance which is of no utilization of picture preparing. The impact of clamors on the picture flag plentifulness and stage is very multifaceted nature. So how to evacuate the clamor and keep the subtleties of picture are the significant undertakings of the picture sifting. Middle channel is a nonlinear strategy for evacuating clamor. Its fundamental thought is to utilize the middle of the area pixel dark incentive rather than the dim an incentive at pixel point. For the odd components, the middle alludes to the extent of the center an incentive in the wake of arranging.

d. Detect Action

A dynamic invigorating of foundation picture by edge differentiate procedure and utilize the power of the foundation subtraction strategy for perceiving the moving item in all respects adequately and precisely.

V. RESULT ANALYSIS

Object Tracking alludes to the way toward following a particular object of intrigue, or different objects, in a given scene. It generally has applications in video and certifiable collaborations where perceptions are made after an underlying object location. Presently, it's vital to self-ruling driving frameworks, for example, self-driving vehicles from organizations like Uber and Tesla. Object Tracking strategies can be partitioned into 2 classifications as indicated by the perception display: generative strategy and discriminative technique. The generative technique utilizes the generative model to portray the clear qualities and limits the remaking mistake to look through the object, for example, PCA.

The discriminative strategy can be utilized to recognize the object and the foundation, its execution is increasingly vigorous , and it bit by bit turns into the principle technique in tracking. The discriminative technique is likewise alluded to as Tracking-by-Detection, and profound learning has a place with this classification. To accomplish tracking-by-identification, we distinguish applicant objects for all edges and utilize profound figuring out how to perceive the needed object from the competitors. There are 2 sorts of essential system models that can be utilized: stacked auto encoders (SAE) and convolutional neural system (CNN). The most mainstream profound system for tracking errands utilizing SAE is Deep Learning Tracker, which proposes disconnected pre-preparing and online tweaking the net. The procedure works this way:

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)
© Copyright: All rights reserved.

A. Off-line unsupervised pre-train the stacked denoising auto-encoder utilizing substantial scale normal image datasets to get the general object representation. Stacked denoising auto-encoder can get increasingly strong component articulation capacity by including commotion in info images and reproducing the first images.

B. Combine the coding some portion of the pre-prepared system with a classifier to get the arrangement organize, at that point utilize the positive and negative examples got from the underlying casing to adjust the system, which can separate the present object and foundation. DLT utilizes molecule filter as the movement model to deliver hopeful patches of the present edge. The order organize yields the likelihood scores for these patches, which means the certainty of their groupings, at that point picks the most astounding of these patches as the object.

C. In the model refreshing, DLT utilizes the method for restricted limit.

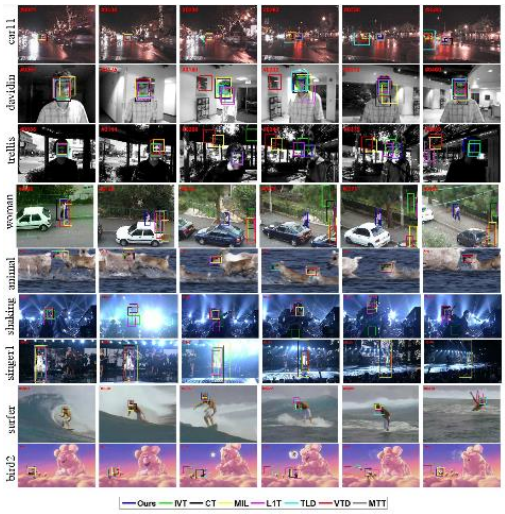


Fig 8 : Result Analysis

Not quite the same as the possibility of FCNT, MD Net uses every one of the successions of a video to follow developments in them. The systems referenced above utilize unessential image information to diminish the preparation request of tracking information, and this thought has some deviation from tracking. The object of one class in this video can be the foundation in another video, so MD Net proposes the possibility of multi-space to separate the object and foundation in each area freely. Also, a space demonstrates a lot of recordings that contain a similar sort of object. As appeared as follows, MD Net is isolated into 2 sections: the mutual layers and the K parts of space explicit layers. Each branch contains a double characterization layer with softmax misfortune, which is utilized to recognize the object and foundation in every area, and the common layers offering to all areas to guarantee the general representation.

