Improved tracking using P-N Learning

A DISSERTATION
SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENT
FOR THE AWARD OF DEGREE
OF

MASTER OF TECHNOLOGY
IN
SIGNAL PROCESSING & DIGITAL DESIGN

Submitted by

NIKITA JAYASWAL
2K14/SPD/10

Under the supervision of

Sh. Rajesh Rohilla
(Associate Professor, Department of ECE)

Department of Electronics & Communication Engineering
Delhi Technological University
(Formerly Delhi College of Engineering)
Delhi-110042
2014 - 2016

Certificate

This is to declare that the dissertation title **"Improved tracking using P-N Learning"** submitted by **Ms Nikita Jayaswal**, Roll. No. 2K14/SPD/10, in partial fulfilment for the award of degree of Master of Technology in Signal Processing & Digital Design at **Delhi Technological University, Delhi**, is a bonafide record of student's own work carried out by him under my supervision and guidance in the academic session 2015-16.

Mr Rajesh Rohilla
Supervisor
Associate Professor
Dep. ECE
Delhi Technological University

Acknowledgement

I express my deep sense of honor and appreciation towards my supervisor **Mr. Rajesh Rohilla, Associate Professor** Department of Electronics and Communication, for his constant support and encouragement which helped in successful completion of my thesis .

I also thankful to **Prof. Prem R. Chadda,** Head of Department (Electronics & Communication Engineering), towards all the faculty and staff of Electronics & Communication Engineering and friends for their ideas and disscussion which were useful in completion of this "**Thesis"** work.

I wish to indebt my deep sense of appreciativeness towards family and friends which were great source of encouragewho bestowed upon me their grace and were source of my inspiratment and easement.

Nikita Jayaswal
M.Tech (SPDD)
2K14/SPD/10

Dedicated to my loving Parents, Guide, Friends

and

the Almighty…

TABLE OF CONTENTS :

LIST OF FIGURES

ABSTRACT

The tracking and detection of target forms the basis of analysis and understanding of visual tracking.In adaptive tracking by detection framework,the object model is known.But due to fixed object model,the cluttered background may generate false positives or some object appearences cannot be detected.Thus a novel approach called Tracking –Learning – Detection(TLD) is used here which is composed of three subtask.The tracking block estimates frame to frame motion.The localisation of appearences observed during tracking is carried out in detection block.Here learning component estimate detector errors and update it to ignore these errors.Using above cascaded approach,computation time is reduced.The real time implementation of Tracking, Learning and Detection is explained and implemented on benchmark sequences.Thus significant improvement is achieved over state of art methods.

CHAPTER 1

INTRODUCTION

1.1 OBJECTIVES : Kalal firstly proposed TLD framework.TLD is a long term tracking framework which explicitly decomposed into tracking ,learning and detection parts.The framework of original TLD is extended for multitarget version but the framework remain general and extensible.Due to general characteristics ,we understand that any state of art algorithm can replace any module in the framework.With the help of extensible characteristics ,any other additional module can be easily added to the framework.Thus tracking and detection accuracy is improved since TLD combines tracking function with detection.

For a given sequences of frames and initialised bounding box in one frame,TLd aims to track the object at every instant.A long term tracker should be able to draw a boundary box when the object of interest comes into view.The tracker should be adaptive to changes in object appearance.Previous state of art methods used tracking by detection framework (the position of object is predicted by detector)where the object model is fixed.Thus detection is failed since background may generate false negatives.This issue is addressed in TLD by a novel method called PN learning.which identify the errors made by detector and correct it to avoid those errors.

1.2 CHALLENGES

The no. of applications in video analysis requires tracking of object or their parts in long video sequences.But there are number of issues which are to be addressed in long term tracking are :

1) OCCLUSION : When the object does not appear for some time or occluded by any other object , then a tracker incorrectly points to nearby object.Thus for long term tracking,a tracker should be able to handle this scenario.

2) BACKGROUND CLUTTER AND IDENTIFICATION : The cluttered environments or other objects of same class may distract long term tracker to correctly distinguish object of interest from other objects of same class.The tracker should be able to handle cluttered background to identify target from surroundings.

3)CHANGES IN APPEARENCE : The target may have appearance variation which adds difficulty in tracking.Throughout the sequence changes in appearance and viewpoint of an object

of interest complicates the tracking process.Since the only information given is a single patch from initial frame which may not be consistent throughout the entire sequence.Thus it should have a mechanism to deal appearance changes.

4)SCALE AND ILLUMINATION CHANGES : The object change its scale and appearance under different illumination.Different illumination causes transition in presence of obect.Prediction of object scale increases tracker susceptibility to failure.

5)NOISE : Blurring,compression etc. are the factors responsible for producing noise causing accuracy to be affected(corrupting the output of tracking algorithm)Thus the tracking algorithm should be able to remove noise.

1.3 CONTRIBUTION : In TLD framework,we follow the approach of Kalal.We are extending the detection cascade of object and make use of different features for enhancing the performance of detector. The points which are used in original Tld algorithm are affected by drift.Thus new feature points must be identified and adjusted with original points for reliable tracking.In addition to improve the detection mechanism ,Brisk features are extracted. The framework is implemented in MATLAB and evaluated on TLD datasets and some other datasets also.

The following work is done :

1) TLD algorithm is evaluated (i.e. challenges of algorithm,usability etc.)

2) The TLD algorithm is compared with other tracking approach.

3) Some improvements over existing TLD algorithm is made using a good set of features.

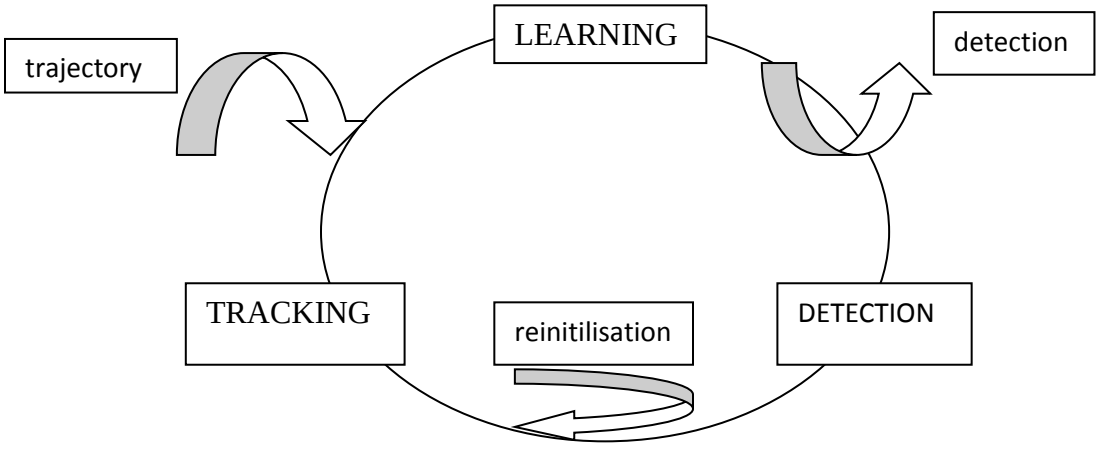


Fig 1 BLOCK DIAGRAM OF OPENTLD

Chapter 2

Tracking

In this recursive method for object tracking,no prior information about the object is needed except for its location in previous frames i.e. it requires an external initialization.By doing manual intervention in the first frame & results of object detection process in subsequent frames,initialization is accomplished.

The Kalal approach for recursive tracking is explained as :

First bounding box is constructed in frame t using an spaced set of points.Using method of Lucas & Kanade , for each of these points , optical flow is estimated. This method is good for points resided on corners but not for points comes under homogeneous regions.The information from two different error measures and Lucas Kanade method is utilized to filter out those tracked points which are likely to be erroneous.These errors measures based on Normalised Cross Correlation(NCC) and Forward Backward (FB)error.For all forward backward error measures , the median is checked.If the median is above a certain threshold then it is considered as an indication of drift else remaining points are used to estimate the new bounding box position in the second frame with the help of transformation model based on changes in scale and translation.

