VISUAL OBJECT TRACKING USING SPATIO TEMPORAL INFORMATION (VOT-STI)

INTERNATIONAL ISLAMIC UNIVERSITY

الجامعة الاسلامية العالمية اسلام آباد

ISLAMABAD

Khizer Mehmood

80-FET/PHDEE/F14

Submitted in partial fulfilment of the requirements for the PhD degree in Electronic Engineering at the Department of Electrical & Computer Engineering

Faculty of Engineering and Technology

International Islamic University,

Islamabad

Supervisor

Prof. Dr. Abdul Jalil

Co-Supervisor

Dr. Ahmad Ali

June, 2022

International Islamic Univ
C... ...AL
L... ...Y
ISL... ...AD

Accession No TH 25755

PhD
006.37
KHV

Computer vision

Machine learning

Image processing -- Digital techniques

Copyright © 2022 by Khizer Mehmood

All rights reserved. No part of the material protected by this copyright notice may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying, recording or by any information storage and retrieval system, without permission of the author.

DEDICATED TO

My Teachers and Family

CERTIFICATE OF APPROVAL

Title of Thesis: Visual Object Tracking using Spatio-Temporal Information (VOT-STI)

Name of Student: KHIZER MEHMOOD

Registration No: 80-FET/PHDEE/F14

Accepted by the Department of Electrical & Computer Engineering, Faculty of Engineering and Technology, International Islamic University (IIU), Islamabad, in partial fulfillment of the requirements for the Doctor of Philosophy degree in Electronic Engineering.

Viva voce committee:

Prof. Dr. Abdul Jalil (Supervisor)
Former Professor, TTS, DEE, FET, IIU Islamabad.

Dr. Ahmad Ali (Co-Supervisor)
General Manager, NESCOM, Islamabad.

Dr. Ihsan ul Haq (Internal)
Associate Professor DECE, FET, IIU Islamabad.

Dr. Rab Nawaz (External-I)
Project Director, NESCOM, Islamabad.

Dr. Abdul Basit (External-II)
Deputy Chief Scientist, MSID, PINSTECH, Islamabad.

Dr. Suheel Abdullah Malik (Chairman, DECE)
Associate Professor DECE, FET, IIU Islamabad.

Prof. Dr. Nadeem Ahmad Sheikh (Dean, FET)
Professor DME, FET, IIU Islamabad.

ABSTRACT

Object tracking is still an intriguing task as the target undergoes significant appearance changes due to illumination, fast motion, scale variations, occlusion, and shape deformation. Background clutter and numerous other environmental factors are other significant constraints that remain a riveting challenge to developing a robust and effective tracking algorithm. In the present study, an adaptive Spatio-temporal context (STC) based algorithm for online tracking is proposed by combining the context-aware formulation, Kalman filter, scale-space tracking, and adaptive model learning rate. To enhance seminal STC-based tracking performance, different contributions are made in the proposed study. First, the context-aware formulation is incorporated in the STC framework to perform better in clutter background situations. Second, a pyramid representation-based scale correlation filter is integrated to overcome the STC's inability to rapid change in target scale by learning appearances induced by variations sampled at a different set of scales. Third, an occlusion detection and handling mechanism are incorporated to avoid the target model from drifting. Occlusion is detected from the peak correlation score of the response map. It continuously predicts the target location during occlusion and passes it to the STC tracking model. After successfully detecting occlusion, an extended Kalman filter is used for occlusion handling. It declines the chance of tracking failure as the Kalman filter continuously updates itself and the tracking model. Further improvement to the model is provided by fusion with average peak to correlation energy (APCE) criteria, automatically updating the target model to deal with environmental changes. Finally, the average difference between consecutive frames is used to update the target model adaptively.

Experimental results on image sequences taken from Temple Color (TC)-128, OTB2013, OTB2015 and UAV123 datasets indicate that the proposed algorithms perform better than various algorithms, both qualitatively and quantitatively.

LIST OF PUBLICATIONS

[1]. **Mehmood, K.;** Jalil, A.; Ali, A.; Khan, B.; Murad, M.; Cheema, K.M.; Milyani, A.H. Spatio-Temporal Context, Correlation Filter and Measurement Estimation Collaboration Based Visual Object Tracking. *Sensors* **2021**, *21*, 2841.

[2]. **Mehmood, K.;** Jalil, A.; Ali, A.; Khan, B.; Murad, M.; Khan, W.U.; He, Y. Context-Aware and Occlusion Handling Mechanism for Online Visual Object Tracking. *Electronics* **2021**, *10*, 43.

[3]. **Mehmood, K.;** Ali, A.; Jalil, A.; Khan, B.; Cheema, K.M.; Murad, M.; Milyani, A.H. Efficient Online Object Tracking Scheme for Challenging Scenarios. *Sensors* **2021**, *24*, 8481.

[4]. **Mehmood, K.;** Chaudhary, N.I.; Khan, Z.A.; Raja, M.A.Z.; Cheema, K.M.; Milyani, A.H. Design of Aquila Optimization Heuristic for Identification of Control Autoregressive Systems. *Mathematics* 2022, *10*, 1749.

[5]. Khan, B.; Jalil, A.; Ali, A.; Alkhaledi, K.; **Mehmood, K.;** Cheema, K.M.; Murad, M.; Tariq, H.; El-Sherbeeny, A.M. Multiple Cues-Based Robust Visual Object Tracking Method. *Electronics* **2022**, *11*, 345.

[6]. Murad, M., Jalil, A., Bilal, M., Ikram, S., Ali, A., Khan, B., and **Mehmood, K.** Radial Undersampling-Based Interpolation Scheme for Multislice CSMRI Reconstruction Techniques. *BioMed Research International*, vol. 2021, Article ID 6638588, 15 pages, 2021.

[7]. Khan, Z.A., Raja, M.A.Z., Chaudhary, N.I., **Mehmood, K.**, He, Y. MISGD: Moving-Information-Based Stochastic Gradient Descent Paradigm for Personalized Fuzzy Recommender Systems. *Int. J. Fuzzy Syst.* **24,** 686–712 (2022).

[8]. Murad, M., Bilal, M., Jalil, A., Ali, A., **Mehmood, K.** and Khan, B., 2020. Efficient reconstruction technique for multi-slice CS-MRI using novel interpolation and 2D sampling scheme. *IEEE Access*, 8, pp.117452-117466

[9]. Khan, B., Ali, A., Jalil, A., **Mehmood, K.,** Murad, M. and Awan, H., 2020. AFAM-PEC: Adaptive Failure Avoidance Tracking Mechanism Using Prediction Estimation Collaboration. *IEEE Access*, 8, pp.149077-149092

[10]. Farooq, Adil and Ali, Umair and Shinwari, Awais Khan and ur Rehman, Ashfaq and Ahmed, Suhail and **Mehmood, Khizer** and Iqbal, Waseem, "SCADA Based Power Management and Distribution System Prototype for Pakistan Scenario," *2018 International Conference on Power Generation Systems and Renewable Energy Technologies (PGSRET)*, 2018, pp. 1-5

The research work presented in this thesis is based on the accepted articles 1 and 2.

ACKNOWLEDGEMENTS

In the name of Allah (Subhanahu Wa Ta'ala), who is the most gracious and merciful. I would like to thank Allah for giving me the strength and patience to complete this research work. Peace and blessings of Allah be upon His last Prophet Muhammad (Sallulah-o-Alaihihe-Wassalam) and all his Sahaba (Razi-Allah-o-Anhu) who dedicated their lives for Dawah and the spread of Knowledge.

I am truly grateful to my supervisor Dr. Abdul Jalil whose inspiration, ideas and efforts make it possible for me to complete my higher studies. He has been a role model for me and many others in teaching, research and other aspects of life. I would also like to thank to co-supervisor Dr. Ahmad Ali for his time, ideas, never-ending support and guidance during my research work.

I offer my sincere thanks to my colleagues Dr. Naveed Ishtiaq Chaudhary, Dr. Zeshan Aslam Khan, Dr. Baber Khan Jadoon, Dr. Athar Waseem, Engr. Fahad Munir, Engr. Muhammad Muzammil and Dr. Khalid Mehmood Cheema for their never-ending support during the last few years. I would like to acknowledge the support of International Islamic University Islamabad, Pakistan for providing me with a full fee waiver during my PhD studies. I am thankful to the administration at the department, and the university level for their kind support.

I am grateful to my father, mother, brother and sister for their love and support throughout my life. I am also very thankful to my wife for her patience, encouragement and prayers during every stage of my PhD degree. Finally, I am thankful to my daughter, whose innocent gestures were the source of inspiration for me.

(Khizer Mehmood)

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

CF	Correlation Filter
CLE	Centre Location Error
CN	Color Naming
CNN	Convolutional Neural Network
DCF	Discriminative Correlation Filter
DPR	Distance Precision Rate
EKF	Extended Kalman Filter
FFT	Fast Fourier Transform
KCF	Kernelized Correlation Filter
KF	Kalman Filter
MOSSE	Minimum Output Sum of Squared Error
STC	Spatio-Temporal Context
VOT	Visual Object Tracking

LIST OF SYMBOLS

A list of symbols in this thesis is given below.

m	Normalization constant
ξ	Shape Parameter
θ	Scale Parameter
ρ	Learning rate
h_t^{sc}	Spatial context model
H^l	1-Dimensional HOG feature extracted from the sample
F_t^l	Correlation filter at tth frame
A_t^l	The numerator of correlation filter at tth frame
B_t	The denominator of correlation filter at tth frame
G_t	2-Dimensional Gaussian Function
y	Response map of correlation filter
x_t	Predicted state at tth frame
P_t	Updated error covariance at tth frame

Chapter 1.

Introduction

In this chapter, the impact of visual object tracking (VOT) in computer vision and image processing is briefly reviewed, along with the need to explore models in developing accurate and robust tracking algorithms. The main contributions are presented in designing a tracking algorithm that can accurately track the target under various challenging scenarios.

1.1 Background

To process visual detail vision system is one of the most perceptual systems. In image processing, different mathematical operations are applied to images. The input to an image processing system can either be an image or video, while its output is either an image or a set of parameters related to the image. VOT aims to predict coordinates of moving target in all frames of image sequence given with only the initial location of target at the first frame. Commonly, a rectangular bounding box is used to describe the shape and size of the target. VOT is applied in several real-world applications with the availability of high-performance, low-cost cameras, as shown in Fig. 1.

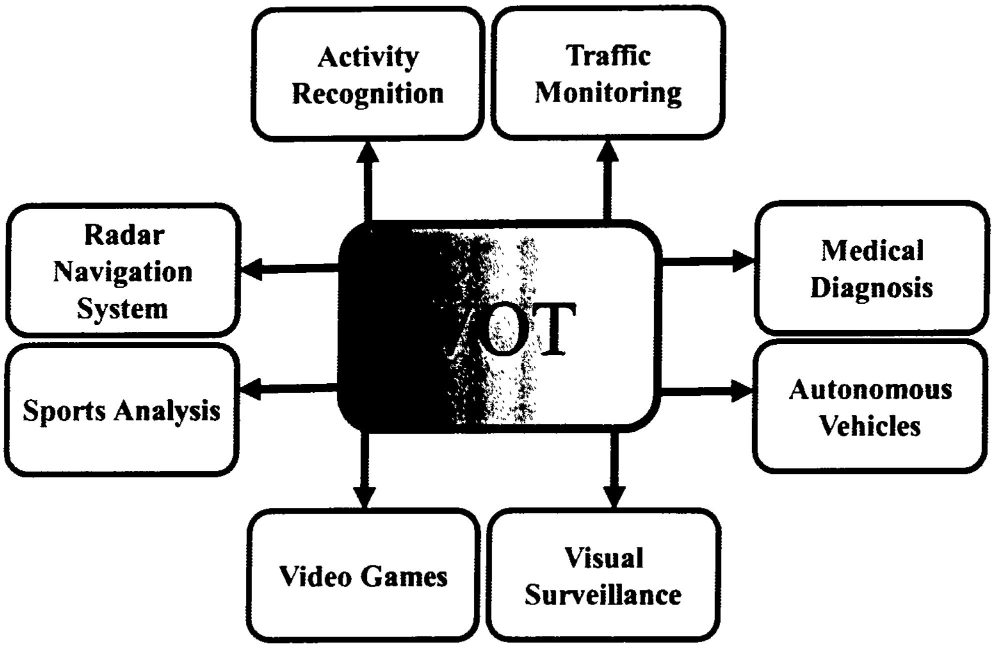


Fig. 1.1 Application areas of visual object tracking

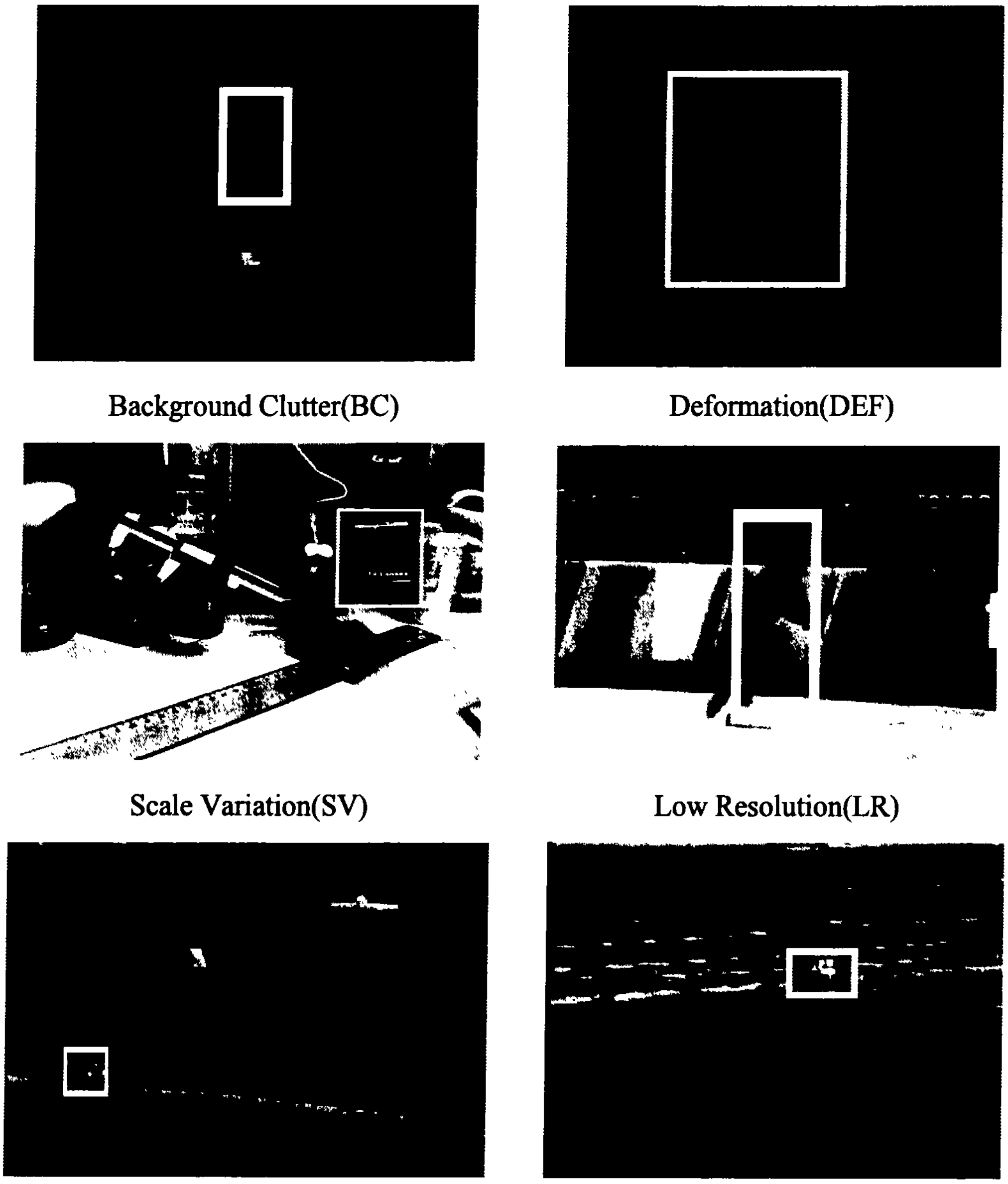
- Activity Recognition: VOT plays a vital role in activity recognition applications such as human action recognition [1], [2], learning activity patterns [3], etc.
- Traffic Monitoring: VOT provides an efficient solution for the management of traffic on highways, such as detection of traffic accidents [4],[5], counting of pedestrians [6], etc.
- Medical Diagnosis: VOT has shown significant progress in diagnosing diseases such as automatic fovea center localization in retinal images [7], vocal tract from magnetic resonance images [8], etc.
- Autonomous Vehicles: VOT plays a vital role in developing autonomous vehicles such as mono camera-based 3D tracking [9], driver assistance systems [10], etc.

- Visual Surveillance: VOT is an essential part of various surveillance and security systems such as foreground object detection and tracking [11], anti-aircraft gun system [12], etc.
- Video Games: VOT is applied in video games for better control, such as gaming AI [13], detection of unknown objects [14], and face tracking in video games [15]. etc.
- Sports Analysis: VOT plays a vital role in providing interesting analysis in various sports such as tracking in basketball games [16], soccer games [17], etc.
- Radar Navigation System: VOT is applied in various navigation systems such as ship's radar [18], maritime radar systems [19], ARPA systems [20], etc.

In recent years, various object tracking has given promising results on different benchmark datasets. Generally, tracking methods are classified into two categories, which are generative tracking and discriminative tracking methods. In generative tracking, the target appearance model is learned and updated by searching for the most similar candidate as the target. In contrast, an online classifier is trained to discriminate target from its background in discriminative tracking. Details of both tracking methods were widely referred in the literature.

1.2 Issues in Visual Object Tracking

In recent years, substantial progress has been made in VOT [21]. However, VOT still faces various challenges, as shown in Fig. 1.2.



Illumination Variation(IV)

Occlusion(OCC)

Background Clutter(BC)

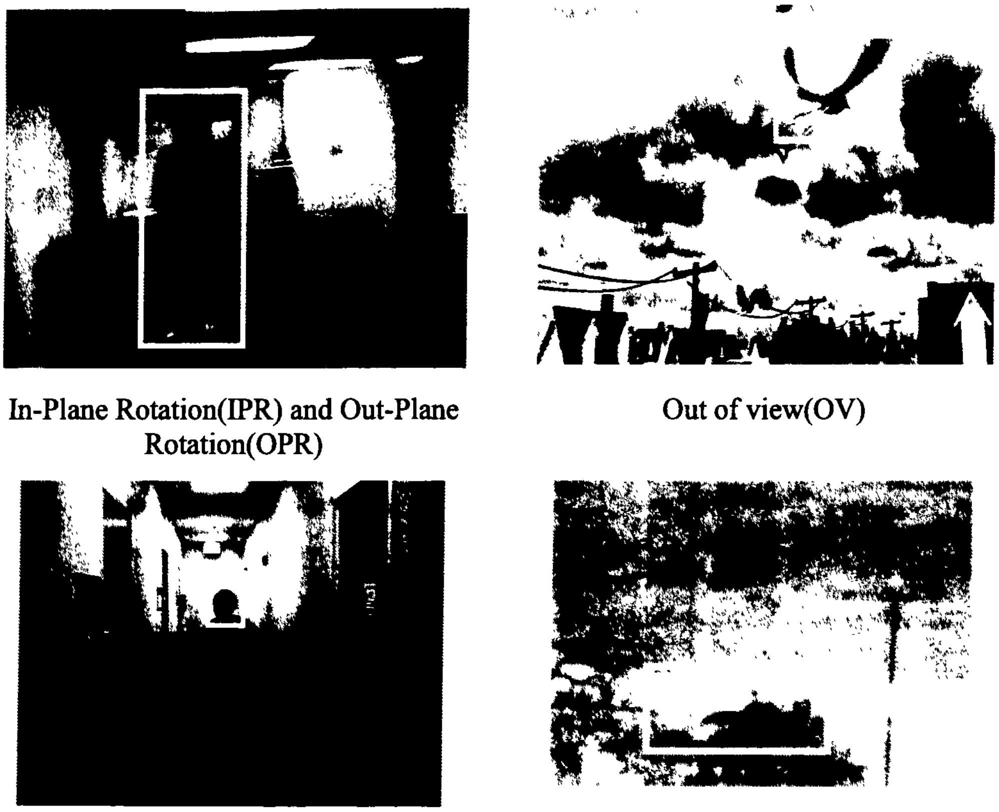
Deformation(DEF)

Scale Variation(SV)

Low Resolution(LR)

Motion Blur(MB)

Fast Motion(FM)



In-Plane Rotation(IPR) and Out-Plane Rotation(OPR)

Out of view(OV)

Fig. 1.2 Issues in visual object tracking

- Illumination Variation: Various target features change considerably due to illumination variations. Therefore, illumination changes around the target of interest need to be handled for robust tracking.
- Occlusion: It is the state of the target in which another object hides it. It is essential to have an occlusion detection and handling module for accurate tracking.
- Background Clutter: Tracker performance decreases significantly when there is a similarity between the target and its background. A tracker must discriminate target from its background for robust tracking.

- Deformation: During tracking, most non-rigid targets change their appearance. Therefore, a tracker should be able to update its motion model when the target undergoes an appearance change.
- Scale Variation: The size of the target changes significantly during tracking. Therefore, the tracker should be able to track the size of the target appearance accurately.
- Low Resolution: Tracker might be lost the target in a video less than 400 pixels around the target of interest. Therefore, a tracker should be able to track under these limitations.
- Motion Blur: The performance of the tracker decreases significantly when the target area is blurred either by the motion of the target or camera movement. Therefore, a tracker should be able to track under these circumstances.
- Fast Motion: During tracking, if the target moves faster than 20 pixels than its ground truth, then the tracker might lose the target. Therefore, a tracker should be able to update its motion model when the target moves faster.
- In-Plane and Out-of-Plane Rotation: One of the challenging issues during tracking is when the target either rotates in the image plane or out of the image plane. Therefore, a tracker should be able to track under these situations.
- Out of View: During tracking, it is possible that some portion of the target leaves the view and then comes back after a few frames. Therefore, a tracker should be able to track under these scenarios.

1.3 Motivation and Objective

VOT is a renowned research problem in computer vision and has been utilized in many applications. With the increasing technological development in mobile cameras, webcams, and closed-circuit televisions (CCTVs). It is a requirement for these systems to develop tracking methods to process information from these videos in real-time. However, it is a challenging task as less information is available related to the target and its context. The main objective of this research is to find appropriate methods for VOT with the best accuracy and robustness when faced with the various challenges mentioned above.

1.4 Problem Statement

Most of the work in the field of VOT is usually based on different assumptions such as structured or non-structured environment, fixed or stationary camera etc. The more issues are in a video sequence the more is difficult to track the object. So, in order to design a robust tracking algorithm it must accurately track the object regardless of the changes in appearance model. The literature survey identifies both generative and discriminative methods, which use observation model of a tracker. Still there are some issues which need to be addressed and best suitable technique is required to be identified for long term real life video sequence.

In spatio temporal information (STI) algorithm a model is learned between the target object and its local surrounding in a scene. The learned model is used to update a spatio temporal context model for the next frame. Tracking in the next frame is formulated by computing a confidence map that integrates the spatio temporal context information, and the best object location can be estimated by maximizing the confidence map. We can address object tracking problems by designing a STI algorithm, which integrates various powerful features to further enhance the tracking process.

1.5 Contributions

This thesis focus on single object online tracking, in which generic tracker is designed for all kind of objects. The contributions of the thesis are presented below:

1. We propose a correlation filter-based context-aware formulation to utilize context information effectively and incorporate it into the Spatio-temporal context framework to deal with effectively in clutter background situations.
2. We propose a scale correlation filter-based pyramid representation mechanism to accurately extract the target without accumulating the scale model's error. We use a combination of Spatio-temporal context and scale correlation filter to achieve accurate object tracking
3. We introduce an effective method of tracking the object can be tracked accurately by utilizing Kalman Filter and response map's peak value to measure the reliability.
4. We also use extended Kalman Filter (EKF) for occlusion handling and the response map's peak value to measure the reliability of the current estimated position. If the tracking result is unreliable, this method can regain the target position to continue tracking.
5. We propose an adaptive learning rate mechanism based on the average peak to correlation energy (APCE) based on the target appearance model. This method can effectively prevent the tracking model from the wrong appearance.

6. We propose an average difference between consecutive frames based adaptive learning rate mechanism to update the model according to change in the environment.
7. Experimental results have been presented on de facto standard videos to show the efficacy of the proposed method over STC [22], DCF_{CA} [23], Modified KCF [24], MACF [25], Modified STC [26], MOSSE_{CA}[23], and AFAM-PEC [27].

1.6 Organization of Thesis

The organization of the thesis is as follows; the background, issues in VOT, motivation, objectives of the research, and contribution of the thesis are given in Chapter 1. Chapter 2 explains the of Spatio-temporal context and correlation filters for visual object tracking systems. The proposed schemes for tracking the target in VOT systems are presented in Chapter 3. The simulation results and discussion for different case studies of Spatio-temporal context are presented in Chapter 4. Finally, conclusions and future directions are given in Chapter 5.

Chapter 2.

