JEAS | Journal of Economics and Administrative Sciences
E-ISSN: 2148-1792 P- ISSN: 1302-2024
Volume 3, Supplement Issue 2
www.jeasweb.org

Tracking targets using learning spatio-temporal contents despite large occlusions

Morteza Dehghani
Master of Computer Science, Artificial Intelligence, Buin Zahra Azad University, Qazvin, Iran.

*Abbas Kochari**
Assistant Professor, School of Mechanics, Electrical Engineering, Electronics, Azad University of Research Sciences,tehran ,Iran.
**Corresponding Author*

ABSTRACT
In this study, a visual tracking system was designed using learning spatio-temporal contents, which is resistant to large occlusions. This tracker uses two types of contents to track, so that temporal content is used to record changes in the target's appearance over time and spatial content to learn the local content around the target. Using temporal content is important because if the tracker faces challenges such as obstructions or brightness changes to find the target during successive frames, the tracker can estimate its probable location based on the time history. Spatial content learns the auxiliary points around the target and the target itself, and is used in the next frames to update the spatio-temporal contents. If the target is obstructed or an object similar to the target is placed in front of the tracker, the spatial content will prevent the tracker from being deflected. In this method, the tracking problem is raised by confidence map and the best target location is calculated based on the maximum likelihood of the locations of the object. Moreover, the proposed method uses Fast Fourier Transform (FFT) for rapid detection and learning. The target scale is obtained by filtering to reduce the noise generated by the estimation error. The proposed tracker was compared with four similar types of trackers where the proposed method was significantly resistant to large occlusions.
Keywords: Tracking, learning spatio-temporal contents, large occlusions, maximum likelihood, FFT, confidence map

Introduction

With the advancement of the computer industry and the increasing speed of computers, some algorithms were developed for automating the video analysis process.

Object tracking is of the most significant and practical issues in machine vision, attracting the attention of researchers nowadays as one of the most up-to-date issues. Target tracking is one of the most important and practical issues in machine vision much studied in the last decade. Tracking is used in video surveillance, content creation, personal communication, human-machine communication robotics and many other fields.

The dramatic increase in the quality and accuracy of video sensors and the tremendous increase in the processing power of computers in recent decades have brought about the development of new algorithms and applications in video tracking. Some requirements like robust, real-time processing must be considered in creating a powerful tracking system. Moreover, the changes that occur in the environment where the object is located, the object itself may change too, for instance, due to changes in brightness.
One of the major challenges of the tracking problem is occlusion. Another object may block part or the entire object so that the object is not visible.

A common method for examining complete occlusion during tracking is to model the motion of the object by linear dynamic models or by nonlinear dynamics. Different algorithms have been proposed with a focus on appearance models.

One of the most significant features of a tracking algorithm is that it can find and continue tracking the original object after it reappears in the frame. The subsequent challenges are the changes in the brightness the appearance of the object. The brightness of the environment may change over time in a video. The appearance of the object may also change somewhat. These two changes mean changes in the properties of the object extracted from the image. If the algorithm only measures the first frame and extracts the properties from it, it will not be able to track the object after changing the features. Another challenge is the existence of similar objects in each frame of the video. If the properties used to select an object are not accurate, then a similar object may be selected instead of the original object. Hence, the necessary care must be taken in feature selection.

As the changes of the target object are continuously spatial and temporal over time, these features can be used in tracking it in a robust way.

Time content information is obtained from the history of changes in appearance and is used in a long-term tracking to prevent tracker deviation. The spatial content model is the key points on and around the target object that forms the auxiliary field after extraction. The auxiliary field generates so much information about the appearance of the target that the target location will be accurately predicted. In the real world, spatio-temporal content information is critical in tracking.

In consecutive frames, the appearance of the target changes gradually, based on which a history of changes in the appearance of the object can be created, and this history, which can include changes in appearance in gesture, scale and brightness, will have more or less effects and limitations in the next area.

Spatial contents of the target are important for information tracker, so it will look for the real goal, clearly showing the significance of spatial content. As target tracking is a continuous physical and psychological process in the spatio-temporal content method, all previous information will be used to predict the next goal situation, and the current appearance of the target is more or less related to its previous appearance (Van et al., 2012).

In the method by Yang et al. (2007), a data mining method was used to extract fragmented areas around the object for tracking called auxiliary objects. To find fixed areas, firstly the key points around the object are extracted to help position the object, and then Surf descriptors are used to show these fixed areas. Thus, large computational operations are needed to indicate and find fixed areas. Furthermore, given the diversity of key points, some fixed areas suitable for determining the position of the object may be lost. In contrast, the proposed algorithm does not have this problem as it considers the local areas around the object as potential fixed areas, and the associated motion between objects and their local contents is learned in successive frames by spatio-temporal contents model that is effectively calculated by FFT (Zhang et al., 2013).

The artistic methods of Quis et al. (2012) and Bolme et al. (2009) and Bolme et al. (2010) have used FFT for efficient calculations and their formulation is based on correlation filters directly obtained by classical signal algorithms. These filters are trained using a large number of samples and then combined to find the most relevant position in the next frame. These filters have been used to obtain more stable results.

In visual tracking, a local content, including a target object and its real-time background, is located within a certain area. Hence, there is a spatio-temporal relationship between local scenes containing the object in successive frames. For instance, if a target object is subjected to a large occlusion, the appearance of the object changes significantly, although the local content containing the object does not change that

much so that the overall appearance remains the same and only a small portion of the content area is blocked. Thus, the existence of local contents in the current frame helps predict the location of the object in the next frame. This content information has recently been used to identify objects. Moreover, the spatial relationship between an object and its local content provides specific information about the configuration of a scene, which helps to distinguish the target from the background in case of many visual changes. The method this study used is a fast and robust algorithm that extracts spatio-temporal contents information. In this method, first, a local content model based on spatial relationships in a scene is learned by solving a deconvolution problem between the target object and the surrounding local background, and then the learned model is used to update the spatio-temporal contents of the next frame. Tracking in the next frame is done by calculating confidence map as a convolutional problem that completes spatio-temporal content information that can formulate the best object location by the estimated maximum confidence map (Zhang et al. 2014).