2.1 CLASSIFICATION :

In classification of tracking algorithm ,we categorize the object state as :

1) **POINTS :** Point tracking is appropriate for when points are used to represent object.It estimates frame to frame motion of object.

2) **PRIMITIVE GEOMETRIC SHAPES :** They are appropriate for tracking rigid object but can be used for non rigid objects also.The most common examples of these shapes are rectangle,eclipse,etc.

3) CONTOUR : The boundary of an object is represented by the contour and the silhouette of an object is defined as the vicinity inside the contour.They are suitable for tracking non rigid shapes.

4) ARTICULATED SHAPE MODELS : Articulated models are made up of body parts connected together with joints.These models are suitable for representing non rigid object motion.Example : human body.

5) SKELETAL MODEL : They are mostly used for recognizing object shapes.The medial axis translation is applied on object silhouette to extract skeleton.

2.2 ESTIMATION OF OPTICAL FLOW :

For Estimation of Optical Flow , specific conditions to be satisfied :

1) Brightness Constancy :- In brightness constancy the same brightness level is maintained between different frames.It is expressed as

$$X(z) = Y(z+d) \qquad (1)$$

Irrespective of location of pixel, whether it is in image I or J image ,it will retain its brightness value.The vector d represented as displacement vector.

2) Time Continuity :- In Time Continuity, the content of image not to be changing very fast with time.This condition is also referred as Temporal Persistence. In Temporal Persistence, the displacement vector is small means that Y(z) can be approximated as

$$Y(z) \cong X(z) + X'(z)d \ , \qquad (2)$$

X’(z) is the gradient of X at location z. An estimate for d is then

$$d = \frac{Y(z) - X(z)}{X'(z)} \qquad (3)$$

3) Spatial Coherence : In spatial coherence ,all the pixels inside a window around a pixel

move coherently .From this assumption , d is found by minimized the term

$$\sum_{(x,y)\epsilon W} (\mathrm{Y(z)} - \mathrm{X(z)} - \mathrm{X'(z)d})^2 \qquad (4)$$

which is least squares minimization.W refers to area around each pixel to be considered.The closed form solution for eqn (4) is

$$\mathrm{I(d)} = \mathrm{e} \qquad (5)$$

where

$$\mathrm{I} = \sum_{z\epsilon W} \mathrm{X'(z)X'(z)^T} = \sum_{z\epsilon W} \begin{matrix} \mathrm{X_x^2(Z)} & \mathrm{X_{xy}(z)} \\ \mathrm{X_{xy}(z)} & \mathrm{X_y^2(Z)} \end{matrix} \qquad (6)$$

&

$$\mathrm{e} = \sum_{(x,y)\epsilon W} (\mathrm{X(z)} - \mathrm{Y(z)})\mathrm{X'(z)} \qquad (7)$$

2.3 ERROR MEASURES :

For increasing robustness of tracker,three criteria should be used to filter out points that were tracked erroneous.

First Criterian :- If I is invertible ,then can be calculated from equation(5).For I to be invertible ,it has two large eigen values (λ_1, λ_2) which is possible when there are gradients in two directions. We use the formula

$$\min(\lambda_1, \lambda_2) > \lambda \qquad (8)$$

The formula given by Shi and Tomas used as first criterian for suitable tracking of points.For the image given in fig(a), we have to correctly track the points and its corners position in fig(b).But due to occlusion , the point 2 tracked at wrong location.It is necessary for the following error measure that the tracking should be reversible.It means that the point l should be tracked back to its original position.But point 2 is tracked at a different position.The following error measure is defined as

$$\epsilon = |q - q''| \quad (9)$$

where q'' is $q'' = LK(LK(q))$ (10)

Equation (10) implies that Lucas Kanade method is executed twice on q.
Normalised Cross Correlation (NCC) is used to compare similarity of two image patches named Q_1 & Q_2.

$$NCC(Q_1, Q_2) = 1/n\text{-}1 \frac{\sum_{x=1}^{n}(Q1(x)-\mu1)(Q2(x)-\mu2)}{\sigma1\sigma2} \quad (11)$$

Where $\mu1, \mu2, \sigma1, \sigma2$ are the means and standard deviations of Q_1 , Q_2.Against homogeneous variation of brightness ,NCC is invariant .

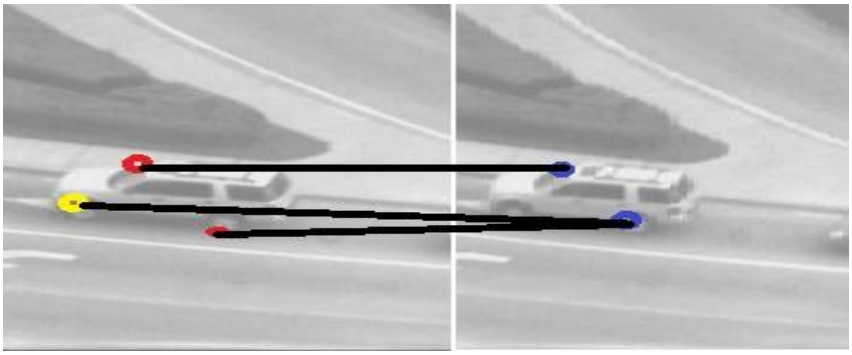


FIG 2: Example of FB error (i.e. some points cannot be retraced to their original position)

2.4 TRANSFORMATION MODEL :

In transformation model, we compute the median med_{NCC} of all similarity measures & the median of all forward backward errors med_{FB}.For the points which has forward – backward error less than med_{FB} and similarity measure larger than med_{NCC}.

For the case of not reliable tracking result, the med_{FB} is greater than predefined $\boldsymbol{\theta}_{FB}$. The points which are left used to compute the translation of bounding box.After that computing the pairwise distances between all points before and after tracking and result in relative increase formed the basis for change in scale.For x-direction translation , median of horizontal translation is computed.Similarly translation is computed for y- direction.

Chapter 3

Detection

In object detection the recursive tracker reinitialisation takes place since tracker does not have an object model to maintain & is unable to recover from failure.Since the location of object in previous frame is used by recursive tracker,to find an object ,the detection mechanism does an exhaustive search.For each input image,several thousand subwindows are evaluated.Thus,this approach is time consuming process.

The detection process employs a sliding window approach.At top of detector cascade,the image is presented which then computes a classification function within each input image.For VGA image (640*480), we employ 50000 to 200000 subwindows depending on initial object size.To find the object of interest each subwindow is tested.The next stage is evaluated only if a subwindow is passed by previous stage.The detector cascade stages are shown below input image.The aim of cacade detector is to filter out non relevant subwindows with lesser computation.

At first stage we need a background model to detect the foreground regions resticting the search space.In variance filter stage , those subwindows are accepted that exhibit a variance greater than a certain threshold.The third stage is ensemble classifier based on random ferns and the last stage comprises nearest neighbor classification and template matching method.Template Matching is based on NCC as similarity measure.If there are overlapping subwindows then non maxima suppression comes with output bounding box.

3.1) Sliding Window Approach :

In sliding window approach , all the subwindows of image are checked, for having object of interest or not since every possible subwindow has a likelihood of containing object of interest.