Literature Review

This chapter provides a comprehensive review of current object tracking methods related to Spatio-temporal context and correlation filters.

2.1 Introduction

VOT has been an active exploration area over the past two decades, during which significant progress has been made [28]–[33]. The performance of tracking methods relies on the appearance model such that it differentiates object from their background under challenging scenarios. The appearance model can be allocated into categories, which are discriminative tracking methods and generative tracking methods. Both methods are widely used in the tracking community.

2.2 Generative Tracking Methods

Generative tracking methods learn target appearance models and search for the highest matching score. These methods achieve good tracking results at the expense of computational cost. Few generative tracking methods are presented below.

Kwon et al. [34] proposed a tracking scheme that utilizes observation, tracking and motion models to construct the target appearance. Ross et al. [35] presented a tracking method that learns low dimensional subspace representation by adapting the appearance change of the target. Mei et al. [36] proposed a sparse representation-based tracking algorithm using the l_1 norm. It dynamically updates target templates and uses non-negativity constraints to filter out clutter. Lara et al. [37] proposed an image descriptor using distribution fields (DFs) representation such that loss of information around pixel value can be prevented. Li et al. [38] proposed a generative tracking method to model target appearance using incremental 3D discrete cosine transform (3D-DCT), which determines the signal dimension and evaluates the likelihood between the target and its background at a low computational cost. Zhang et al. [39] proposed a weighted multiple instance learning (WMIL) based tracking method. It utilizes a probability function that uses a large weight near the target location and optimizes the likelihood function. Oron et al. [40] proposed a probabilistic model of object variation over time by calculating the earth mover's distance using the cost of moving pixel and color change.

2.3 Discriminative Tracking Methods

Discriminative methods treat tracking as a binary classification problem that can discriminate targets from their background by training online classifiers. These methods perform favorably well. However, tracking might get affected when training data is small. Few discriminative tracking methods are presented below.

Grabner et al. [41] proposed an object tracking method that uses online AdaBoost feature selection such that the classifier is adapted to appearance change. Moreover, depending upon the background, the most discriminative features are selected. Kalal et al. [42] proposed a tracking method that exploits positive and negative structural constraints to enforce labeling on unlabeled data. Hare et al. [43] proposed a kernelised structured output support vector machine (SVM) learned online for adaptive tracking. To prevent the unbound growth of the support vector, a budgeting mechanism is also incorporated to run it at a high frame rate. Henriques et al. [44] exploits the Circulant matrices' properties and uses Fast Fourier Transform (FFT) for detection.

Zhang et al. [45] proposed a tracking algorithm to address the target's drift problem and appearance change. They employed non-adaptive random projections to preserve the image and sparse matrix structure to extract features of the appearance model. Classification is done by using a naïve Bayes classifier in the compressed domain.

2.4 Spatio-Temporal Context Tracking

In visual object tracking, the target is characterized by objects around the target present in the current frame. The area which is present around the target is called context. In the context around the target, various temporal and spatial relationships exist in continuous frames. STC tracking algorithm is based on the Bayesian framework to find the target location based on background knowledge accurately. It formulates the task of finding an object center by maximizing the confidence map in every frame. For every current frame target location is represented by x^* with its features defined as $X^c = \{y(i) = (I(i), i) | i \in \Omega_c(x^*)\}$ where $I(i)$ is the image grey scale value at location i while $\Omega_c(x^*)$ is the context around the target center x^*. It is shown in Fig. 2.1.

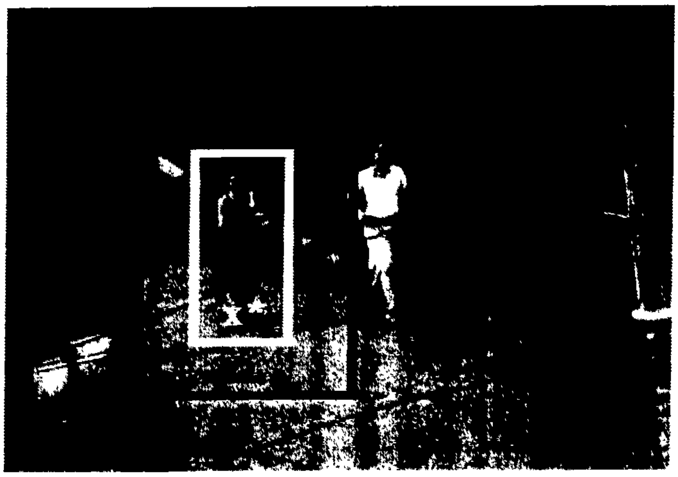


Fig. 2.1 Spatial relationship between object and its context.

A confidence map of the target location is described in (2.1).

$$
\begin{aligned}
y(x) &= P(x|j) = \textstyle\sum_{y(i)\in X^c} P(x, y(i)|j) \\
&= \textstyle\sum_{y(i)\in X^c} P(x, y(i)|j)\, P(y(i)|j)
\end{aligned}
\qquad (2.1)
$$

where j is the target, $P(y(i)|j)$ is context prior model that represents the features of context appearance. $P(x, y(i)|j)$ is a spatial context model that formulates spatial relation between object location and its information of context. It is used to identify and resolve various uncertainties for different image measurements. The goal in this tracking problem is to train the spatial context model $P(x, y(i)|j)$.

2.4.1 Confidence Map

Confidence map function $y(x)$ is presented in (2.2).

$$y(x) = P(x|j) = re^{-|\frac{x-x^*}{\alpha}|^{\xi}} \qquad (2.2)$$

where r is the normalization constant, α is the scale parameter, and ξ is the shape parameter. The problem of location ambiguity frequently occurs in object tracking. An appropriate selection of shape parameters can resolve this problem and helps learn the spatial context model. Setting ξ>1 results in over smoothing of the confidence map near the center, thereby increasing location ambiguities. However, if ξ<1, it generates a sharp peak response due to which few positions are activated while learning spatial context. Due to these issues, STC uses ξ=1.

2.4.2 Context Prior Model

To learn the spatial context model, the context prior model needs to be calculated first. Then, it is modeled using the image intensity function to represent the target appearance and Gaussian weighted function mentioned in (2.3) and (2.4).

$$P(y(i)|j) = I(i)\,\omega_{\gamma}(i - x^{*}) \tag{2.3}$$

$$\omega_{\gamma} = de^{-|\frac{x-x^{*}}{\sigma^{2}}|2} \tag{2.4}$$

where d is a normalization constant that restricts (2.4) to range between 0-1, and σ is scale representation. The closer the context location i is to the current target location x^{*} larger weight should be set to predict the target location in the next frame.

2.4.3 Learning Spatial Context Model

The conditional probability function defines the spatial context model is presented in (2.5).

$$P(x, y(i)|j) = h^{sc}\,(x - i) \tag{2.5}$$

Solving (2.5) for spatial context.

$$= h^{sc}(x - i)\,I(i)\,\omega_{\gamma}(i - x^{*})$$

$$= \mathrm{h}^{\mathrm{sc}}(x) \otimes (\mathrm{I}(x)\,\omega_{\gamma}(x - x^{*})) \tag{2.6}$$

where ⊗ is a convolution operator in (2.6). Fast Fourier Transform (FFT) is used and calculated as presented n (2.7) to improve calculation speed.

$$F(y(x)) = F(h^{sc}(x)) \odot F(I(x)\omega_{\gamma}(x - x^{*})) \tag{2.7}$$

where F is the FFT operation and ⊙ denotes element-wise multiplication. Solving (2.7) for spatial context model.

$$h^{sc}(x) = F^{-1}\left(\frac{F\left(re^{-|\frac{x-x^{*}}{\alpha}|\xi}\right)}{F((I(x)\omega_{\gamma}(x - x^{*})))}\right) \tag{2.8}$$

where F^{-1} denotes inverse FFT in (2.8). The spatial context model h^{sc} learns relatively spatial relations between different pixels in the Bayesian framework.

2.4.4 Model Update

In the STC model, tracking is considered a detection task. Target is initialized position at the first frame. At t^{th} frame, the STC model $H_{t+1}^{stc}(x)$ can be updated using the spatial context model $h_t^{sc}(x)$. Then the target center position x_{t+1}^{*} of the (t + 1) frame can be attained by computing the extreme of the confidence map given in (2.9).

$$x_{t+1}^{*} = arg_{x\in\Omega_c(x_t^*)} max\, y_{t+1}(x) \tag{2.9}$$

The confidence map $y_{t+1}(x)$ at t+1 frame can be calculated as described in (2.10).

$$y_{t+1}(x) = F^{-1}\left(F\left(H_{t+1}^{stc}(x)\right) \odot F\left(I_{t+1}(x)\omega_\gamma(x - x_t^*)\right)\right) \tag{2.10}$$

Here, H_{t+1}^{stc} derives from spatial context h_t^{sc} and can reduce noise caused by abrupt appearance changes of I_{t+1}. STC model can be updated as mentioned in (2.11).

$$\mathbf{H_{t+1}^{stc} = (1 - \rho)\, H_t^{stc} + \rho h_t^{sc}} \tag{2.11}$$

where ρ is the learning rate and h_t^{sc} is the spatial context model computed in (2.8). The complete flow of the algorithm is presented in Fig. 2.2.

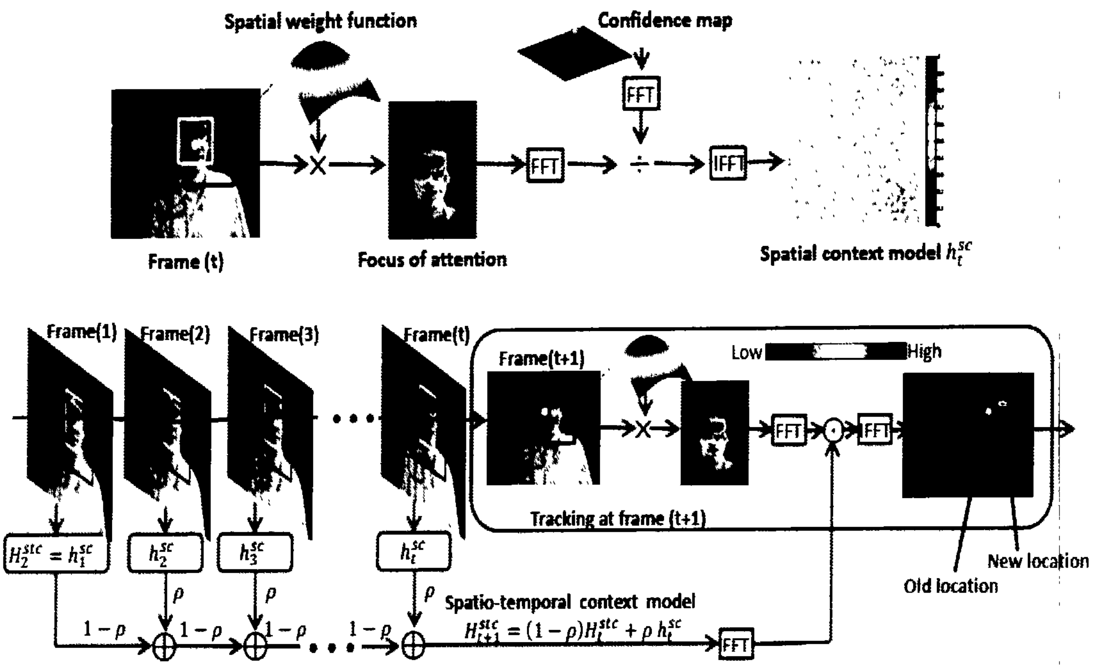


Fig. 2.2 Learning and Detection using STC [22]

2.4.5 Improvements in STC

STC algorithm achieves favorable tracking results. However, significant improvements have been made in these in recent years few of them are presented below.

Zhang et al. [46] proposed an adaptive STC model for online tracking by incorporating a histogram of oriented gradients (HoG) features and color naming (CN) features in the STC framework. They also used the average difference between adjacent frames to adjust the learning rate when the model is updated. Wang et al. [47] proposed an improved tracking model that combines STC with a convolutional neural network (CNN) to extract online CNN's deep characteristics without training. Wan et al. [48] and Li et al. [49] proposed a motion vector-based mechanism for predicting target position under motion is incorporated in the STC framework to improve the STC scale. It also combined a scale correlation filter with STC to extract different scale samples around the target and used the HoG operator to form a pyramid of scale characteristics. Tian et al. [50] proposed an enhanced STC tracker to address occlusion by incorporating patch-based occlusion detection mechanisms into the STC framework. Chen et al. [51] proposed an improved STC tracker to address occlusion by incorporating a Kalman filter to predict target location in case of occlusion.

Munir et al. [52] proposed a modified STC tracker to address occlusion by incorporating a Kalman filter to predict the target location in case of occlusion and implemented it for real-time eye-tracking applications. Cui et al. [53] proposed an amended STC tracker to address the limitation of full occlusion. They incorporated an occlusion detection mechanism which consists of three stages during which motion and template update information is stored and used when the target is occluded. Yang et al. [54] proposed an enhanced STC tracker to address occlusion by incorporating the PSR-based occlusion feedback mechanism for model and scale updates in the STC framework. Yang et al. [55] proposed an improved STC tracker to address occlusion by incorporating a Kalman filter to predict the target location and using Euclidean distance to detect occlusion. Qi et al. [56] proposed an improved STC algorithm by incorporating a context-aware correlation filter in the STC framework. Wei et al. [57] proposed an enhanced STC algorithm. They included a bag of multiple models (BMM) for target appearance in the STC framework and used the Bayesian Kalman filter for tracking.

2.5 Correlation Filter

The tracking framework of the correlation filter is shown in Fig. 2.3. During tracking correlation filter is trained from the first frame in which the target location is initialized. Then for each following frame target location is detected. Afterward, features are extracted, and a cosine window is applied for boundary effects. Fast Fourier Transform (FFT) is used to compute the vector. A response map of the target function can be obtained by calculating inverse FFT. The target's position in the next frame corresponds to the coordinates of the maximum value in the response map. Afterwards, appearance is estimated by extracting and updating the correlation filter.

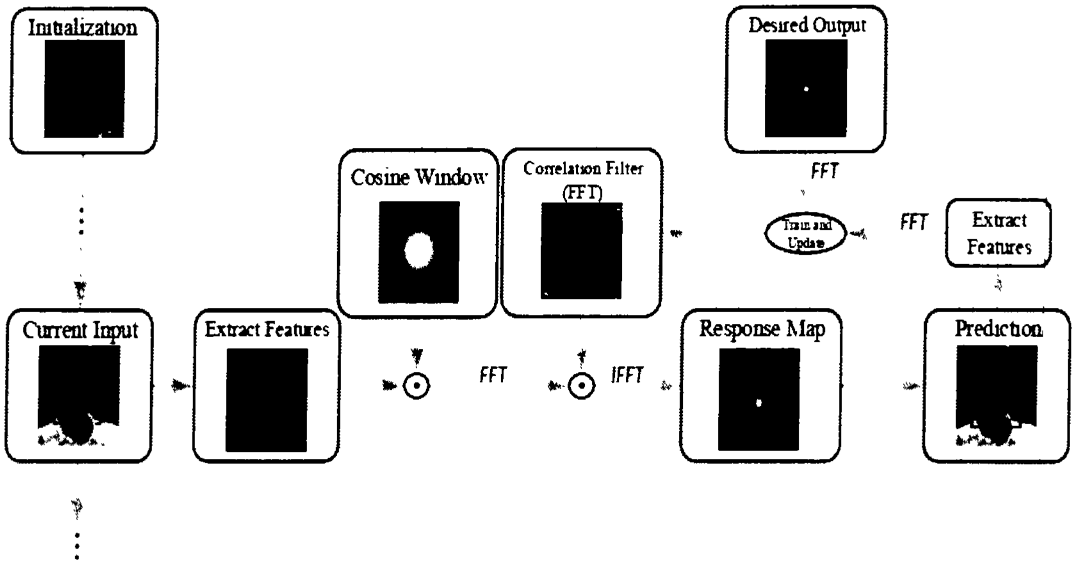


Fig. 2.3 Correlation filter tracking framework [32]

2.5.1 Kernelized Correlation Filter

Correlation filters use a sampling method to discriminate the target position from the region of interest in consecutive frames at a low computational cost. It models all possible translations of the target in the search window as circular shifts and concatenates them to form a square matrix A_0. Thus, it facilitates computing the Fourier domain solution to the ridge regression problem presented in (2.12).

$$\min_{w} \|A_0 w - y\|_2^2 + \lambda_1 \|w\|_2^2 \qquad (2.12)$$

In (2.12), the learned correlation filter is denoted by vector w. Square matrix A_0 contains all circular shifts of image patch and regression target y is vectorized image of 2D Gaussian. Let $x(j)$ be the j^{th} component of vector x, and its conjugate is x^*. Then, its Fourier transform $F^H x$ is $\hat{x}$. (2.12) can be solved using (2.13).

$$X = Fdiag(\hat{x})F^H \text{ and } X^T = Fdiag(\hat{x}^*)F^H \qquad (2.13)$$

Convex in (2.12) is complex and has a unique global minimum. Equating its gradient to zero leads to a closed-form filter solution as given in (2.14).

$$w = (A_0^T A_0 + \lambda_1 \mathrm{I})^{-1} A_0^T y \tag{2.14}$$

As A_0 is circulant, (2.14) can be diagonalized, and its solution in the Fourier domain is given in (2.15).

$$\widehat{w} = \frac{\hat{a}_0^* \odot \hat{y}}{\hat{a}_0^* \odot \hat{a}_0 + \lambda_1} \tag{2.15}$$

The target location is the same as the location of maximum response when (2.15) is convolved with the search window for the next frame. Finally, the detection formula is given in (2.16).

$$r_p(w, \mathrm{Z}) = \mathrm{Z}w \leftrightarrow \hat{r}_p \odot \widehat{w} \tag{2.16}$$

where Z is the search window circulant matrix.

2.5.2 Improvements in CF and KCF

Both minimum output sum of squared error (MOSSE) [58] and KCF [59] achieve favorable tracking results. However, significant improvements have been made in these in recent years few of them are presented below.

Ahmed et al. [60] proposed a real-time correlation-based tracking framework by utilizing an open-loop control strategy so that the target is always at the center of the frame. Moreover, a video stabilization method was incorporated to eliminate the vibration at low computational cost. Ma et al. [61] proposed a long-term correlation filter tracker (LCT) that decomposed the tracking problem into the estimation of translation and scale and redetected the target by online training of a random fern classifier. Mueller et al. [23] proposed a context-aware framework for correlation filter trackers by reformulating original optimization problem for single and multidimensional features in primal and dual domains. Masood et al. [62] proposed a tracking framework that uses a maximum average correlation height (MACH) filter for detection and a proximal gradient algorithm-based particle filter for tracking. Khan et al. [27] proposed an improved tracking algorithm based on LCT. They incorporated the Kalman filter in the LCT framework for occlusion handling and the PSR of the response map for occlusion detection. Ali et al. [63] proposed a tracking algorithm that combines mean-shift tracker, Kalman filter, and correlation filter heuristically. It updates the template based on the change in the appearance model of the target and computes similarity for each forthcoming frame based on the current frame similarity value.

Shin et al. [24] proposed an improved KCF based tracking algorithm. They incorporated a module to detect tracking failure, a mechanism for re-tracking in multiple search windows, and an analysis of motion vectors for deciding on search window in the KCF framework. Zhang et al. [25] proposed a motion-aware correlation filter (MACF) that predicts the position and scale of the target in the next frame by utilizing instantaneous motion estimation. Farahi et al. [64] proposed a probabilistic Kalman filter (PKF) to improve target estimation. They incorporated a different stage in the Kalman filter to refine estimated positions by constructing an observed trajectory-based probabilistic graph, further refined by applying the Viterbi algorithm. Khalkhali et al. [65] proposed an improved Kalman filter-based vehicle tracking method. They incorporated situation assessment and motion history graph in Kalman filtering framework for estimation of target

Danelljan et al. [66] proposed a tracker based on correlation filters for translation and scale in image scale pyramid representation. Ma et al. [67] proposed a fast and accurate scale estimation method by incorporating average peak to correlation energy (APCE) in multi-resolution translation filter. Ruan et al. [68] proposed an online scale adaptive tracker by formulating target insight correlation filters between the target and its context by using target likelihood map weights of the image. Li et al. [69] incorporated a scale adaptive tracking method in the KCF framework. They addressed the issue of fixed template size in KCF and included HoG and CN features. Bibi et al. [70] modify the KCF tracker by maximizing posterior distribution over the scales grid and updating the filter by fixed point optimization. Danelljan et al. [71] proposed a tracker based on discriminative scale space tracking by utilizing a scale correlation filter for sampling target appearance at different scales. It also uses various strategies for reducing computational cost. Yin et al. [72] modified the scale adaptive with multiple features (SAMF) tracker by using APCE-based rate of change between consecutive frames to control scale size. Khan et al . [73] proposed a multi cue robust object tracking framework by incorporating scale invariant features in modified KCF framework. Lu et al .[74] combined KCF and Fourier–Mellin transform to deal with rotation and scale variation of the target. Yuan et al. [75] proposed a scale adaptive object tracking algorithm. They extracted features from different layers of ResNet to produce a response map which is further fused in the AdaBoost algorithm. Moreover, an occlusion detection technique based on the number of effective local maxima (NELM) and scale correlation filter is incorporated in their framework to handle scale variations.

2.6 Summary

This chapter presented a detailed overview of VOT methods based on STC and CF. A comprehensive review of STC and CF was given along with the standard procedures to identify the target position for the VOT system. Finally, the state-of-the-art tracking algorithms based on STC, CF, and KCF are discussed.

Chapter 3.

Proposed Tracking Methods

3.1 Introduction

This chapter presents the design of algorithms based on the Spatio-temporal context presented for visual object tracking systems. Two tracking schemes are proposed, namely:

1. Context-aware and occlusion handling mechanism for VOT
2. Correlation filter and measurement estimation collaboration for VOT

Details of both tracking schemes are presented below.

3.2 Context-aware and occlusion handling mechanism for VOT

In this section, the proposed tracker is introduced in detail. First, a context-aware object tracking model is investigated. Second, the Kalman filter-based motion estimation model is discussed. Third, the average difference of consecutive frames-based model update scheme is presented. Finally, the tracker will be mentioned in Algorithm 1. Fig. 3.1 shows the flowchart of the proposed algorithm.

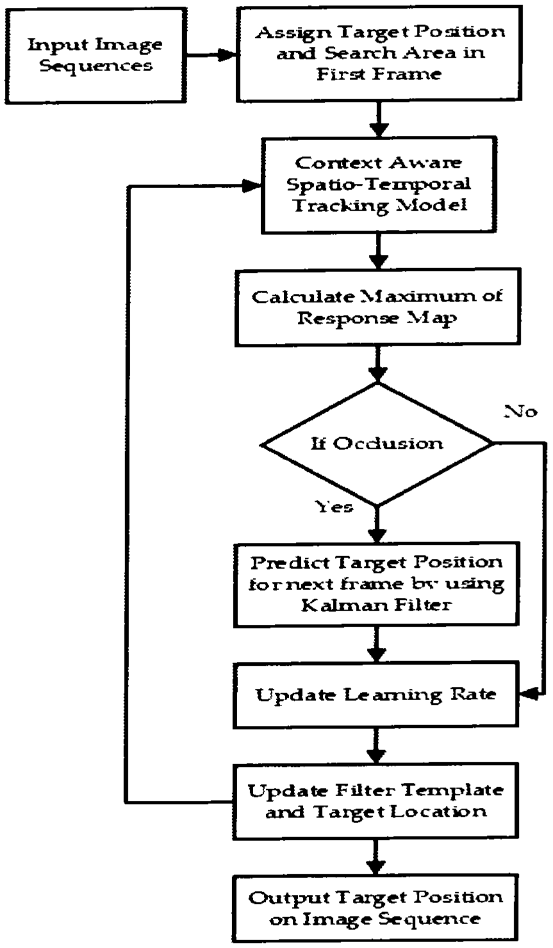

Input Image Sequences
Assign Target Position and Search Area in First Frame
Context Aware Spatio-Temporal Tracking Model
Calculate Maximum of Response Map
If Occlusion
No
Yes
Predict Target Position for next frame by using Kalman Filter
Update Learning Rate
Update Filter Template and Target Location
Output Target Position on Image Sequence

Fig. 3.1 Execution Mechanism of case study 1

3.2.1 Context-Aware Tracking Framework

As information of context around the target elevates the tracking performance. Therefore, it is added to the context-aware correlation filter solution as given in (3.1).

$$\min_{w} \quad \|A_0 w - y\|_2^2 + \lambda_1 \|w\|_2^2 + \lambda_2 \sum_{i=1}^{k} \|A_i w\|_2^2 \tag{3.1}$$

It should be noted that there are other probable selections for incorporating context terms. However, it leads to constrained convex optimization requiring an iterative relatively slow solution. When the position for the current frame is computed by STC, then the filter w is trained and the background term A_i is as small as possible. The objective function can be rewritten by forming a new data matrix $B \in \mathbb{R}^{(k+1)n\times n}$ which consists of target and context patches as given in (3.2).

$$f_p(\mathrm{w}, \mathrm{B}) = \quad \|\mathrm{Bw} - \bar{y}\|_2^2 + \lambda_1 \|w\|_2^2 \tag{3.2}$$

$$\text{where } \mathrm{B} = \begin{bmatrix} \mathrm{A_0} \\ \sqrt{\lambda_2}\,\mathrm{A_1} \\ \vdots \\ \sqrt{\lambda_2}\,\mathrm{A_k} \end{bmatrix} \text{ and } \bar{y} = \begin{bmatrix} \mathrm{y} \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Like the correlation filter, the function in (3.2) is convex and minimized by setting the gradient to zero. It is presented in (3.3).

$$w = \quad (\mathrm{B^T B} + \lambda_1\, \mathrm{I})^{-1}\mathrm{B^T}\bar{y} \tag{3.3}$$

Similar to (2.12), using (2.13) to determine Fourier domain closed-form solution as described in (3.4).

$$\hat{\mathrm{w}} = \frac{\hat{\mathrm{a}}_0^* \odot \hat{\mathrm{y}}}{\hat{\mathrm{a}}_0^* \odot \hat{\mathrm{a}}_0 + \lambda_1 + \lambda_2 \sum_{\mathrm{i}=1}^{\mathrm{k}} \hat{\mathrm{a}}_\mathrm{i}^* \odot \hat{\mathrm{a}}_\mathrm{i}} \tag{3.4}$$

The target window and its position are updated according to (3.4). Furthermore, the confidence map and STC model in (2.9) and (2.11) are updated based on the target position.

