[[1]](#footnote-1)

Visual Tracking of Multiple Objects Using Shallow Convolution Feature

Nagaraja N S1, Dr.Vasudeva2

*1Srinivas School of Engineering,Mukka,Mangaluru, 2 SMVITM, Bantakal,Udupi,Karnataka*

*Abstract*— Traditional object tracking algorithms use manual extraction of features. This is difficult to cope with the challenges of occlusion, rotation and deformation. A shallow convolutional network is used in this paper to extract features for tracking. This combines hard negative mining technology and bounding box regression to refine the target location. The effectiveness of proposed tracker compared with tracking algorithms is discussed.

*Index Terms*—Object tracking, feature extraction, shallow convolutional network

# INTRODUCTION

C

OMPUTER VISION is a field that deals with how computers can be made to gain high-level understanding from digital images or videos. A video is a group of basic structural units, such as scene, shot and frame associated with audio data. A frame is defined as a single picture shot of movie camera, led by many successive frames for seamless video. In other words a video is a temporal sequence of image frames. Each frame exists independently of each other. Video coding generally encode the temporal relationship between the image frames. Any entity of interest within a scene that can be analyzed, depending on the requirements of the application, is referred to as an object. A set of objects are sub images in each frame of a sequence of image frames. Objects are embedded in the background within the image in a frame.

Intelligent video surveillance systems allow the computer to automatically locate, recognize, track and monitor the people, vehicles, and other objects by automatic analysis of image sequences which is captured by cameras. These systems which involve automatic detection and tracking of moving objects, and the scene analysis attracted a great deal of interest among researchers.

The most fundamental processes in understanding video contents is Visual object tracking. The object tracking is done on the sequence of images obtained from a video. Object tracking is the process of estimating trajectory or path of an object in consecutive frames of video. It involves of finding the location and dynamic configuration of one or more moving objects in each frame of a video. The objective is to establish correspondence of objects and object parts between consecutive frames at subsequent time instants by analyzing selected object features. It provides cohesive temporal data about moving objects.

The challenges in object tracking arise due to changes in object appearance, such as its deformation, scale variations, fast motion, and inplane and out-of-plane rotation, and due to a dynamic environment, such as illumination variations, occlusion, and background clutter.

This work addresses the problem of tracking a object in an accurate and robust manner, with arbitrary motion and no prior knowledge other than its position in the first video frame. Even after tremendous research accurate, robust and efficient tracking remains challenging due to tough environmental variations.

.

# previous works

Current trackers can mainly be categorized into either generative or discriminative approaches.

Generative trackers treat the tracking process as finding the candidate most similar to the target object. These methods are mostly based on templates. Template based approaches directly use the appearance of the pixels without extracting features, and optimize a similarity measure between a template and a captured image, based on the Newton method or its variants, to determine the pose of the plane. These types of approaches usually suffer from perturbation factors such as illumination changes, partial occlusions and fast motions.

On the other hand, the discriminative trackers equate tracking as a binary classification problem in order to distinguish the target from the background. Unlike generative methods, the information of both the target and its background is used simultaneously. These utilize projected gradient to facilitate multiple kernels in finding the best match during tracking under predefined constraints. And further a set of weak classifiers are combined into a strong one for robust visual tracking.

Boosting based discriminative appearance models have been widely used in visual object tracking because of their powerful discriminative learning capabilities. The key idea with this type of tracking is how to exploit the boosting classifier. Because the framework uses a boosting classifier, it can be categorized as a discriminative approach, and thus the computation time should be fast and feature extraction that can be computed rapidly is required. For example, the Adaboost classifier is implemented in an object tracking algorithm in [1] and a boosting classifier has been used in the learning of multiple instances for object tracking [2]. A discriminative approach using a boosting classifier has a limitation related to the region used for searching the target object.