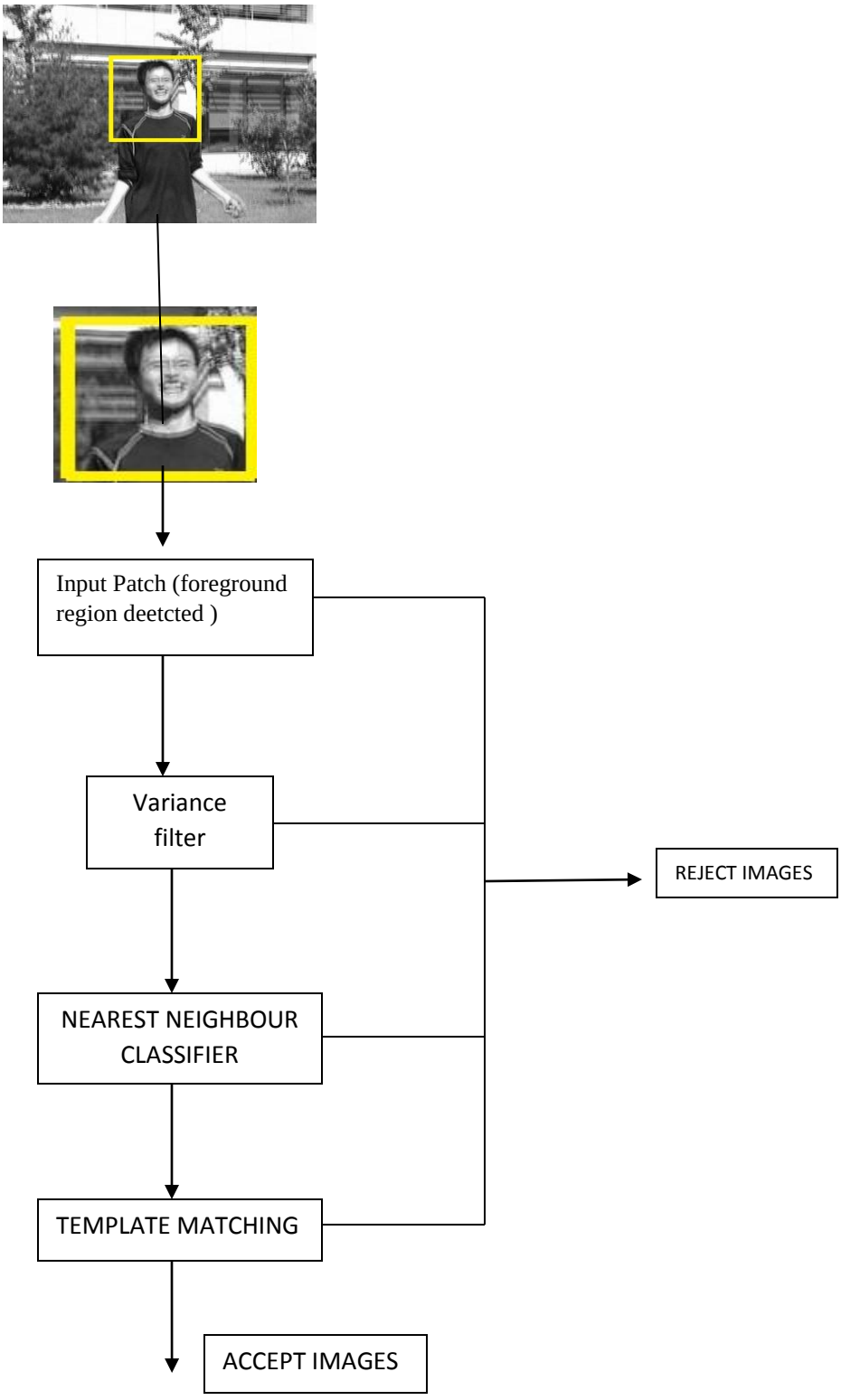


FIG: 3 SLIDING-WINDOW APPROACH FOR OBJECT DETECTION

3.2) Foreground Detection :

To detect foreground region in video stream,we perform background subtraction.From the fig..given below,there is a background image B_{bg} given and the image B in which the detection of object is to be performed.Next,calculate the absolute difference between both images.

$$B_{absdiff} = | B_{bg} - B | \quad (12)$$

Then applying thresholding of 16 pixels to $B_{absdiff}$ which results in binary image $B_{binary.}$

$$B_{binary}(x,y) = \begin{cases} -1, & \text{if } Babsdiff(x,y) > 16 \\ 0, & otherwise \end{cases} \quad (13)$$

In binary image, the connected white pixels referred as components.Then applying labeling algorithm to compute the area and smallest BBox the blob fits into , calculating labels in a single pass.

3.3) Variance Filter :

This mechanism works by rejecting patches in a image subwindows with variance lower than σ^2_{min}(threshold).By variance,we refer to the measure for uniformity in an image patch.

This kind of variance filter easily rejects uniform background regions but fails to differentiate between well structured objects.The mechanism for computing variance is shown below.The variance σ^2 is given as

$$\sigma^2 = 1/n \sum_{i=1}^{n} (yi - \mu)^2 \quad (14)$$

Where n referred to number of pixels in image and μ is

$$\mu = 1/n \sum_{i=1}^{n} yi \quad (15)$$

By taking advantage of integral images since they share same pixel value,it will be faster to calculate variance.

$$\sigma^2 = \frac{1}{n} X''(B) - [\frac{1}{n} X'(B)]^2 \quad (16)$$

where

$$X'(B) = \sum_{i=1}^{n} x_i \quad (17)$$

$$X''(B) = \sum_{i=1}^{n} x_i^2 \quad \text{,B –> Bounding Box}$$

3.4) Ensemble Classifier :

This classification method used in third stage of detection cascade is referred to as random fern classification.The classifier decision is depending on comparing intensity value of several pixels.A probability measure P_{pos} is computed for each subwindow & if the probability measure is lesser than the threshold value, then the subwindow is rejected .Although , the method is slower as compared to variance filter, but fast as compared to classification method using SIFT feature.The method of feature calculation is as follows.From the fig. shown, a sample image is to be classified.Below this image,there are four boxes shown in which black and white dots are referred to as pair of pixels in the original image .The dots position are drawn out of uniform distribution at the start and remain constant . It is now tested for each of the boxes whether in the subwindow,the pixel at white dot position is brighter than the pixel at the black dot position.

This can be mathematically expressed as

$$y_i = \begin{cases} 0, & if\ I(x_{i,1}) < I(x_{i,2}) \\ 1, & otherwise \end{cases} \quad (18)$$

$x_{i,1}$ and $x_{i,2}$ specify two random locations.The above comparison is invariant against constant variations of brightness.The result of these comparison are concatenated into a binary order.The value at i^{th} bit of a no. is determined by i^{th} feature.

From the obtained feature value , the probability P(y=1/F) is retreived, where y = 1 indicates that the subwindow has a positive class label.

3.5) Nearest Neighbour Classifier & Template Matching :

The last stage of cascade detector uses template matching and 1NN (nearest neighbour) classification.

All patches are resized to 15* 15 patches.The Normalised Correlation Coefficient (NCC) is used for comparing two patches Q_1 & Q_2.

$$ncc(Q_1, Q_2) = 1/n\text{-}1 \sum_{x=1}^{n} \frac{(Q1(x)-\mu 1)(Q2(x)-\mu 2)}{\sigma_1 \sigma_2} \quad (19)$$

NCC value lies between -1 and 1.When the two patches are similar, NCC value is closer to 1.The distance between two patches is defined in such a way that yields values between 0&1.

$$d(Q_1,Q_2) = 1 - 1/2((Q1,Q2)+ 1) \quad (20)$$

The templates are maintained for both positive and negative class as P^+ and P^- respectively.The sequence Multi Face Turning is used for learning positive and negative templates.For an image patch given with an unknown class label, the distances to the positive class and negative class is computed as

$$d^+ = \underbrace{min}_{Qi \epsilon Q+} d(Q_0,Q_i) \quad , \quad d^- = \underbrace{min}_{Qj \epsilon Q-} d(Q_0,Q_i) \ ,$$

Both distance fuse together into single value as

$$p^{+} = d^{-} / d^{-} + d^{+} \quad (21)$$

Above equation represents the confidence ($\mathbf{p}^{+}$) whether belongs to the positive class .If (p^{+}) confidence value is greater than $\boldsymbol{\theta}^{+}$,it implies that patch belongs to positive class.

3.6) Non Maximal Suppression

Post cascade detector stage, more than one subwindow can be considered as possible candidate for object.According to Blaschko, it is difficult to select the subwindow with higher confidence, since other local maxima may be ignored.Thus non maximal suppression strategies are employed to identify relevant local maxima.In non maximal suppression , cluster detections are used on the basis of their spatial overlap.The bounding boxes are averaged for each of the cluster and compressed into a single result.