3.2.2 Kalman Filter Based Motion Estimation Model

The Kalman Filter is an optimal filter that minimizes the difference between true and estimated states. It consists of four processes which are 1) Initial guess of the state vector and state error covariance, 2) Forward time step propagation of state vector and state error covariance, 3) Estimation of Kalman gain based on state error covariance and measurement noise covariance, 4) Update state vector and state error covariance based on estimated output and Kalman gain [76]. The constant velocity motion model is used due to its simplicity and effectiveness in describing the target's motion. It consists of two stages which are prediction and correction.

3.2.2.1 Kalman Filter Prediction

During this state, uncertainty about the target is determined by both state and covariance prediction. The current system state can predict position based on the previous state. Similarly, covariance is calculated by multiplying the covariance matrix from the last iteration by the state transition matrix and adding process noise Q. The prediction equations are described in (3.5) and (3.6).

$$X_t = AX_{t-1} + Bu_{t-1} \tag{3.5}$$

where X_t is the state target vector, A is the state transition matrix, and Bu_{t-1} is noise.

$$S_t = AS_{t-1}A^T + Q \tag{3.6}$$

where S_t is the predicted error covariance and Q is the covariance of the process noise

3.2.2.2 Kalman Filter Correction

The position of the target obtained from STC is used as a measurement value Y_t.By combining it with the predicted result, Kalman gain can be calculated as described in (3.7).

$$K_{t-1} = S_{t-1}H^T (HS_{t-1}H^T + R)^{-1} \tag{3.7}$$

where R is the measurement noise covariance. The estimate is updated by combining it with old estimate and measurement as given in (3.8).

$$X_{t+1} = X_t + K_{t-1}(Y_t - HX_t) \tag{3.8}$$

The difference $(Y_t - HX_t)$ is called measurement innovation or residual. It reflects the discrepancy between predicted measurement HX_t and actual measurement Y_t. Error covariance is calculated using (3.9).

$$S_{t+1} = (I - K_t H) S_t \tag{3.9}$$

where S_{t+1} is the updated error covariance, H is matrix related to the measurement of the state and K_t is the updated Kalman gain.

3.2.3 Occlusion Detection

When the target undergoes occlusion, the STC model is updated incorrectly, thereby losing the target. To detect occlusion, the maximum value of the target map is used, which changes its value with the situation of the target state. If the target is occluded, then the value of the response map is small. However, when the target reappears then its value increases. The value of the response map determines whether the target is tracked by STC or by the Kalman filter. For a given input image sequence, the first confidence map is calculated in the frequency domain, and then the Spatio-temporal model is learned for tracking. If the target is severely occluded, the next frame Kalman filter will predict the position and update STC using a feedback loop. The filter template for context-aware is updated accordingly. Kalman filter prediction can be updated as observation of the target position marked for next frame.

3.2.4 Adaptive Learning Rate

During object tracking, target motion changes in each image sequence frame. Therefore, it is necessary to update the target model correctly. In STC, the learning rate is fixed, making it evitable to different appearances in the environment. So, to make it adaptive, an average difference of two consecutive frames-based mechanisms is incorporated [46]. It is given in (3.10).

$$er = \frac{\sum_{i,j}^{M,N} \left| I_{ij}^{n} - I_{ij}^{n-1} \right|}{M * N} \tag{3.10}$$

where I_{ij} is the pixel value, and $M * N$ is the size of an image. The learning rate is adjusted as given in (3.11).

$$\rho = \begin{cases} 0.005, & er < 1.2 \\ 0.075, & 1.2 \leq er < 10 \\ 0.1, & er > 10 \end{cases} \tag{3.11}$$

Value of learning rate ρ is assigned based on er using (3.11).

Algorithm 1: Proposed Tracker at time step t

```
Input: Image Sequence of n Frames. Position of Target at First Frame.
Output: Target Positions in each frame for Image Sequences.
for frame 1 to n frames.
    Calculate context prior model using (2.3).
    Calculate the confidence map using (2.10).
    Calculate the target center.
    Calculate the maximum of the response map.
    if response map<threshold
        new position=Kalman prediction
    end
    Estimate position for next frame using (3.5).
    Estimate error covariance using (3.6).
    Calculate Kalman gain using (3.7).
    Update estimate via measurement using (3.8).
    Update error covariance using (3.9).
    Calculate the average difference between consecutive frames using (3.10).
    Adjust learning rate using (3.11).
    Update filter template using (3.4).
    Update context prior model on Kalman prediction using (2.3).
    Update spatial context model using (2.8).
    Update the Spatio-temporal context model using (2.11).
    Draw a rectangle on the target in each frame.
end
```

3.3 Correlation filter and measurement estimation collaboration for VOT

In this section, the proposed tracker will be discussed. First, a correlation filter-based adaptive scale scheme is discussed. Second, an extended Kalman filter-based occlusion handling mechanism is investigated. Third, an adaptive learning rate scheme is presented. The execution scheme of the proposed tracker is shown in Fig. 3.2. In each image sequence, the target of interest's location is initialized manually on the first frame from the given ground truth. Afterward, the target confidence map is calculated. Next, sample patches of a different set of scales are estimated from the confidence map of STC. Then, the maximum value of the response map is calculated. The extended Kalman filter is activated if the response map's value is less than the fixed threshold. Kalman filter will predict the location of the next frame and update the tracking model during this entire period. Once the response map's value exceeds the fixed threshold, the Kalman filter is deactivated. Afterwards, the learning rate is updated, and the target entire tracking model is updated based on the calculated position.

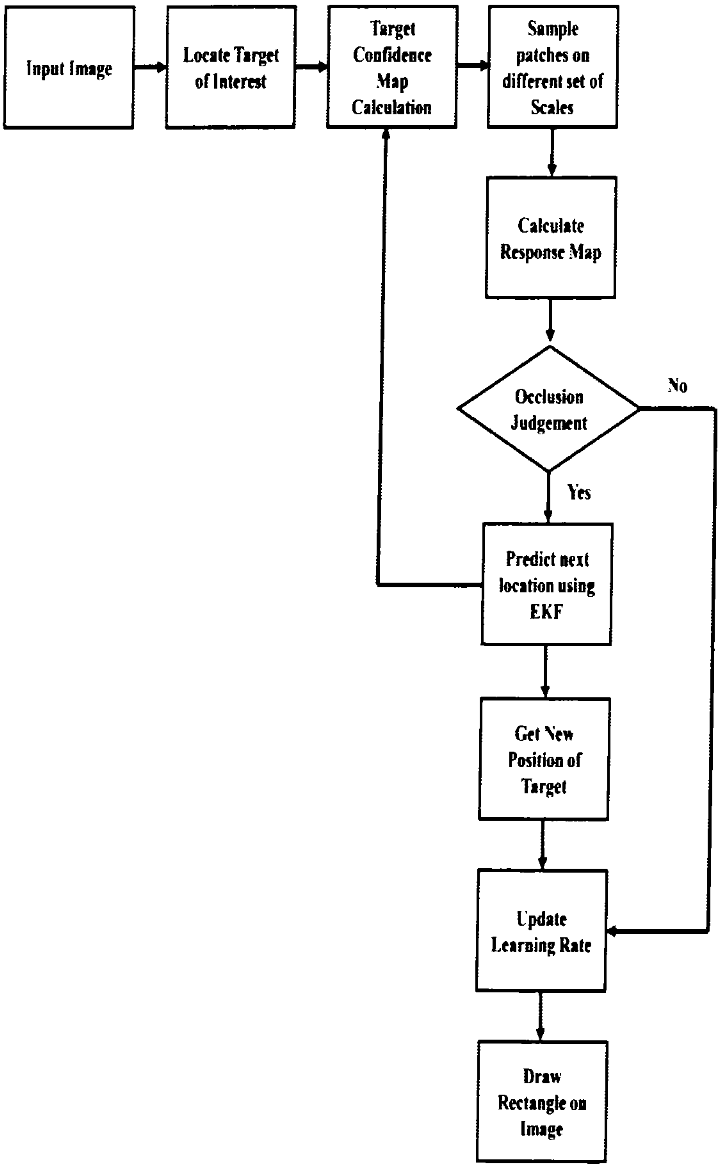

Input Image
Locate Target of Interest
Target Confidence Map Calculation
Sample patches on different set of Scales
Calculate Response Map
Occlusion Judgement
No
Yes
Predict next location using EKF
Get New Position of Target
Update Learning Rate
Draw Rectangle on Image

Fig. 3.2 Execution mechanism of case study 2

TH-25755.

3.3.1 Scale Space Tracking

Discriminative correlation filters are widely used in visual object tracking. A scale correlation filter-based tracking model is used to estimate the target scale. It first extracts different scale samples around the target position; then, the HOG feature pyramid sample is extracted from the location. For finding an optimal correlation filter, the cost function given in (3.12) needs to be minimized.

$$\varepsilon = \left\|\sum_{l=1}^{d} h^l * f^l - g\right\|^2 + \lambda \sum_{l=1}^{d} \left\|h^l\right\|^2 \tag{3.12}$$

where g is desired output, λ is the regularization term, * is the circular convolution operator, h is the HOG features after extracting from the sample, l indicates l-dimensional HOG features, g indicates two-dimensional Gaussian function, d indicates the total dimension of HOG features, and f is the correlation filter. The solution of (3.12) in the frequency domain is given in (3.13).

$$H^l = \frac{\bar{G}F^l}{\sum_{k=1}^{d} \overline{F^k}F^k + \lambda} \tag{3.13}$$

where H^l is 1-D HOG feature extracted from sample, λ is the regularization term and $\bar{G}$ is 2-D Gaussian function. An optimal filter can be obtained by minimizing output error over training patches. However, it is not suitable for online tracking because of computational cost. For efficient tracking numerator and denominator of correlation filter H^l are updated separately as given in (3.14) and (3.15).

$$A_t^l = (1 - \gamma)A_{t-1}^l + \gamma\bar{G}_t F_t^l \tag{3.14}$$

$$B_t = (1 - \gamma)B_{t-1} + \gamma \sum_{k=1}^{d} \overline{F_t^k} F_t^k \tag{3.15}$$

where γ is the learning rate. By maximizing the correlation score target state can be determined as given in (3.16).

$$y = F^{-1}\left\{\frac{\sum_{l=1}^{d} \bar{A}^l Z^l}{B + \lambda}\right\} \tag{3.16}$$

where Z^l denote HOG features extracted from prediction

3.3.2 Extended Kalman Filter

Within the visual object tracking research area, EKF is widely used to estimate the system. The target location problem can be viewed as an estimation problem, providing measurement-based prediction. For the current estimate, EKF linearizes the nonlinear equations. Afterward, EKF is applied to that linearized model [76]. Thus, EKF involves two steps which are prediction and correction. During prediction, state and covariance estimates are computed for the current frame using (3.17) and (3.18).

$$x_t^- = A\hat{x}_{t-1} + Bu_t \tag{3.17}$$

$$P_t^- = AP_{t-1}A^T + Q \tag{3.18}$$

where x_t^- is the state target vector, Bu_t is noise, A is process Jacobian, Q is process noise covariance and P_t^- is the predicted error covariance. During the correction, Kalman gain K_t is calculated. It balances prior estimation uncertainty and measurement noise as presented in (3.19).

$$K_t = P_t^- J_H^T (J_H P_t^- J_H^T + R)^{-1} \tag{3.19}$$

where J_H is measurement Jacobian and R is measurement noise. State estimate is updated using prior estimate and error between measurement and predictive measurement as given in (3.20).

$$x_t = \hat{x}_t^- + K_t(z_t - J_H\hat{x}_{t-1}^-) \tag{3.20}$$

The difference $(z_t - J_H\hat{x}_{t-1}^-)$ is called measurement innovation or residual. It reflects the discrepancy between predicted measurement $J_H\hat{x}_{t-1}^-$ and actual measurement z_t.Posteriori estimation of variance is given in (3.21).

$$P_t = (I - K_t J_H)P_t^- \tag{3.21}$$

where P_t is the updated error covariance, J_H is matrix related to the measurement of the state and K_t is the updated Kalman gain.

3.3.3 Occlusion Detection

When the target undergoes occlusion, the STC model is updated incorrectly, thereby losing the target. The maximum value of the target map is used to detect occlusion, which changes its value with the target state's situation. If the target is occluded, then the value of the response map is small. However, when the target reappears then its value increases. The value of the response map determines whether the target is tracked by improved STC or by EKF. For a given input image sequence first confidence map is computed in the frequency domain. If the target is severely occluded, EKF will predict the position and update improved STC using a feedback loop for the next frame.

3.3.4 Adaptive Learning Rate

The model is updated adaptively using average peak to correlation energy (APCE) [77]. It is defined in (3.22).

$$\mathrm{APCE_t} = \frac{|\mathrm{f_{max}} - \mathrm{f_{min}}|^2}{\mathrm{mean}\left(\sum_{\mathrm{w,h}}\left(\mathrm{f_{w,h}} - \mathrm{f_{min}}\right)^2\right)} \tag{3.22}$$

where $\mathrm{f_{max}}$ is maximum response value,$\mathrm{f_{min}}$ is minimum response value and $\mathrm{f_{w,h}}$ is the corresponding row and column value of the response map. APCE specified the degree of fluctuation between response maps and detected targets. (3.23) gives expression of model update.

$$\begin{cases} \mathrm{bz_t} = \dfrac{\mathrm{APCE_t}}{\mathrm{APCE_0}} \\ \gamma_t = \gamma_0\,, \quad \mathrm{bz_t} > \mathrm{bz_0} \\ \gamma_t = \gamma_0 . \mathrm{bz_t}\,, \quad \text{otherwise} \end{cases} \tag{3.23}$$

where $APCE_t$ is the value at t-th frame,$APCE_0$ is the value at the initial frame and bz_0 is the threshold to decide the learning rate.Algorithm 2 is presented below.

Algorithm 2: Proposed Tracker at time step t

```
Input: Image Sequence of n Frames. Position of Target at First Frame.
Output: Target Position in each frame for Image Sequences.
for frame 1 to n frames.
Calculate context prior model using (2.3).
Calculate the confidence map using (2.10).
Calculate the target center.
Calculate translation correlation using (3.16).
Calculate the maximum value of the response map
if response map<threshold
    new position=Kalman prediction
end
Calculate Kalman gain using (3.19).
Estimate position for next frame using (3.20).
Estimate error covariance using (3.21).
Calculate APCE using (3.22).
Update model using (3.23).
Calculate scale correlation using (3.16).
Update translation and scale model using (3.14) and (3.15).
Update context prior model using (2.3).
Update spatial context model using (2.8).
Update the Spatio-temporal context model using (2.11).
Calculate the target position for each frame.
Draw a rectangle on the target in each frame.
end
```

In case study 1, STC is modified by incorporating KF, context aware formulation and average of difference based learning rate. In case study 2, STC is modified by incorporating EKF, discriminative scale filter context and APCE based learning rate mechanism. The only similarity between both case studies is the use of peak response value for occlusion detection.

3.4 Summary

First, the building blocks of proposed tracking mechanisms are presented in this chapter. Then, the design of proposed trackers based on STC, i.e., context-aware and occlusion handling mechanism for VOT and correlation filter and measurement estimation collaboration for VOT, are presented.

Chapter 4.

Results and Discussion

4.1 Introduction

In this chapter, the results of both proposed tracking algorithms are presented. The results given in this chapter are based on our published research works. To evaluate the performance of the proposed trackers both qualitatively as well as quantitatively, extensive experiments were conducted on image sequences selected from Temple Color (TC)-128 [78], OTB2013 [21], OTB2015 [79], and UAV123 [80] datasets. The proposed trackers are compared quantitatively with existing tracking methods based on distance precision rate (DPR) and centre location error (CLE). CLE is defined as the Euclidean distance calculated between the tracker and the ground truth of target. The calculation formula is mentioned in (4.1).

$$\mathrm{CLE} = \sqrt{(x_{\mathrm{i}} - x_{\mathrm{gt}})^2 + (y_{\mathrm{i}} - y_{\mathrm{gt}})^2} \tag{4.1}$$

where $(x_{i,}, y_i)$ are positions calculated by tracking algorithm and (x_{gt}, y_{gt}) are ground truth values. DPR is the percentage of frames when the distance threshold is greater than the estimated CLE. MATLAB software version 2018b is used for simulations on a intel core i3 processor with 16 GB Ram.

4.2 Case Study 1: Context-aware and occlusion handling mechanism for VOT

4.2.1 Quantitative Analysis

DPR comparison is given in Table 4.1. In sequences (Cardark, Cup, Jogging-1, Juice, and Man) proposed tracker outperforms MOSSE$_{CA}$, STC, MACF, and DCF$_{CA}$. In sequences (Carchasing_ce3 and Plate_ce2), all tracking methods have similar performance. In sequence, Busstation_ce2 proposed has slightly less precision value. However, the proposed has a higher mean value than other tracking methods.

Table 4.1. Distance precision rate at threshold of 20 pixels.

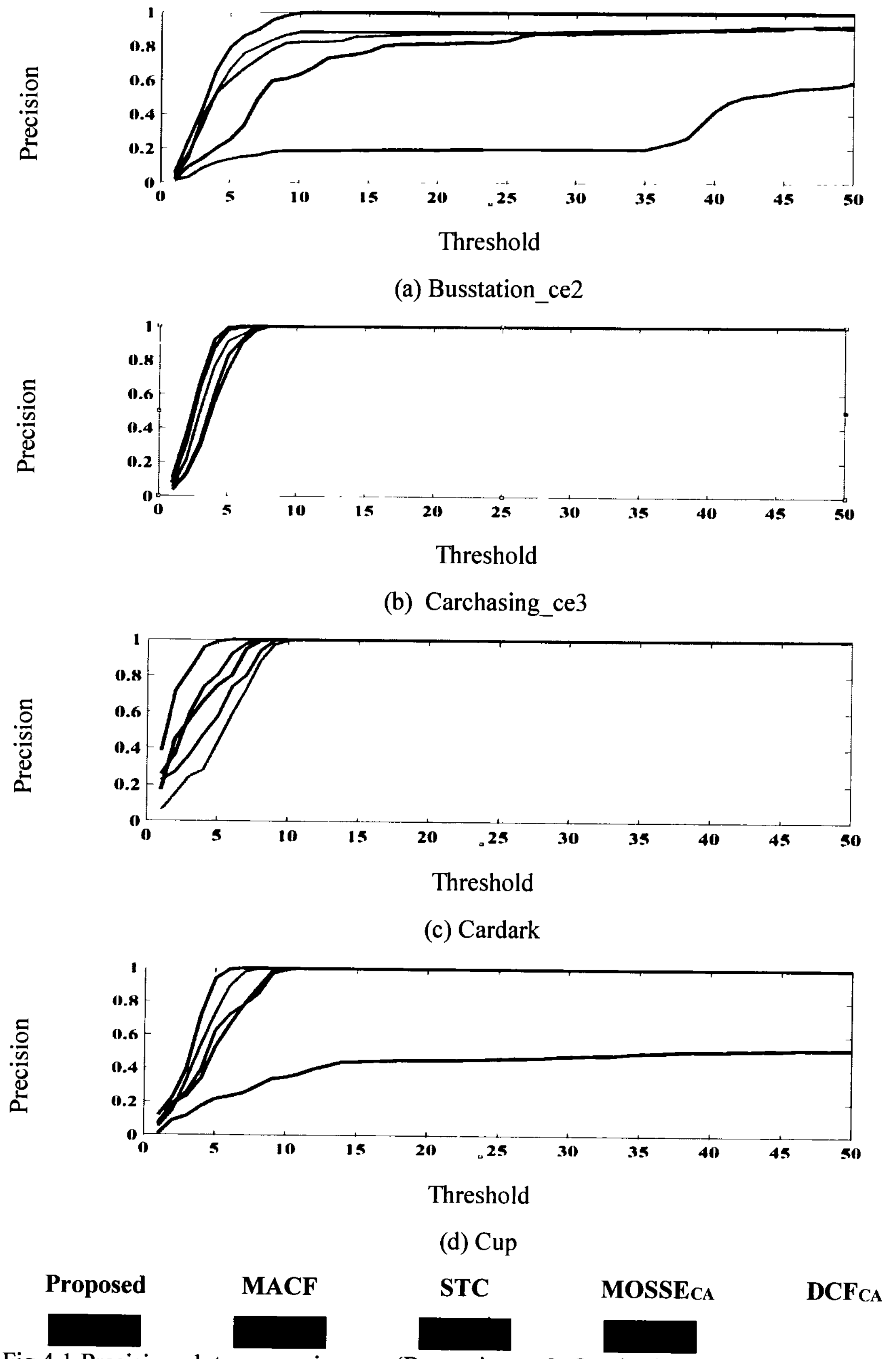
Sequence	Proposed	STC	MACF	MOSSE$_{CA}$	DCF$_{CA}$
Busstation_ce2	0.878	0.194	**1**	0.820	0.886
Carchasing_ce3	**1**	1	1	1	1
Cardark	**1**	1	1	1	1
Cup	**1**	1	1	0.452	1
Jogging-1	**0.996**	0.228	0.231	0.231	0.231
Juice	**1**	1	1	1	1
Man	**1**	1	1	1	1
Plate_ce2	**1**	1	1	1	1
Mean Precision	**0.984**	0.803	0.904	0.813	0.890

The average centre location error comparison is given in Table 4.2. In sequences (Busstation_ce2, Cup, Jogging-1, and Man) proposed tracker outperforms STC, MOSSE$_{CA}$, MACF, and DCF$_{CA}$. The sequences (Carchasing_ce3, Cardark, Juice, and Plate_ce2) proposed has slightly high error values. However, the proposed has the lowest mean error and has the lowest mean error of other tracking methods.

Table 4.2. Average Centre Location Error.

Sequence	Proposed	STC	MACF	MOSSE$_{CA}$	DCF$_{CA}$
Busstation_ce2	10.86	78.25	**3.58**	14.50	9.71
Carchasing_ce3	3.90	3.55	**2.39**	2.61	3.05
Cardark	4.09	2.83	**1.67**	3.15	5.11
Cup	4.63	4.84	**3.11**	95.87	3.85
Jogging-1	**8.40**	5010	94.93	115.98	89.44
Juice	4.63	5.08	**0.91**	3.71	1.92
Man	**1.32**	1.49	1.73	1.72	2.23
Plate_ce2	2.58	2.34	**1.62**	1.77	1.83
Mean Error	**5.05**	638.55	13.74	29.91	14.64

The precision plots are shown in Fig. 4.1 and Fig. 4.2. These plots provide frame by frame precision in entire image sequences. Since precision gives the mean value of the entire image sequence, the tracker might get the drift for few frames but then again tracks the target correctly. Therefore, these plots were presented to show the efficacy of the tracking method. Various challenges were present in sequences such as occlusion, illumination variations, background clutter, etc. The sequences (Carchasing_ce3, Cardark, Cup, Jogging-1, Juice, Man, and Plate_ce2) reported has the highest precision in the entire sequence. In sequence, Busstation_ce2 reported has slightly low precision.



