Fast Visual Tracking Using Spatial Temporal Background Context Learning

Asif Mukhtar[1]
[1]Department of Telecom Engineering,
Institute of Communication Technology,
University of Engineering and
Technology, Peshawar Pakistan
Email: asif.nufast [AT] gmail.com

Arslan Majid[2]
[2]Departments of Electronics Engineering,
Iqra University Islamabad, Pakistan

Kashif Fahim[3]
[3]Department of Telecom Engineering,
Institute of Communication Technology,
University of Engineering and
Technology, Peshawar, Pakistan

Abstract—**Visual Tracking by now has gained much provenience among researchers in recent years due to its vast variety of applications that occur in daily life. Various applications of visual tracking include counting of cars on a high way, analyzing the crowd intensity in a concert or a football ground or a surveillance camera tracking a single person to track its movements. Various techniques have been proposed and implemented in this research domain where researchers have analyzed various parameters. Still this area has a lot to offer. There are two common approaches that are currently deployed in visual tracking. One is discriminative tracking and the other one is generative tracking. Discriminative tracking requires a pre-trained model that requires the learning of the data and solves the object recognition as a binary classification problem. On the other hand, generative model in tracking makes use of the previous states so that next state can be predicted. In this paper, a novel tacking based on generative tracking method is proposed called as Illumination Inavariant Spatio Temporal Tracker (IISTC). The proposed technique takes into account of the nearby surrounding regions and performs context learning so that the state of the object under consideration and its surrounding regions can be estimated in the next frame. The learning model is deployed both in the spatial domain as well as the temporal domain. Spatial domain part of the tracker takes into consideration the nearby pixels in a frame while the temporal model takes account of the possible change of object location. The proposed tracker was tested on a set of 50 images against other state of the art four trackers. Experimental results reveal that our proposed tracker performs reasonably well as compared with other trackers. The proposed visual tracker is both efficiently with respect to computation power as well as accuracy. The proposed tracker takes only 4 fast Fourier transform computations thus making it reasonably faster. The proposed trackers perform exceptionally well when there is a sudden change in back ground illumination.**

Keywords- Visual Tracking, Context Learning, Spatial Temporal, Confidence Map, Fast Fourier Transform

I. INTRODUCTION

Visual Tracking is defined as keeping check of an object, person or a thing in a video or in sequences of images when the position is moved from one frame to another [1]. A tracker assigns appropriate label of the objects that are tracked in a snap shot of a scene or a video. Tracking of objects that are in motion or in a scenario where the background is constantly changing is often a tedious task due to changes in object structures, loss of information when a 3-dimensional world information is mapped on to a 2-dimensional image plane and variation in illumination intensities or complex object motion. In most of the algorithms that are based on different tracking methodology, the basic assumption is that the motion of object is smooth in nature. This assumption means that there are no changes that are abrupt in nature [1]. Normally, previous information which in mathematical terms is referred to prior knowledge consists sizes of the objects, number of objects and their corresponding shapes and appearances are required.

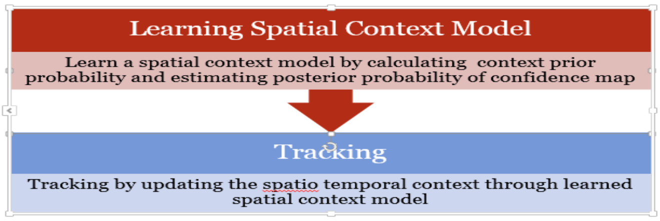
It's been one of the most exciting and thriving research areas in the space of modern Computer Vision and has many applications like augmented reality, surveillance and object detection and recognition [1]. One of the major issues that are usually encountered while tracking an object is to tackle the variations in appearances and exterior conditions of the target particularly due to variation in the shape or geometry or pose of the object. Many a times, it often happens that object appearance may suddenly change shape such that the tracker cannot keep check of the target object. This problem has been addressed by many researchers in recent years; yet it still remains a challenge as far as efficiency and accuracy is concerned [1].

One of the major tasks in Visual Tracking is predicting the target in the incoming frames. There may be cases that an object may experience geometric changes, intensity changes, scale changes, rotations, fast motion, and noise in the images. As a result of these changes, object may lose its trajectory and the tracker may drift away. It is imperative that the prediction of the targeted subject in the immediate next frame must be carried out. Prediction of the object is usually solved by formulation of a probabilistic model where the probability of the object is calculated in the next frame. In related research field like computer vision, pattern recognition and artificial intelligence, the process of prediction is the formulation of learning and classification problem. There have been two types of methods that deal with the classification difficulties. One is generative approach and the other one is discriminative approach. The primary difference between these two approaches is that generative learning method calculates the Joint Probability Distribution p(x,y) of the available data whereas the Discriminative Method calculates the Conditional Probability Dstribution p(y|x) of the data. Basically, a generative algorithm simulates that how the data was actually generated while categorizing a certain signal while on the other hand, the discriminative algorithm simply categorizes a given signal [2].

Based on the taxonomy of machine learning, Visual Tracking can be classified into approaches. These approaches are generative tracking and discriminative tracking [3].