The proposed algorithm has the advantages of both production and differentiation methods. On the one hand, the content including the target and the background are located in its neighborhood, making the proposed method have the advantages of differentiation models, and on the other, the content includes all its target and background: our method has the advantages of production models.

As the sequence of images has spatio-temporal characteristics, the content information of a moving object has both spatial and temporal properties. The spatial contents of the object include the local texture and other background objects in the current frame, and the temporal content of the object contains the history of the object. In visual tracking, a local content including a target object and its real-time background is located within a given area, so there is a spatio-temporal relationship between the local scenes containing the object in successive frames. Thusly, this information can be used in robust target tracking. The study tried to enhance the tracking performance of the object in the face of large occlusions by learning the spatio-temporal contents.

The assumption of the study is that using spatio-temporal algorithm enhances the problem of large occlusions and the calculation of the maximum likelihood obtains the best location of the object. Additionally, it allows FFT calculation, fast learning and object detection.

Based on the above regarding object tracking, the study tries to answer the question of whether learning spatio-temporal contents can solve the problem of large occlusions in tracking target or not.

Materials and Methods

The research method was theoretical-applied model.

Ten different video sequences were used to test the performance of the proposed tracker against large occlusions according to Table 1. Moreover, video sequences in outdoor and indoor environments were used

Table 1: Sequence specifications

Sequence	The frames with occlusions	The number of frames outdoor	The number of frames indoor
Following the teddy bear	25		1336
Following the can	15		291
Following the pedestrian	4	252	
Movement of the book against the girl's face	65		892
Movement of the book against the boy's face	33		812
Following the footballer	40	362	
Following the student in the class	15		297
The man's head moving against the woman's face	138		500
Following the doll tiger	5		354
Following the female runner	8	207	

JPEG format video images were extracted from the link below.

https://sites.google.com/site/trackerbenchmark/benchmarks/v10

Three parameters are considered for data analysis:

A- Execution of 500 frames per second in MATLAB software to execute the proposed method

B- Using diagrams to determine the accuracy of the object location, where the problem of object ambiguity is solved by selecting the parameter β

C- Using probabilistic formulas to quickly track and obtain the exact location of the object: The first step in the proposed algorithm is to calculate the spatio-temporal relationships between the object and its local content. In the second stage, the algorithm focuses on areas that need to be analyzed in detail, thus effectively reducing the side effects of background distortions leading to more robust results. In the third step, the problem of object ambiguity is examined using confidence map with an appropriate prior distribution. Thus, more stability and more accurate efficiency are obtained for visual tracking, and the problem of scale matching is reached.

Problem formulation

The tracking problem is formulated by calculating a confidence map that estimates the likelihood of the object location:

$$C(x)= p(x|o) \quad (1)$$

Here, $x \epsilon R^2$ is the location of the object and on the presence of the object in the scene. Below is the spatial content information obtained. Figure 1 shows the graphical model of the spatial content. X* is the location of the object is in the current frame (e.g., the coordinates of the center of the tracked object).

The content feature set is displayed as $X^C=\{C(z)=I(z)|z \epsilon \Omega_c(x^*)\}$, where I (z) shows the intensity of the image at location z, and $\Omega_c(x^*)$ the content of the neighborhood x*.

With the probability P (x, c (z) | o) the object likelihood function can be represented by the following formula:

$$C(x)= p(x|o) =\sum_{c(z)\epsilon X^c} P(x, c(z)|o)=\sum_{c(z)\epsilon X^c} P(x|c(z), o)\, P(c(z)|o) \quad (2)$$

Here, c (x) is the confidence map of the location of object x and Xc the content of the location of the object x.

The main task here is to learn P(X|c(z),o), so that it is a bridge for the empty space between the location of the object and its spatial content.

Spatial content model: The conditional probability function P (x | c (z), o) in Equation 2 is defined as follows:

$$P(x|c(z),o)=h^{sc}(x-z) \quad (3)$$

Here, $h^{sc}(x-z)$ is a function based on the relative distance and direction between the location of object x and the location of local content z, thus the spatial relationship between an object and its spatial content is encoded.

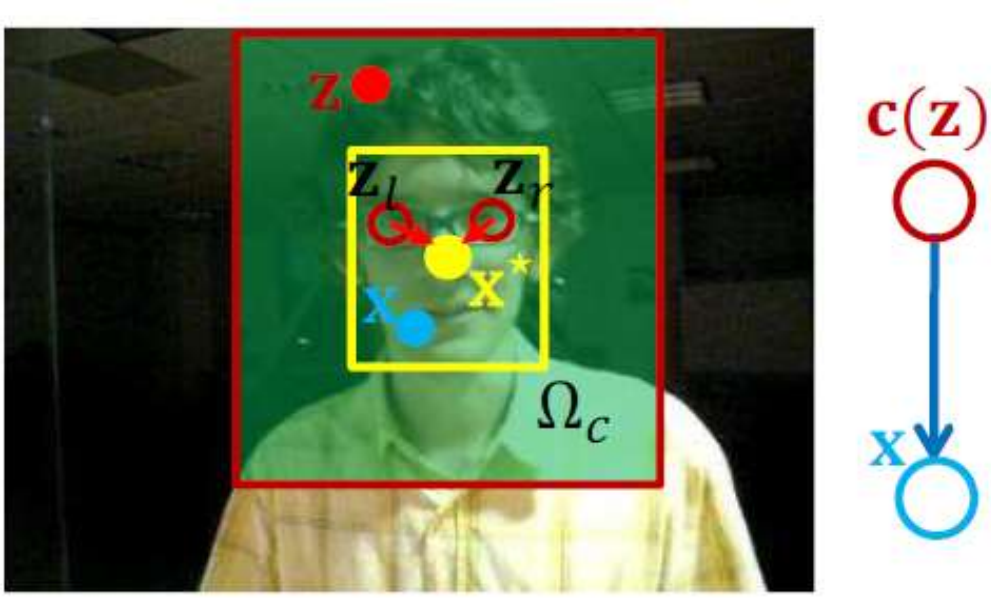

Figure 1: Graphic model of the spatial relationships between the object and the local content around it

