Journal of Harbin Engineering University
ISSN: 1006-7043
Vol 44 No. 7
July 2023

Robust Composite Tracker of Objects in Video Images with Several Features and Reliable Design using Hierarchical Convolution Features

Shadi Shanesazzadeh[1], Karim Mohammadi[1]
[1] School of Electrical Engineering, Iran University of Science and Technology, Tehran, Iran,

Abstract- Backdrop modeling is crucial in machine vision and image processing, such as automatic video surveillance and human-machine interaction (HCI). It involves modeling the background, eliminating pictures from the background image, and removing shadows from the final image. Moving objects leave a shadow behind them in the picture, which is represented as a backdrop image due to ambient noise and variations in brightness. Shadow motion detection is essential for recognizing objects in video streams, as shadow points are often misclassified as object points, resulting in segmentation and tracking problems. Many techniques for modeling shadows have been developed, but algorithms continue to be tested under various illumination settings. Block-based models have lately been adopted due to these circumstances, splitting the picture into equal blocks and determining movement by a collection of blocks in the most recent frames. This work presents a novel model based on PCA + LDA and hierarchical convolution features to overcome the difficulties in block-based techniques. The proposed technique considers the distinction between fixed and moving items based on the type of peripheral object.

Keywords- Hierarchical Convolution Features, Object Tracker, Video Images, PCA, LDA.

Introduction

Detecting a moving object in sequential images from research areas is challenging [1, 2]. The use of visual and image processing systems to detect and track objects in different natural environments is highly complex. The overall goal is to identify motion in the image in three different ways. In the first praxis, low-level applications, in the second praxis, intermediate processing, and in the third praxis, high-level processing is performed [3, 4]. From the first praxis, the identification of scene components and segmentation, from the second praxis, object identification, and tracking, and from the third praxis, string analysis of images and scene analysis can be enumerated, which is the most important part of motion identification. The input video is converted to a series of images (or consecutive frames). These images provide information about the scene in sequence. In general, the scene is divided into two parts: fixed and dynamic or variable. The fixed part of the scene consists of the parts that do not move, the background, and the variable part of the scene includes the moving parts, the objects. Conventional motion detection models are based on subtracting the image from the background, which requires constant updating of the background model. In this regard, methods such as light flux [4] and temporal subtraction [1, 5] are common approaches to background subtraction. But the optical flux model has a lot of computational load and can not be used in practical conditions. In contrast, frame subtraction (FDM) models [6] are easy to implement and can be quickly calculated. However, the weakness of these methods is that the motion detection in them is not done accurately and there is always noise in the image [6]. In the background subtraction (BSM) model, the values of different pixels change over time. Appropriate algorithms should take this change into account in modeling. Therefore, background modeling is very difficult despite issues such as changing the speed of objects, changing brightness, obstruction, and overlap. Recently, block-based models have been highly developed in modeling and motion estimation [7].

Among the standard tracking algorithms, the discriminative correlation filter (DCF) [5–7] framework-based tracking algorithm has significant benefits and has been quickly deployed and improved. Bolme et al. [8] employed the correlation filter framework, which used the minimal output sum of square error (MOSSE) method, to considerably enhance tracking speed. However, the MOSSE tracker's tracking precision was insufficient to fulfill the real demand. Henriques proposed the circulant structure of tracking-by-detection with kernels (CSK) algorithm [9, 10], which used the diagonalization of the circulant matrix in the calculation process to simplify the calculation of nuclear regression, so the target tracking speed was greatly improved, as was the tracking accuracy.

However, when the target scale grows larger, the convolution computation for extracting target features and training filters grows larger, resulting in a drop in

target tracking speed. The kernel correlation filter method (KCF) [11] was an enhancement of the CSK technique that tracked the target using the histogram of oriented gradients (HOGs) and enhanced tracking accuracy. The HOG characteristics were extracted to recognize the item and improve tracking accuracy. Galoogahi et al. [12] proposed a backdrop-aware correlation filter (BACF) based on HOG features that efficiently controls an object's diversity in both foreground and background.

Liu et al. [13] investigated a patch-based tracking approach based on multi-CF models. Noise effects might be successfully adjusted by combining numerous sections. Danelljan et al. published a DSST technique based on MOSSE in 2014 that employed HOG characteristics to generate a scale pyramid for target scale estimate [14]. However, when the goal size increased, so did the convolution computation in training, resulting in a drop in tracking speed.