II. RESERCH METHOD

The tracking problem in the proposed methodology is based on the algorithm suggested by Shang Fe et al [4]. In tracking formulation, the object under consideration takes account of the contribution from its neighboring regions in a frame of a sequence. This method is usually referred to as context modeling. The tracking algorithm intended for the target actually suffers if tracker loses the trajectory of the object. It is difficult to recover and resume the correct tracking response. However, with the context modeling approach, it is possible in predicting the possible position (or location) of that particular target object in immediate coming frame of a sequence. Figure 2.1 shows the overall proposed methodology stages.



Learning Spatial Context Model

Learn a spatial context model by calculating context prior probability and estimating posterior probability of confidence map

Tracking

Tracking by updating the spatio temporal context through learned spatial context model

Figure 2.1: Spatial Temporal Context Tracker Stages.

As seen in the figure 2.1 the proposed method is broken into two major working blocks. The blocks are

- Learning the spatial context model using probabilistic learning
- Tracking based on the spatial model from the previous continuous learning of the frames

During the tracking formulation, region having a local context which being consisted of a targeted subject and its instant next background surrounding in a pre-defined fixed region that is determined. Thus, there exists a very staunch relationship between all the corresponding images and encompassing the target object in successive frames both in spatial domain and also in its temporal properties. For instance, the target can undergo a very heavy opaque object as seen in the occlusions. This type of impediment very drastically changes the appearance of the object. But when considering the local context target object may not change its appearance and a very small portion of the target may be occluded. In this case the tracker is efficiently able to handle changes in appearance such that only a small region of the context region gets occluded thereafter. This assists in discriminating the target from its background upon many changes in its appearance. Lately, numerous methods [4]– [5] exploit the information extraction based on the context and the background regions to facilitate Visual Tracking with pretty good and successful results in terms of efficiency and accurate estimations. However, they do have a cost as these approaches require high computational loads for example a very large chunk of memory must be reserved for the saving of learned data for extraction of features in the training as well as tracking phases. This kind of approach is partially feasible to a very small extent.

Figure 2.2 is shows the flow of algorithm developed for learning the spatial context as given. In the initial stage, the deconvolution method is deployed so that image can be decomposed based on their spatial properties. The background which is local in this case is separated from the spatial context model and we will get an initial estimate of the background scene. The context model that was extracted will be successively passed to update its surrounding regions. Learning of the context model is achieved by this process. After that spatial model would then be updated accordingly to also include the temporal region properties in the next in coming frame of the video. A base calculation function known as a confidence map function merges the spatial as well as temporal properties in the frame. This can be achieved by performing convolution with the first frame and the updated frame of spatial temporal region. Thus, we can set a boundary for the tracker process. In order to search for the finest location as closest to the target, it is necessary that the confidence map should be maximized.

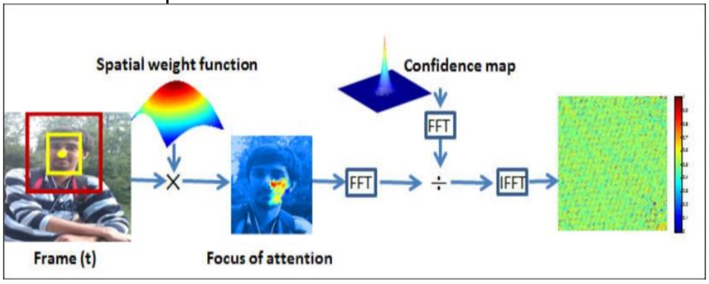


Figure 2.2 Learned spatial context at frame (t)

For tracking problem, a confidence map needs to be found at the start. The location likelihood function of the object can be found by confidence map estimation. The object likelihood is given by equation

$$m(x) = P(x|o)$$

(2.1)Where x ∈ R^2 specifies at which location the object is present in the present frame and “o” is the presence of the concerned object in the scene. Following to this, the spatial context information has used the likelihood function as computed in equation (4.1) to get details of the spatial context information. The graphical model representation of the flow is specified in Figure 2.3.

The tracking object center coordinates are denoted by in the existing frame. Thus, we now have the certain location of object at this specified point. The feature set of the context region is defined by the relation as X_c = {c(z) = (I(z), z)|z ∈ Ωc(x^*)} where intensity of image is specified by I(z) at location, z is the neighborhood of location x^* and the neighborhood the location x^* is specified by the symbol Ωc(x^*). And if the joint probability function P(x, c (z) |o), then this probability function of the target object in equation can be further decomposed by successive summation of its probability function which are given in equations 2.1 2.2 and 2.3.

$$m(\mathrm{x}) = P(\mathrm{x}|o)=\mathrm{P}_{\mathrm{c(z)}\in X_c}P(\mathrm{x},\mathrm{c(z)}|o) \qquad (2.2)$$

$$=\mathrm{P}_{\mathrm{c(z)}\in X_c}P(\mathrm{x}|\mathrm{c(z)},o)P(\mathrm{c(z)}|o) \qquad (2.3)$$

The relationship in spatial domain among its context and the object location is represented by the Conditional Probability Function P(x|c(z), o). The big benefit of this probability function is that it can resolve any ambiguities when there are different settings with different image measurement parameters. The prior probability P(c(z)|o) is a prior context, in our case the probability which takes account of the context works out to find the local context's appearance. Our basic task in all of this is to basically estimate the posterior probability P(x|c(z),o) which is learning function so that difference of the context location and concurrently the target location can be minimized. The closer or smaller the vale, the better the result of tracker.