Proposed tracking algorithm: The proposed method assumes that the target location in the first frame is initialized manually or by object detection algorithms. In the t-th frame, we learn the spatial content model $h_t^{sc}(x)$ used to update the spatio-temporal content model $h_{t+1}^{stc}(x)$ and is used to identify the location of the object in the t + 1-th frame. When the t + 1-th frame is reached, the local content area Ωc (x*) is picked up based on the e_t^* tracking location in the t-th frame, and the corresponding content feature set is created as follows.

$$X_{t+1}^c=\{c(z)=(I_{t+1}(z),z)|z \in z\epsilon\Omega_c(x_t^*)\} \quad (4)$$

Then the location of the object in the t + 1 frame is determined by the new maximum confidence map.

$$X_{t+1}^* \; argmax c_{t+1}(x) \quad (5)$$
$$X\in \Omega_c(x_t^*)$$

Here, Ct + 1 (x) is shown as follows:

$$c_{t+1}(x) = f^{-1}(f(H_{t+1}^{stc}(x))) \odot f(I_{t+1}(x)\, w_{\sigma_t}(x\text{-}x^*)) \quad (6)$$

Updating spatio-temporal content: The spatio-temporal content model is updated as follows:

$$H_{t+1}^{stc} = (1-\rho)H_t^{stc} + \rho H_t^{sc} \quad (7)$$

Here, ρ is a learning parameter and h_t^{sc} is the spatial texture model calculated by Equation 4 in the t-frame. One way of filtering is the time when it can easily be displayed in the frequency domain:

$$H_w^{stc} = f_w h_w^{sc} \quad (8)$$
$$H_w^{stc} \triangleq \int H_w^{stc} e^{-jwt} dt$$

FFT is H_t^{stc} and similar to h_w^{sc}. F_w time filter is formulated as follows:

$$F_w = \frac{\rho}{e^{-jw} - (1-\rho)} \quad (9)$$

Here, j shows an imaginary unit. F_w is a low-pass filter used for easy validation. Thus, the proposed spatio-temporal content model can effectively filter the output image noise introduced by the appearance variables, resulting in more robust results.

Results

In this study, the proposed tracker that has ten various video sequences with different scenarios has been compared and examined:

the sequence of girl's face blocked by a book with 892 frames, the sequence of can blocked by shrub leaves with 291 frames, the sequence of pedestrian blocked by a tree with 252 frames, the sequence of boy's face blocked by a book with 812 frames, the sequence of footballer blocked by another player with 362 frames, the sequence of student blocked by his classmates with 297 frames, the sequence of woman face blocked by man with 500 frames, the sequence of female runner blocked by an electric light pole with 308 frames, the sequence of teddy bear blocked by a table with 500 frames, and the sequence of doll tiger blocked by a shrub with 354 frames

In all of them, the proposed tracker significantly accomplished target tracking successfully compared to other trackers.

Proposed tracker with four types of trackers (Table 1) - Zhang et al. (2012) multi-task tracking (MTT), Rose et al. (2008) incremental visual tracker (IVT), Bolme et al. (2009) circulant sparse tracker (CST) and Bolme et al. (2010) Minimum Output Sum Of Square (MOS). Ten video sequences (teddy bear, can, pedestrian, girl's face block, boy's face block, soccer player, class, woman and man, female runner and doll tiger) have been compared as follows. MTT: in this type of tracker, object tracking is formulated in a particle filtering framework as a privacy multifunction learning problem, where tracking is shown as a multifunctional structure. Thus, linear particle models are a list of patterns dynamically learned and updated.

This tracker detects the target object in the sequence of a teddy bear in Figure 2, which is marked in blue. In frame (a), in frame (b), the object is partially detected by this tracker, but when the object is blocked by a large occlusion and when the target object reappears in frame (c), the tracker loses and deflects.

In the can sequence in Figure 3, where the MTT is specified in blue; First in frame (a), the target object is identified in frame (A) and then in frame (B), where the target moves and is located behind the bush, this tracker detects it based on the upper part of the target. However, after being blocked by the shrub and then leaving the occlusion, the target is lost from the view of this tracker and deviates in tracking it.

In the pedestrian sequence in Figure 4, this tracker detects the target in frame (a) and moves with the target object in frame (b), but after the target is blocked by the tree, the tracker loses it and focuses on the tree.

In the sequence of blocking the girl's face by the book in Figure 5, MDD tracker detects the girl's face in frames (A) and (B). However, when the target is placed behind the book, the tracker misses the main target and focuses on the book, so that in frame (C), the book is assumed as the target instead of identifying the girl's face as the target and deviates towards it.

In the boy sequence in Figure 6, MTT tracker in frame (a) assumes the boy's face as the target and pursues it in frame (b), but when in this frame the target disappears from the tracker's view due to large occlusion. In frame (C), the tracker may be diverted to the book and mistakenly assumes it is the target instead of identifying the target.

In the footballer sequence in Figure 7, where MTT is shown by yellow, the player is first identified by the tracker as the target in frame (A). Then by the target object moving and its being blocked behind the other player in frame (B), the tracker loses the target, so that in frame (C) it assumes another instead of pursuing the main target.

In the classroom sequence in Figure 8, where the tracker is shown in blue, the tracker identifies an individual as the target in frame (A) and due to rapid movement and subsequent occlusion of the target in the frame (B), the tracker is left out of the target identification, clearly seen in frame (C).