The Kalman filter technique was employed in [15, 16] to anticipate the state of the target, decide if the target was occluded, and mark the target that was still occluded afterwards. Due to the dynamic tracking environment, the target may have varied deformations, severe occlusion, and other difficulties throughout the long-term target tracking process, which may cause tracking failure. The ability to swiftly resume the target tracking capability is critical for long-term target monitoring.

Zhang et al. [17] defined the descriptors for rotation and scale normalization and merged color and texture characteristics to conduct optimal similarity matching on the descriptors in the front and rear frames candidates for target tracking. Yuan et al. [18] created a target-focusing convolutional regression model for the visual object tracking task to place more emphasis on the target sample than on the background samples. The target-focusing loss function may effectively balance the percentage of positive and negative data and avoid the appearance model from overfitting to the background samples.

Ma et al. [19] used an online random fern classifier to redetect objects in the event of a tracking failure. To address the drawbacks of using a single feature to represent the target, certain tracking approaches based on multiple feature fusion were developed [20–23], which might increase the algorithm's resilience to some extent.

Deep learning has outperformed traditional image processing approaches in terms of accuracy in face identification and image detection as it has expanded fast in the field of machine vision in recent years. Deep learning, unlike previous approaches, does not require manual feature creation and instead models the human visual perception system and abstract expressions using the characteristics of the original image. By pretraining deep CNNs, extracting the last three layers of convolutional features, and learning adaptive correlation filters, Ma et al. [24] enhanced tracking accuracy and resilience.

Qi et al. [25] concentrated on a hierarchical CNN-based tracking framework (HDT) that made full use of multiple characteristics and employed an adaptive Hedge method to hedge these trackers into a stronger one. Valmadre et al. [26] encoded the DCF learner as an end-to-end differentiable CNN layer and tracking target. Despite their effectiveness, all of these approaches are either hampered by a higher computing cost or yield inadequate tracking performance.

To tackle the challenges in block-based approaches, a novel model based on PCA + LDA and hierarchical convolution features is used in this study. Based on the kind of peripheral object, the proposed approach distinguishes between stationary and moving things. Instead of simply removing pixels, the picture of each frame is separated into equal blocks and differentiated between the blocks in this work. Of course, edge information and certain moving pieces may be eliminated during this subtraction. Segmentation information can also be utilized to overcome this problem and completely separate moving portions. In this work, after presenting several object detection methods such as PCA and LDA, a combined PCA + LDA approach is provided to identify the motion of a moving object in pictures, and the results of applying this algorithm to one of the legitimate datasets are compared to prior methods.

Hierarchical Convolutional Features

To remove the drawbacks of thinness and slowness, the suggested approach combines both as regulators in a single objective function termed LDA. The suggested learning algorithm's input data is a sequence of cubes with a time dependency, and each cube in this sequence is indicated by t. The model is trained to learn the integration properties of $p^{(t)}$ from $x^{(t)}$ data by tackling the issue of minimizing the following finite cost function:

$$\min_{W} \sum_{t=1}^{N-1} \left\| p^{(t)} - p^{(t+1)} \right\|_1 + \gamma \sum_{t=1}^{N} \left\| p^{(t)} \right\|_1 \quad (1)$$
$$Subject\ to\ WW^T = I$$

The properties $p^{(t)}$ are obtained by converting the following data from $x^{(t)}$:

$$p^{(t)} = \sqrt{H(Wx^{(t)})^2} \quad (2)$$

Where $W \in \mathbb{R}^{k\times n}$ are the weights that connect the input data to the simple neurons, and $H \in \mathbb{R}^{m\times k}$ are the weights that connect the simple neurons to the neurons. The merging cells (usually H is fixed) n, k, and m are the input dimensions, the number of simple neurons, and the number of merging neurons, respectively. The limit of normal orthogonality guarantees a variety of features and prevents feature reproduction. The structure of this model is shown in Figure 1. In this cost function, minimizing the second expression increases the thinness of the learned features. An ISA network can be described by a two-layer network, consisting of nonlinear quadratic and quadratic root functions in the first and second layers, respectively. W weights are learned in the first layer and H weights are constant in the second layer and define the subspace structure of the first layer neurons. In fact, each hidden layer neuron in the second layer is obtained by merging neurons in a small neighborhood. Interior search algorithm (ISA) and Slow Feature Analysis (SFA) algorithms have normal orthogonal constraints to prevent the reproduction of features and the generation of various features. Applying this limitation causes disadvantages such as preventing learning of more than complete representations as well as slowing down the optimization process, especially for large-sized data such as video sequences. By applying this condition to the proposed cost function (1), we will have the following unlimited cost function, which can be minimized by any general and unlimited optimization algorithm:

$$\min_W \sum_{t=1}^{N} \left\| x^{(t)} - W^T W x^{(t)} \right\|_2^2 + \lambda \sum_{t=1}^{N-1} \left\| p^{(t)} - p^{(t+1)} \right\|_1 + \gamma \sum_{t=1}^{N} \left\| p^{(t)} \right\|_1 \quad (3)$$