Now, the Conditional Probability Function
P(x|c(z),o) in (2.3) is defined as
$P(\mathrm{x}|\mathrm{c(z)},o)=h^{sc}(\mathrm{x}-\mathrm{z})$ (2.4)

Where h^{sc}(x − z) is such a function that formulates a comparative relationship of the distance along with its d=orientation in the frame between target location at x and the context of local region which is present at z. This formulates a spatial relationship between two regions that is the object and its spatial context which is background in our case. As the function is calculating different spatial information along in its local context region and the target objects it is helpful in isolating the smaller objects that are present in its nearby regions. As an example, presented in figure 2.3 we can see a person face with two eyes. While tracking an eye which is based only on appearance (denoted by z_l). Now in this case if we are to track eye based on appearance denoted by the tracker may pick the wrong object in this case the other eye as bot the eyes and the background region surroundings have similar appearances. This may result in problem when the searching region is having a large background or if an object is having a very fast motion. In the proposed algorithm however both of the eyes are at similar locations or distances. Their relative locations are also different to one another. Thus, the spatial relationship of both eyes will not be same meaning they have different relative context as shown in Figure 4.3 So we can assume that the function $h^{sc}(x - z)$ must be a non-radially symmetric function. $h^{sc}(x-z) \neq h^{sc}(|x-z|)$.

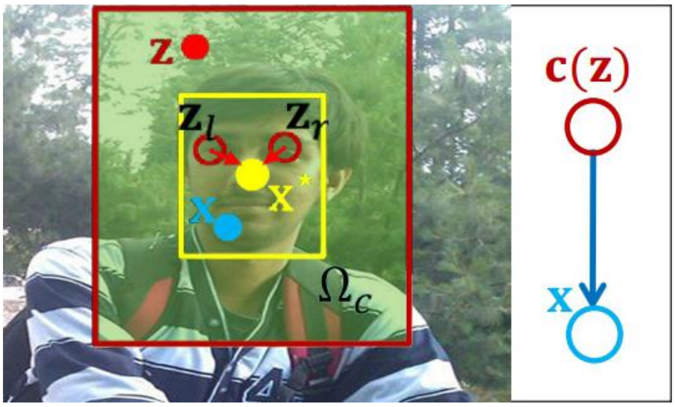


Figure 2.3: Pictorial portrayal depiction of spatial association.

A. *Context Prior Model*

The prior context probability found in equation (2.5) can be presented in terms of image and its kernel mask or convolution kernel by following relation

$$\mathrm{P}(\mathrm{c(z)}|\mathrm{o}) = \mathrm{I(z)w\sigma(z - x)} \quad (2.5)$$

Where the intensity of the image at location z denotes the appearance of the context model and ω_σ $(\cdot)$ is a weighted function or kernel mask which is given by the exponential relation

$$\mathrm{w}_\sigma(\mathrm{z - x*}) = \mathrm{ae}^{-|\mathrm{z-x*}|2\ /\sigma2} \quad (2.6)$$

where alpha is constant or normalization factor that will keep the range of probability P(c(z)|o) in (2.5) to basically range from 0 - 1 so that 1 can easily satisfy all the conditions of the probability. The Scaling factor is denoted by the parameter σ. The relationship in the mentioned equation is specifying an interesting concept. As one can see if the location of context z is to the currently track the target's location x^* , is closer to location's target, it somehow is more necessary to know about the state of the incoming frame. This means prediction is a must in the preceding frame and the weight may shift the tracker's focus. For example, in the technique used in any locality sensitive histogram [6] the pixels that in in close region have larger values. As far as we go, their contribution decreases exponentially. Similarly, on these lines the closer the object is to the target, the larger value of the weight function should be set to get the best possible result in terms of accuracy. This method is by and large very much different that are presented in other context-based methods as in [7]-[8]. This method used the weighted spatial function to get the correspondence of context at different location as seen in figure 2.3. The spatial weight function highlights the value and importance of context it presents. The trackers [7] and [8] simply used the sampling techniques adopted in spatial domain to get the focus of context. Their main task was to get the object detail that is more concentrated to the closer of the object center. In that way they sampled more context locations.

B. *Confidence Map*

In order to find the object location on basis of its confidence map we can map the location using its probability model. The Object location in terms of confidence map can be formed by the following relation

$$m(x)=p(x|o\ ) = be^{-|x-x^{*}/\alpha||^{\beta}} \quad (2.7)$$

here scale parameter is specified by α and shape parameter is β. b in the equation is the constant of normalization which will keep it in range of 0 to 1.

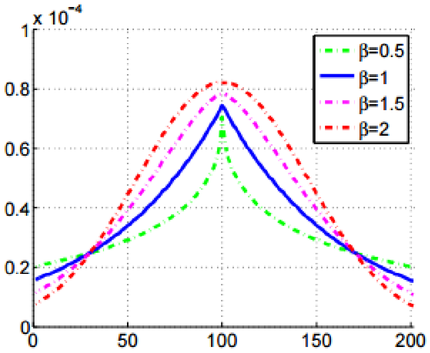


Figure 2.4 Confidence map with diverse estimations of β.

.