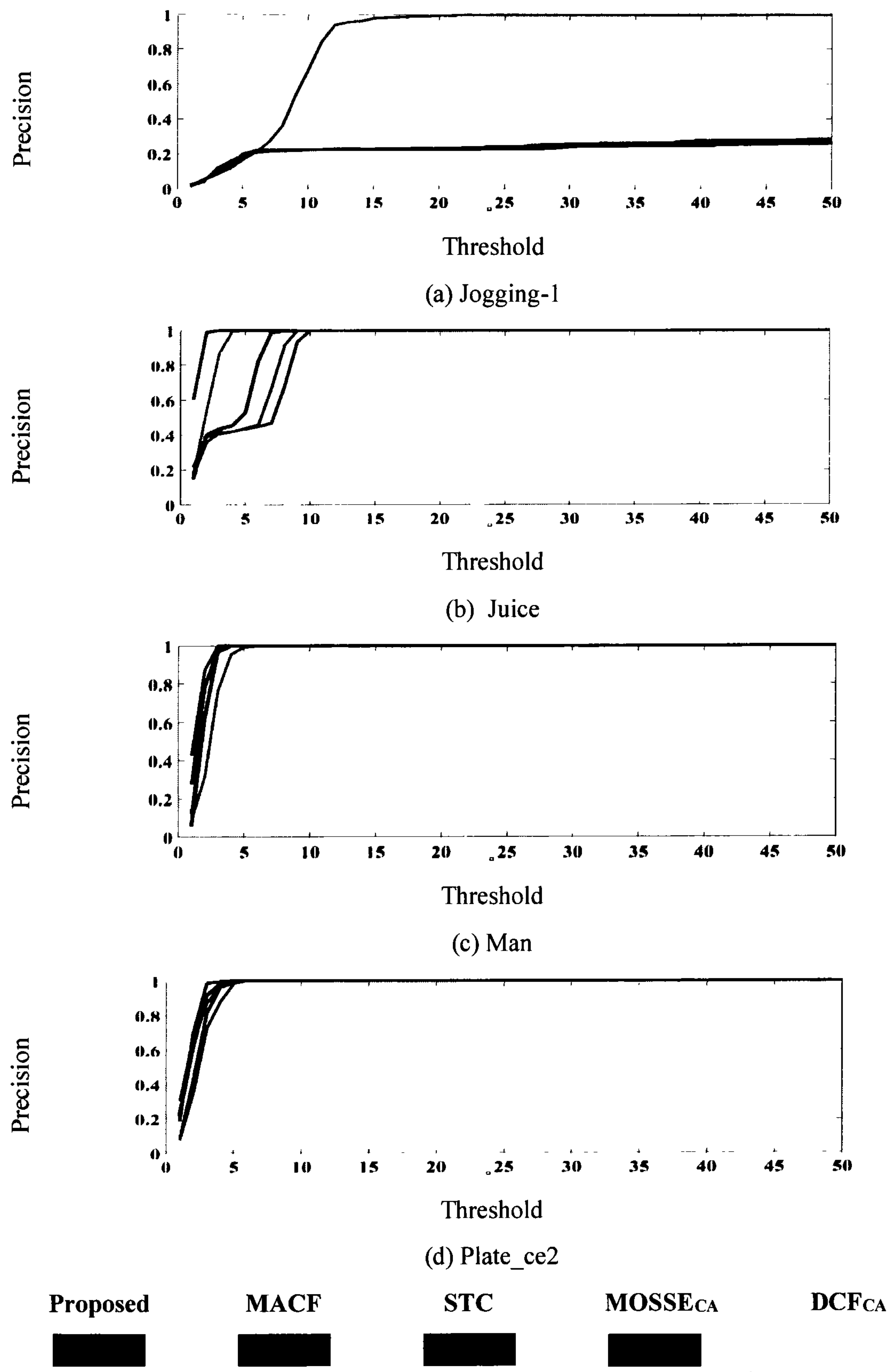
(a) Busstation_ce2

(b) Carchasing_ce3

(c) Cardark

(d) Cup

Fig 4.1 Precision plots comparison on (Busstation_ce2, Carchasing_ce3, Cardark, and Cup) sequences.



(a) Jogging-1

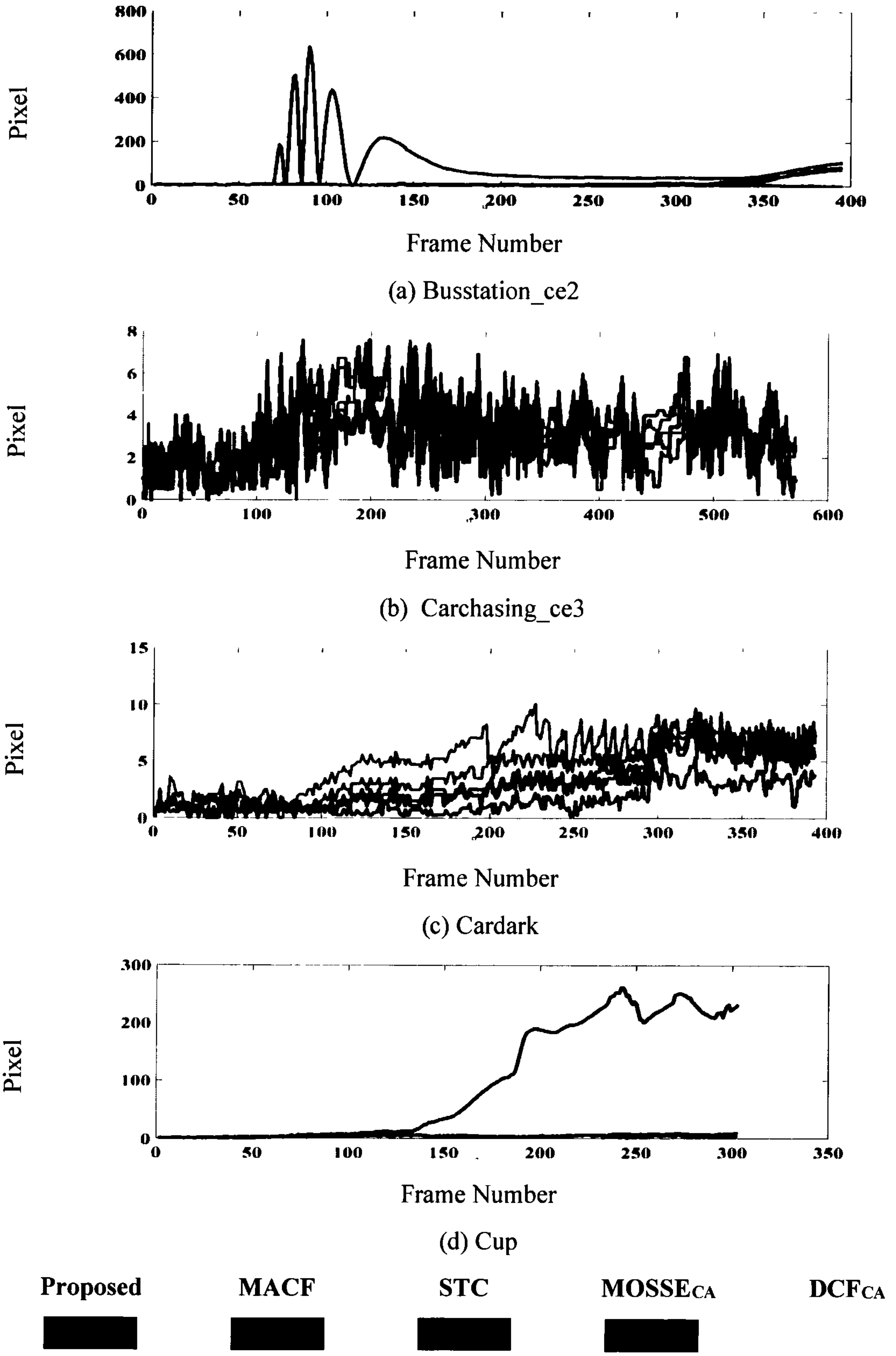
(b) Juice

(c) Man

(d) Plate_ce2

Fig 4.2. Precision plots comparison on (Jogging-1, Juice, Man and Plate_ce2) sequences.

The location error plots are shown in Figs. 4.3 and Fig. 4.4. These plots provide frame-by-frame errors in entire image sequences. Since average centre location gives mean error of entire image sequence. It is possible that the tracker might get the drift for few frames but then again tracks the target correctly. Therefore, these plots were presented to show the effectiveness of the tracking method. Various challenges were present in sequences such as occlusion, illumination variations, deformation, etc. In sequences (Busstation_ce2, Cup, Jogging-1, and Man) the lowest error was reported in the entire sequence. In sequences (Carchasing_ce3, Cardark, Juice, and Plate_ce2) a slightly high error was reported.



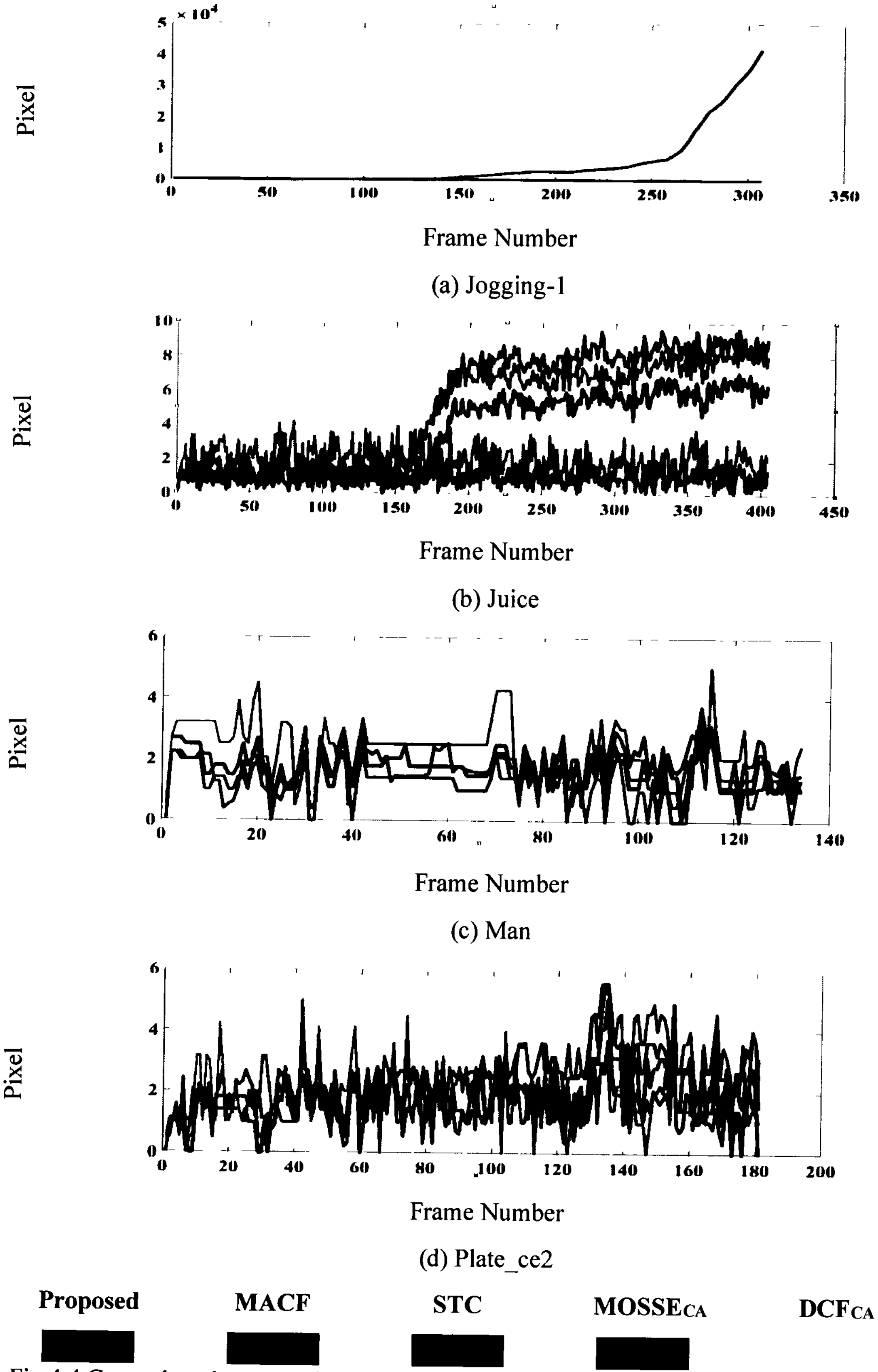
(a) Busstation_ce2

(b) Carchasing_ce3

(c) Cardark

(d) Cup

Fig 4.3 Centre location error (in pixels) comparison on (Busstation_ce2, Carchasing_ce3, Cardark, and Cup) sequences.



(a) Jogging-1

(b) Juice

(c) Man

(d) Plate_ce2

Fig 4.4 Centre location error (in pixels) comparison on (Jogging-1, Juice, Man, and Plate_ce2) sequences.

4.2.2 Experimental Results

Qualitative results of proposed tracking with four state-of-the-art trackers over eight image sequences are shown in Fig. 4.5. It involves various challenges such as partial or full occlusions, illumination variations, background clutter, etc. DCF_{CA} and MOSSE_{CA} contain similar tracking components as our approach, i.e., correlation filtering and context-aware formulation. However, the correlation filter in MOSSE_{CA} and DCF_{CA} is not robust to blur motion in (Cup), illumination variations in (Man, Cardark), and occlusions in (Jogging-1, Busstation_ce2). In (Carchasing_ce3 and Plate_ce2), where the target undergoes scale variations, both MOSSE_{CA} and DCF_{CA} have similar performance with the proposed target tracking. With the joint instantaneous motion model and Kalman filter in discriminative scale space tracking frame, MACF performs better on various challenging sequences. However, MACF tends to drift when the target undergoes occlusion and fails to recover from tracking failures (Jogging-1). Although STC can estimate scale, it does not performs well in motion blur (Juice) and scale variations (Cup). This is because STC only uses intensity features and estimate scale from the response map of a single translation filter. Moreover, it does not effectively deal with occlusion (Jogging-1, Busstation_ce2) as there is no occlusion handling mechanism to deal with this issue. Moreover, its target model is updated on a fixed learning rate, making it vulnerable to the background environment.

The proposed tracker performs well in all these challenging sequences. This performance can be attributed to three reasons. First, the context-aware formulation in the STC framework is incorporated, making it less sensitive to illumination variation(Cardark and Man) and motion blur (Juice, Man, and Cup). Second, the incorporation of occlusion detection based on the response map and occlusion handling using the Kalman filter makes it effective towards for partial or full occlusion(Jogging-1, and Busstation_ce2). Third, the fusion of adaptive learning rate in the model update of the tracking model effectively deals with scale variation and fast motion (Plate_ce2).

4.2.3 Discussion

We discuss several observations from experimental and quantitative analysis. First, the context-aware formulation in the correlation filter outperforms trackers without this formulation. This can be attributed to the correlation filters regress all circular shifts of the target appearance model. Second, trackers with occlusion detection and handling module outperform trackers without these modules. Again, this can be attributed to that occlusion detection, and the handling mechanism does not lead the tracker to drift. Third, trackers with adaptive learning rate mechanisms perform better than fixed learning rates. Again, it is because it copes tracking model with the changes in the environment.

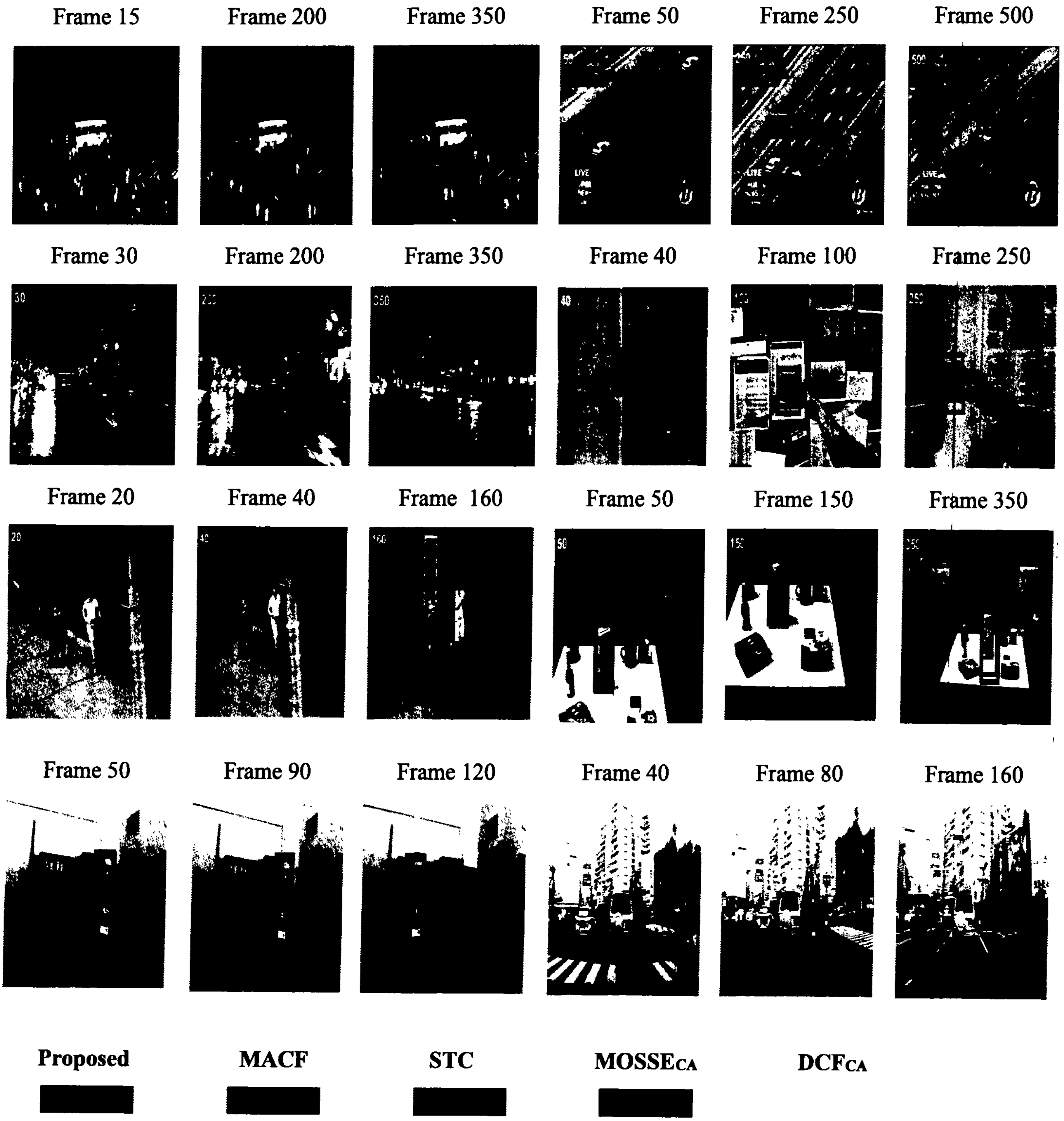


Fig. 4.5 Qualitative comparison on TC-128, OTB2013 and OTB2015 datasets

4.3 Case study 2: CF and measurement estimation collaboration for VOT

4.3.1 Quantitative Analysis

DPR comparison is given in Table 4.3. In sequences (Baby_ce, Car9, Carchasing_ce4, Crossing, Jogging2, Ring_ce, Singer1, Tennis_ce2, and Tennis_ce3), the proposed tracker outperforms Modified KCF, STC, MACF, and $\mathrm{DCF}_{\mathrm{CA}}$. In sequences (Building3, Carchasing_ce3, Cardark, Cup, Juice, Man, Plate_ce2, and Sunshade), all tracking methods have similar performance. In sequences (Bike3, Busstation_ce2, Car4, Girl2, Guitar_ce2, Human3, Jogging1, Skating2, and Walking2), the proposed has slightly less precision value. However, the proposed has a higher mean value than other tracking methods.

Table 4.3. Distance precision rate at the threshold of 20 pixels.

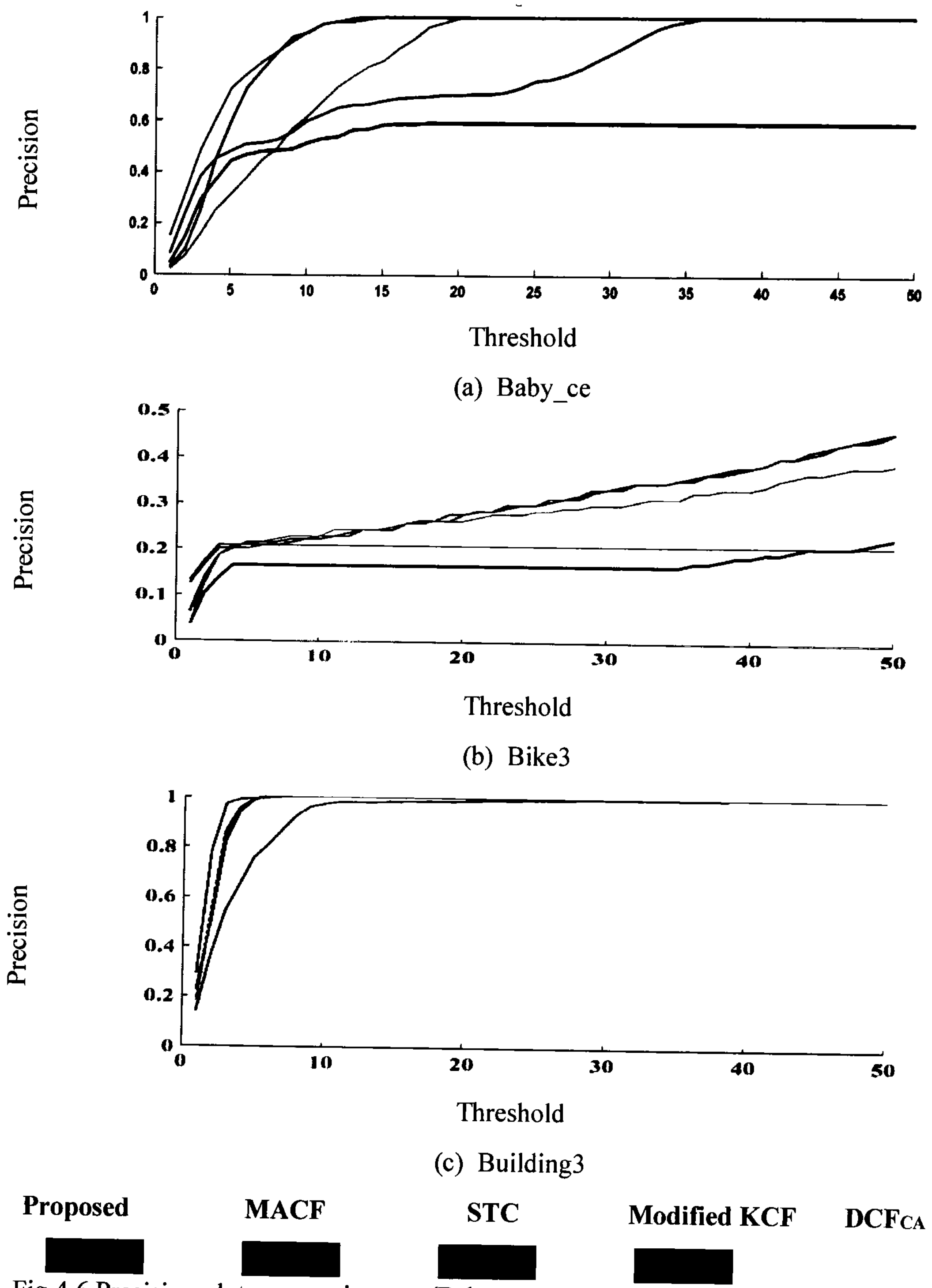
Sequence	Proposed	Modified KCF	Modified STC	STC	AFAM-PEC	MACF	DCFCA
Baby_ce	**1**	0.591	0.591	0.699	0.456	1	0.997
Bike3	0.206	0.166	**0.296**	0.275	0.124	0.275	0.262
Building3	**1**	1	1	1	1	1	1
Busstation_ce2	0.238	0.889	0.878	0.194	**0.927**	1	0.886
Car4	0.997	0.998	0.452	0.991	0.98	**1**	0.998
Car9	**0.988**	0.362	0.976	0.201	0.917	0.988	0.424
Carchasing_ce3	**1**	1	1	1	1	1	1
Carchasing_ce4	**1**	0.400	0.556	0.995	0.199	1	0.717
Cardark	**1**	1	1	1	1	1	1
Crossing	**1**	1	0.575	0.533	1	1	1
Cup	**1**	1	1	1	1	1	1
Girl2	0.830	0.591	0.372	0.262	**0.940**	0.097	0.071
Guitar_ce2	0.568	0.505	0.108	0.524	0.524	0.524	**0.581**
Human3	0.302	0.006	0.018	0.088	**0.795**	0.005	0.006
Jogging 1	0.879	0.993	**0.996**	0.228	0.973	0.231	0.231
Jogging 2	0.980	0.945	0.228	0.185	**0.990**	0.166	0.160
Juice	**1**	1	1	1	1	1	1
Man	**1**	1	1	1	1	1	1
Plate_ce2	**1**	1	1	1	1	1	1
Ring_ce	**1**	0.905	0.129	1	1	1	1
Singer1	**1**	0.815	1	1	1	1	0.843
Skating2	0.074	**0.302**	0.076	0.023	0	0.014	0.423
Sunshade	**1**	1	0.228	1	1	1	1
Tennis_ce2	**1**	0.656	0.101	0.652	1	1	1
Tennis_ce3	**1**	0.098	0.186	0.691	0.108	0.107	0.108
Walking2	0.694	0.408	0.934	0.442	0.722	**1**	0.564
Mean Precision	**0.837**	0.717	0.604	0.653	0.794	0.746	0.703
Variance	0.087	0.110	0.148	0.133	0.110	0.157	0.133

The average centre location error comparison is given in Table 4.4. In sequences (Baby_ce, Car4, Carchasing_ce4, Cardark, Crossing, Plate_ce2, Singer1, Tennis_ce2 and Tennis_ce3) the proposed tracker outperforms Modified KCF, STC, MACF and DCF$_{CA}$. In sequences (Bike3, Building3, Busstation_ce2, Carchasing_ce3, Cup, Girl2, Guitar_ce2, Human3, Jogging1, Jogging2, Juice, Man, Ring_ce, Skating2, Sunshade, and Walking2), the proposed tracker has a slightly high error value. However, the proposed tracker has the lowest mean error than of the other tracking methods.