Template based generative appearance model is discussed in [3]. An adaptive probabilistic tracking algorithm that updates the models using an incremental update of Eigen basis is proposed in Paper [4]. Sparse representation has been applied to various computer vision tasks. The 𝑙1 minimization tracker that uses the low resolution target image along with trivial templates as dictionary elements is proposed in [5]. The occlusion problem was addressed through a set of trivial templates. A tracking algorithm using the structural local sparse appearance model was proposed in [6], which exploits both partial information and spatial information of the target based on an alignment-pooling method. A two-view sparse representation based algorithm was presented in [7] where the tracked objects are sparsely represented by both templates and candidate samples in the current frame. A multi-feature joint sparse representation for object tracking to encode more information is presented in [8].

 Hybrid methods exploit the complementary advantages of the previous two approaches.

In [9] for tracking two different models are used, where the target appearance is described by low-dimension linear subspaces and a discriminative classifier is trained to focus on recent appearance changes.

A sparse collaborative tracking algorithm is presented in [10] which exploit both holistic templates and local patches. A hybrid model for object tracking, where the target is represented by different appearance manifolds is developed in [11]. The tracking method in [12] integrated the structural local sparse appearance model and the discriminative classifier with a support vector machine.

The generative model in [13] employs the shallow feature learning strategy to account for occlusion and the discriminative model adopts the deep feature learning strategy to effectively separate the foreground from the background. A hybrid tracking method by the combination of discriminative global and generative multi-scale local models is proposed in [14].

# Proposed tracker

In this section, we introduce the framework of tracker which consists of candidate target generation, feature extraction model, tracking strategy and model updating. A. Candidate target generation: In the tracking process, the target shifting between adjacent frames is so small that the Gauss sampling can be performed around the target position of the previous frame. Given the target location xt-1 in the previous frame, we assume the locations in frame T are subject to a Gaussian distribution, which is denoted as:

**p(xt|xt-1) = N(xt,xt-1,Є) xt = (x, y, σ)**

where x, y, σ is respectively as the center coordinates and the scale change of the bounding box, ϵ is a diagonal covariance matrix that indicates the variances of the location parameters. By the above method, the selected target generated around the previous frame is sent into the convolution network to extract features, which provide the selected sample for the following target tracking.

## Feature Extraction Model:

This employs convolutional neural network (CNN). The architecture of a CNN usually consists of several layers, including convolutional layers, normalization layers, and pooling layers. Moreover, convolutional layers usually consist of several layers: from a shallow layer to the deepest layer. Although the deepest layer provides the best results for image classification, in this research, we used a shallow layer because it provides more favorable information than the deepest layer for object tracking owing to the fact that the information from a shallow layer of a pretrained CNN only requires small number operations as compared to the deepest layer. Based on this fact, the information from a shallow layer can still represent the input, which will be more useful for the case of object tracking. The framework of the proposed method is represented in Figure 1

##

Figure 1 framework of proposed method

The proposed network is only five layers, less than the commonly used target detection network layer, but enough to meet the needs of the target tracking task. In the initialization process, we use a certain number of positive and negative samples to train the FC layers. The positive samples are generated in a small area around the previous frame generation targets, negative samples are generated in the whole range of pictures, so the number of negative convolutional features *f-(X)* is much greater than the number of positive features *f+(X)*. The two fully connected layers (FC) are respectively connected with a ReLUs layer and a dropout layer. Object tracking is considered to be a two classification task that distinguishes between the target and the background, so the latest FC layer connects a binary classifier.

We sample N target candidates X1…XN around the previous frame and put them into the network to obtain their features f(X).

## Tracking strategy

Tracking strategy In the initialization process, we use a certain number of positive and negative samples to train the FC layers. The positive samples are generated in a small area around the previous frame generation targets, negative samples are generated in the whole range of pictures, so the number of negative convolutional features f minus (X) is much greater than the number of positive features f plus (X).