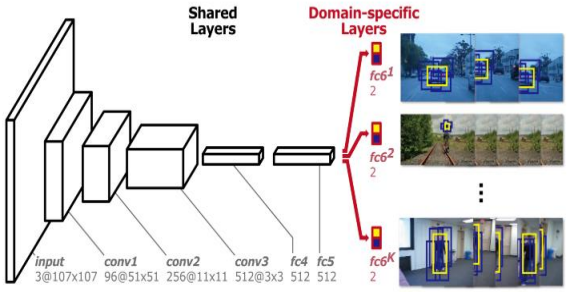


Fig 9 : Convolutional Neural Network

As of late, profound learning analysts have attempted diverse approaches to adjust to highlights of the visual tracking task. There are numerous bearings that have been investigated : applying other system models, for example, Recurrent Neural Net and Deep Belief Net, planning the system structure to adjust to video preparing and start to finish getting the hang of, improving the procedure, structure, and parameters, or notwithstanding consolidating profound learning with customary techniques for PC vision or methodologies in different fields, for example, Language Processing and Speech Recognition.

VI. CONCLUSION

In this paper, we have proposed action-driven method using deep convolutional neural networks for visual object tracking. The proposed system is controlled by a framework called flask, using Gaussian blur algorithm which finds the target of object by sequential actions. The action-driven tracking techniques make significant approaches to reduce the complexity of tracking. In addition, RL makes it possible to use partially labelled data, which could greatly contribute to the build of training data with some effort.

REFERENCES

1. H. Grabner, C. Leistner, and H. Bischof, “Semi-supervised on-line boosting for robust tracking,” in Proc. Eur. Conf. Comput. Vis., 2008, pp. 234–247.
2. B. Babenko, M.-H. Yang, and S. Belongie, “Robust object tracking with online multiple instance learning,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 33, no. 8, pp. 1619–1632, Aug. 2011.
3. Z. Kalal, K. Mikolajczyk, and J. Matas, “Tracking-learning-detection,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 34, no. 7, pp. 1409–1422, Jul. 2012.
4. D. S. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui, “Visual object tracking using adaptive correlation filters,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2010, pp. 2544–2550.
5. M. Danelljan, G. Häger, F. Khan, and M. Felsberg, “Accurate scale estimation for robust visual tracking,” in Proc. Brit. Mach. Vis. Conf., Nottingham, U.K., Sep. 2014.
6. J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, “High-speed tracking with kernelized correlation filters,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 37, no. 3, pp. 583–596, Mar. 2015.
7. Z. Hong, Z. Chen, C. Wang, X. Mei, D. Prokhorov, and D. Tao, “MUltistore tracker (MUSTer): A cognitive psychology inspired approach to object tracking,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2015, pp. 749–758.
8. J. Choi, H. Jin Chang, J. Jeong, Y. Demiris, and J. Y. Choi, “Visual tracking using attention-modulated disintegration and integration,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2016, pp. 4321–4330.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)
© Copyright: All rights reserved.