In the male-female sequence in Figure 9, where the MTT tracker is shown as orange in frame (A), the woman's face as the target is identified by this tracker, and in frame (B) where the woman's face is blocked by the man's head, it loses the main target and instead considers the male face as the target, evident in the following frames (B), (C), and as a result the tracker is diverted to another object instead of the main target.

In the sequence of the woman runner in Figure 10, this tracker is shown in blue in frame (A). After the woman runner is shown as the target in a white T-shirt, the tracker detects her. When the light pole in the foreground in frame (B) causes an occlusion, the tracker will not be able to track the target in the next frame, frame (C), due to the loss of the target because of occlusion and will be diverted to another person who is running next to the target.

In the sequence of the doll tiger in Figure 11, where this tracker is shown in blue, after being identified as the target for the tracker in frame (A), it continues to track until in frame (B), the target is hidden from view by shrub branches and this occlusion causes the tracker in frame (c) to be diverted to the branches instead of identifying and tracking the target, and then cannot track the target.

However, CST uses a rotational tracking structure with kernel detection for tracking, and the method provides a theoretical framework for analysis of the results of in-depth sampling in detection tracking, and thus a set of solutions for online training, detection and nonlinear kernel calculations are presented. In this task, using the theory of rotational matrices, a link has been obtained for Fourier analysis, which allows learning and diagnosis very quickly by FFT. This can be done quickly in the space of nuclear machines by linear classification. Gaussian and polynomial kernels have been used for training and diagnosis. In the tracking method, links are created by this tracker for Fourier analysis, creating the use of FFT to accelerate the synthesis of information in all windows by preventing their duplication. A critical component in tracking is classification detection. Each frame is a collection of samples gathered around the estimated position of the target. Samples close to the target are labeled positively, and samples much farther from the target are labeled negatively. Classifier updates allow the samples to be adapted over time. Only a limited number of random samples are used given computational limitations. In this task, the classifier is taught to all the samples, called deep sampling. This tracker is designed for high level localization.

CST tracker detects the target object in the sequence of the teddy bear in Figure 2, marked in yellow in frame (A). In frame (B), where the target object moves rapidly behind the table, the target is tracked by this tracker, but in Frame (C) will not be able to fully detect the object after the object is blocked.

In the cans sequence in Figure 3, this tracker is shown in yellow. First, in frame (A), the can is detected as a target by the tracker and then in frame (B), it detects it; however, when the target object in frame (C) is blocked by the leaves of the bush, this tracker loses it in frame (D), will not be able to track it.

In the pedestrian sequence in Figure 4, where this red-colored tracker is marked, first in frame (A), the target object is detected by the tracker, and in frame (B), where the target is blocked by a tree, the tracker loses focus and in the frame (C), it loses the target.

In the girl sequence in Figure 5, CST identifies the girl's face as the target, but in frame (C), where the book moving on the girl’s face obstructs the target, it loses focus on the target and deviates toward the book. Thus, the result of this deviation is clearly seen in frame (C).

In the boy sequence in Figure 6, this tracker functions similarly to the female face block sequence in Figure 4-5. In the football player sequence in Figure 7, this tracker in frame (A) detects the target object. However, the occlusion that occurs in frame (B), it misses the target so that in frame (C), it is diverted to another player.

In the classroom sequence in Figure 8, the tracker detects a person wearing a black shirt as the target in frame (A), then loses the target in the next frame due to rapid movement, and then blocks that target as in the frame (C), this issue is observed.

In the male-female sequence in Figure 9, the tracker is shown in blue. In frame (A), the girl's face is identified as the target, then the target face is blocked by another person's head in frame (B). This tracker

deviates from the main target, assumes the second person's head as the target, and detects that this is shown in frame (C).

In the female runner sequence in Figure 10, where the tracker is shown in yellow, the woman in the white T-shirt is identified as the target by the tracker in frame (A). After a light pole blocks the target object in frame (B), this tracker loses it and deflects towards another person who is running next to it in frame (C).

In the doll tiger sequence in Figure 11, where the tracker is marked in red, the doll tiger is identified as the target by this tracker. After the target object is blocked by the shrub leaves in frame (B), the tracker loses the target and will not be able to correctly track the target in the frame (C).

IVT uses an incremental learning method to track, where low-dimensional subspaces are learned, and the changes in the target's appearance are effectively adapted online. In this tracking method, model updating based on incremental algorithms for principal component analysis (PCA) consists of two features: one method for correct updating of the sample mean and a forgetting factor to ensure the modeling power of previous observations is appropriate. Both methods are used to increase tracking power. The tracking problem is formulated as an inferential problem in the context of the Monte Carlo Markov chain and the particulate filter for propagating the distribution of samples recorded over time. Firstly, this algorithm does not need training images of the target object before starting the target task. Secondly, the sampling method uses a particle back filter so that samples are distributed over time. Thirdly, both the average samples and the specific basics are properly updated as new data.

MOS tracker performs tracking using adaptive correlation filters, and these filters have been modified for this purpose, so that they are trained online and used in an adaptive way for visual tracking. Filter-based trackers model the appearance of an object using filters applied to training sample images. Initially, the target is selected based on an object-focused tracking window in the first frame. The tracking and filtering tasks are trained together from this point. In the next frame, a correlation filter under a search window tracks the target. The location corresponding to the maximum output value of the filter shows the correlation of the new target position. An on-line update is done based on the new location then.

IVT incremental learning trackers and MOS correlation filters, in ten sequences of teddy bear, can, pedestrian, girl, boy, soccer player, class, woman and man, female runner and doll tiger, perform similarly to CST and MTT trackers and will not be able to track goals properly against large occlusions.

The proposed LOSTC tracker uses spatio-temporal content learning to track the target object quickly. It learn this by solving a deconvolution problem model of the spatial content between the target object and the surrounding local background based on the spatial relationships in a scene, and then use the learned spatial content model to update a spatio-temporal model in the next frame. Tracking in the next frame is formulated by calculating a confidence map as a convolutional problem that completes spatio-temporal content information and the best location of the object is estimated by confidence map.