Table 4.4. Average Centre Location Error.

Sequence	Proposed	Modified KCF	Modified STC	STC	AFAM-PEC	MACF	DCFCA
Baby_ce	**3.93**	81.38	37.25	12.02	40.42	4.89	8.67
Bike3	131.33	123.70	86.37	81.97	**75.20**	83.58	87.73
Building3	2.02	1.96	1.79	**1.50**	1.97	3.76	2.02
Busstation_ce2	79.33	10.74	10.86	78.25	6.2	**3.58**	9.71
Car4	**2.83**	2.93	229.81	4.28	5.04	3.08	2.66
Car9	5.94	210.36	13.69	205.7	3.48	**3.08**	255.42
Carchasing_ce3	2.68	3.04	3.90	3.55	2.52	**2.39**	3.05
Carchasing_ce4	**2.06**	26.73	112.99	2.92	140	2.24	16.68
Cardark	**1.03**	6.04	3.21	2.83	3.35	1.67	5.11
Crossing	**1.23**	2.24	27.05	34.06	4.71	1.64	2.20
Cup	2.82	4.02	4.63	4.84	**2.48**	3.11	3.85
Girl2	30.79	98.93	101.77	200.5	**9.32**	137.62	356.78
Guitar_ce2	19.20	59.91	168.93	29.72	19.12	**19.03**	16.06
Human3	66.31	249.60	348.38	210.8	**15.2**	308.41	257.99
Jogging 1	18.34	**3.72**	8.40	5010	3.87	94.93	89.44
Jogging 2	5.46	**4.74**	43.04	104.02	5.09	148.98	148.33
Juice	2.16	1.96	4.63	5.08	2.42	**0.91**	1.92
Man	2.00	2.36	**1.32**	1.49	2.20	1.73	2.23
Plate_ce2	1.23	1.79	2.58	2.34	**1.21**	1.62	1.83
Ring_ce	1.56	5.21	69.55	**1.30**	1.71	1.80	1.68
Singer1	**2.50**	12.84	6.58	5.76	7.22	3.34	12.65
Skating2	142.18	78.67	69.79	106.33	200.4	277.60	**46.90**
Sunshade	4.91	4.54	68.68	4.99	4.57	**4.20**	4.84
Tennis_ce2	**5.51**	31.21	133.69	16.92	5.66	5.74	5.69
Tennis_ce3	**5.78**	97.29	79.58	40.73	91.1	90.95	90.72
Walking2	45.15	32.09	11.94	13.83	46.34	**4.81**	22.33
Average Error	**22.63**	44.54	63.48	237.91	26.95	46.72	56.02

The precisions plots are shown in Fig. 4.6 to Fig. 4.14. Table 4.3 provides the mean precision value of the tracker in the entire image sequence. However, the tracker might get the drift for few frames and then recover itself. Therefore, these plots presented to review tracker performance during the whole image sequence. Various challenges were present in sequences such as occlusion, scale variations, deformation, etc. In sequences (Baby_ce, Carchasing_ce3, Car4, Cardark, Carchasing_ce4, Crossing, Cup, Jogging1, Jogging2, Guitar_ce2, Man, Plate_ce2, Ring_ce, Singer1, Sunshade, Tennis_ce2, and Tennis_ce3), the proposed tracker has the highest precision in the entire sequence. In sequences (Bike3, Building3, Busstation_ce2, Car9, Girl2, Human3, Juice, Skating2, and Walking2), the proposed tracker has slightly low precision.

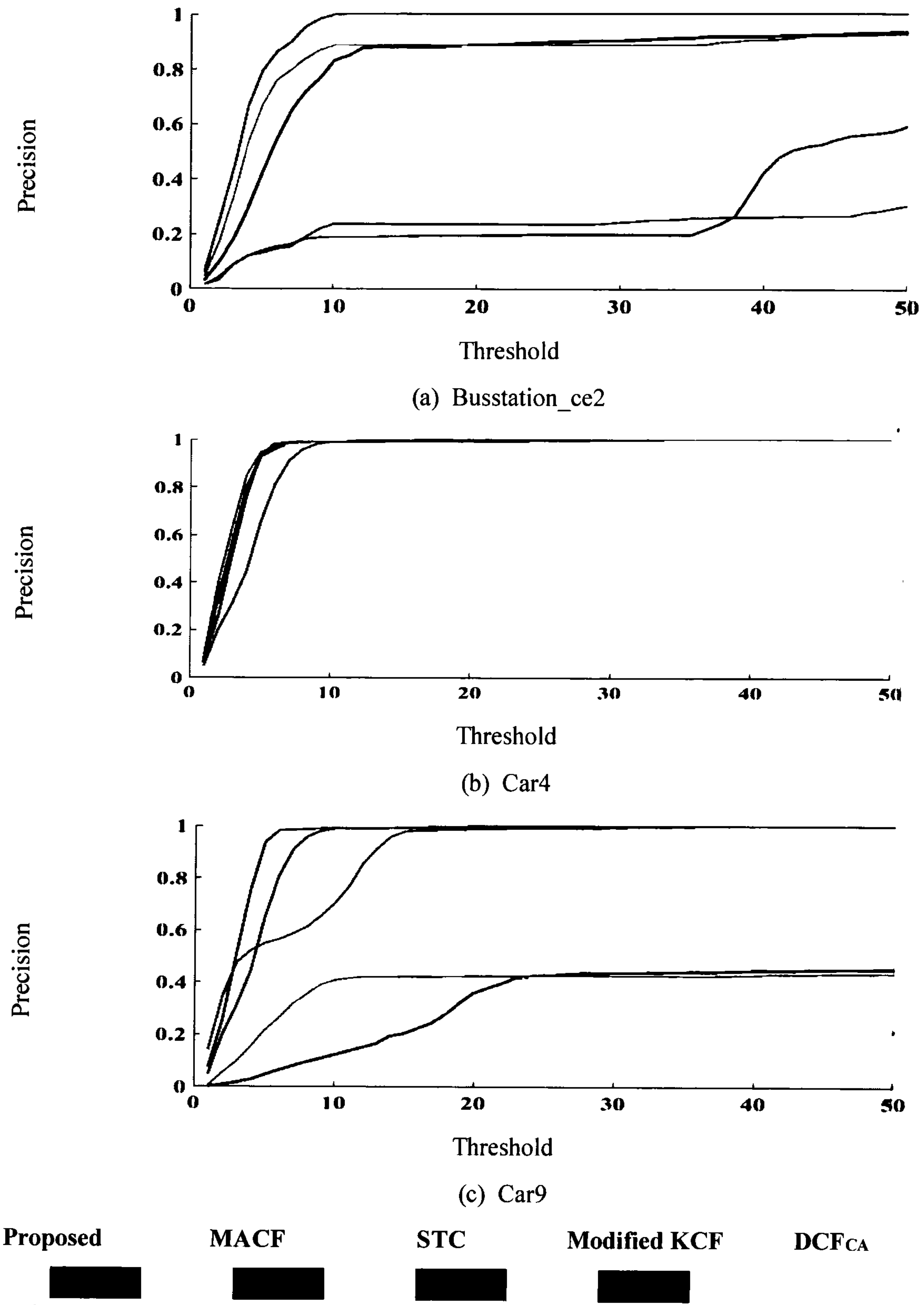


(a) Baby_ce

(b) Bike3

(c) Building3

Fig 4.6 Precision plots comparison on (Baby_ce, Bike3, and Building3) sequences.

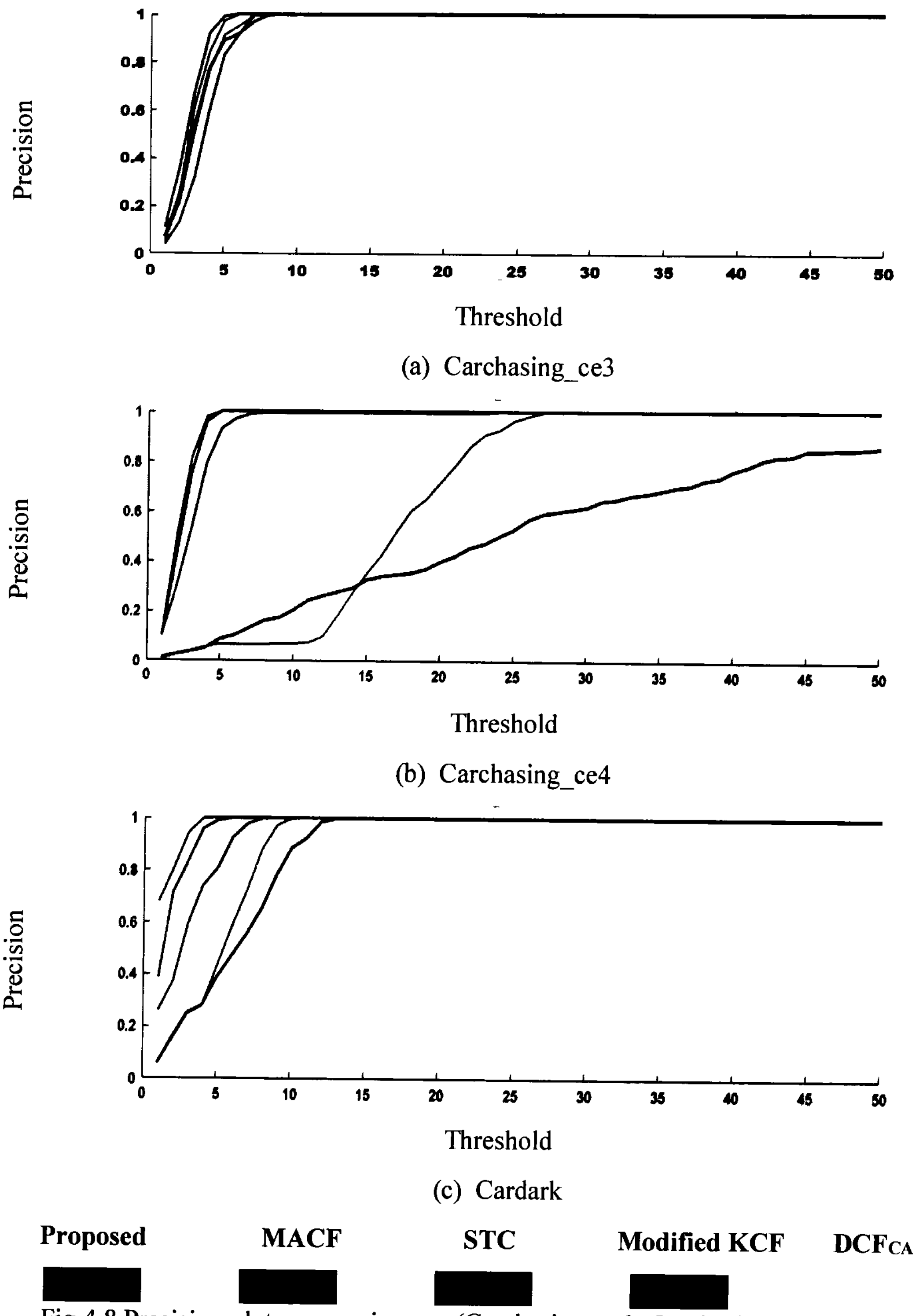


(a) Busstation_ce2

(b) Car4

(c) Car9

Fig 4.7 Precision plots comparison on (Busstation_ce2, Car4, and Car9) sequences.

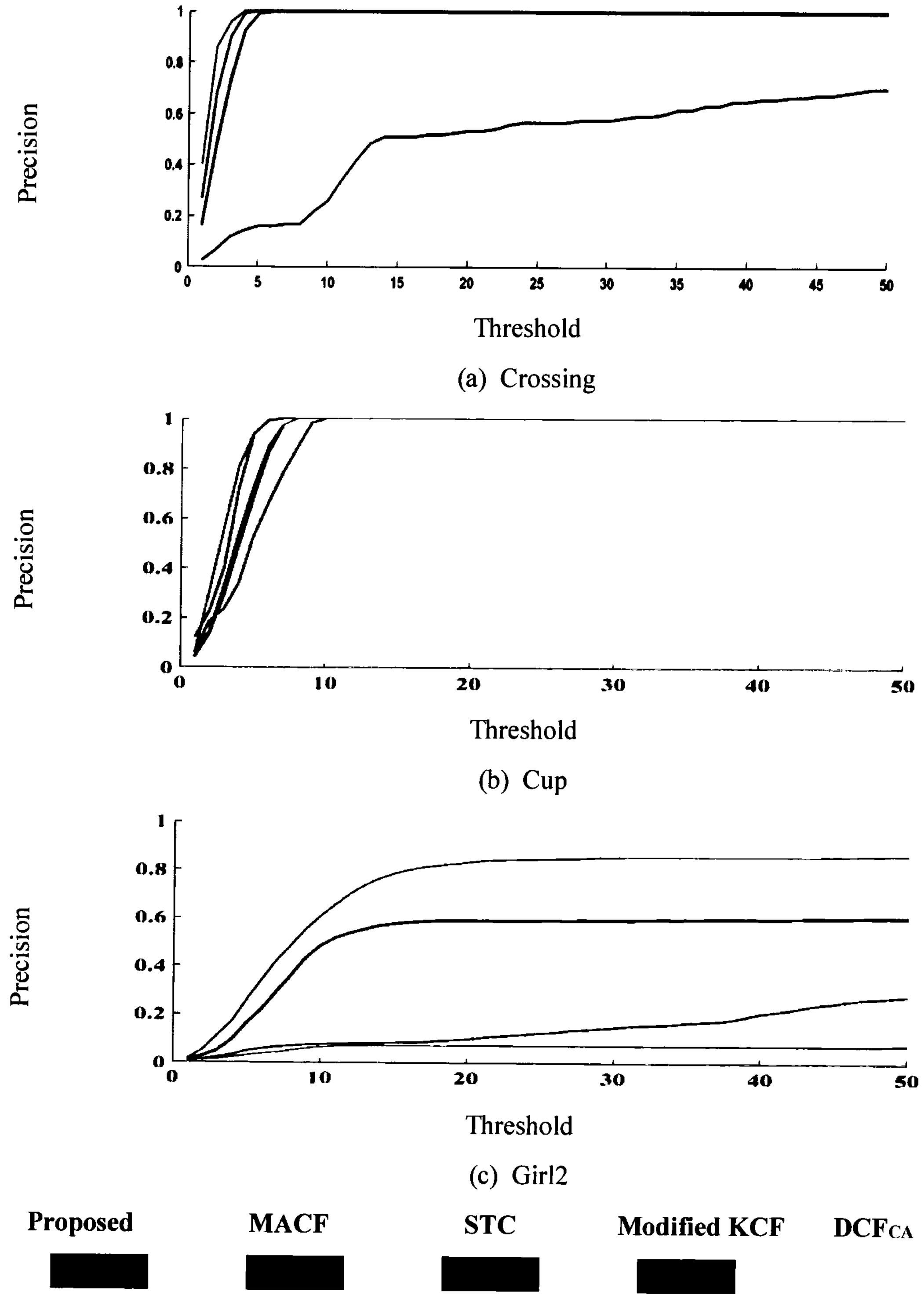


(a) Carchasing_ce3

(b) Carchasing_ce4

(c) Cardark

Fig 4.8 Precision plots comparison on (Carchasing_ce3, Carchasing_ce4, and Cardark) sequences.

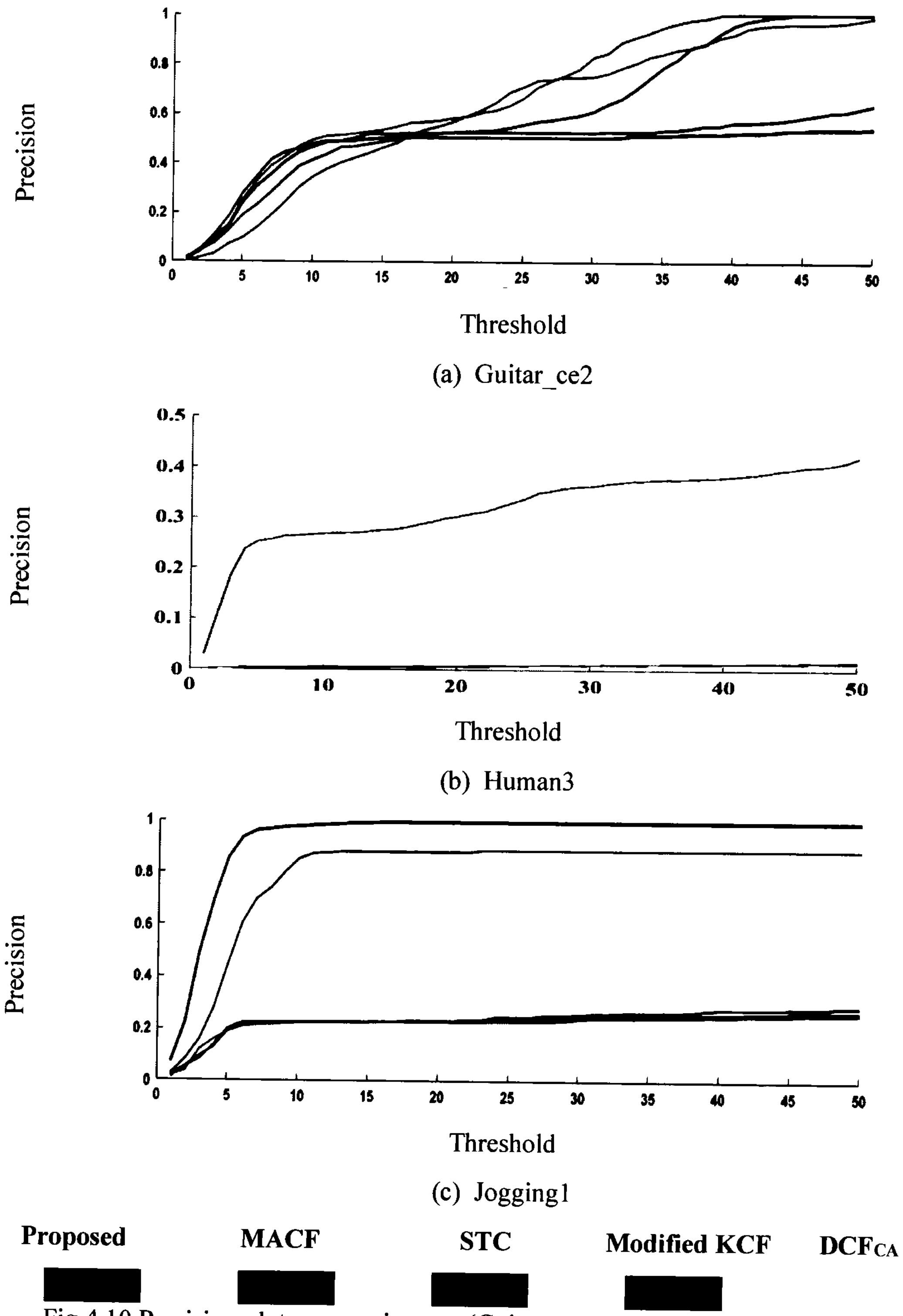


(a) Crossing

(b) Cup

(c) Girl2

Fig 4.9 Precision plots comparison on (Crossing, Cup, and Girl2) sequences.

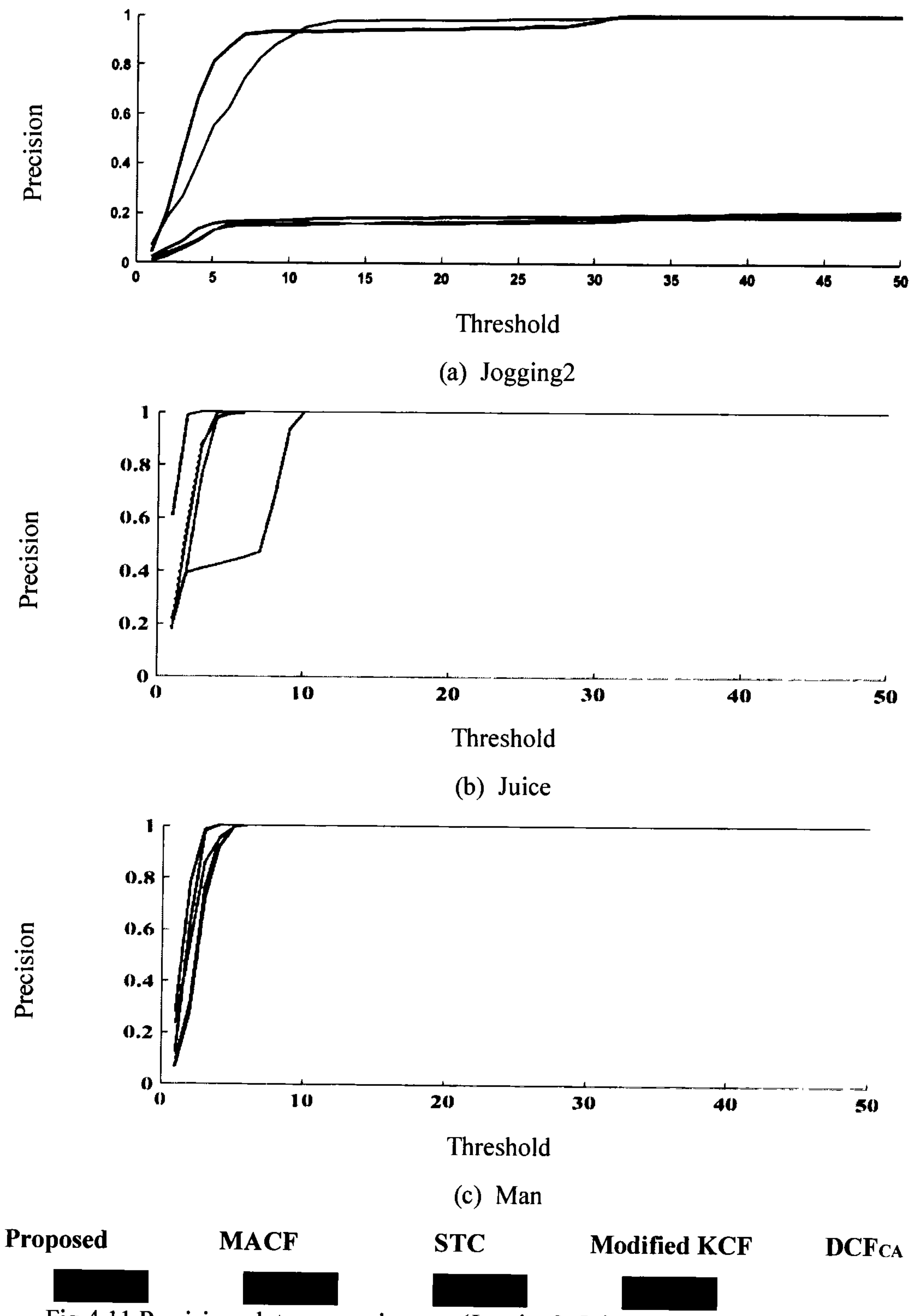


(a) Guitar_ce2

(b) Human3

(c) Jogging1

Fig 4.10 Precision plots comparison on (Guitar_ce2, Human3, and Jogging1) sequences.