Which is equal to

$$=\sum_{z\in\Omega} hsc(x-z)I(z)w\sigma(z-x^{*}) \quad (2.8)$$

$$= h^{sc}(x)\otimes(I(x)\,w_{\sigma}(x\text{-}x^{*})) \quad (2.9)$$

$$=F(be^{-|\ x-x\ ^{*}\ \alpha\ |\ \beta}\ ) = F(h^{\ sc}(x))\ \ O\ \ F(I(x)w_{\sigma}(x - x^{*}\ )) \quad (2.10)$$

Where the FFT function is denoted by F and ⨀ is the element-wise dot product. So, we can express the following relation in terms of Fourier Coefficients.

$$h^{sc}(x) = F^{-1}F(be\ -|\ x\text{–}x\ ^{*}/\ \alpha\ |\ \beta\ )/\ F(I(x)w_{\sigma}(x - x\ ^{*})) \quad (2.11)$$

Whereas, the inverse FFT function is denoted by $F^{-1.}$

Figure 2.5 is an illustration of the computed confidence maps, spatial weight functions and learned context models of David, bolt, fish and singer sequences.

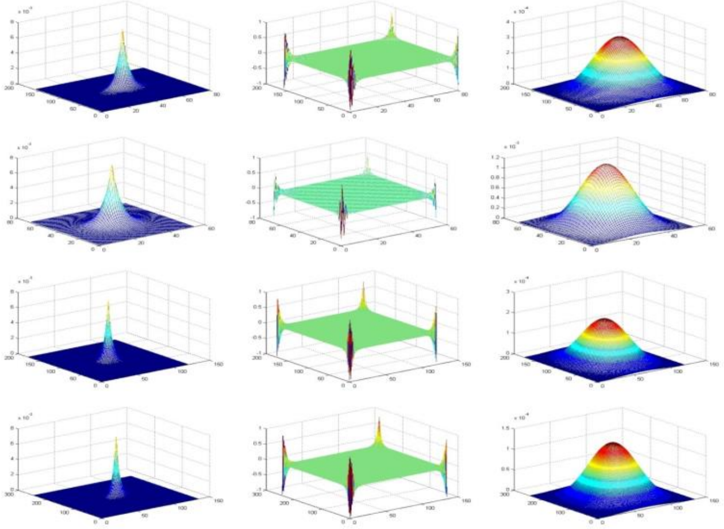


Figure 2.5 Computed confidence maps, spatial weight functions and learned context models of David, bolt, fish and singer sequences.

A. C. *Tracking algorithm*

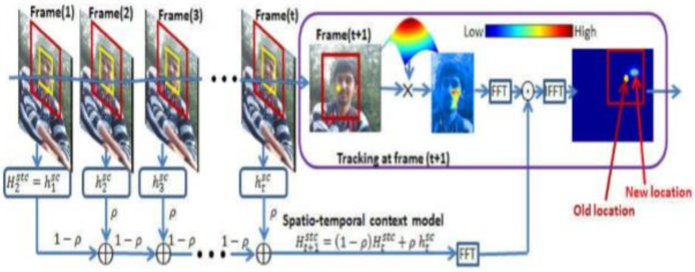


Figure 2.6 Flow of Tracking Algorithm

This acquired info will be used in the updation of the spatial temporal context model being used to ensure the update in the spatio-temporal context model H_{t+1}^{stc} . After that, H_{t+1}^{stc} will be applied and tested to carefully predict the location of the object (t+1)-th frame. Upon arrival of the (t+1)-th frame, we localize its context region $\Omega_c(x_t^{*})$ which will be cropped out and based on the location that is tracked x_t^{*} at the t-th frame. At this point, a subsequent context's feature set shall be obtained $X_{t+1}^{c} = \{c(z) = (I_{t+1})(z), z\ |z\epsilon\Omega_c(x_t^{*})\ \}$. The targeted object location x_{t+1}^{*} in the (t+1)-th frame will be then found by maximizing and get the best out of the confidence map.

$$x_{t+1}^{*} = \arg\max x \in \Omega c(x_t^{*})\, m_{t+1}(x) \quad (2.12)$$

Where $x \in \Omega_c(x_t^{*})$

Where $m_{t+1}(x)$ in terms of context model and image with its weighted function by the equation

$$m_{t+1}(x) = H^{\ stc}{}_{t+1}(x) \otimes (I_{t+1}(x)w_{\sigma t}\ \ (x - x^{\ *}{}_{t})) \quad (2.13)$$

B. D. *Update of Spatio-Temporal Context*

The context model which is spatio-temporal is updated by,

$$H^{stc}{}_{t+1} = (1-\rho)H_t^{stc} + \rho h^{sc}{}_t \quad (2.14)$$

Where learning parameter is ρ and h sc t is spatial context model which is calculated in equation 2.11 at the **t-th** frame. Temporal filtering or masking is a convolution procedure. Such a procedure can easily be performed in frequency domain using normal multiplication such that $H_{\omega}^{stc} = F_{\omega} h_{\omega}^{sc}$ where

$$H^{stc}{}_{\omega} = F_{\omega} h^{sc}{}_{\omega} \quad (2.15)$$

is the transformation of H_{ω}^{stc} which is Fourier temporal transformation. This function is similar to h_{ω}^{sc} in spatial or time domain. The filtering term in F_{ω} Fourier domain is being formulated as

$$F_{\omega} = \rho / e^{j\omega} - (1 - \rho) \quad (2.16)$$

Scale update is of vital importance as the frames change subsequently According to (2.11), the objective area in the present frame is calculated by amplifying the confidence map extracted out of the weighted setting district encompassing that past target area. In any case, the size of the target frequently changes after some time. Along these lines, the parameter of scale σ in the weight work wσ (2.5) ought to be refreshed as needs be.