From figure (), for computing the overlap between 2 bounding boxes , the formula from the PASCAL challenge is used.

$$\text{Overlap} = \frac{B1 \cup B2}{B1 \cap B2} \quad (23)$$

Where B1 = area of first Bounding Box , B2 = area of second Bounding Box , I = area of intersection of 2 Bounding Boxes

The overlap is bounded between 0 & 1.

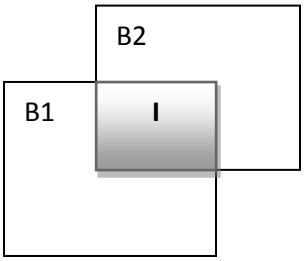


FIG 4: Overlap measure between bounding boxes B1 & B2

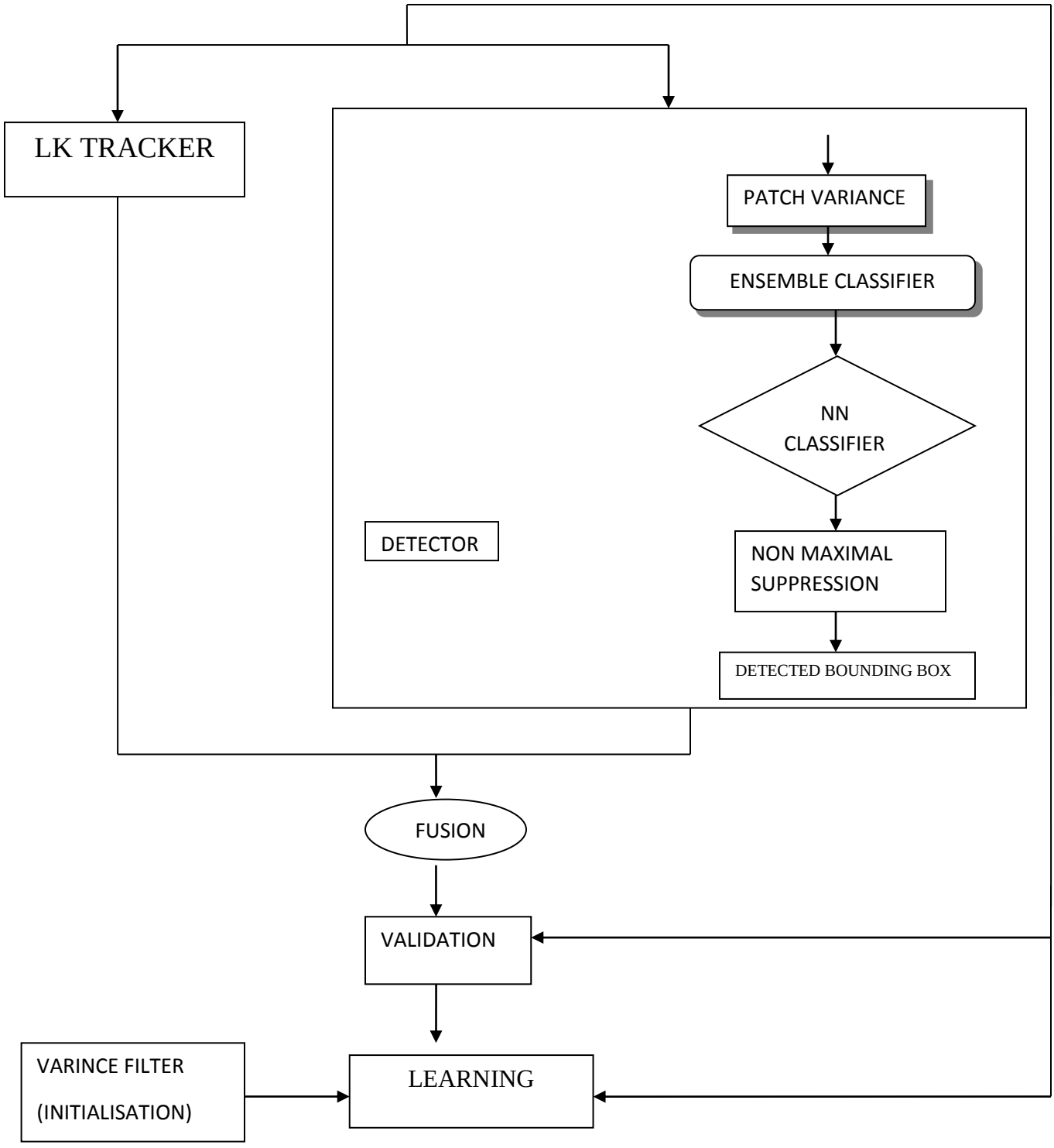


FIG 5 : ARCHITECTURE OF TLD

Chapter 4

Learning

4.1 Introduction

When we process an image , both tracker and detector run simultaneously.While processing , the variance filter threshold are not adaptive but the ensemble classifier are trained online.In learning step , the template update issue is solved with the use of certain criteria (rules) which are to be fulfilled to give final results. It assess classifier and eliminates those examples that have been classified in inconsistency with the two constraints and improves the training set with the revised samples in a repeated process.

4.2 PN Learning

PN learning is a kind of online learning strategy whose purpose is to improve the performance of detector in video stream.PN learning uses semi-supervised learning that has both labeled examples as well as unlabelled data. Semi-supervised learning method uses the information embedded in training data as supervising information.The class distribution takes place by unlabelled data and updates the classifier from the class separation used as a training set.

It is not necessary to update classifier in each frame however only when some criteria is fulfilled.Thus Saves a lot of computational cost.From the current frame of video stream ,the detection error is evaluated and is used to update the classifier which avoids the similar error in the upcoming frames.
It comprises of 4 parts :- A classifier,training set,supervised learning and PN experts.

For binary classification , the training data is extracted from unlabelled data by structural constraints.The two constraints in P/N learning are

P constraint analyses false negatives.It means that the examples are labelled negative.Similarly **N- constraint** analyses false positives.

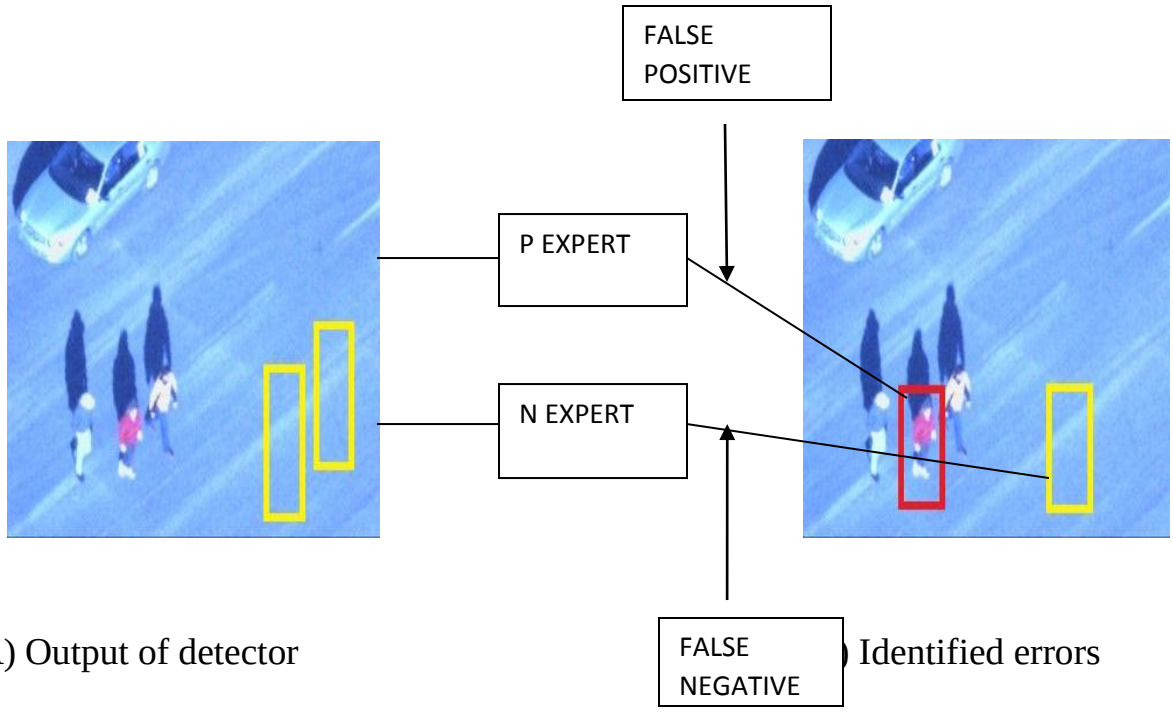


Fig 6 : Illustration of P-N experts: A) Output of Pedestrian2 detector ,B)Errors detected by PN experts

4.3 Formalisation : From the figure 7,X_s represents unlabelled data.The unlabelled data is classified using classifier and accredits label Y_s to X_s .The misclassified examples X_d (labels Y_d) are identified with the help of structural constraints.Now the examples are added to the training set which results in updation of classifier.