9. N. Wang, S. Li, A. Gupta, and D.-Y. Yeung. (2015). "Transferring rich feature hierarchies for robust visual tracking."[Online].Available:https://arxiv.org/abs/1501.04587
10. S. Hong, T. You, S. Kwak, and B. Han. (2015). "Online tracking by learning discriminative saliency map with convolutional neural network." [Online]. Available: https://arxiv.org/abs/1502.06796
11. H. Nam and B. Han. (2015). "Learning multi-domain convolutional neural networks for visual tracking." [Online]. Available: https://arxiv.org/abs/1510.07945
12. O. Russakovsky et al., "ImageNet large scale visual recognition challenge," Int. J. Comput. Vis., vol. 115, no. 3, pp. 211–252, Dec. 2015.
13. Y. Wu, J. Lim, and M. H. Yang, "Object tracking benchmark," IEEE Trans. Pattern Anal. Mach. Intell., vol. 37, no. 9, pp. 1834–1848, Sep. 2015.
14. M. Kristan, et al., "The visual object tracking VOT2014 challenge results," in Proc. Eur. Conf. Comput. Vis. Workshops, Berlin, Germany, Mar. 2014, pp. 191–217.
15. K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. (2014). "Return of the devil in the details: Delving deep into convolutional nets." [Online]. Available: https://arxiv.org/abs/1405.3531
16. R. Tao, E. Gavves, and A. W. Smeulders. (2016). "Siamese instance search for tracking." [Online]. Available: https://arxiv. org/abs/1605.05863
17. R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," Mach. Learn., vol. 8, nos. 3–4, pp. 229–256, 1992.
18. S. Yun, J. Choi, Y. Yoo, K. Yun, and J. Young Choi, "Action-decision networks for visual tracking with deep reinforcement learning," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2017, pp. 1349–1358.
19. H. Yang, L. Shao, F. Zheng, L. Wang, and Z. Song, "Recent advances and trends in visual tracking: A review," Neurocomputing, vol. 74, no. 18, pp. 3823–3831, Nov. 2011.
20. A. W. M. Smeulders, D. M. Chu, R. Cucchiara, S. Calderara, A. Dehghan, and M. Shah, "Visual tracking: An experimental survey," IEEE Trans. Pattern Anal. Mach. Intell., vol. 36, no. 7, pp. 1442–1468, Jul. 2014.
21. M. Kristan et al., "The visual object tracking vot2015 challenge results," in Proc. IEEE Int. Conf. Comput. Vis. Workshops, Dec. 2015, pp. 1–23.
22. J. Choi, H. J. Chang, S. Yun, T. Fischer, Y. Demiris, and J. Y. Choi, "Attentional correlation filter network for adaptive visual tracking," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jul. 2017, pp. 4828–4837.
23. M. Danelljan, G. Häger, F. S. Khan, and M. Felsberg, "Convolutional features for correlation filter based visual tracking," in Proc. IEEE Int. Conf. Comput. Vis. Workshops, Dec. 2015, pp. 58–66.
24. M. Danelljan, A. Robinson, F. Shahbaz Khan, and M. Felsberg, "Beyond correlation filters: Learning continuous convolution operators for visual tracking," in Proc. ECCV, 2016, pp. 472–488.
25. N. Wang and D.-Y. Yeung, "Learning a deep compact image representation for visual tracking," in Proc. Adv. Neural Inf. Process. Syst., 2013, pp. 809–817.
26. H. Li, Y. Li, and F. Porikli, "Robust online visual tracking with a single convolutional neural network," in Proc. Asian Conf. Comput. Vis., 2014, pp. 194–209.
27. H. Li, Y. Li, and F. Porikli, "DeepTrack: Learning discriminative feature representations online for robust visual tracking," IEEE Trans. Image Process., vol. 25, no. 4, pp. 1834–1848, Apr. 2016.
28. L. Wang, W. Ouyang, X. Wang, and H. Lu, "Visual tracking with fully convolutional networks," in Proc. IEEE Int. Conf. Comput. Vis., Dec. 2015, pp. 3119–3127.
29. D. Held, S. Thrun, and S. Savarese. (2016). "Learning to track at 100 FPS with deep regression networks." [Online]. Available: https://arxiv.org/abs/1604.01802 V. Mnih et al., "Recurrent models of visual attention," in Proc. Adv. Neural Inf. Process. Syst., 2014, pp. 2204–2212.

© Copyright: All rights reserved.