Low computational multifaceted nature is a one of a prime normal for the proposed calculation in amongst which just 6 FFT activities are included for handling one casing including taking in the spatial setting model and registering the certainty delineate. The computational multifaceted nature for registering each and every FFT is just O (MN log (MN)) for the neighborhood setting locale of M ×N pixels, along these lines bringing about a quick technique.

III. EXPERIMENT RESULTS

A. TRACKING DATASETS AND SEQUENCES

The proposed (IISTC) tracking algorithm was tested on 50 challenging sequences available online on Visual Tracking benchmark. The 4 trackers were compared with are the Locality Sensitive Histogram tracker [6], Ensemble tracker [7], Fast Compressive Tracking [8] and Struck [9] method. For compared trackers, the parameters from the source codes were used for best results. Matlab version 2013 was used for the proposed implementation of the tracker, on a Pentium Core I5 system GHz CPU with 8 GB RAM.

Figure 3.1: Sample data set of the sequences used in evaluation of the trackers.

B. EXPERIMENTAL SETUP

At first, the context region's size is set to twice and double in size of the object in concern. The parameter σ is initially set to σ =1, where "h" and "w" denote the height along with width of the respective initial tracking rectangle.

The parameters, the spatial weight and the confidence mapping functions are actually set to α = 2.25 and β = 1 respectively. And the parameter for learning "ρ" for tracking is set to 0.075. The scale parameter "st" is initially set at 1, and λ = 0.25, the learning parameter. The total number of frames required for updating the relevant scale being set to n = 10. To lessen impacts of enlightenment that is changes in illumination change, every pixel intensity value in the nearby region in the setting district is standardized by subtracting the normal power of that local context region. At that point, the power in nearby region requires a "Hamming Window" to diminish the recurrence impact by the intensity variation while utilizing the FFT.

C. EXPERIMENT RESEULTS

All of the video frames are in gray scale and the two of metrics were used to quantitatively assess the suggested algorithm with the other mentioned trackers. Those two metrics are Success Rate (SR) and Central Location Error (CLE)

1) *Success Rate*

This metric is calculated on the basis of the obtained scores. Basically, the Success rate is calculated as the ratio of total number of frames being correctly tracked by total frames in a said sequence. For this, the Tracked frames and the bounding box values are evaluated with the manual labeled ground truth. Ground truth values are the expected values of the tracker to be achieved. If the score obtained is equal to 0.5 or greater than it, then the frame is considered a success otherwise not. For this, the overlap rate has to be calculated between the bounded box of tracker and on the other side the bounding box of the ground truth. In Figure 3.2 the red bounding box is the basically the ground truth box while the blue one is the box calculated by the tracker. Overlap rate is determined as a ratio of output of tracker's bounding box output and the values of ground truth.

Table 3.1 is the listing of the results as far their success rates are concerned. Rate of success is calculated as total number of corrected frames prediction over a total number of frames in all tracking sequences. The tracker which gave the best results is donated by Red bold letter while the ones giving second best performance are denoted by the Bold blue font.

TABLE 5.1: SUCCESS RATE (SR) RESULTS OF THE METHOD PROPOSED AND THE EVALUATED TRACKERS.

Sequence	Total Frames	Proposed Method	Struck	FCT	EBT	LSHT
Basketball	725	710	708	600	400	690
Bolt	350	323	340	320	300	298
Boy	602	570	400	389	200	303
Car4	659	659	600	323	400	500
Car Dark	393	393	380	383	100	393
Car Scale	252	40	89	47	63	100
Coke	291	189	280	192	230	212
Couple	140	90	100	82	20	45
Crossing	120	120	110	86	100	113
David	770	770	740	690	700	632
David2	537	520	530	400	130	686
David3	252	240	243	200	120	223
Deer	71	23	16	28	40	63
Dog1	1350	983	1210	1280	920	1012
Doll	3872	1873	1013	1700	1091	1600
Dudek	1145	1011	980	200	406	1032
Faceocc1	892	700	720	681	403	391
Faceocc2	812	812	800	703	600	281
Fish	476	476	460	443	210	400
Fleetface	707	493	393	400	381	450
Football	362	350	200	100	187	129
Football1	81	80	28	23	18	29
Freeman1	326	150	143	183	190	110
Freeman3	474	200	90	184	63	19
Freeman4	297	93	53	82	90	71
Girl	500	470	300	188	173	283
Ironman	166	110	100	115	93	91
Jogging	307	300	183	200	115	123
Jumping	313	309	200	103	140	87
Lemming	1336	983	900	920	732	900
Liquor	1741	1000	983	900	903	803
Matrix	100	50	41	73	10	19
Mhyang	1490	1490	1490	1490	3890	1490
Motor Rolling	164	80	29	73	53	13
Mountain Bike	228	228	200	103	193	220
Shaking	365	300	310	306	303	293
Singer1	351	351	351	351	340	330
Singer2	366	183	200	153	140	193
Skating1	400	200	230	180	191	142
Skiing	81	81	50	38	41	29
Soccer	392	100	18	29	93	27
Subway	175	38	67	83	100	46
SUV	945	900	910	830	700	631
Slvster	1345	1293	873	300	393	730
Tiger1	354	340	310	323	330	329
Tiger2	365	330	300	340	310	330
Trellis	569	560	289	183	140	469
Walking	412	412	300	189	173	363
Waling 2	500	500	500	438	200	103
Woman	297	580	183	181	273	171
Successful Frames		21506	19984	17088	14718	17689
Average SR		94.24	80.12	71.14	59.01	70.92