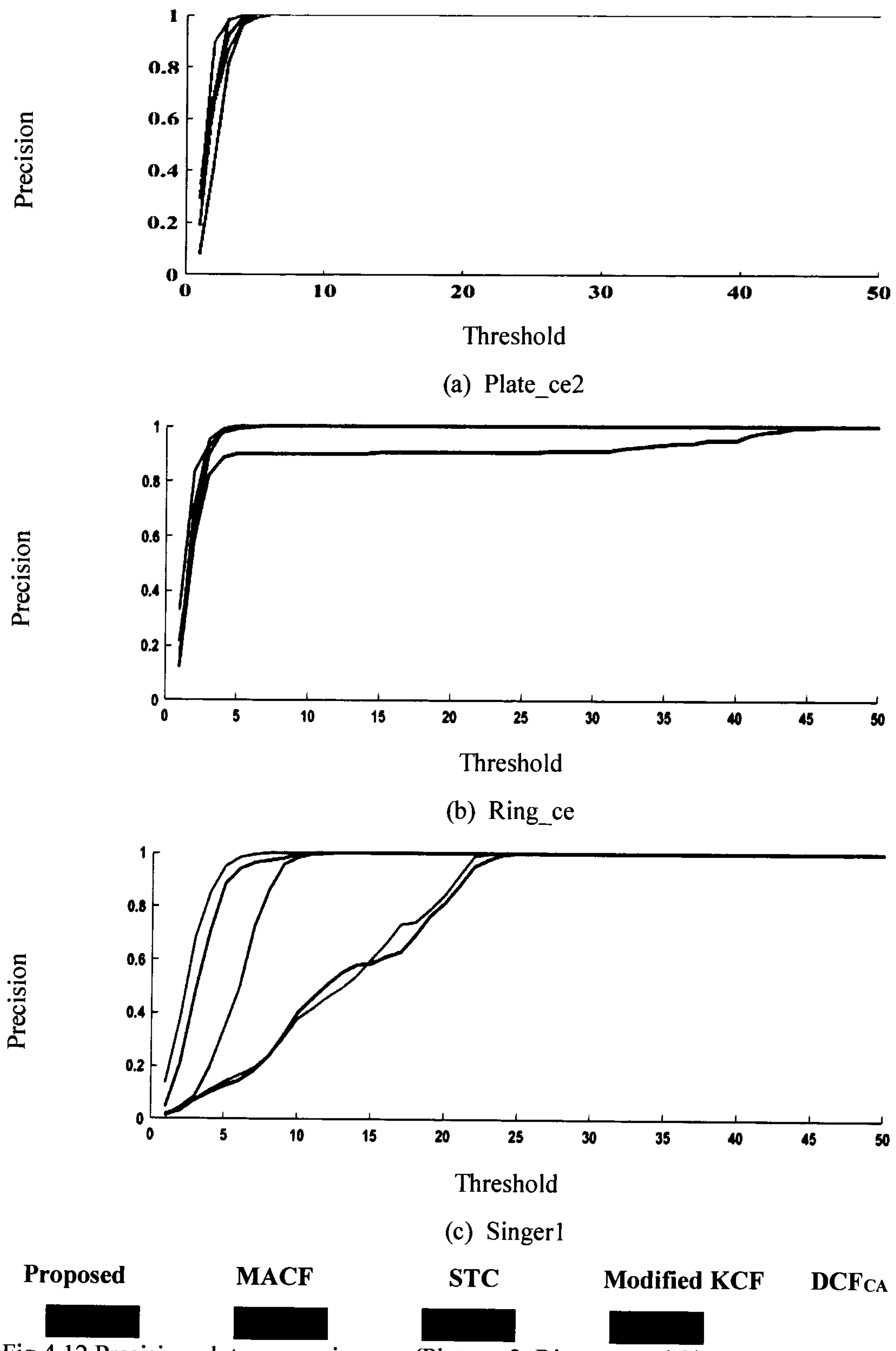


(a) Jogging2

(b) Juice

(c) Man

Fig 4.11 Precision plots comparison on (Jogging2, Juice, and Man) sequences.

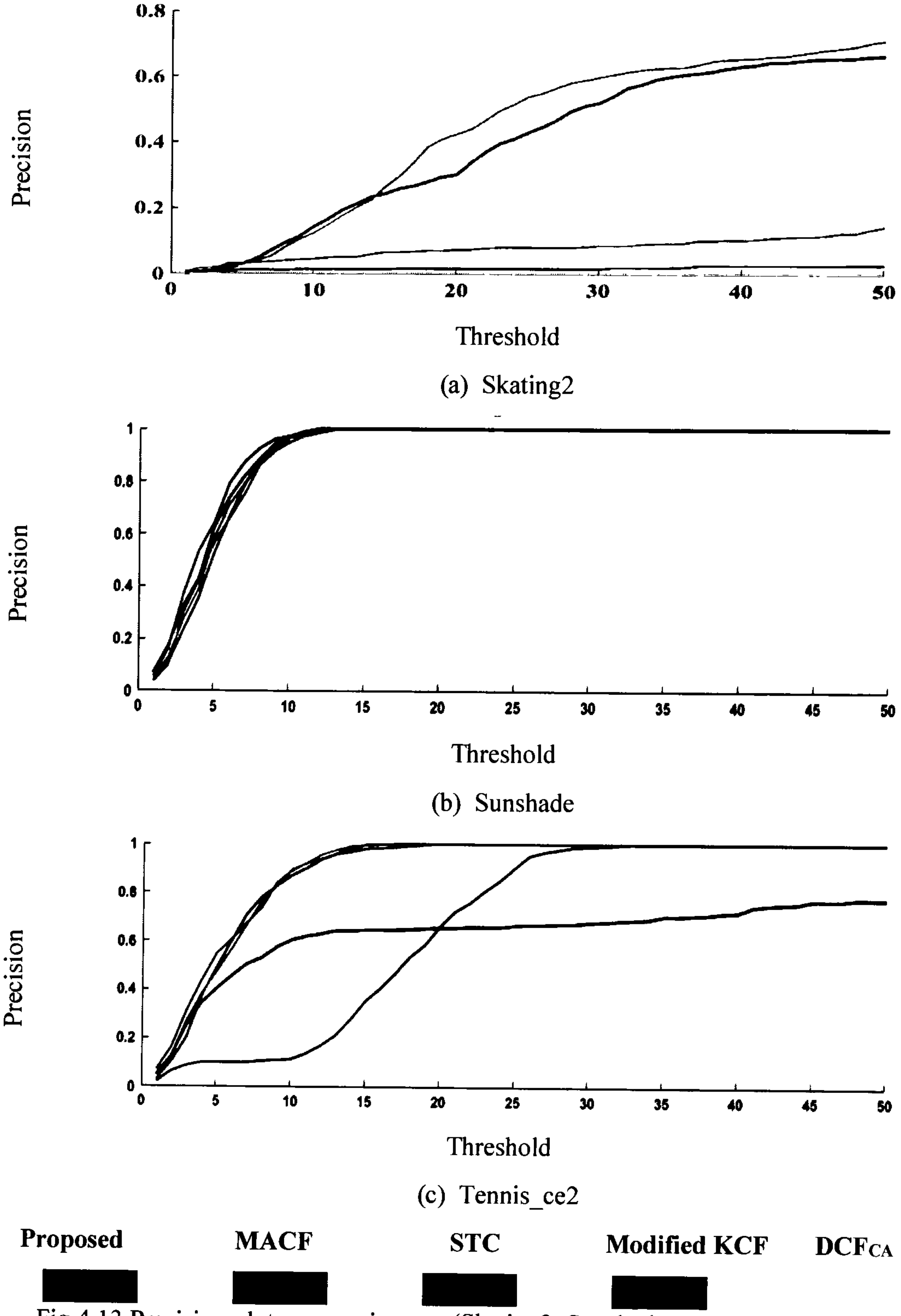


(a) Plate_ce2

(b) Ring_ce

(c) Singer1

Fig 4.12 Precision plots comparison on (Plate_ce2, Ring_ce, and Singer1) sequences.

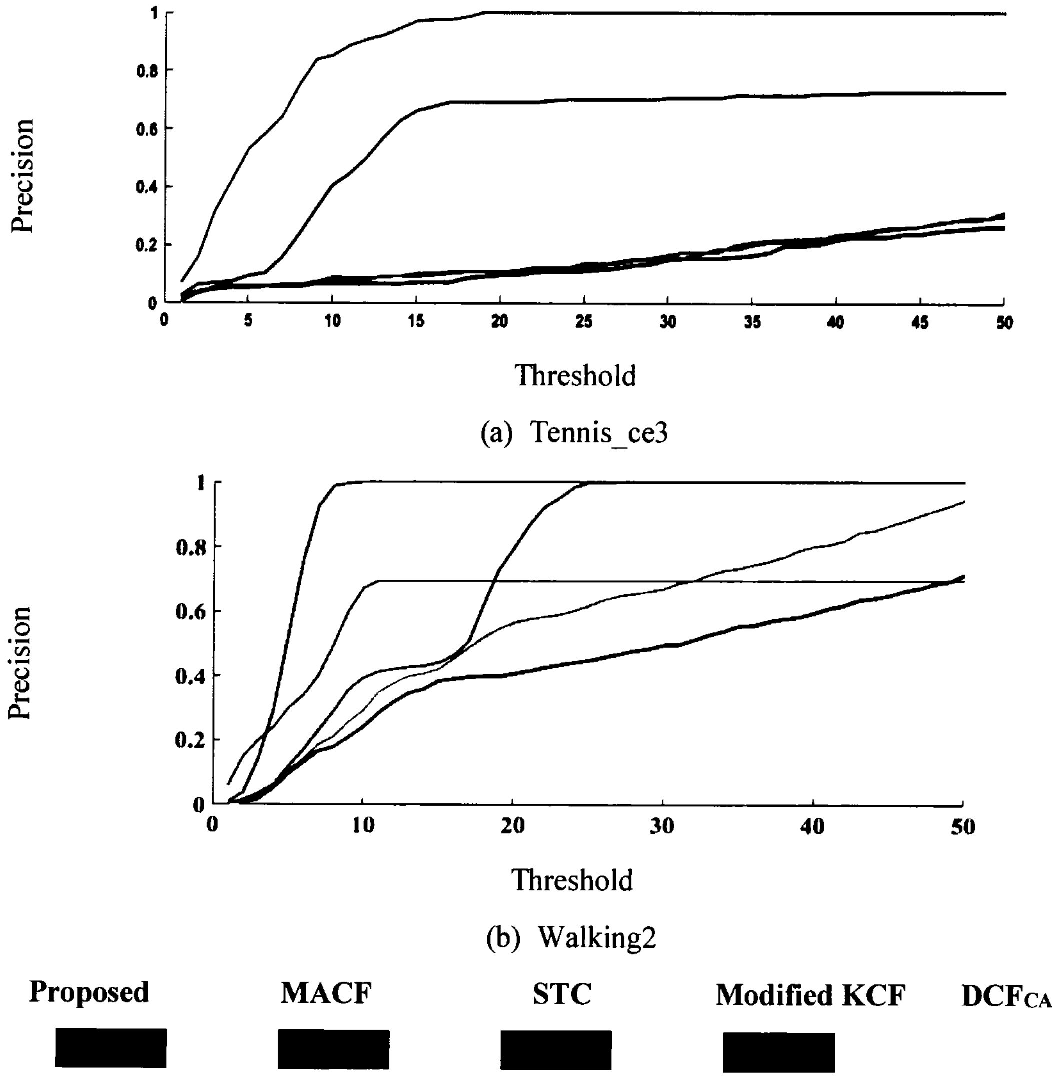


(a) Skating2

(b) Sunshade

(c) Tennis_ce2

Fig 4.13 Precision plots comparison on (Skating2, Sunshade, and Tennis_ce2) sequences.

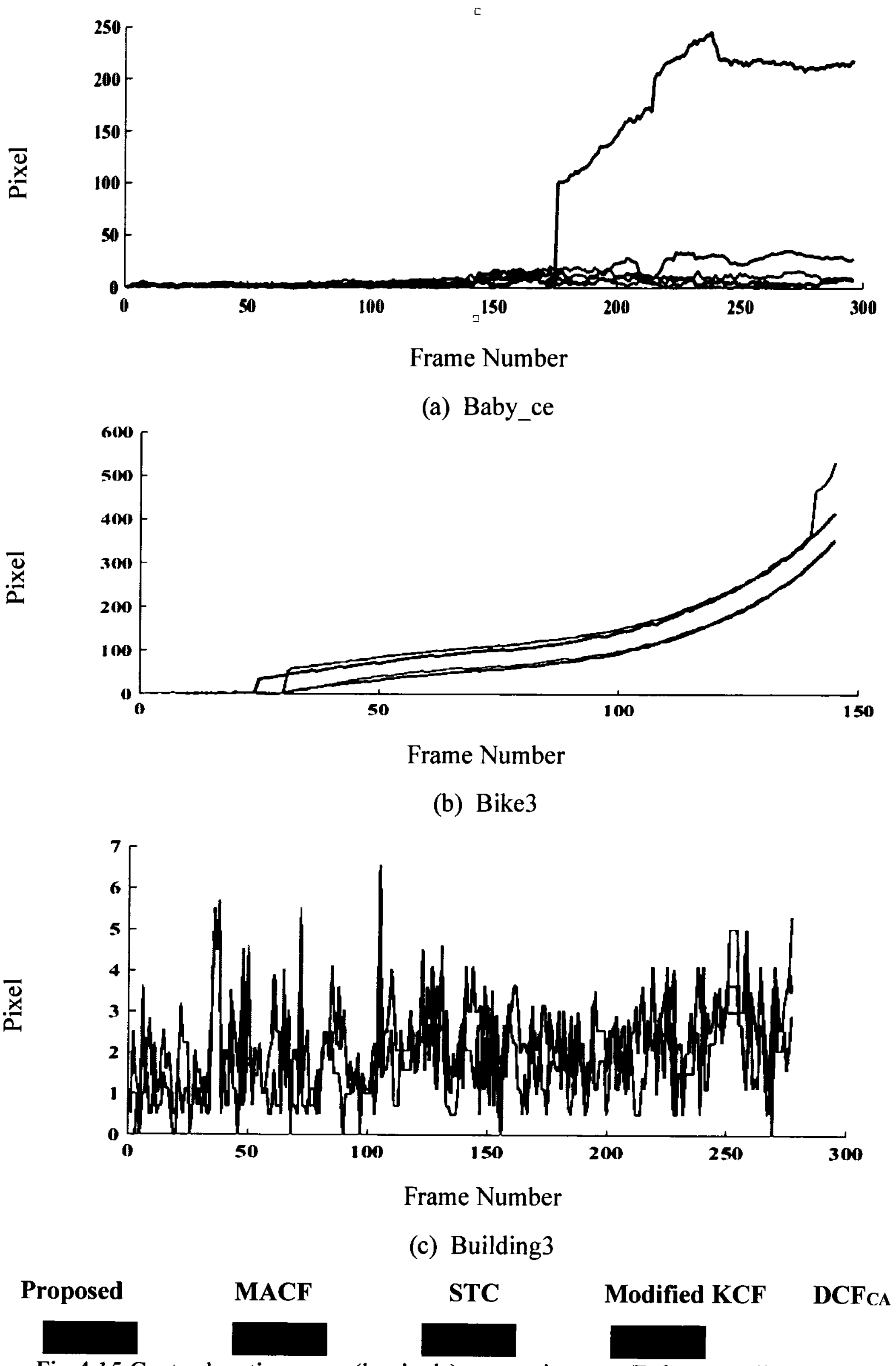


(a) Tennis_ce3

(b) Walking2

Fig 4.14 Precision plots comparison on (Tennis_ce3 and Walking2) sequences.

The location error plots are shown in Fig. 4.15 to Fig. 4.23. In Table 4.4 average centre location is calculated for each image sequence. It gives an idea about tracker performance, but it does not entirely incorporate all information necessary to review tracker performance. A possible scenario exists in object tracking when a tracker might drift for few frames in a sequence resulting in a high error value. However, when the tracker recovers from drift and starts tracking the target accurately, the error will be low during those frames, but its average value will be high. Therefore, these plots are presented to review tracker performance on each frame. The proposed tracker performs consistently for sequences (Baby_ce, Car4, Car9, Cardark, Crossing, Carchasing_ce3, Carchasing_ce3, Cup, Guitar_ce2, Juice, Jogging2, Ring_ce, and Tennis_ce3) over the entire duration. In sequences (Girl2, Human3, Skating2, and Walking2), the tracker drift between the frames but recovers after few frames. For most of the frames in these sequences the proposed tracker accurately tracks the target. However, when the tracker got drifted, then the accumulative error was high for these sequences. In sequences (Bike3, Building3, Busstation_ce2, Jogging1, Man, Plate_ce2, Singer1, Sunshade, and Tennis_ce2), the proposed method has similar performance with compared trackers.

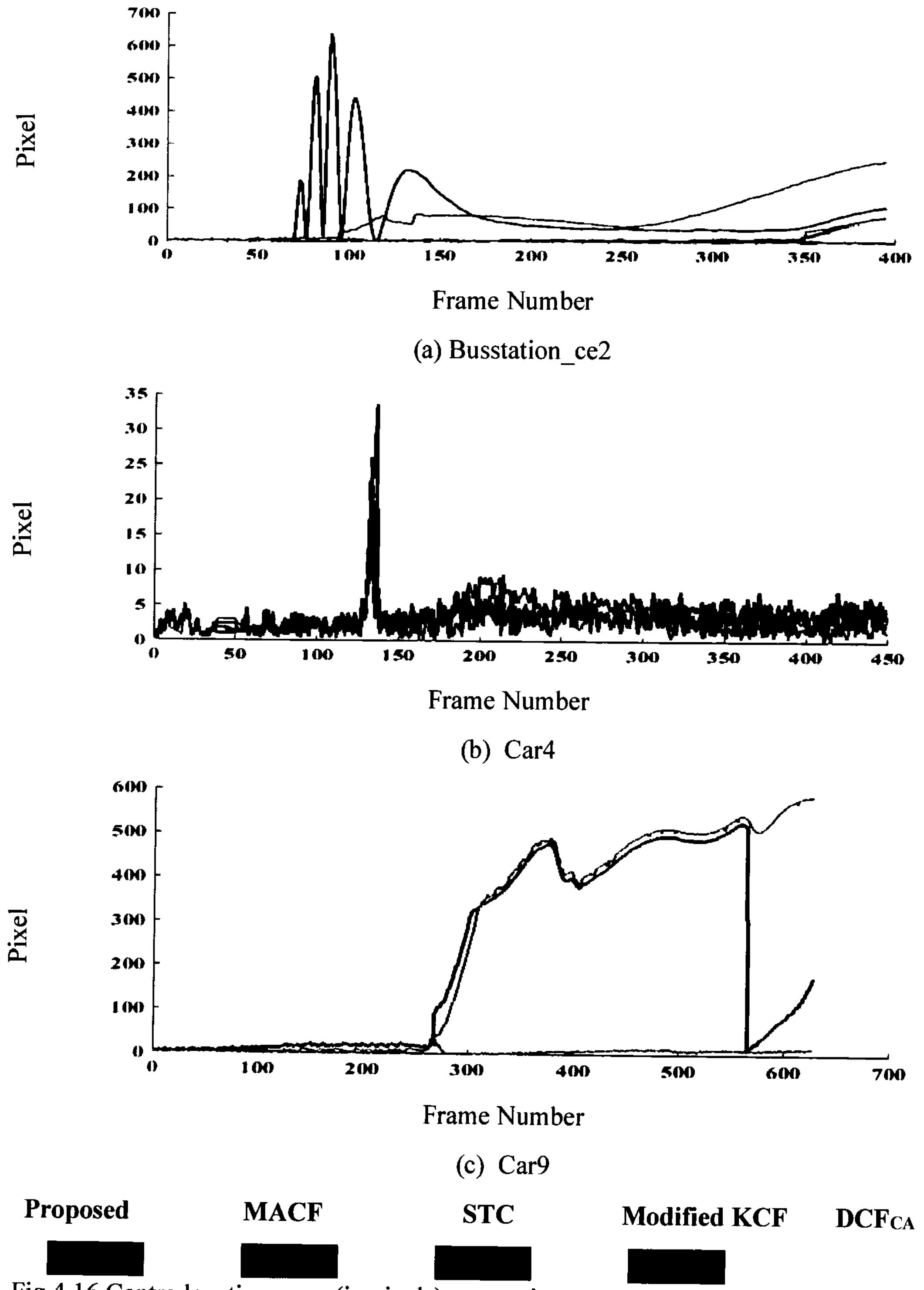


(a) Baby_ce

(b) Bike3

(c) Building3

Fig 4.15 Centre location error (in pixels) comparison on (Baby_ce, Bike3, and Building3) sequences.

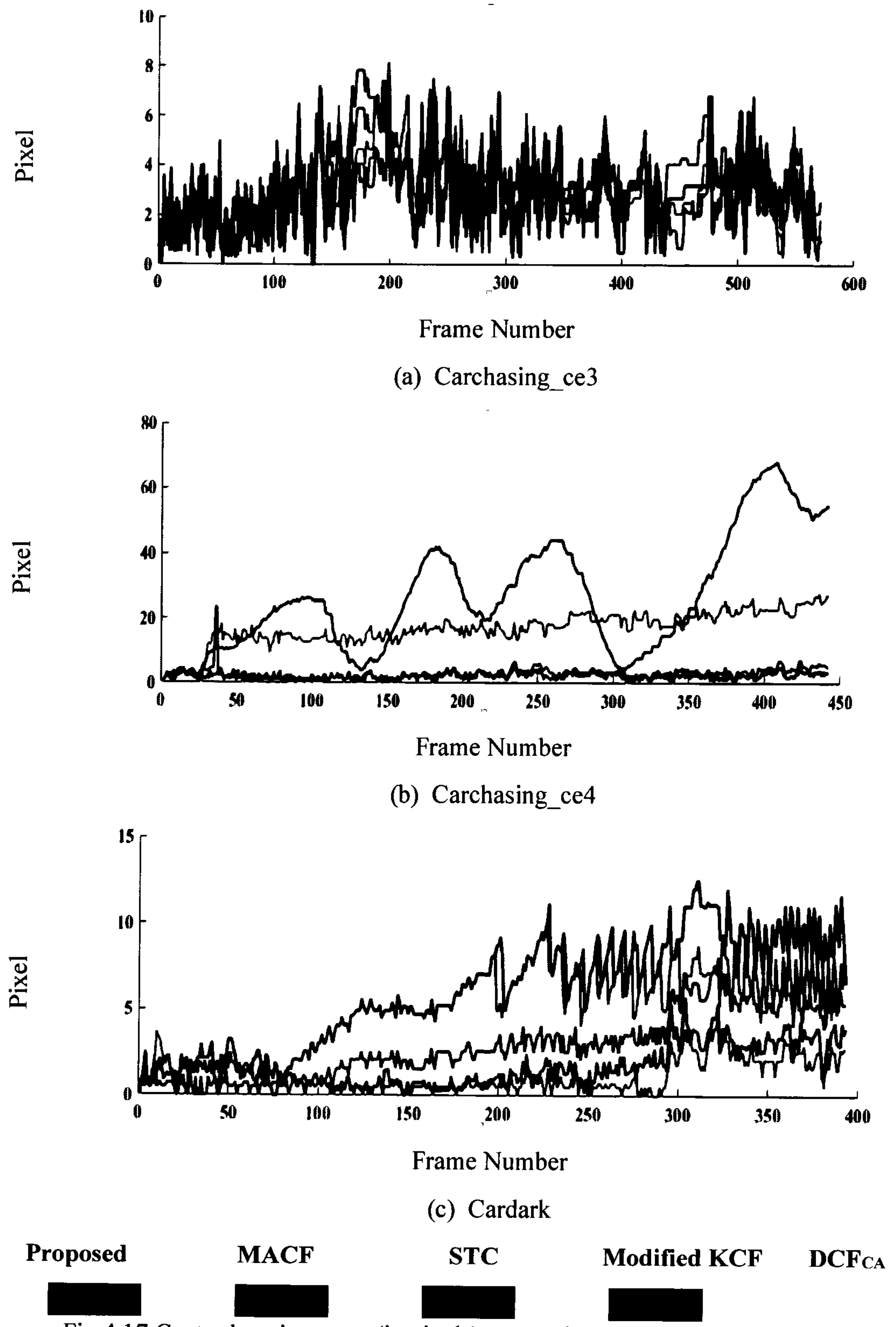


(a) Busstation_ce2

(b) Car4

(c) Car9

Fig 4.16 Centre location error (in pixels) comparison on (Busstation_ce2, Car4, and Car9) sequences.

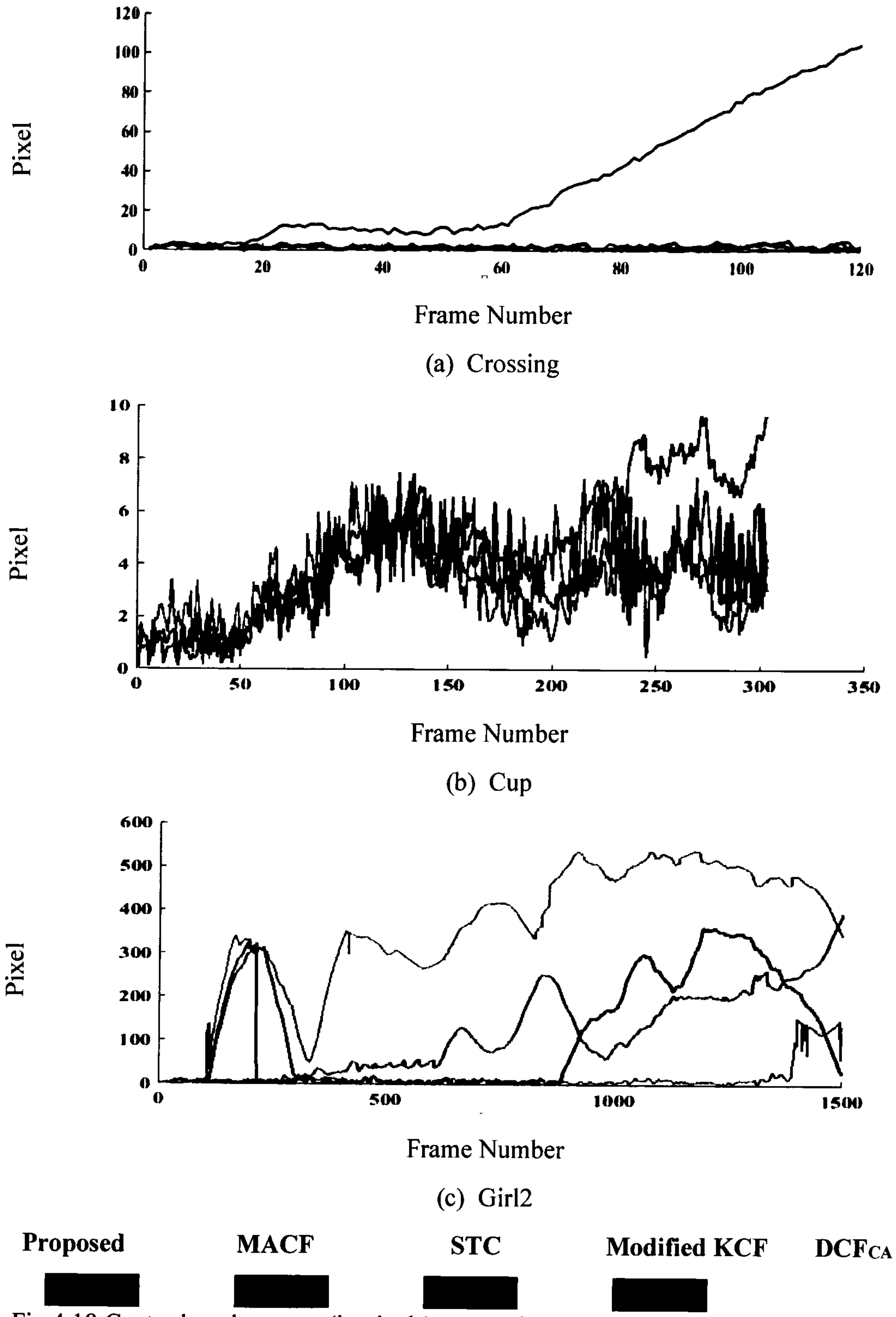


(a) Carchasing_ce3

(b) Carchasing_ce4

(c) Cardark

Fig 4.17 Centre location error (in pixels) comparison on (Carchasing_ce3, Carchasing_ce3, and Cardark) sequences.