Where the first expression is the reconstruction condition that encodes the data x^t using the vector matrix multiplication as $z^{(t)}= Wx^{(t)}$ and then the result using another forward pass reconstructs $x^{(t)} =W^Tz^{(t)}$. This term can also be interpreted as an automatic encoder reconstruction cost. In this cost function, λ and γ are add-on parameters that specify the importance of being slow and thin, respectively. In order to develop the LDA algorithm into high-dimensional inputs, convolutional neural network architecture has been used. In this model, PCA and LDA have been used as subunits of the observer learning algorithm. This architecture, called DL-LDA, is shown in Figure 1.

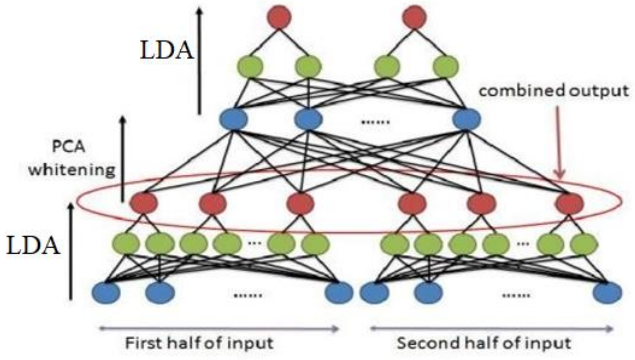

Figure 1. Proposed DL-LDA architecture

The instruction of this proposed model is as follows: First, we teach the LDA algorithm on small input components. We then convolution the trained network with a larger area of input. The combination of convolution phase responses is then used as input for next-layer training, with the next layer, another LDA algorithm, using PCA as its preprocessing step to clear and reduce data dimensions.

Blocking for motion simulation

Data platforms that are large in size, along with the capabilities they have and the opportunities they provide, create many computational challenges. One of the problems with large data sets is that most of the time all

the properties of the data are not critical to finding the knowledge that lies in the information. Processing on all pixels of an image, while increasing the processing accuracy, is equally effective in reducing processing speed and time. For this reason, processing on a batch of selected pixels seems more logical. In such a way that the image is divided into equal parts and the processing is done on them in groups. To do this, methods are proposed that filter less valuable data and process it on higher value data. That is, they process the image by reducing the size of the data. In many areas, data reduction has been a significant issue. These methods transform a multidimensional space into a space with smaller dimensions. In fact, by combining the values of existing attributes, they create fewer attributes so that these attributes have all (or a large part) of the information contained in the original attributes.

The problem of reducing the dimensions of the data can be expressed mathematically as follows: We have a p-dimensional random variable $X = (x_1,..., x_k)^T$. The k-dimensional variable $S =(s_1,...., S_k)^T$ must be found in such a way that firstly it is $k \leq p$, and secondly S has the contents of X according to a specific criterion. Linear methods try to obtain each of these k components from the linear composition of the original component p.

(4) $S_i=w_{i,1} + \dots w_{i,p} x_{p'}$ for $i=1 , \dots , k$, or $S=W_x$,

Where $W_{k \times p}$ is a matrix of linear mapping weights.

Tables, Figures, Equations

Figure 2 summarizes the PCA + DL-LDA combination recognition system:

(5) $Y=\varphi T x$

(6) $Z= Wy^T x$

In the first expression, Φ is the PCA transfer function that maps the x image to the space below the y-faces. Wy is the best DL-LDA transfer on the PCA subspace, which transfers the PCA coefficients to the LDA subspace. Wy transfer is the linear recording of a combination of image space to the final LDA space.