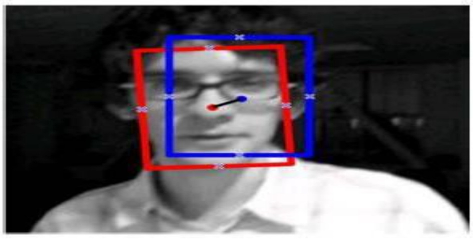


Figure 3.2: Overlapping of the manually labeled ground truth and the tracker box.

Overall Success rate of the trackers compared with the proposed one is shown in figure 5.3 in the form of a bar chart. From the picture it could be inferred that the suggested IISTC tracker beats the four trackers used in the evaluation.

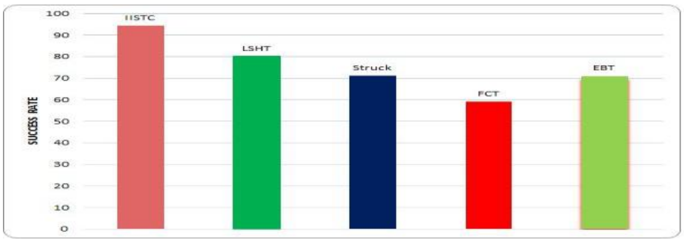


Figure 3.3 Bar graph plot of the success rate of the trackers.

2) *Central Location Error (CLE)*

Central location Error is basically the calculation of the Euclidean distance which exists amid the central locations of target being tracked while the ground truth being labeled manually. Central location error can be computed by the equation 3.2. Where initial coordinates along x, y axis of ground truth bounding box and which represents initial coordinates along x, y axis of tracker bounding box. Then average CLE over all existing frames in a sequence gets determined to summarize the overall performance made for that particular sequence. Table 3.2 lists the average central location error of all the sequences used in the evaluation phase.

The bar graph of the average central location error of all the sequences is shown in figure 3.4.It can be seen from the figure that CLE value of the proposed method is around 65 pixels which is comparatively less than the trackers that were used in quantitative evaluation.

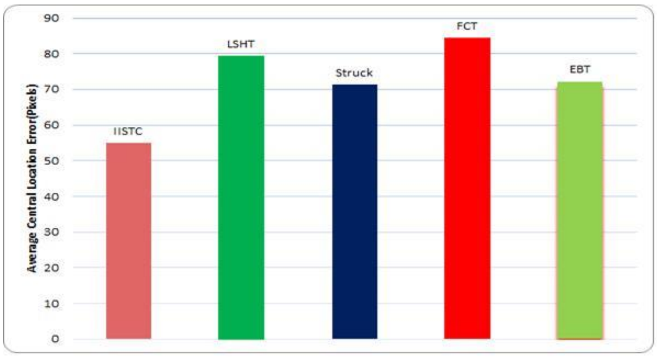


Figure 3.4: Bar graph plot of Central Location Error in Pixels.

D. QUALITATIVE EVALUATION

Qualitative evaluation was done on the mentioned sequences based on their different challenging factors as described in chapter 2.

1) *Sequences having illumination variation*

It was observed that in sequences like Car, Trellis, David and Basketball the object appearances changes due to casting shadows and lightening variations. The basketball sequences did not produced expected results as desired due to sudden illumination changes. The Struck and LSHT methods performed much better in this sequence. In Trellis and David indoor sequences the target objects experienced gradual illumination change. In David Indoor sequence the object was at first completely in dark region. It was noted that the other mentioned trackers only used a subset of image sequences and not the whole image sequence. In evaluation of David Indoor and Trellis some image frames were cut from the evaluation phase. The Struck and the EBT tracker completely discarded the frames in when the patient was in a darker region. The LST performed much better in this regard. In David indoor sequence the proposed method at initial stage drifted away but recovered in latter stages and hence majority of is tracking was a success. In Trellis sequence the object was first in a brighter region and at the very end of the sequence suddenly goes under darker region. The proposed method in this regard performed well. The reason is that the proposed method of update works well when the illumination invariant filtering process is deployed. The illumination Invariant filter enables the tracker to handle gradual illumination changes efficiently. This can be attributed to the proposed feature mechanism which in insensitive to gradual illumination changes. In Figure 3.5 some frames of Car, Trellis, David and basketball sequence are shown.

Table 3.2: Values of Central location error (CLE) in terms of its pixel values and frames per second (FPS). Red text styles show about the best execution and the blue textual styles demonstrate all the second best ones. The aggregate assessed image sequences are 29491.