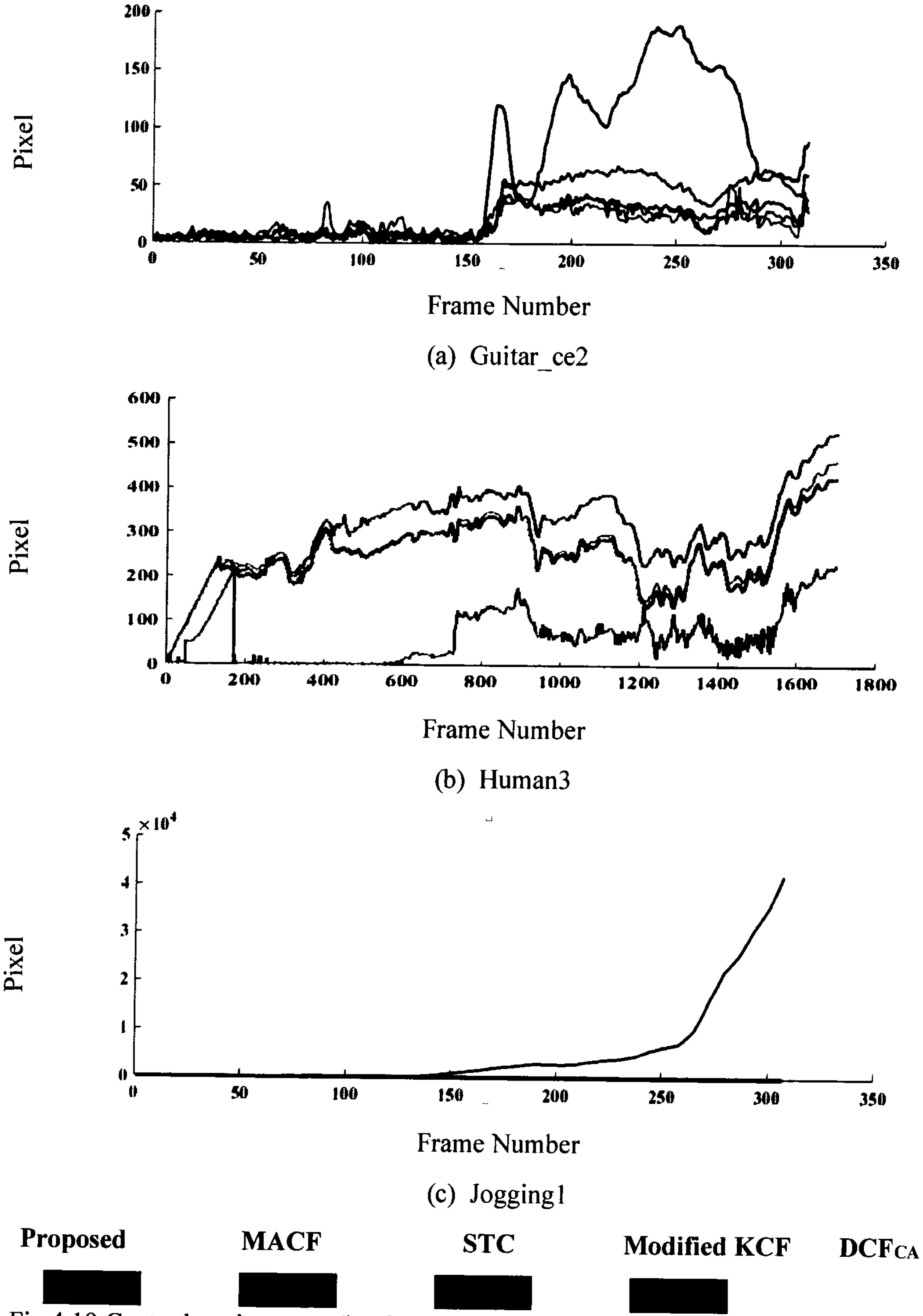


(a) Crossing

(b) Cup

(c) Girl2

Fig 4.18 Centre location error (in pixels) comparison on (Crossing, Cup, and Girl2) sequences.

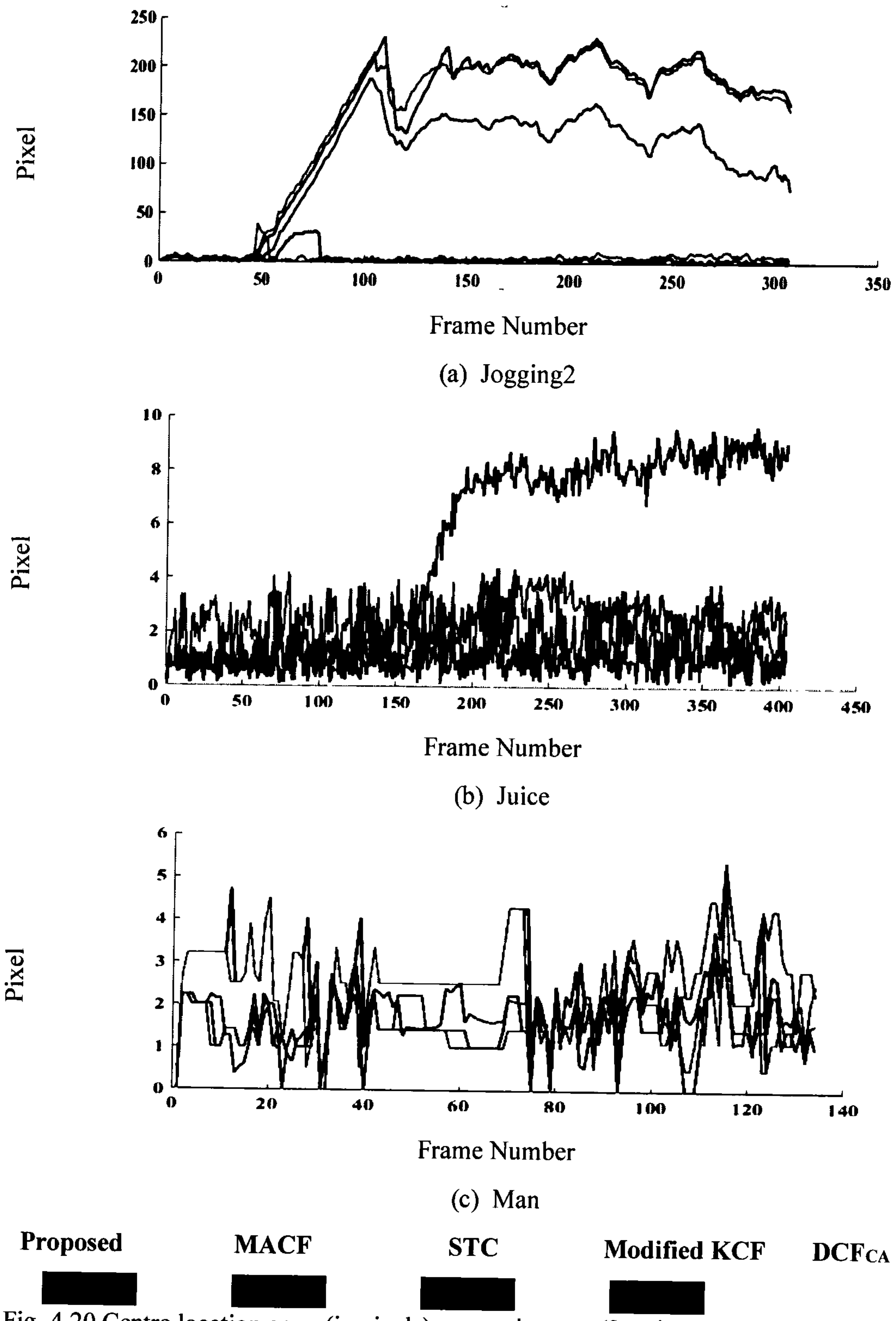


(a) Guitar_ce2

(b) Human3

(c) Jogging1

Fig 4.19 Centre location error (in pixels) comparison on (Guitar_ce2, Human3, and Jogging1) sequences.

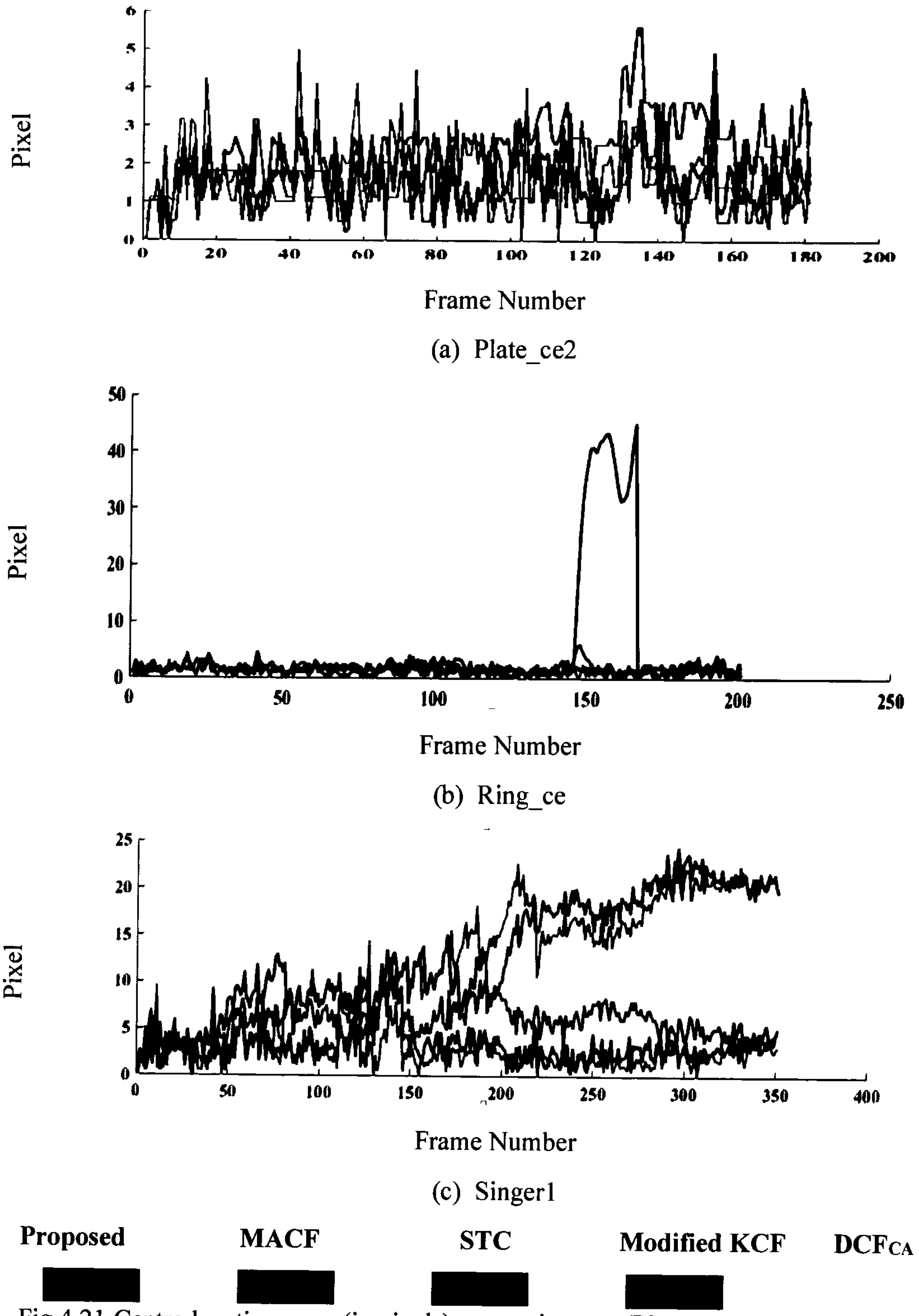


(a) Jogging2

(b) Juice

(c) Man

Fig. 4.20 Centre location error (in pixels) comparison on (Jogging2, Juice, and Man) sequences.

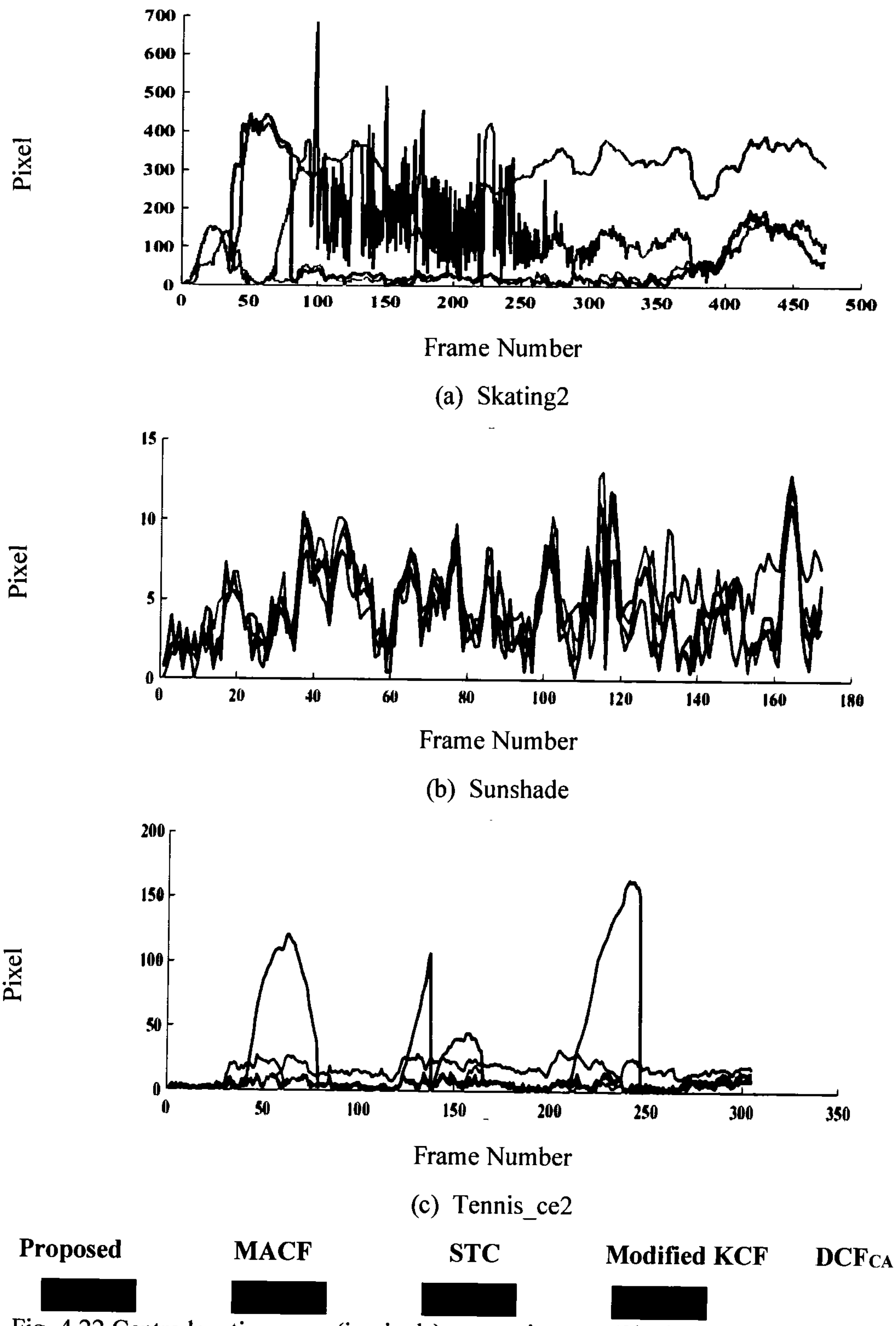


(a) Plate_ce2

(b) Ring_ce

(c) Singerl

Fig 4.21 Centre location error (in pixels) comparison on (Plate_ce2, Ring_ce, and Singerl) sequences.

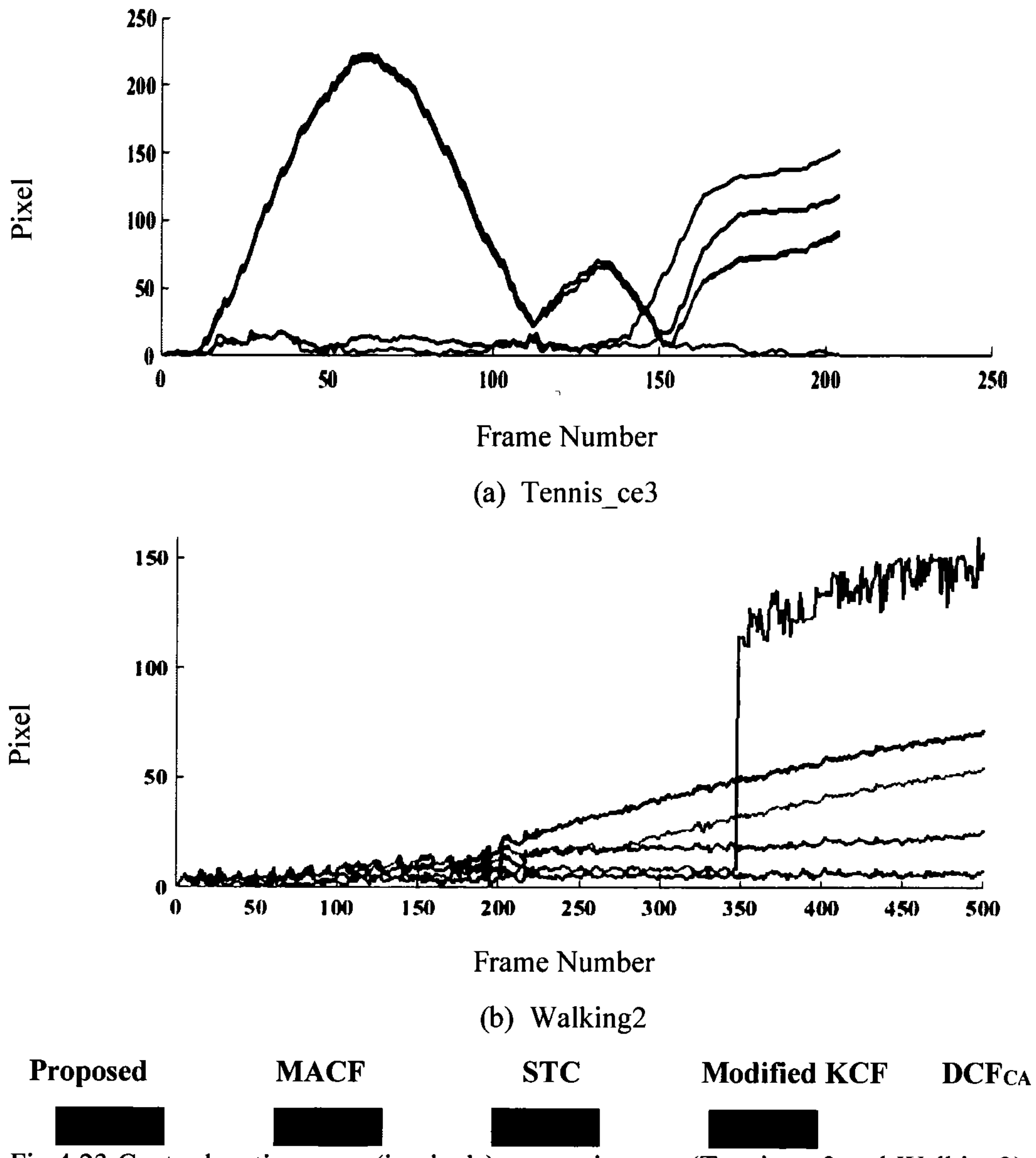


(a) Skating2

(b) Sunshade

(c) Tennis_ce2

Fig. 4.22 Centre location error (in pixels) comparison on (Skating2, Sunshade, and Tennis_ce2) sequences.



(a) Tennis_ce3

(b) Walking2

Fig 4.23 Centre location error (in pixels) comparison on (Tennis_ce3 and Walking2) sequences.

4.3.2 Qualitative Analysis

Fig. 4.24 depicts the proposed tracking qualitative results with four state-of-the-art trackers over twenty-six image sequences involving various challenges such as partial or full occlusions, scale variations, background clutter, etc. MACF contains a similar tracking component as our approach, i.e., scale correlation filter and Kalman filter. Even though MACF performs favorably well in sequences involving scale variations, it does not deal effectively with sequences involving occlusions (Girl2, Human3, Jogging1, Jogging2, and Skating2). STC uses intensity features and response of a single translation filter to estimate scale. This makes STC a comparatively fast tracker; however, there is no occlusion detection or handling mechanism due to which its tracking results are affected in sequences (Busstation_ce2, Girl2, Human3, Jogging1, and Jogging2). Moreover, due to only one translation filter, its tracking results are also affected (Car9, Crossing, and Tennis_ce3). $\mathrm{DCF_{CA}}$ contains correlation filtering combined with the context-aware formulation. However, it is not robust in occlusions, scale variations, and deformation challenges. Therefore, $\mathrm{DCF_{CA}}$ does not perform well in sequences (Car9, Carchasing_ce4, Girl2, Human3, Jogging1, Jogging2, Skating2, and Tennis_ce3). Modified KCF performs significantly well in sequences involving occlusions. However, it does not perform well in scale variation sequences (Baby_ce, Car9, Carchasing_ce4, Guitar_ce2, Ring_ce, Singer1, Tennis_ce2, and Tennis_ce3).

It can be seen that the proposed tracking method outperforms other trackers in these sequences. In sequences (Baby_ce, Car4, Carchasing_ce4, Crossing, Cup, Jogging1, Jogging2, Guitar_ce2, Plate_ce2, Ring_ce, Singer1, Tennis_ce2 and Tennis_ce3) the proposed method can accurately track target for entire image sequences. In sequences (Bike3, Busstation_ce2, Girl2, Human3, Skating2, and Walking2) tracker cannot accurately for the entire sequence. All trackers have similar performance in sequences (Building3, Carchasing_ce3, Cardark, Juice, Man, and Sunshade).