For P-constraint , the highly overlapping patches with the final output comes under positive examples (overlap measure is less than 20%).Similarly for N-constraint ,all the patches that are not overlapping with the final output comes under negative examples (overlap measure is less than 20%) . If the overlap is present for atleast 60% , we assume that bounding box B highly overlapped with B_t .But for bounding box B to not overlap with B_t ,the overlap should be smaller than 20%.

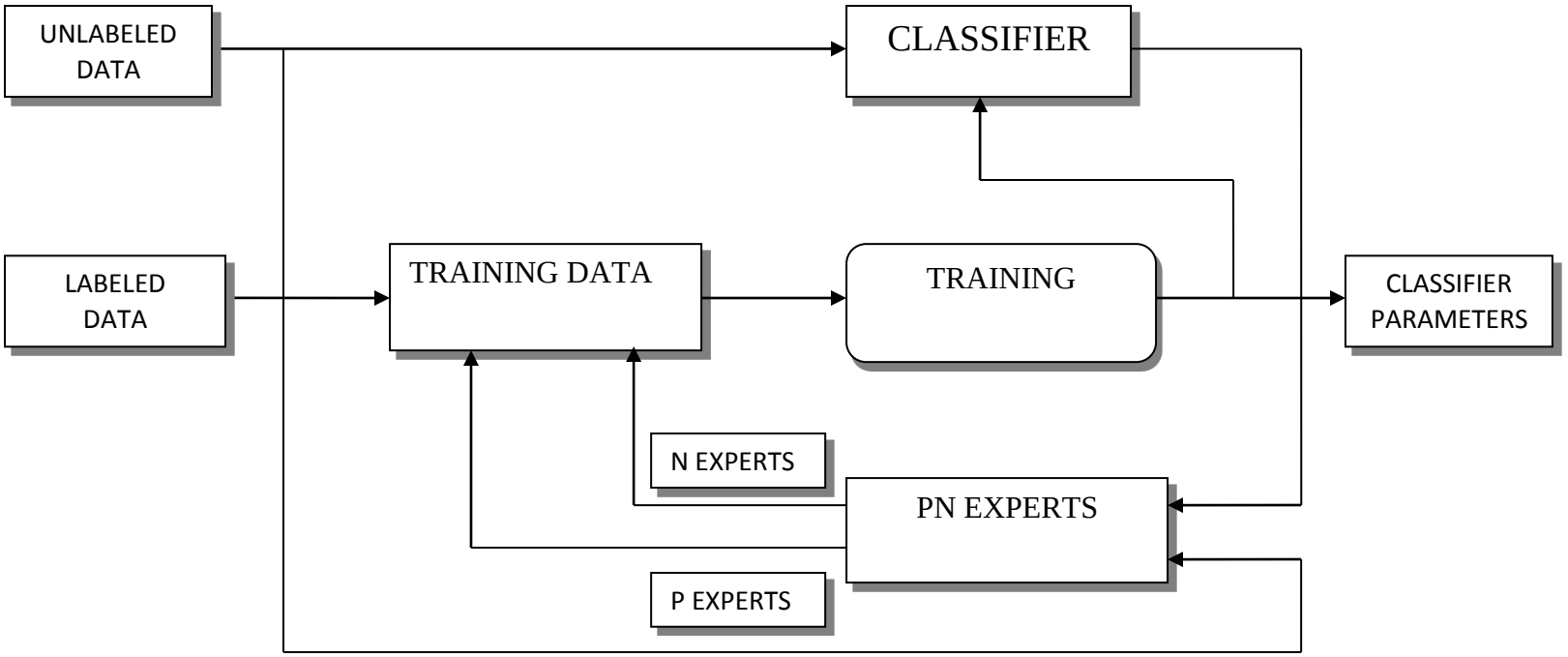


Fig 7: The block diagram of the P-N learning

4.4 Stability

The stability of learning affected by errors caused by PN experts .Thus for analysis of purpose , it is supposed that ground truth labels of Xu are known. Now , we can evaluate the errors made by the classifier .Initially a classifier is used for classification of unlabeled data after then the classification is corrected on the basis of PN experts output . The no. of False Negative (FN) and False Positives(FP) defines the classifier performance .Thus reducing these errors is the main aim of PN learning. The ouput of P experts are correct positives examples and false positive examples.Thus classifier changes negative examples to positive.

Similarly the ouput of N experts are $c_N^-(k)$ which are correct (negatives based on ground truth) and $f_N^-(k)$ which are false (negatives based on ground truth).Thus classifier output changes as

$$n^-(k) = c_N^-(k) + f_N^-(k) \qquad (24)$$

In next iteration, the no. of false positives and false negative errors becomes

$$P_F(k+1) = P_F(k) - c_N^-(k) + f_N^+(k) \quad (25)$$

$$N_F(k+1) = N_F(k) - c_N^+(k) + f_N^-(k) \quad (26)$$

If $c_N^-(k) > c_f(k)$, false positive decreases i.e. the examples correctly relabelled to negative is greater than the examples that were incorectly relabelled to positive .Similarly to, $P_F(k+1)$, False negatives also decreases if $c_N^+(k) > f_N^-(k)$.

QUALITY MEASURES :

The quality of PN experts is defined by 4 quality measures.

P precision : It is defined as ratio of number of corrected positives examples to the number of all positives examples output (P expert).

$$P^+ = \frac{c_n^+}{c_n^+ + f_n^+} \quad (27)$$

P recall : Indicate False negative errors.It is ratio of the corrected positive examples to the False negatives.

$$R^+ = \frac{c_n^+}{FN} \quad (28)$$

N precision : It is ratio of corrected negative examples to the number of all positive example output (N expert) .

$$P^- = \frac{c_n^-}{c_n^- + f_n^-} \quad (29)$$

N recall : Indicate False Positive errors i.e. number of corrected negative examples to the number of all false positives.

$$R^- = \frac{\mathbf{c_n^-}}{\mathbf{FP}} \tag{30}$$

From above quality measures i.e. The correct and false examples output at iteration k have the form.

$$c_n^+(k) = R^+FN(k)\ ,\ f^+(k) = \frac{(1-P^+)}{P^+} R^+FN(k) \tag{31}$$

$$c_n^-(k) = R^-FP(k)\ ,\ f^-(k) = \frac{(1-P^-)}{P^-} R^-FN(k) \tag{32}$$

$$FP(k+1) = (1 - R^-)\,FP(k) + \frac{(1-P^+)}{P^+} R^+FN(k) \tag{33}$$

$$FN(k+1) = \frac{(1-P^-)}{P^-} R^-FN(k) + (1 - R^+)\,FN(k) \tag{34}$$

$$\text{State vector} \quad \vec{x}\,(k) = [\ FP(k) \qquad FN(K)]^T \tag{35}$$

$$M = \begin{bmatrix} (1 - R^-) & \frac{(1-P^+)}{P^+} R^+ \\ \frac{(1-P^-)}{P^-} R^- & (1 - R^+) \end{bmatrix} \tag{36}$$

Rewriting the above equation as $\vec{x}(k+1) = M\,\vec{x}(k)$

From above recursive equation, the state vector $\vec{x}$ converges to 0 if transition matrix eigen values λ_1 , λ_2 are less than 1.