Sequence	Proposed Method	Struck	FCT	EBT	LSHT
Basketball	26	31	53	100	22
Bolt	31	43	67	123	29
Boy	63	71	82	131	54
Car4	83	78	123	65	123
Car Dark	40	67	89	77	45
Car Scale	98	45	78	87	43
Coke	89	123	91	98	101
Couple	78	67	87	86	89
Crossing	89	76	23	27	72
David	16	25	72	23	45
David2	34	23	62	42	29
David3	54	74	23	87	65
Deer	59	23	76	112	98
Dog1	13	34	86	58	8
Doll	33	56	29	54	47
Dudek	16	59	63	12	24
Faceocc1	18	22	17	19	30
Faceocc2	22	19	36	41	29
Fish	52	86	56	51	40
Fleetface	129	168	120	354	151
Football	31	11	62	24	37
Football1	27	18	23	87	28
Freeman1	66	61	24	78	117
Freeman3	32	34	62	23	34
Freeman4	117	120	73	181	112
Girl	43	58	67	45	65
Ironman	19	67	60	22	15
Jogging	23	98	13	65	32
Jumping	34	65	27	82	129
Lemming	56	92	123	188	120
Liquor	12	14	17	78	23
Matrix	17	57	32	19	42
Mhyang	10	32	76	13	8
Motor Rolling	67	87	53	32	123
Mountain Bike	12	23	14	76	121
Shaking	35	65	12	45	33
Singer1	24	21	31	37	23
Singer2	18	25	46	23	12
Skating1	27	57	98	110	65
Skiing	34	78	12	32	48
Soccer	70	96	13	45	87
Subway	183	123	176	18	123
SUV	69	437	23	123	72
Slvster	22	134	176	32	27
Tiger1	28	12	11	34	19
Tiger2	33	45	76	75	23
Trellis	10	19	48	69	92
Walking	17	11	12	19	27
Waling 2	19	15	17	38	65
Woman	10	12	43	12	21
Average CLE	54.95	79.42	71.32	84.3	72.17

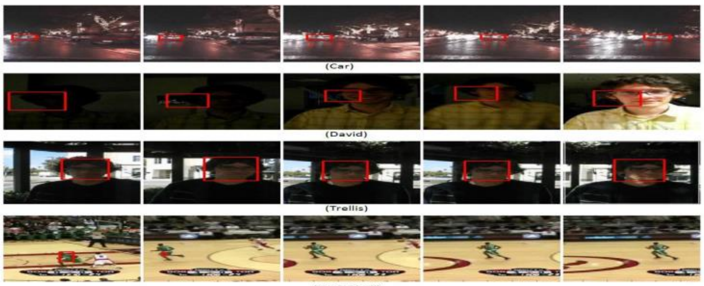


Figure 3.5: Screen shots of Car, David, Trellis and Basketball sequences

2) *Sequences having Out of Plane Rotation and Abrupt Motion*

The target objects in Girl, David 2, Football, Dudek, Mhyang, Mountain bike and Freeman sequences. The tracker performed well in Girl sequence despite the tracker having 360 turn. This is due to the fact that the learned context model helps in prediction of the incoming target in the next frame. But by seeing the very end of that sequence in which when the face of the girl is obstructed and blocked by the man in front of it, the tracker lost its trajectory. It was also observed that all the other trackers failed to track target object in this regard at the end. Dudek and Mhyang sequences experienced in and out of plane rotations. The proposed tracker performed correct in this regard. In mountain bike sequence except the Stuck and the proposed method correctly tracked at the sequence's ending. The free man sequences suffer from occlusions as well. The proposed tracker performed correctly in two of three freeman sequences. But in Freeman 3 sequence the tracker is drifted away due to severe occlusions by the people and the objects present. In Football sequence the target object also suffers from occlusions but since the learned context model helps it to track correctly despite some heavy occlusions at the end of the sequence. In figure 2.6 screen shots of girl, Dudek, Mountain bike and Football are presented that undergo plane rotations.

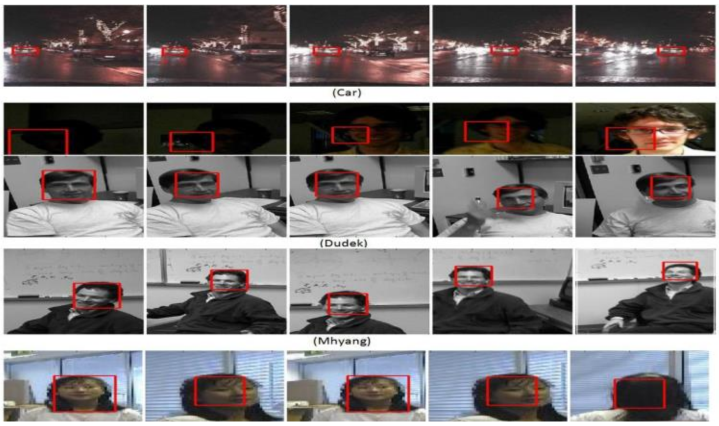


Figure 5.6: Screen shots of Mhyang, Dudek, Girl and Mountain Bike Sequences.