In the teddy bear sequences, this tracker is marked in red, after the exact location of the bear is detected by the tracker in frame (A), the target object moves and then in frame (B), it is blocked by other objects. However, the tracker does not lose it with the help of spatial content and the apparent history or time content, so the tracking continues after the occlusion in frame (C) is removed. The suggested tracker in the rest of the sequences - can, pedestrian, girl, boy, soccer player, class, man and woman, female runner and doll tiger - with the help of spatial content as well as using temporal contents, which is a history of changes in appearance, performs tracking successfully after a part of the tissue is covered by a large occlusion.

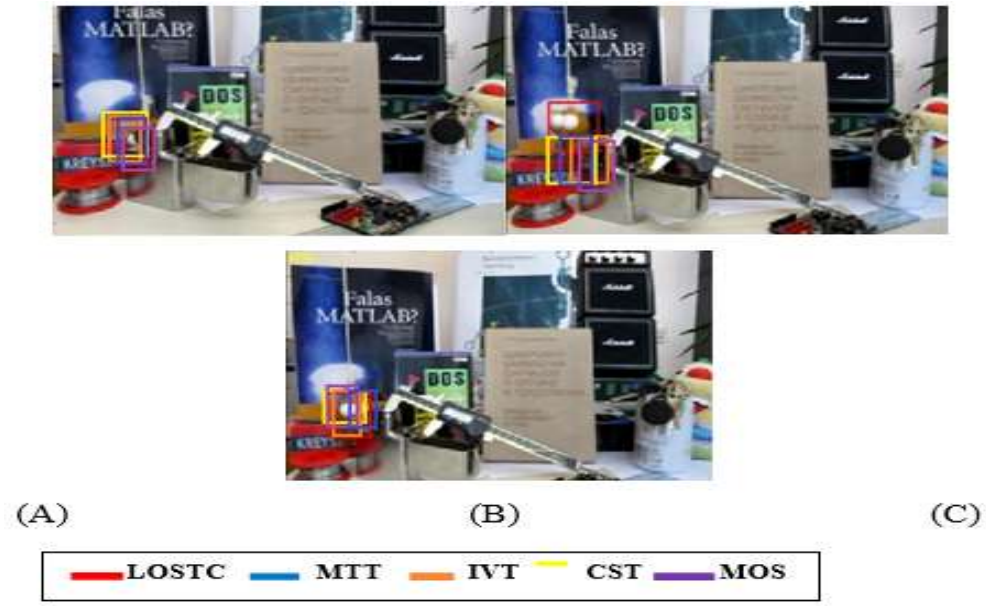

Figure 2: Video sequence of the teddy bear movement

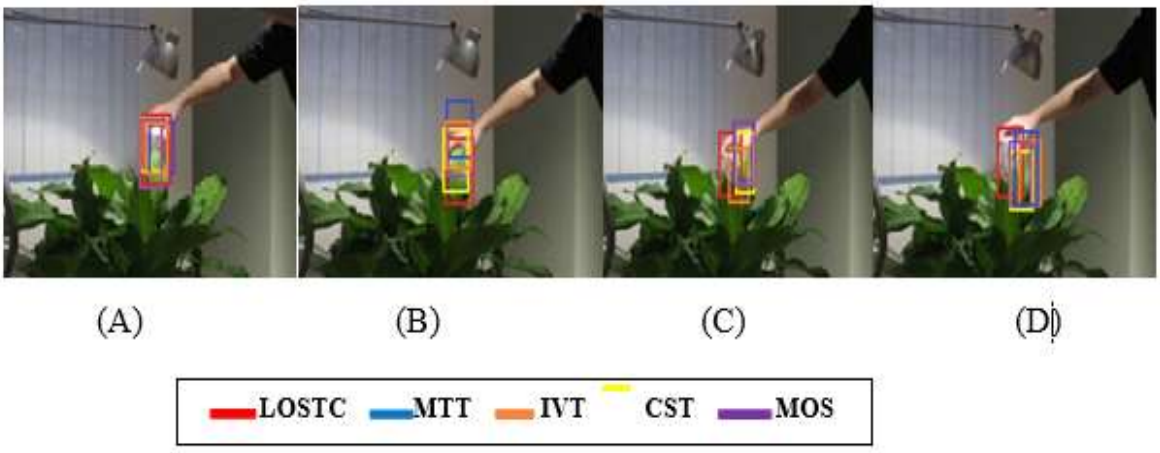

Figure 3: Video sequence of can movement

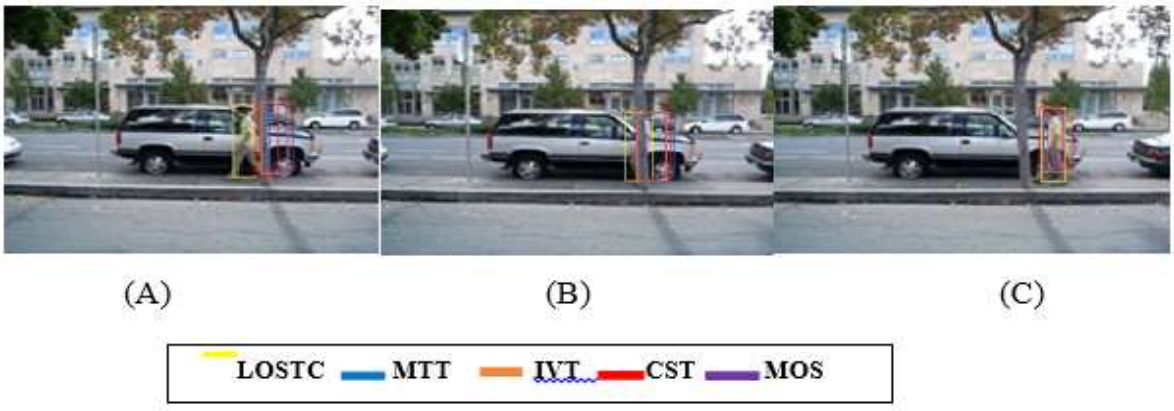

Figure 4: Video sequence of pedestrian movement on the street

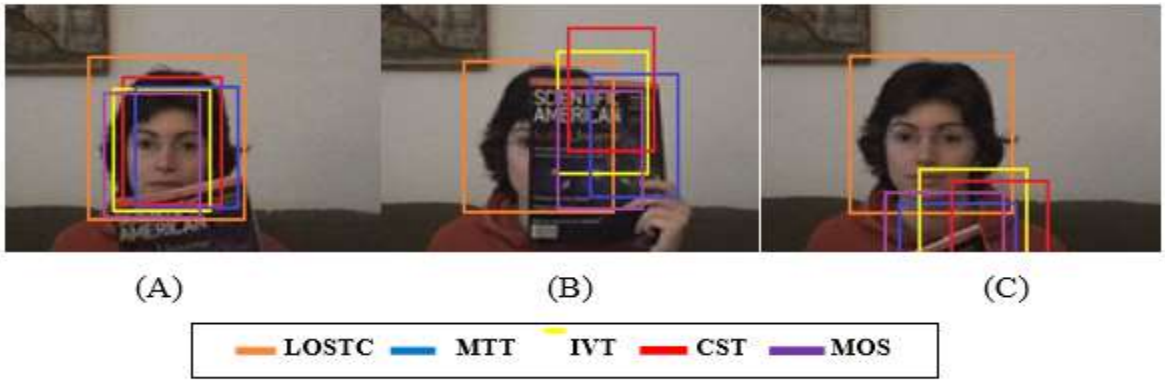

Figure 5: Video sequence of the book moving in front of the girl's face

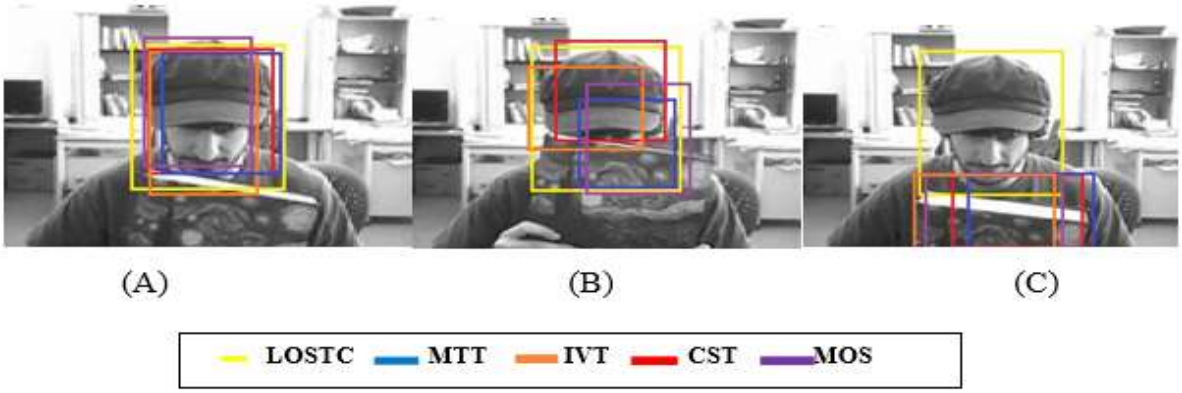

Figure 6: Video sequence of the book moving in front of the boy's face.

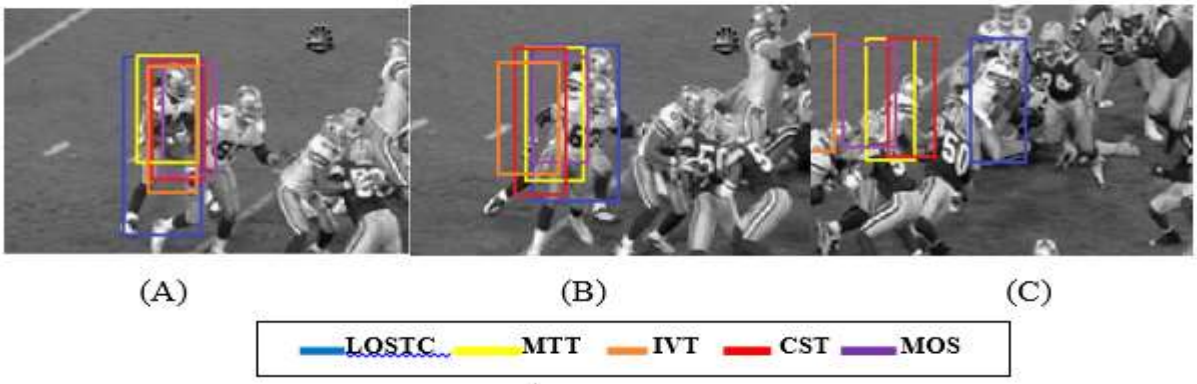

Figure 7: Video sequence of a football player

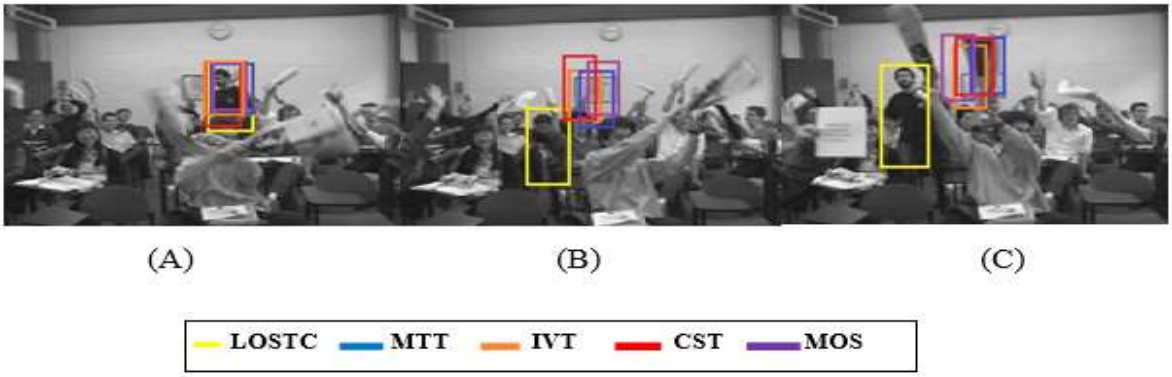

Figure 8: Video sequence of student movement in the classroom.