4.5 Fusion & Validity

In fusion , we combine the result of recursive tracker T_r and detections T_d into a final result Tb.On the basis of number of detections and their confidence values P_{Td}^{+} and confidence of tracking results P_{Tr}^{+} , we take the decision.

The confidence P_{Tr}^{+} is obtained from template matching method implemented on tracking result.If the confidence value of detector from exactly one result is higher than the value of confidence obtained from recursive tracker , then final result constitutes the response of detector .This is known as reinitialistion of recursive tracker.But if the recursive tracker is not reinitialised and tracker estimates a result .Then the recursive tracker result constitutes final output. Since the detector yields more than one detection or there is exactly one detection which has confidence value less than the confidence value of recursive tracker.

For the remaining cases , no bounding box is present which implies that the object is not present in the ongoing frame or scene.The final output is valid for both of the conditions where the recursive tracker is not reinitialised and for all other cases , the result is not considered.

1) The final output is also valid for the conditions when the previous result was considered valid and if the recursive tracker has confidence value greater than $\boldsymbol{\theta}^{+}$.

2) The final output is also valid for the conditions when the previous result was considered valid and if the recursive tracker has confidence value greater than $\boldsymbol{\theta}^{-}$.

   The threshold $\boldsymbol{\theta}^{+}$ refers that the output comes under positive class and threshold $\boldsymbol{\theta}^{-}$ the output comes under negative class.

Chapter 5

Tracking-Learning-Detection

5.1 FRAMEWORK

As shown in below figure,TLD framework has four components : tracker, learning, detector and integrator. TLD distinguishes two modes: (i) initialization, and (ii) run-time.

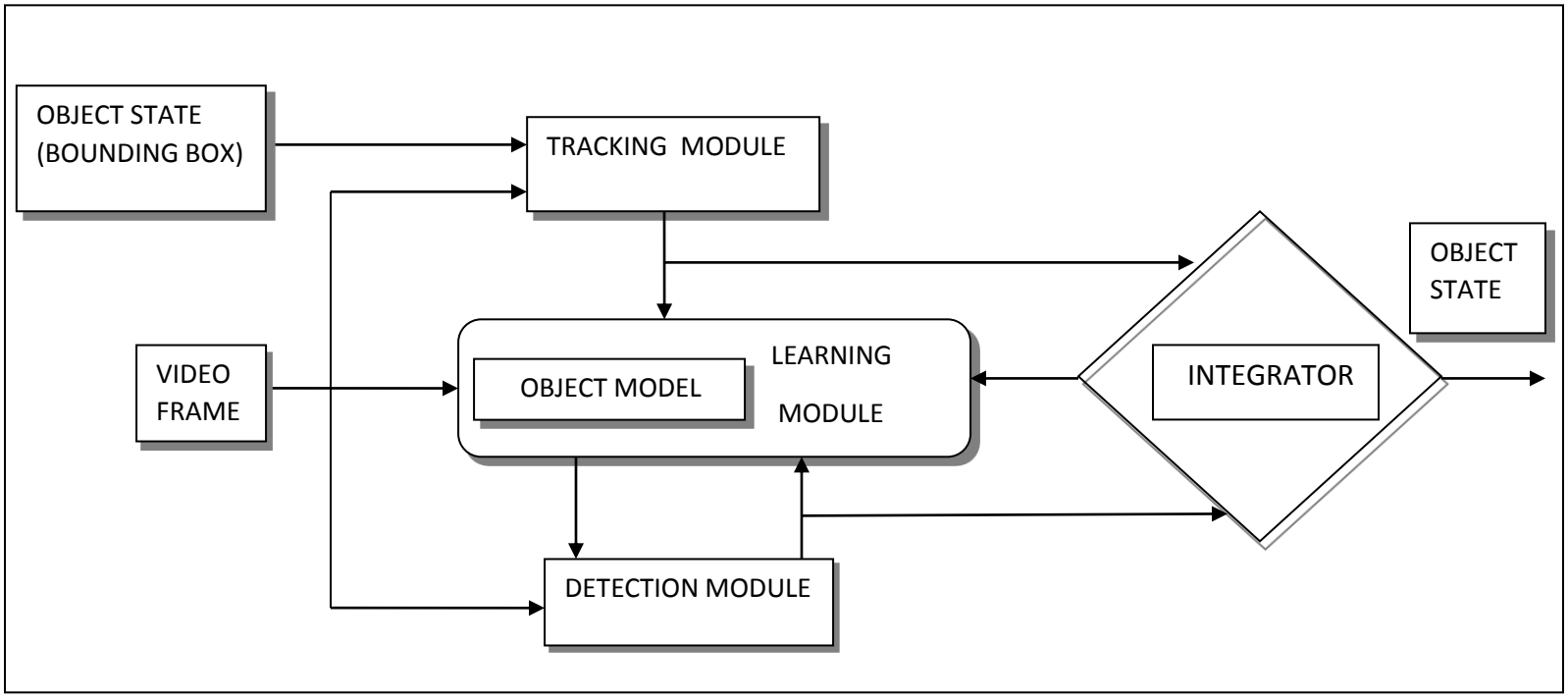


Fig8: Block diagram of the TLD framework.

5.1.1 : INITIALIZATION : In initialization model , the first frame is acquired and its object state is represented by a bounding box. The framework also accept images describing the object from multiple scenes or background images for which the object is absent. Thus following operations are :

1) **Tracker initialisation** : It comprises of fixing the initial state(i.e. template extraction).

2) **Object model initialization** : The examples of object and background are inserted into the object model.

3) **Detector initialization** : The object detector is trained to localize the appearences in the object model.The resulting detector is referred to as initial detector.
   Now,the TLD framework is ready for frame by frame processing of video stream.

5.1.2 : RUN TIME : In run time mode , after acquiring video frame ,the framework sends it parallerly to its all the components except integrator.depends on its previous state, the tracker estimates the motion of object and gives a single hypothesis.Similar to tracker , the detector also gives a number of hypotheses regarding location of target.The output of both tracker and detector goes to the integrator which combines them into the final output.After then the tracker ,detector and integrator evaluation is carried out in learning block.The learning block assess error and revises the detector to tune out these errors.

5.2 IMPLEMENTATION

5.2.1 Object Representation

A) **Object State :** The state of object at any instant is represented by boundary box.The aspect ratio is fixed for bounding box(initial bounding box).The parameter such as location and scale are considered.The parameter such as location and scale are considered. The similarity between two bounding box is evaluated using overlap measure.The overlap measure is specified as a ratio of intersection and union of two bounding boxes.

B) Object Appearance : Given the object bounding box, the image patch P is sampled.After then the patch P is resampled to normalized resolution about aspect ratio.The similarity measure between two patches is given by

$$S(P_i, P_j) = 0.5(NCC(P_i, P_j) + 1)$$

NCC : Normalised Correlation Coefficient.The similarity measure ranges from 0 to 1.

C) Object trajectory : The trajectory of an object is defined by sequence of object states .The trajectory splits up since the object may not be detected every time.

5.2.2 Object model

The data structure represented by object appearance and its vicinity noted so far.The positive patches (P_1^+ , P_2^+,……. P_M^+) and negative patches (P_1^- , P_2^-,……. P_n^-) form together object model.

P_1^+ : It represent first positive patch in the collection.
P_1^+ : It represent last positive patch in the collection.

Given an arbitrary patch P , there are several similarity measures defined in object model M,which indicates the resemblance of patch with the object appearences.

1) **Resemblance With Positive Nearest Neighbour** :

$$S^+(P,M) = \max_{P_i^+ \epsilon M} S(P, P_i^+)$$

2) **Resemblance With Negative Nearest Neighbour :**

$$S^-(P,M) = \max_{P_i^+ \epsilon M} S(P, P_i^-)$$

The other two measures are Relative and Conservative Similarity.