3) *Sequences having Change of Scale*

The target objects in Fish, Car Scale, Dog, Doll, Singer 1, Singer 2 and Dollar sequences observed the scale of change in the frames. The tracker performed well in Fish sequence. The fish sequence also experienced the back ground clutter. The proposed method, Struck and LSHT performed well in this regard. FCT method lost its track at some point in the sequence. This is due to the fact that the learned context model helps at predicting incoming target object in the subsequent frame to come. The sequence of car scale also experienced scale changes as well as occlusion effect. It was noted that in this particular sequence that all the trackers along with the proposed tracker failed in tracking the target till the end of sequence. The Jeep in the sequence occlusion due to trees and as a result the trackers lost track of the object. In this case it was determined which tracker has the least error location. The LSHT tracker in this regard came out on top followed by Struck and then the proposed method. The Dog and the Doll sequences are the largest data sequences in the entire data set. It was observed that not all the frame sequence was evaluated in the EBT and Struck tracker. Thus for an un biased comparison all the frame sequence was used for comparison. It was observed that although the proposed tracker performed correctly up to the task still yet, at the end of the doll sequence, the tracker was failed in adapting the correct scale. The bounding box did not fully encompass the target object. Still none the fact the tracker performed well as compared to other trackers. Only the proposed and LSHT performed correctly in these two sequences. In Singer 1 there was change of scale as the camera zoomed out at the sequence's ending. The tracker though tracked the object correctly till the end still but it also suffered from scale tracked phenomena. In latter frames the size of the woman singer is smaller but perhaps due to illumination variation. In this particular the Struck and the LSHT method performed till the end. EBT and FCT trackers failed in this regard. Just like in Singer 1 there was change of scale in singer 2 sequence along with the object deformation and abrupt motion The proposed tracker did not fully tracked the object correctly till the end In this particular sequence only the LSHT method performed good till at end. The proposed tracker did not fully track the object till the end. EBT and FCT trackers failed in this regard. The Dollar sequence suffered from the cluttered due to fact both the background and the object had similar texture.

IV. CONCLUSION

This paper articulates simple and straightforward, nonetheless quick and powerful calculations which utilize spatio-temporal setting data for Visual Tracking. Two of the nearby setting models entailing spatial setting and the spatio-temporal setting models are suggested which proved robust to appearance ambiguities offered by impediment, changes in brightness & posture varieties. This paper also propsed an unequivocal scale adjustment plot being abled to adjust targeted scale varieties successfully. The Fast Fourier Transform (FFT) calculation was utilized for learning and identification, bringing about an effective target following strategy. Numerous investigations with best in class calculations on testing successions exhibited that the proposed technique accomplishes positive outcomes regarding precision, robustness, and computational speed alike.

Form the given test sequences it was determined that the given proposed method of tracking was efficient when compared with other trackers. The algorithm exploiting the back ground information helps in making decision about predicting the next in coming frame target and hence by comparison with the previous frame the value of confidence can be computed. This methodology mostly helped in those cases where the bounding box drifts from the target object. It was found that the method proves itself robust to sudden illumination fluctuations and changes when occurred. The success rate and the central location error shows that the overall efficiency has improved by 16 percent when compared to LSH tracker.

Low computational multifaceted nature is basically the one prime normal for the suggested calculation in which just the 6 FFT activities are included for handling one casing including and taking in the spatial setting model (9) and registering the certainty delineate). The calculation multifaceted nature for registering every FFT is just O (MN log(MN)) for the neighborhood setting locale of M ×N pixels, along these lines bringing about a quick technique.

REFERENCES

[1] E. Maggio, and A. Cavallar, “Video Tracking: Theory and Practice”, John Wiley & Sons , 2011.

[2] V. Vapnik, “Statistical Learning Theory”, John Wiley & Sons, 1998.

[3] Y.Andrew, and M. Jordan, “On Discriminative vs. Generative Classifiers: A comparison of logistic Regression and Naive Bayes”, Neural Information Processing Systems, 2001.

[4] H. Shengfeng, Y. Qingxiong, L. Rynson, J. Wang and M. Yang, “Visual Tracking via Locality Sensitive Histograms”, Computer Vision and Pattern Recognition, June, 2013.

[5] M.Danellian, G.Hager, K.Fahad, and M.Felsbeg, “Accurate Scale Estimation for Robust Visual Tracking”,British Machine Vision Conference, 2014.

[6] H. Shengfeng, Y. Qingxiong, L. Rynson, J. Wang and M. Yang, “Visual Tracking via Locality Sensitive Histograms”, 2013, Computer Vision and Pattern Recognition, June.

[7] J. Kwon, and K. M. Lee. “Tracking by Sampling Trackers”, International Conference on Computer Vision, 2011.

[8] K. Zhang, L. Zhang, and M.-H. Yang, “Fast Real-time Compressive Tracking”, European Conference on Computer Vision, 2012

[9] S. Hare, A. Saffari, and P. H. S. Torr. “Struck: Structured Output Tracking with Kernels”, International Conference on Computer Vision, 2011.

[10] P. Viola, and M. Jones, “Robust Real Time Object Detection”, International Journal of Computer Vision, 2004.

[11] Y. Freund, and R. Schapire , “A Short Introduction to Boosting ”, Journal of Japanese Society for Artificial Intelligence, 1999.

[12] R. Kalman, and R. Emil, "A New Approach to Linear Filtering and Prediction Problems", Transactions of the ASME Journal of Basic Engineering,1960.

[13] Y. Wu, J. Lim and M. Yang, “Online Object Tracking: A Benchmark”, IEEE Conference on Computer Vision and Pattern Recognition, 2013.