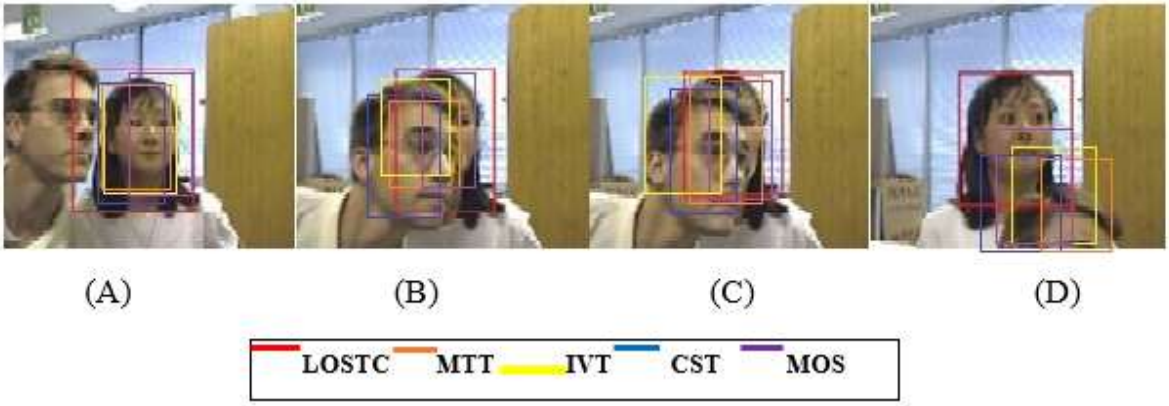

Figure 9: Video sequence of the head movement in front of a woman's face.

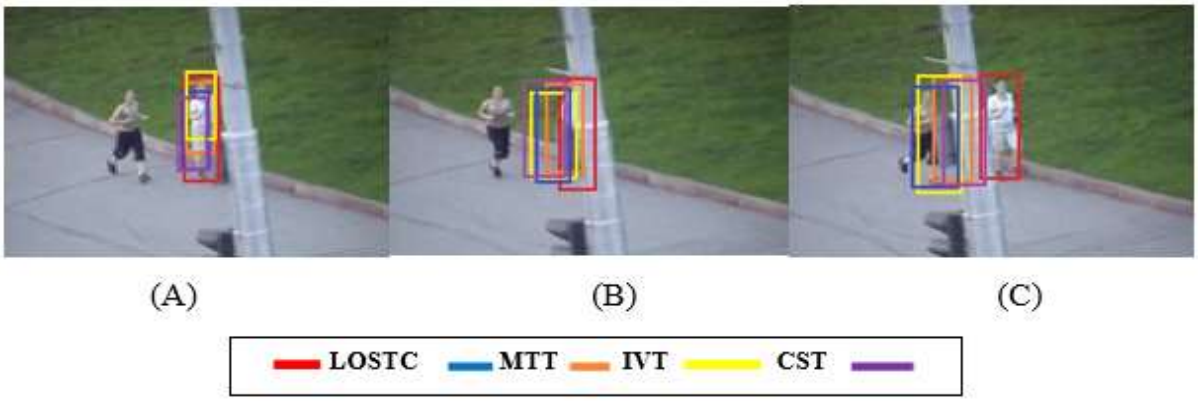

Figure 10: Video sequence of a female runner

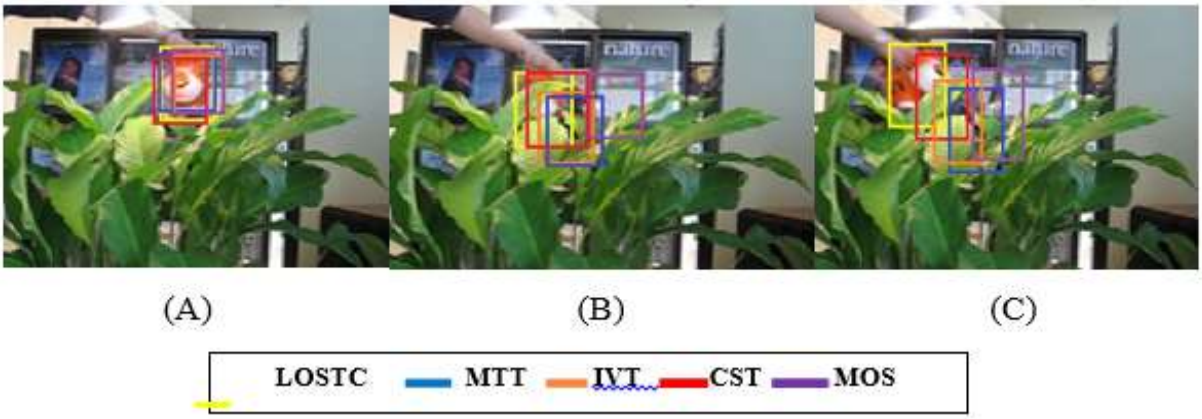

Figure 11: Video sequence of doll tiger movement

In a comparison between the proposed and the other four trackers, the success percentage of the target tracking in the blocked frames was compared with our tracker. For instance, in the teddy bear sequence, where the target is blocked at 25 frames, MTT could detect the target at 5 frames and in MOS tracker at 7 frames. IVT tracker could detect the target in 10 frames and the CST track in 4 frames. Nevertheless, the proposed tracker could track the target in 22 frames. This success has been interestingly noted for the proposed tracker in other sequences too. The average success rate of the proposed tracker is 30%, which is significantly higher than other trackers (Table 2).