3) Relative Similarity : It ranges from 0 to 1.

$$S^{r} = \frac{s^{+}}{s^{+} + s^{-}}$$

4) **Conservative Similarity :** It also ranges from 0 to 1.

$$S^{\theta} = \frac{s^{\Theta}}{s^{\Theta} + s^{-}}$$

In nearest neighbor , if the relative similarity measure is greater than $\boldsymbol{\theta}_{nn}$,patch P is classified as positive . With the help of parameter $\boldsymbol{\theta}_{nn}$, the nearest neighbor classifier is adaptive toward either recall or precision.The confidence of classification represented by margin as $S^{r}(P,M) - \Theta'_{nn}$.

MODEL UPDATE : In model update, the nearest neighbor classifier classifies the patch and added to collection in case of incorrect classification.This leads to the decrease in amount of accepted patches but the decision boundary becomes coarse.Therefore the strategy is changed by adding those patches for which classification margin is less than λ.If value of λ increases , more patches are accepted for decision boundary to become better.

Object tracker and detector are already discussed in Chapter 2 and Chapter 4 .

INTEGRATOR : The purpose of integrator is to combine the output coming from the tracker and detector response into a single response .If there is no response coming from both tracker and detector ,the object is considered as not visible.Otherwise the output is bounding box with a maximum confidence score.

CHAPTER 6

RESULTS

The framework is evaluated for both single target tracking and multitarget tracking.For performance evaluation ,the standard metrics are employed i.e. Recall and Precision.The C++ opencv files are implemented as MEX files of MATLAB under MSVisual StuDIO 2012,MATLAB 2015,Windows 32 bit.

6.1 EVALUATION PROTOCOL :

The overlap measure is used to compare the algorithm output and ground truth values.Using this measure the translations are penalizes in both directions and scale.The frames of sequence are classified as one of the five cases based on the overlap measure.For a given threshold ω , if overlap measure is greater than ω ,the result is classified as True Positive.The result is classified as False Negative, if ground truth exists but there is no output comes from algorithm.The opposite holds for False Positive.But ,if overlap measure is lower than ω, the result is categorized as False Negative and False Positive.True Negatives exists for a case if neither algorithm output nor ground truth exists.

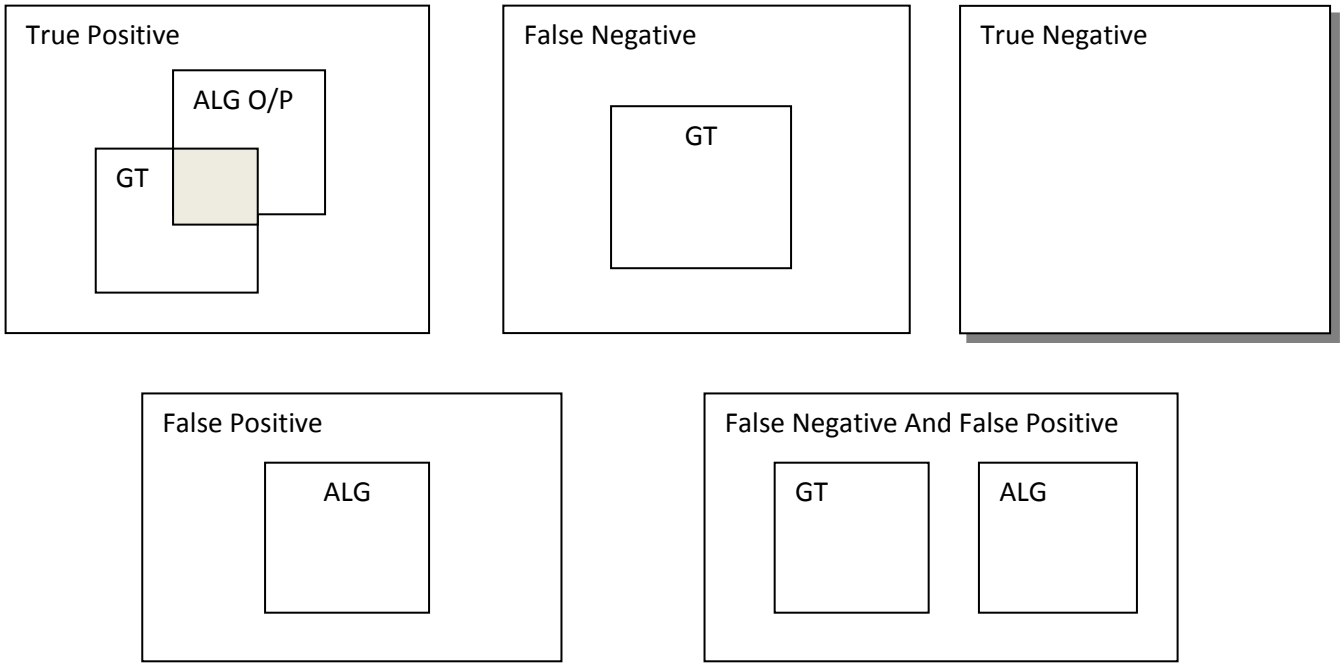


Fig9 :Output cases when detection output is compared to ground truth

6.2 SEQUENCES :

FIG 10 : Dataset used for evaluation of TLD

SEQUENCES	TOTAL NO. OF FRAMES	IMAGE SIZE	MOVING CAMERA	SCALE & ILLUMINATION CHANGE	SIMILAR OBJECTS	OCCLUSION (Partial)
1. Car	945	320*240	yes	no	yes	yes
2. Pedestrian 1	140	320*240	yes	no	yes	no
3. Pedestrian 2	338	320*240	yes	no	no	no
4. Pedestrian 3	184	320*240	yes	no	yes	yes
5. Jumping	313	320*240	yes	no	no	no
6. Carchase	9928	290*217	yes	yes	yes	no
7. Motocross	2665	320*240	yes	yes	yes	yes
8. Panda	3000	312*233	yes	yes	yes	no
9. David	761	320*240	yes	yes	no	yes
10. Volkswagen	8576	640*240	yes	yes	yes	yes

Fig. 11 : Properties of sequences used for evaluation

6.3 COMPARED ALGORITHMS :

1) **Multiple instance learning :** In this algorithm ,the approach of online boosting is used but improvement over drift update is used.

2) **CoGD (Co-trained generative and discriminative) :** It is a kind of particle filter in which co-training of pair of classifiers is performed.

3) **Beyond semi supervised online boosting :** In BSOB , simultaneously 3 classifiers are trained for better adaptability and stability .

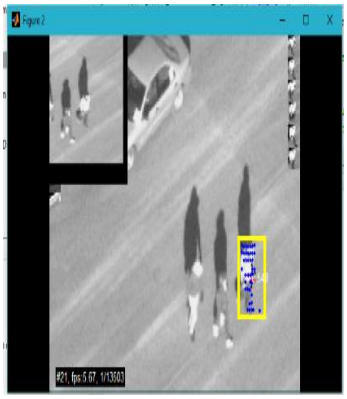
SEQUENCES	BSOB	MIL	CoGD	Our approach
1. Car	0.2	0.72	0.96	0.94
2. Pedestrian 1	0.1	0.69	1	1
3. Pedestrian 2	0.04	0.11	0.81	0.91
4. Pedestrian 3	0.62	0.68	0.92	0.99
5. Jumping	0.12	0.87	1	1
6. Carchase	0.19	0.07	0.08	0.61
7. Motocross	0.0	0.01	0.42	0.78
8. Panda	0.25	0.3	0.09	0.55
9. David	0.21	0.09	1	1
10. Volkswagen	0.0	0.06	0.11	0.84