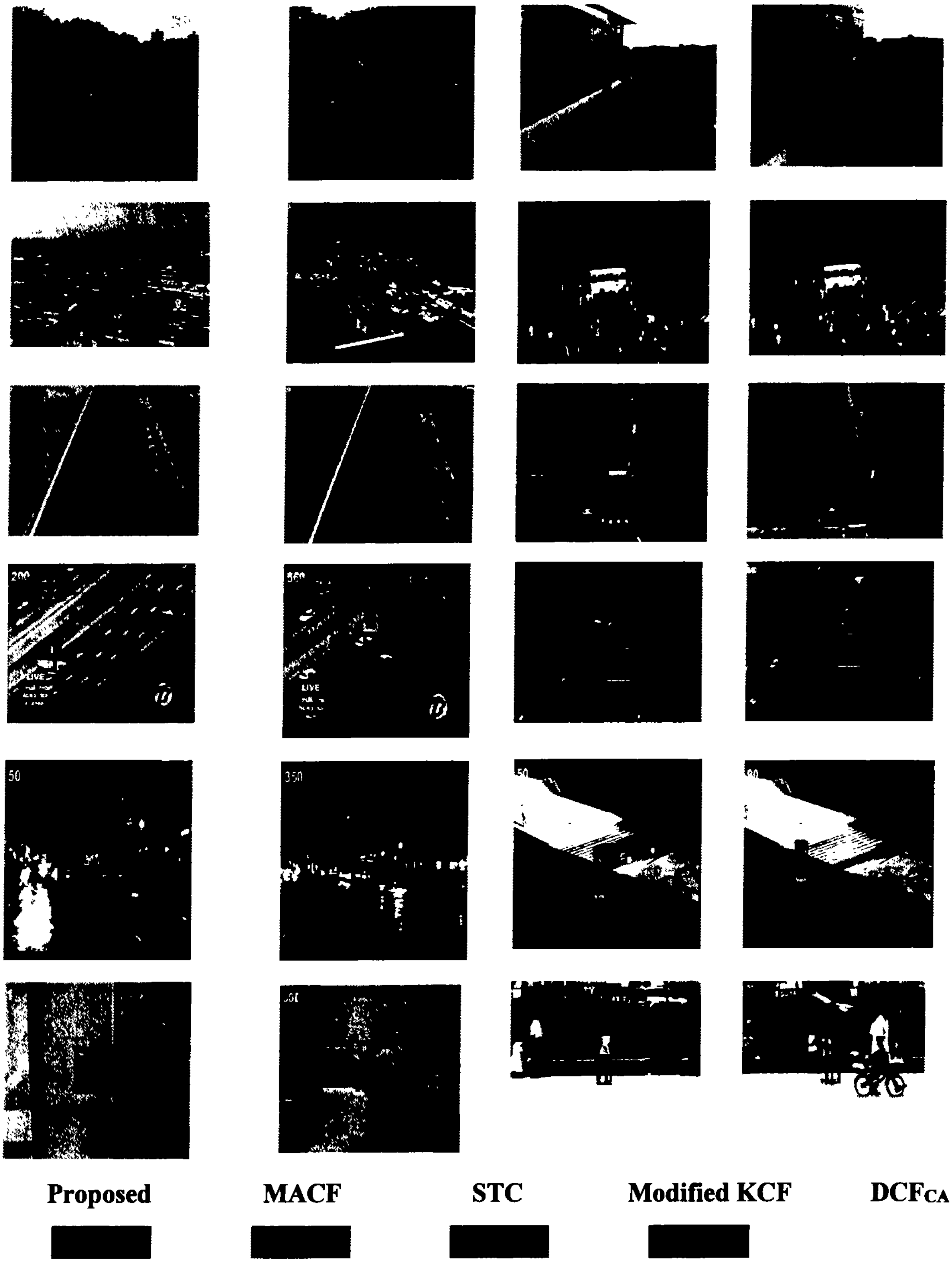


Fig. 4.24 Qualitative comparison on TC-128, OTB2013, OTB2015, and UAV123 datasets.(Continued)

Fig. 4.24 Qualitative comparison on TC-128, OTB2013, OTB2015, and UAV123 datasets

4.3.3 Discussion

It can be seen from Figure 4.24 that the proposed tracking method outperforms other trackers in these sequences. We discuss several observations from performance analysis. This performance can be strengthened for three reasons. First, the scale correlation filter is incorporated in the STC framework making it deal effectively better than the STC scale. This scale filter learns target appearance on different scales, making it better to track the target accurately under scale variation scenarios. It can be seen in sequences (Baby_ce, Car4, Car9, Carchasing_ce3, Carchasing_ce4, Plate_ce2, and Ring_ce) that the proposed tracker deals better with scale variation of the target. Second, incorporating of an extended Kalman filter makes it robust to handle occlusions. When the target undergoes partial or full occlusions, then EKF predicts the target state and updates the tracking model. The sequences (Girl2, Jogging1, and Jogging2) can be seen that the proposed method can effectively handle the target's occlusion. Third, the fusion of APCE based adaptive learning rate further elevates the tracking performance in illumination variations, motion blur, and clutter background challenges. It can be seen in sequences (Building3, Cardark, Crossing, Cup, Guitar_ce2, Juice, Man, Singer1, Sunshade, Tennis_ce2, and Tennis_ce3) tracker can accurately follow the target. The tracker's appearance model can cope with changes in the environment by utilizing information in each frame.

Even though the proposed tracker performs significantly better than various trackers, there are few sequences (Bike3, Busstation_ce2, Human3, Skating2, and Walking2) in which the tracker does not track the target accurately. In the Bike3 sequence, the tracker fails due to fast movement combined with scale variation. In Skating2 tracker fails due to the deformation of the target. In (Busstation_ce2, Human3, and Walking2) tracker fails due to occlusions, fast motion, and motion blur. The limitations can be addressed by working in few directions, such as developing a better occlusion detection and handling mechanism, extending the aspect ratio adaptability, and incorporating context-aware formulation.

4.4 Summary

This chapter presents simulation results of proposed tracking algorithms on VOT datasets were presented in various graphical illustrations. From both case studies, it can be concluded that the CF and measurement estimation collaboration-based tracking mechanism performs better than other proposed tracking algorithms.

Chapter 5.

Conclusion and Future Work

In this chapter, conclusions for proposed tracking schemes are presented. In addition, some research directions were also stated for the researchers interested in working in VOT.

5.1 Conclusion

Conclusions from this study are:

- In the first case study, an adaptive Spatio-temporal context (STC) based algorithm for online tracking is presented, which combines the context-aware formulation, Kalman filter, response map-based occlusion detection, and average difference based adaptive model update in the STC framework. As a result, the algorithm performs better in scenarios such as full occlusion, illumination variation, deformation, and background clutter compared than various algorithms with the achievement of efficient performance in datasets.
- The context-aware formulation can be efficiently applied to deal with background clutter issues. The maximum value of the response map can be used to detect occlusions. Afterward, a Kalman filter can be applied for occlusion handling. The model update can also be related to the target's motion, as the STC model is updated on a fixed learning rate, making it vulnerable to target motion. Based on target motion, the tracking model should be updated adaptively.

- The second case study gives insight into the robust tracking algorithm based on STC by incorporating a scale correlation filter based on pyramid representation for adaptive scale estimation, extended Kalman filter for occlusion handling, and APCE criteria for the adaptive learning rate of the tracking model. Experimental results indicate that the proposed tracking algorithm performs better than the various state-of-the-art qualitatively and quantitatively.
- A correlation filter-based discriminative scale mechanism is incorporated in the Spatio-temporal context model, making it robust and effective in scenarios such as clutter background, illumination variation, scale variations, and fast motion. The adaptive learning rate mechanism is based on APCE between consecutive frames. It is fused into the framework to update the tracking model according to the target's shape and motion. If the model is updated on a fixed learning rate, it does not cope with the target's shape, losing it in the subsequent frames.
- The extended Kalman filter aspect, utilized in case study 2, is when the target undergoes occlusion. The condition to decide whether the target is occluded is based on the response map's maximum value. In both case studies, Kalman filter is applied, but a mechanism is also devised for its activation in the STC framework, making it better both qualitatively and quantitatively than various trackers.

5.2 Future Work

Research directions for future development in VOT are:

- The proposed tracking algorithms can be investigated by establishing neural network-based algorithms [18], [19] to improve robustness and tracking accuracy.

- This study is considered a step further to design new occlusion detection and occlusion handling mechanisms for tracking algorithms.
- The current framework can be extended to context-aware and target adaptation formulation, incorporating more features to learn target appearance and extending the aspect ratio adaptability.
- One may explore this framework by combining heuristics and fractional order algorithms for model update.
- One can explore other methods for making learning rate adaptive by using confidence of squared response map (CSRM) [25], APCE based degree indicator [81], channel features [82], global optimization methods [83]–[86] and deep learning based methods [87]–[89].

BIBLIOGRAPHY

[1] L. Yao, Y. Liu, and S. Huang, “Spatio-temporal information for human action recognition,” *EURASIP J. Image Video Process.*, vol. 2016, no. 1, pp. 1–9, 2016.

[2] J. K. Aggarwal and L. Xia, “Human activity recognition from 3d data: A review,” *Pattern Recognit. Lett.*, vol. 48, pp. 70–80, 2014.

[3] C. Stauffer and W. Grimson, “Learning Patterns of Activity Using Real-Time Tracking,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 747–757, 2000, doi: 10.1109/34.868677.

[4] F. Jiansheng, “Vision-based real-time traffic accident detection,” in *Proceeding of the 11th World Congress on Intelligent Control and Automation*, 2014, pp. 1035–1038.

[5] B. Tian, Q. Yao, Y. Gu, K. Wang, and Y. Li, “Video processing techniques for traffic flow monitoring: A survey,” in *2011 14th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, 2011, pp. 1103–1108.

[6] O. Masoud and N. P. Papanikolopoulos, “A novel method for tracking and counting pedestrians in real-time using a single camera,” *IEEE Trans. Veh. Technol.*, vol. 50, no. 5, pp. 1267–1278, 2001.

[7] W. Zhou, C. Wu, X. Yu, Y. Gao, and W. Du, “Automatic fovea center localization in retinal images using saliency-guided object discovery and feature extraction,” *J.*

Med. Imaging Heal. Informatics, vol. 7, no. 5, pp. 1070–1077, 2017.

[8] M. J. M. Vasconcelos, S. M. R. Ventura, D. R. S. Freitas, and J. M. R. S. Tavares, "Towards the automatic study of the vocal tract from magnetic resonance images," *J. Voice*, vol. 25, no. 6, pp. 732–742, 2011.

[9] A. Kuramoto, M. A. Aldibaja, R. Yanase, J. Kameyama, K. Yoneda, and N. Suganuma, "Mono-camera based 3d object tracking strategy for autonomous vehicles," in *2018 IEEE Intelligent Vehicles Symposium (IV)*, 2018, pp. 459–464.

[10] U. Handmann, T. Kalinke, C. Tzomakas, M. Werner, and W. von Seelen, "Computer vision for driver assistance systems," in *Enhanced and Synthetic Vision 1998*, 1998, vol. 3364, pp. 136–147.

[11] S. H. Oh, S. Javed, and S. K. Jung, "Foreground object detection and tracking for visual surveillance system: A hybrid approach," in *2013 11th International Conference on Frontiers of Information Technology*, 2013, pp. 13–18.

[12] A. Ali, H. Kausar, and I. K. Muhammad, "Automatic visual tracking and firing system for anti aircraft machine gun," in *6th International Bhurban Conference on Applied Sciences & Technology*, 2009, pp. 253–257.

[13] Z. Chen and D. Yi, "The game imitation: Deep supervised convolutional networks for quick video game AI," *arXiv Prepr. arXiv1702.05663*, 2017.

[14] X. Ge, J. Renz, and P. Zhang, "Visual detection of unknown objects in video games using qualitative stability analysis," *IEEE Trans. Comput. Intell. AI Games*, vol. 8, no. 2, pp. 166–177, 2015.

[15] S. Wang *et al.*, “Face-tracking as an augmented input in video games: enhancing presence, role-playing and control,” in *Proceedings of the SIGCHI conference on Human Factors in computing systems*, 2006, pp. 1097–1106.

[16] P. K. Santhosh and B. Kaarthick, “An automated player detection and tracking in basketball game,” *Comput. Mater. Contin.*, vol. 58, no. 3, pp. 625–639, 2019.

[17] P. Zhang, L. Zheng, Y. Jiang, L. Mao, Z. Li, and B. Sheng, “Tracking soccer players using spatio-temporal context learning under multiple views,” *Multimed. Tools Appl.*, vol. 77, no. 15, pp. 18935–18955, 2018.

[18] W. Kazimierski and J. Łubczonek, “Verification of marine multiple model neural tracking filter for the needs of shore radar stations,” in *2012 13th International Radar Symposium*, 2012, pp. 554–559.

[19] W. Kazimierski, “Proposal of neural approach to maritime radar and automatic identification system tracks association,” *IET Radar, Sonar Navig.*, vol. 11, no. 5, pp. 729–735, 2016.

[20] A. Stateczny, “Neural manoeuvre detection of the tracked target in ARPA systems,” *IFAC Proc. Vol.*, vol. 34, no. 7, pp. 209–214, 2001.

[21] Y. Wu, J. Lim, and M.-H. Yang, “Online object tracking: A benchmark,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 2411–2418.

[22] K. Zhang, L. Zhang, Q. Liu, D. Zhang, and M. H. Yang, “Fast visual tracking via dense spatio-temporal context learning,” in *Lecture Notes in Computer Science*

(including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2014, vol. 8693 LNCS, no. PART 5, pp. 127–141, doi: 10.1007/978-3-319-10602-1_9.

[23] M. Mueller, N. Smith, and B. Ghanem, "Context-aware correlation filter tracking," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1396–1404.

[24] J. Shin, H. Kim, D. Kim, and J. Paik, "Fast and robust object tracking using tracking failure detection in kernelized correlation filter," *Appl. Sci.*, vol. 10, no. 2, p. 713, 2020.

[25] Y. Zhang, Y. Yang, W. Zhou, L. Shi, and D. Li, "Motion-aware correlation filters for online visual tracking," *Sensors*, vol. 18, no. 11, p. 3937, 2018.

[26] K. Mehmood *et al.*, "Context-Aware and Occlusion Handling Mechanism for Online Visual Object Tracking," *Electronics*, vol. 10, no. 1, p. 43, 2021.

[27] B. Khan, A. Ali, A. Jalil, K. Mehmood, M. Murad, and H. Awan, "AFAM-PEC: Adaptive Failure Avoidance Tracking Mechanism Using Prediction-Estimation Collaboration," *IEEE Access*, vol. 8, pp. 149077–149092, 2020.

[28] A. Ali *et al.*, "Visual object tracking—classical and contemporary approaches," *Frontiers of Computer Science*, vol. 10, no. 1. pp. 167–188, 2016, doi: 10.1007/s11704-015-4246-3.

[29] A. W. M. Smeulders, D. M. Chu, R. Cucchiara, S. Calderara, A. Dehghan, and M. Shah, "Visual tracking: An experimental survey," *IEEE Trans. Pattern Anal.*

Mach. Intell., vol. 36, no. 7, pp. 1442–1468, 2014, doi: 10.1109/TPAMI.2013.230.

[30] X. Li, W. Hu, C. Shen, Z. Zhang, A. Dick, and A. Van Den Hengel, "A survey of appearance models in visual object tracking," *ACM Transactions on Intelligent Systems and Technology*, vol. 4, no. 4. 2013, doi: 10.1145/2508037.2508039.

[31] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: A survey," *ACM Computing Surveys*, vol. 38, no. 4. p. 45, Dec. 25, 2006, doi: 10.1145/1177352.1177355.

[32] Z. Chen, Z. Hong, and D. Tao, "An Experimental Survey on Correlation Filter-based Tracking," 2015. Accessed: Feb. 02, 2021. [Online]. Available: https://arxiv.org/abs/1509.05520.

[33] M. Fiaz, A. Mahmood, S. Javed, and S. K. Jung, "Handcrafted and deep trackers: Recent visual object tracking approaches and trends," *ACM Computing Surveys*, vol. 52, no. 2. Association for Computing Machinery, p. 43, May 01, 2019, doi: 10.1145/3309665.

[34] J. Kwon and K. M. Lee, "Visual tracking decomposition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010, pp. 1269–1276, doi: 10.1109/CVPR.2010.5539821.

[35] D. A. Ross, J. Lim, R.-S. Lin, and M.-H. Yang, "Incremental learning for robust visual tracking," *Int. J. Comput. Vis.*, vol. 77, no. 1–3, pp. 125–141, 2008.

[36] X. Mei and H. Ling, "Robust visual tracking and vehicle classification via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 11, pp. 2259–2272, 2011.

[37] L. Sevilla-Lara and E. Learned-Miller, “Distribution fields for tracking,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2012, pp. 1910–1917, doi: 10.1109/CVPR.2012.6247891.

[38] X. Li, A. Dick, C. Shen, A. Van Den Hengel, and H. Wang, “Incremental learning of 3D-DCT compact representations for robust visual tracking,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 4, pp. 863–881, 2012.

[39] K. Zhang and H. Song, “Real-time visual tracking via online weighted multiple instance learning,” *Pattern Recognit.*, vol. 46, no. 1, pp. 397–411, 2013, doi: 10.1016/j.patcog.2012.07.013.

[40] S. Oron, A. Bar-Hillel, D. Levi, and S. Avidan, “Locally orderless tracking,” *Int. J. Comput. Vis.*, vol. 111, no. 2, pp. 213–228, 2015.

[41] H. Grabner, M. Grabner, and H. Bischof, “Real-Time Tracking via On-line Boosting,” *Proc. Br. Mach. Vis. Conf.*, vol. 1, pp. 1–10, 2006, doi: 10.5244/C.20.6.

[42] Z. Kalal, J. Matas, and K. Mikolajczyk, “P-N learning: Bootstrapping binary classifiers by structural constraints,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010, pp. 49–56, doi: 10.1109/CVPR.2010.5540231.

[43] S. Hare *et al.*, “Struck: Structured Output Tracking with Kernels,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 10, pp. 2096–2109, 2016, doi: 10.1109/TPAMI.2015.2509974.

[44] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, “Exploiting the circulant

structure of tracking-by-detection with kernels," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2012, vol. 7575 LNCS, no. PART 4, pp. 702–715, doi: 10.1007/978-3-642-33765-9_50.

[45] K. Zhang, L. Zhang, and M.-H. Yang, "Fast compressive tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 10, pp. 2002–2015, 2014.

[46] Y. Zhang, L. Wang, and J. Qin, "Adaptive spatio-temporal context learning for visual tracking," *Imaging Sci. J.*, vol. 67, no. 3, pp. 136–147, 2019.

[47] H. Wang, P. Liu, Y. Du, and X. Liu, "Online convolution network tracking via spatio-temporal context," *Multimed. Tools Appl.*, vol. 78, no. 1, pp. 257–270, 2019.

[48] H. Wan, W. Li, and G. Ye, "An improved spatio-temporal context tracking algorithm," in *2018 13th IEEE Conference on Industrial Electronics and Applications (ICIEA)*, 2018, pp. 1320–1325.

[49] W.-G. Li and H. Wan, "An improved spatio-temporal context tracking algorithm based on scale correlation filter," *Adv. Mech. Eng.*, vol. 11, no. 2, p. 1687814019825903, 2019.

[50] J. Tian and Y. Zhou, "Real-time patch-based tracking with occlusion handling," in *International Conference on Neural Information Processing*, 2014, pp. 210–217.

[51] P. Chen and M. Yang, "STC Tracking Algorithm Based on Kalman Filter," 2016.

[52] F. Munir, A. Jalil, and M. Jeon, "Real time eye tracking using Kalman extended

spatio-temporal context learning," in *Second International Workshop on Pattern Recognition*, 2017, vol. 10443, p. 104431G.

[53] Z. Cui, J. Yang, S. Jiang, J. Li, and Y. Gu, "Robust spatio-temporal context for infrared target tracking," *Infrared Phys. Technol.*, vol. 91, pp. 263–277, 2018.

[54] X. Yang, S. Zhu, D. Zhou, and Y. Zhang, "An improved target tracking algorithm based on spatio-temporal context under occlusions," *Multidimens. Syst. Signal Process.*, vol. 31, no. 1, pp. 329–344, 2020.

[55] H. Yang *et al.*, "Combining spatio-temporal context and Kalman filtering for visual tracking," *Mathematics*, vol. 7, no. 11, p. 1059, 2019.

[56] Q. Fang, H. Zhou, and L. Zhang, "Spatio-Temporal Context Tracking Algorithm Based on Correlation Filtering," in *Journal of Physics: Conference Series*, 2019, vol. 1213, no. 3, p. 32010.

[57] X.-G. Wei, S. Zhang, and S.-C. Chan, "A novel visual object tracking algorithm using multiple spatial context models and Bayesian Kalman filter," in *2015 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2015, pp. 1034–1037.

[58] D. S. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui, "Visual object tracking using adaptive correlation filters," in *2010 IEEE computer society conference on computer vision and pattern recognition*, 2010, pp. 2544–2550.

[59] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 583–596, 2014.

[60] J. Ahmed, A. Ali, and A. Khan, “Stabilized active camera tracking system,” *J. Real-Time Image Process.*, vol. 11, no. 2, pp. 315–334, 2016.

[61] C. Ma, X. Yang, C. Zhang, and M.-H. Yang, “Long-term correlation tracking,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 5388–5396.

[62] H. Masood, S. Rehman, A. Khan, F. Riaz, A. Hassan, and M. Abbas, “Approximate Proximal Gradient-Based Correlation Filter for Target Tracking in Videos: A Unified Approach,” *Arab. J. Sci. Eng.*, vol. 44, no. 11, pp. 9363–9380, 2019.

[63] A. Ali, A. Jalil, J. Ahmed, M. A. Iftikhar, and M. Hussain, “Correlation, Kalman filter and adaptive fast mean shift based heuristic approach for robust visual tracking,” *Signal, Image Video Process.*, vol. 9, no. 7, pp. 1567–1585, 2015.

[64] F. Farahi and H. S. Yazdi, “Probabilistic Kalman filter for moving object tracking,” *Signal Process. Image Commun.*, vol. 82, p. 115751, 2020.

[65] M. B. Khalkhali, A. Vahedian, and H. S. Yazdi, “Vehicle tracking with Kalman filter using online situation assessment,” *Rob. Auton. Syst.*, vol. 131, p. 103596, 2020.

[66] M. Danelljan, G. Häger, F. Khan, and M. Felsberg, “Accurate scale estimation for robust visual tracking,” 2014.

[67] H. Ma, Z. Lin, and S. T. Acton, “FAST: Fast and Accurate Scale Estimation for Tracking,” *IEEE Signal Process. Lett.*, vol. 27, pp. 161–165, 2019.

[68] W. Ruan, C. Liang, Y. Yu, J. Chen, and R. Hu, "SIST: Online scale-adaptive object tracking with stepwise insight," *Neurocomputing*, vol. 384, pp. 200–212, 2020.

[69] Y. Li and J. Zhu, "A scale adaptive kernel correlation filter tracker with feature integration," in *European conference on computer vision*, 2014, pp. 254–265.

[70] A. Bibi and B. Ghanem, "Multi-template scale-adaptive kernelized correlation filters," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2015, pp. 50–57.

[71] M. Danelljan, G. Häger, F. S. Khan, and M. Felsberg, "Discriminative scale space tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 8, pp. 1561–1575, 2016.

[72] X. Yin, G. Liu, and X. Ma, "Fast Scale Estimation Method in Object Tracking," *IEEE Access*, vol. 8, pp. 31057–31068, 2020.

[73] B. Khan *et al.*, "Multiple Cues-Based Robust Visual Object Tracking Method," *Electronics*, vol. 11, no. 3, p. 345, 2022.

[74] H. Lu, D. Xiong, J. Xiao, and Z. Zheng, "Robust long-term object tracking with adaptive scale and rotation estimation," *Int. J. Adv. Robot. Syst.*, vol. 17, no. 2, p. 1729881420909736, 2020.

[75] Y. Yuan, J. Chu, L. Leng, J. Miao, and B.-G. Kim, "A scale-adaptive object-tracking algorithm with occlusion detection," *EURASIP J. Image Video Process.*, vol. 2020, no. 1, pp. 1–15, 2020.

[76] S. Ting Goh, S. A. Zekavat, and O. Abdelkhalik, "An introduction to Kalman filtering implementation for localization and tracking applications," *Handb. Position Locat. Theory, Pract. Adv. Second Ed.*, pp. 143–195, 2018.

[77] M. Wang, Y. Liu, and Z. Huang, "Large margin object tracking with circulant feature maps," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4021–4029.

[78] P. Liang, E. Blasch, and H. Ling, "Encoding color information for visual tracking: Algorithms and benchmark," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5630–5644, 2015.

[79] Y. Wu, J. Lim, and M. Yang, "Object Tracking Benchmark," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1834–1848, 2015, doi: 10.1109/TPAMI.2014.2388226.

[80] M. Mueller, N. Smith, and B. Ghanem, "A benchmark and simulator for uav tracking," in *European conference on computer vision*, 2016, pp. 445–461.

[81] K. Mehmood *et al.*, "Efficient Online Object Tracking Scheme for Challenging Scenarios," *Sensors*, vol. 21, no. 24, p. 8481, 2021.

[82] C. S. Asha and A. V Narasimhadhan, "Adaptive learning rate for visual tracking using correlation filters," *Procedia Comput. Sci.*, vol. 89, pp. 614–622, 2016.

[83] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proceedings of ICNN'95-international conference on neural networks*, 1995, vol. 4, pp. 1942–1948.

[84] J. H. Holland, “Genetic algorithms,” *Sci. Am.*, vol. 267, no. 1, pp. 66–73, 1992.

[85] L. Abualigah, D. Yousri, M. Abd Elaziz, A. A. Ewees, M. A. A. Al-qaness, and A. H. Gandomi, “Aquila Optimizer: A novel meta-heuristic optimization Algorithm,” *Comput. Ind. Eng.*, vol. 157, p. 107250, 2021.

[86] L. Abualigah, A. Diabat, S. Mirjalili, M. Abd Elaziz, and A. H. Gandomi, “The arithmetic optimization algorithm,” *Comput. Methods Appl. Mech. Eng.*, vol. 376, p. 113609, 2021.

[87] S. K. Pal, A. Pramanik, J. Maiti, and P. Mitra, “Deep learning in multi-object detection and tracking: state of the art,” *Appl. Intell.*, vol. 51, no. 9, pp. 6400–6429, 2021.

[88] M. Zhai, L. Chen, G. Mori, and M. Javan Roshtkhari, “Deep learning of appearance models for online object tracking,” in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018, p. 0.

[89] P. Li, D. Wang, L. Wang, and H. Lu, “Deep visual tracking: Review and experimental comparison,” *Pattern Recognit.*, vol. 76, pp. 323–338, 2018.