Due to the imbalance of positive and negative samples, the SGD method easily suffers from a drift problem . We use hard negative mining technique which selects a certain number of representative negative samples with top M negative scores to balanced proportion of positive and negative samples. As the learning proceeds and the network become more discriminative, the classification becomes more distinctive

# tracking by bayesian inference

Object tracking can be treated as a Bayesian inference task . Given the observations of target Zt = {z1, z2, ..., zt} up to time t, the current target state st can be obtained by the maximum a posteriori estimation via:

 **………..** (1)

where denotes the i-th sample of the state . The posterior probability can be recursively computed by the Bayesian theorem via

 …….. (2)

where and denote the dynamic model and observation model, respectively. The dynamic model describes the temporal correlation of the target states in consecutive frames, and the motion of the target between consecutive frames is modeled by an affine transformation. The state transition is formulated by random walk,

i.e.,

**p(st|st−1) = N(st : st−1, ∑),** where

 **st = {αt , βt , µt , νt}**

 denote the x, y translations, scale and aspect ratio at time t, respectively.

 is a diagonal covariance matrix whose elements are the variances of the affine parameters. The observation model p(zt|st) estimates the likelihood of observing zt at state st . In this paper, the collaborative likelihood of the c-th candidate is defined as:

and the candidate with the maximum likelihood value is regarded as the tracking result.

Object Tracking: Video is collection of frames. The negligible time gap between frames makes the stream of photos looks like flow of scenes. When designing algorithm for video processing. Videos are classified into two classes. Video stream is an ongoing process for video analysis. Video sequence is video of fixed length. All the consecutive frames are obtained prior to processing of current frame. Motion is distinct factor that differentiates video form frame. Object properties and action can be realized by noticing only sparse points in the image.

# Algorithm

Input: The first frame of the image and the pre-trained convolution network (w1:w5)

1) Sample X+ and negative sample X- are Generates positive using Eqn (2). Fine-tune the FC layer W4, W5.

2) Train a bounding box regression model.

3) From 2 to N

4) Generate candidate target samples X with Equation.1, and send them into the convolution network to extract features.

5) Estimate the target scale and position with the pre-trained bounding box regression model.

7) Tracking model update: when the target confidence is less than 0.5

8) Fine-tune full connectivity layer (W4, W5) and bounding box regression model.

9) End

10) End

 **Output:** tracking results.

#  multiple object tracking

Simple Online Real Time Tracking (SORT) SORT is a realistic approach to achieve Multi Object Tracking (MOT). SORT is a framework that has Kalman filtering has its crux. Image by image data association is achieved by Hungarian method over an association metric like appearance that measures bounding box overlap.

# flow chart for multiple object tracking



# performance metrics

The trained must be evaluated for its performance on unseen data called as test dataset. The choice of performance metrics will influence the analysis of algorithms. This helps in identifying reasons for misclassifications so that it can be corrected by taking necessary measures.

1. Confusion Matrix: It gives prediction information of various objects for binary classification
2. Accuracy: Accuracy measure is calculated by using formula .
3. Precision, Recall: Precision is the percentage of classification results that are relevant. Recall is the percentage of total relevant results that are classified correctly by algorithm.

The formulas to calculate these metrics are

The detected objects are bounded with bounding box. Tracking is performed on frames of the videos to identify objects in the successive frames using SORT. The evaluation metrics like True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN) thereby accuracy, Precession and Recall are calculated.

OTB is a popular tracking benchmark that contains a large number of annotated videos with substantial variations. OTB video sequence contains Illumination Variation, Scale Variation, Deformation, Blur and other 11 kinds of challenges. Our tracker will be evaluated in the all 11 challenges. OTB provides three evaluation metrics: one pass evaluation (OPE), time robustness evaluation (TRE) and spatial robustness evaluation (SRE).