Fig 12:F-score of TLD, BEYOND SEMIBOOST,MIL,C0GD

6.3 RESULTS

CARCHASE

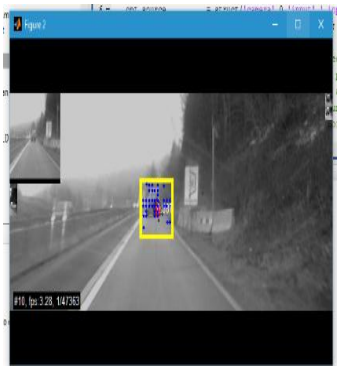
MOTOCROSS



PEDESTRIAN 2

PEDESTRIAN 3

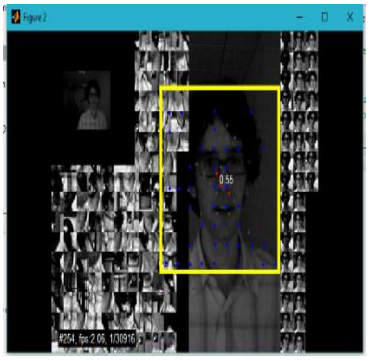
PANDA



VOLKSWAGEN

CAR

JUMPING

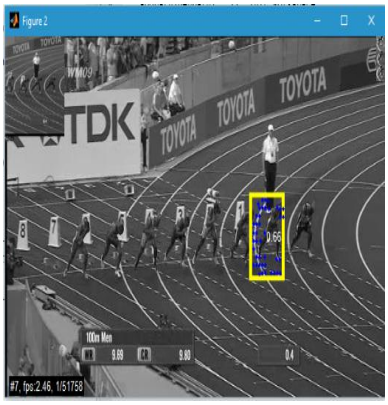


DAVID

PEDESTRIAN 1

CARDARK

FISH



JOGGING

FIG 13 : TRACKING RESULTS OF FRAMEWORK ON TLD AN VOTR DATASETS

The algorithm is also performed on PETS dataset .Here sequence of view_003 of Pets 2009 dataset is used .

Frame 6 **Frame 28** **Frame 127**

FIG 14 : TRACKING RESULTS OF FRAMEWORK ON PETS VIEW 003 SEQUENCES

Here the described algorithm is also used for multi object tracking but it becomes fail in case of occlusion or similar appearance changes.

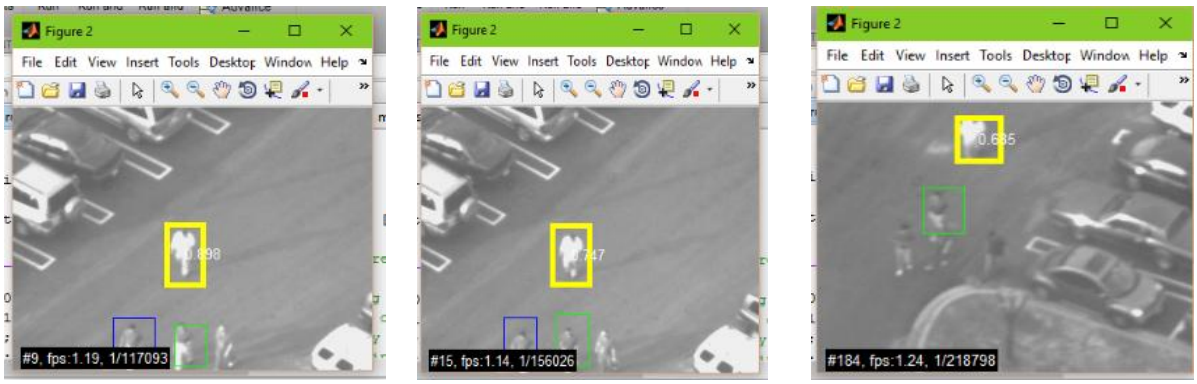


Fig 15 : Detection output of TLD for tracking multiple object

CHAPTER 7

CONCLUSION :

In this work ,we are extending the Kalal approach of Open TLD.Using Opentld as a benchmark for tracking object,showing that using feature descriptor improve the detection process.For single target sequences in which occlusion is present ,our approach outperforms adaptive tracking and detection process.Thus for single target ,TLD performs well but in case of multiple targets,it need some improvements.This method(for multiple targets) becomes fail when group of multiple object overlap with each other.In future we can work on improvement in learning process so that tracking is improved for multiple targets also.

The major problem occurs in multiple object tracking is that the detector unable to discriminate between objects which has similar appearance.The shortcomings of this framework are also associated with bounding box.In bounding box,the rectangular area covering the object may also contain background region.Thus object of interest may not correctly recognised when it will be detected in a another background.Thus to distinguish object from background for learning,segmentation approach can be used in the future.In the above approach of TLD for object tracking ,training is performed in detector only.Thus the same errors are caused by the tracking component.We can extend this approach by giving training to the tracking component also.

REFERENCES

1. H. Wang, Q. Xiao, Q. Ye and X. Wang, "Cross Camera Object Tracking in High Resolution Video Based on TLD Framework," Multimedia Big Data (BigMM), 2015 IEEE International Conference on, Beijing, 2015, pp. 264-267.

2. Jianbo Shi and C. Tomasi, "Good features to track," Computer Vision and Pattern Recognition, 1994. Proceedings CVPR '94., 1994 IEEE Computer Society Conference on, Seattle, WA, 1994, pp.593-600.

3. S. Avidan, "Ensemble Tracking," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, no. 2, pp. 261-271, Feb. 2007.

4. Z. Kalal, J. Matas and K. Mikolajczyk, "Online learning of robust object detectors during unstable tracking," Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on, Kyoto, 2009, pp. 1417-1424.

5. R. T. Collins, Yanxi Liu and M. Leordeanu, "Online selection of discriminative tracking features," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 10, pp.1631-1643,Oct.2005.

6. B. Babenko, M. H. Yang and S. Belongie, "Visual tracking with online Multiple Instance Learning," Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, Miami, FL, 2009, pp. 983-990.

7. Z. Kalal, K. Mikolajczyk and J. Matas, "Tracking-Learning-Detection," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 7, pp. 1409-1422, July 2012.

8. Y. Wu, J. Lim and M. H. Yang, "Online Object Tracking: A Benchmark," Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, Portland, OR, 2013, pp. 2411-2418.

9. E. Park, H. Ju, Y. M. Jeong and S. Y. Min, "Tracking-Learning-Detection Adopted Unsupervised Learning Algorithm," Knowledge and Systems Engineering (KSE), 2015 Seventh International Conference on, Ho Chi Minh City, 2015, pp. 234-237.

10. Q. Yu, T. B. Dinh, and G. Medioni,"Online Tracking and Reacquisition Using Co-trained Generative and Discriminative Trackers" ,2008, pages 678–691, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg.

11. D. Jepson, D. J. Fleet and T. F. El-Maraghi, "Robust online appearance models for visual tracking," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25, no. 10, pp. 1296-1311, Oct. 2003.

12. Babenko, M. H. Yang and S. Belongie, "Visual tracking with online Multiple Instance Learning," Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, Miami, FL, 2009, pp. 983-990.

13. W. Hailong, W. Guangyu and L. Jianxun, "An improved tracking-learning-detection method," Control Conference (CCC), 2015. 34th Chinese, Hangzhou, 2015, pp. 3858-3863.

14. Z. Kalal, K. Mikolajczyk and J. Matas, "Forward-Backward Error: Automatic Detection of Tracking Failures," Pattern Recognition (ICPR), 2010 20th International Conference on, Istanbul, 2010, pp. 2756-2759.

15. A. W. M. Smeulders, D. M. Chu, R. Cucchiara, S. Calderara, A. Dehghan and M. Shah, "Visual Tracking: An Experimental Survey," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 36, no. 7, pp. 1442-1468, July 2014.

16. Shuai Cheng, Guangwen Liu, Junxi Sun,“Robust and Fast Tracking-Learning-Detection,” csic-15. Advances in Computer Science Research,volume 16,ISSN 2352-538x.

17. C. Sun, S. Zhu and J. Liu, "Fusing Kalman filter with TLD algorithm for target tracking," Control Conference (CCC), 2015 34th Chinese, Hangzhou, 2015, pp. 3736-3741.

18. Yang Hua, Karteek Alahari, Cordelia Schmid.”Occlusion and Motion Reasoning for Long-term Tracking”, ECCV 2014 - European Conference on Computer Vision, Sep 2014, Zurich,Switzerland. Springer, 2014.

19. Sheetal Balsaraf , Uday Joshi ,”Implementation of Generic Object Tracker based on TLD Framework, using Generic tools ,” International Journal of Computer Applications (0975 – 8887) Volume 79 – No 16, October 2013.