Table 2: The success rate of the proposed tracker in large occlusions

Sequence	MTT	MOS	IVT	CST	LOSTC
Following the teddy bear	5	7	10	4	22
Following the can	11	10	13	9	15
Following the pedestrian	1	3	2	1	4
Movement of the book against the girl's face	23	20	31	35	50
Movement of the book against the boy's face	12	20	31	25	32
Following the footballer	23	22	26	27	32
Following the student in the class	10	11	8	9	14
The man's head moving against the woman's face	50	60	55	99	120
Following the doll tiger	2	3	2	4	5
Following the female runner	3	5	6	4	7
Average success	14	17	19	22	30

Discussion and Conclusion

Target tracking is one of the topics attracting the attention of researchers in machine vision. Accurate goal tracking despite natural and unnatural challenges is an interesting subject. Now, if the tracked target is a human or the equipment and devices made by human intelligence, the subject becomes even more attractive as it is a nonlinear model that may occur in their intelligent movement. Tracking targets is a pattern of tracking the target by human visual system generalized to tracking devices and tools over time. In this field of machine vision, there have been wide studies and various methods have been proposed by

researchers, each of which has its advantages and disadvantages like Kalman filter, particle filter, using texture, auxiliary objects, kernels, correlation filters, spatio-temporal content, and so on.

Each of these methods has been a tool helping address a variety of tracking challenges like background clutter, brightness changes, sudden target rotation, occlusion, and so on. Here, linear models have been developed for tracking like Kalman and Kalman filters, and other methods have used nonlinear models like particle filters, which have been more successful than the previous method. Some methods have used online updates in the models and other methods have used offline updates, each of which has its own advantages and disadvantages. Thus, online updates have been more practical than offline. Some of these tracking algorithms have used production methods to model the target object where the first frame, the target object is precisely specified and the model is created based on the information extracted from it. In the next frame, the template with the highest degree of compatibility with the model is selected as the target, and some other differentiation methods are used to model the object, whose general structure is similar to the generative methods, except that background information is used to create the model. The proposed method has benefited from both production and differentiation methods. On the one hand, the content includes the target and the background in its neighborhood, which consequently makes our method have the advantages of differentiation models, and, the context includes all goals and backgrounds on the other. Our method has the advantages of generative models.

The study examined various trackers with the highest similarity to the proposed tracker in terms of application like: text-aware visual tracking, tracking using learning auxiliary objects, using deflectors and auxiliaries in the texture, selected space for accurate visual tracking, robust visual tracking using incremental learning, visual tracking using background distribution, robust visual tracking through privacy multitasking, real-time tracking through online amplification, structured output of the tracker with kernels, using rotational tracking structure by means of detection using kernels, experimental study of texture in object recognition, visual tracking of the object using adaptive correlation filters, tracking using spatio-temporal structured content, each of which were analyzed. After that the proposed tracker was compared with four types of trackers (MTT, IVT, MOS and CST) in ten different scenarios (ten sequences of teddy bear, can, pedestrian, girl, boy, soccer player, class, man and woman, female runner and doll tiger).

Two significant points exist in the proposed method:

A) Spatial relationships between various categories of objects used that help inference to increase accuracy

B) The temporal contents used to summarize the evidence and the continuous tracking of the object and the history of the object in general

In the method used in the study, the spatial content model between the target object and the surrounding local background is learned based on the spatial relationships in a scene by solving a deconvolution problem, then from this model to improve a spatio-temporal content model in the frame. Tracking in the next frame is formulated by calculating a confidence map as a convolutional problem that complements spatio-temporal content information. Then the best location of the object is estimated by the maximum confidence map. Ultimately, based on estimated confidence map, a scale matching method is proposed that ultimately provides an accurate and effective tracking. The proposed method uses content information for tracking and reaches robust and fast results. In this approach, a scale update scheme is presented to deal with the scale changes of the target object. The difference between this and similar methods is that the target is not lost to the tracker under large occlusions, or large occlusions will not cause the tracker to deviate or jump from the main target to something other than the target. The reason for that is using both temporal and spatial contents. In addition, the speed of movement or movement of the target or changes in scale and rotation in different light conditions will not prevent the tracker from deviating. Given the comparison of this tracker with similar trackers in various sequences and in various environments and its success compared to similar trackers, one can claim that the proposed tracker is significantly more efficient than similar trackers.

Based on the obtained results, tracking of offending vehicles on the roads, tracking of target people in the streets, tracking enemy planes by smart missiles in electronic warfare, using in the vision system of drones and the vision system of robots are suggested.

References

[1] Wen.L, Cai. Z, Lei. Z, Yi.D, and Li.S, 2012, "online spatio-temporal structure context learning for visual tracking," in ECCV, pp. 716–729.
[2] Yang. M, Yuan. J, and Wu.Y, 2007, "Spatial selection for attentional visual tracking," in CVPR, pp. 1–8.
[3] Zhang. K, Zhang.L, Hsuan Yang.M, Liu. Q, and Zhang. D, 2014, " Fast Tracking via Spatio-Temporal Context Learning, "in ECCV, pp.127-141.
[4] Bolme. D. S, Draper. B. A, and Beveridge.J. R, 2009, "Average of synthetic exact filters," in CVPR, pp. 2105–2112.
[5] Bolme. D. S, Beveridge. J. R, Draper. B. A, and Lui.Y. M, 2010, "Visual object tracking using adaptive correlation filters, in CVPR, pp. 2544–2550.
[6] https://sites.google.com/site/trackerbenchmark/benchmarks/v10
[7] Henriques. J, Caseiro. R, Martins.P, and Batista. J, 2012, "Exploiting the circulant structure of tracking-by-detection with kernels," in ECCV, pp. 702–715.