The expected precision and success plot of OPE, TRE and SRE are shown in figure [15]







# Conclusion

The inclusion of Artificial Intelligence to solve Computer vision tasks has outperformed the image processing approaches of handling the tasks. The CNN model trained to on dataset for single object detection has high validation accuracy because of huge amount of data on which it is trained from each class. Results of performance metrics is totally dependent on image data set used. Further objects are detected in video based on region of interest. Multiple objects are detected and tracked on different frames of a video

References

 [1] H. Grabner, M. Grabner, and H. Bischof, “Real-time tracking via on-line boosting,” in Proceedings of the British Machine Vision Conference (BMVC ’06), vol. 1, pp. 6.1-6.10, BMVA Press, Edinburgh, UK, September 2006.

[2] B. Babenko, M.-H. Yang, and S. Belongie, “Robust object tracking with online multiple instance learning,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 33, no. 8, pp. 1619–1632, 2011.

[3] Jepson, A. D., Fleet, D.J., EI-Maraghi, and TF: ‘Robust online appearance models for visual tracking’. Proc, IEEE Conf Computer Vision and pattern recognition (CVPR),Kauai, HI, USA, December 2001,pp. 415-422

[4] Lim ,J., Ross, D.,Lin,R.S.,Yang, M.H.: ‘Incrimental learning for visual tracking’,in Weiss,Y.,Bottuou,L.,(Eds.), ‘Advances in neural informa,tion processing systems,(MTI Press,2005),pp.793-800

[5] C. Bao, Y.Wu, H. Ling, and H. Ji, “Real time robust l1 tracker using accelerated proximal gradient approach,” in IEEE Conference on Computer Vision and Pattern Recognition, 2012.

[6] Jia, X.; Lu, H.; Yang, M.H. Visual Tracking via Adaptive Structural Local Sparse Appearance Model. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence, RI, USA, 16–21 June 2012; pp. 1822–1829. 13.

[7] Wang, D.; Lu, H.; Bo, C. Online Visual Tracking via Two View Sparse Representation. IEEE Signal Process. Lett. 2014, 21, 1031–1034. 14.

[8] Hu, W.; Li, W.; Zhang, X.; Maybank, S. Single and Multiple Object Tracking Using a Multi-Feature Joint Sparse Representation. IEEE Trans. Pattern Anal. Mach. Intell. 2015, 37, 816–833. [CrossRef] [PubMed]

[9] Yu, Q.; Dinh, B.T.; Medioni, G. Online Tracking and Reacquisition Using Co-trained Generative and Discriminative Trackers. In Proceedings of the European Conference on Computer Vision (ECCV), Marseille, France, 12–18 October 2008; pp. 678–691

[10] Zhong, W.; Lu, H.; Yang, M.H. Robust Object Tracking via Sparse Collaborative Appearance Model. IEEE Trans. Image Process. 2014, 23, 2356–2368. [CrossRef] [PubMed]

[11] Zhou, T.; Lu, Y.; Di, H. Locality-Constrained Collaborative Model for Robust Visual Tracking. IEEE Trans. Circuits Syst. Video Technol. 2015, 27, 313–325. [CrossRef]

[12] Dou, J.; Qin, Q.; Tu, Z. Robust Visual Tracking Based on Generative and Discriminative Model Collaboration. Multimed. Tools Appl. 2016. [CrossRef]

[13] Bohan Zhuang n , Lijun Wang, Huchuan Lu, Visual tracking via shallow and deep collaborative model. Neurocomputing, Elsevier,2016

[14] Zhiguo Song, Jifeng Sun and Jialin Yu, Object Tracking by a Combination of Discriminative Global and Generative Multi-Scale Local Models,MDPI, Information 2017, 8,43. 2017

[15] Wei wang , Mingquan Shi , Weiguang Li1, Object tracking with shallow convolution feature, IEEE, 2017

[16] Mohana, HV Ravish Aradhya,Object Detection and Tracking using Deep Learning and Artificial Intelligence for Video Surveillance Applications, International Journal of Advanced Computer Science and Applications, Vol. 10, No. 12, 2019

1. Nagaraja N S

Associate Professor,Srinivas School of Engineering,

Mukka,Mangaluru,Karnataka-574164

E-mail :nagaraj\_n\_s@yahoo.com [↑](#footnote-ref-1)