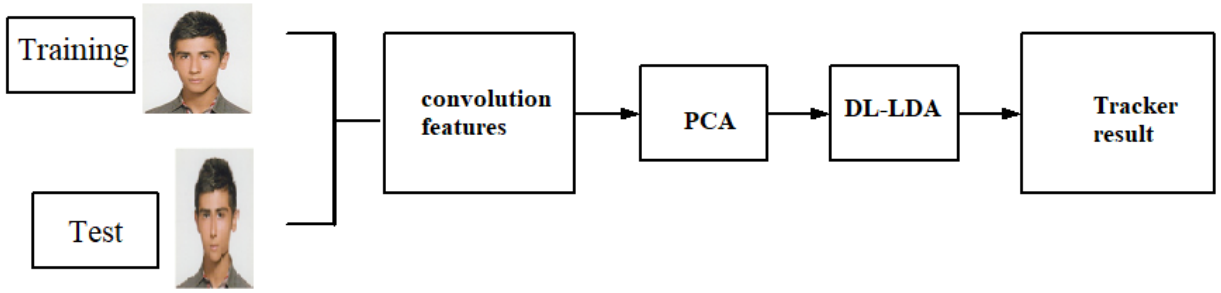

Figure 2. The proposed tracker system based on convolution features a PCA and DL-LDA

Therefore, the PCA + DL-LDA hybrid model has been used to evaluate the motion or tracking of objects in the proposed approach. Because the LDA method distinguishes between still and moving images based on the type of peripheral objects. Therefore, it is expected that in the proposed model, with the help of the same approach, the fixed parts will be extracted from the moving part properly. First, different features of the image for motion extraction will be introduced and finally, the results of the algorithm will be presented.

Color space models have been used to produce photometric still images and play a very important role in motion detection. An overview of some color space models is described in Table I. In Table I, C1, C2, and C3 represent the first, second, and third colors, respectively. Also, R represents red, G represents green, and B represents blue. In addition, the other symbols are defined as follows: H color, S color saturation, V value, I intensity, Y value, I color phase, Q squaring, and Cb and Cr pigments. Among the color models, in this article, we use the standard RGB color space and the Y-based gray model. Because RGB color space is very common and is used in many common applications to detect motion. Meanwhile, PCA and LDA processing are performed entirely on the gray color space and the images are evaluated with their gray values. The Y calculation relation is the same as the YIQ relation in the fourth row of Table I. After calculating the gray values of the pixels, their mean values in the blocks are calculated and then subtracted from their mean. The PCA algorithm is then applied to it and the LDA algorithm is applied to its output. Also, the transfer matrix is calculated in the initial frames, and in the subsequent frames, new data is obtained by multiplying the gray data in the transfer matrix. Finally, by calculating the Euclidean distance between the new and original data, the degree of similarity of the data to the background and the fixed part is obtained.

Table I. Overview of color space

Model	Convert from RGB space
HSI	If G≥B; H = $\cos^{-1}$[(R – 0.5G – 0.5B)] If B>G; H = 360 – $\cos^{-1}$[(R – 0.5G – 0.5B)] S = 1 – (min(R,G,B)/((R + G + B)/3)) I = (R + G + B) / 3
HSV	If G≥B; H = cos-1[(R – 0.5G – 0.5B)] If B>G; H = 360 – cos-1[(R – 0.5G – 0.5B)] S = 1 – (min(R,G,B)/ max(R,G,B)) V = max(R,G,B)/255
$C_1C_2C_3$	C1 = tan-1 (R/max(R,G,B)) C2 = tan-1 (G/max(R,G,B)) C3 = tan-1 (B/max(R,G,B))
YIQ	Y = 0.299*R + 0.587*G + 0.114*B I = 0.596*R – 0.275*G – 0.321*B Q = 0.212*R – 0.523*G + 0.311*B
YC_bC_r	Y = 0.257*R + 0.504*G + 0.098*B + 16 I = -0.148*R – 0.291*G + 0.439*B + 128 Q = 0.439*R – 0.368*G - 0.071*B + 128

In this paper, the square model is used to find movable objects in a robust manner. The dimensions of the window and the partial blocks of the image are selected in such a way as to cover the dimensions of the moving objects. For example, as shown in Figure 3, the motion is limited to a rectangular frame, and only elements outside the rectangular box are used in the previous frame. In this example, a simple rectangular model is used to separate the moving target from the background. Here, the background includes all the elements of the image except the moving object. For this purpose, all fixed objects are considered part of the background. With the useful extraction of a fixed model from the scene, motion information can be extracted more accurately. This section demonstrates the background modeling method by principal component analysis (PCA). At the beginning of the modeling process, the PCA model is calculated using all existing frames. When multiple frames are available, only the most recent ones are used. PCA model updates for each new frame are also performed in the modeling process. Summarizing these cases, the main steps of the PCA algorithm are summarized in Table II.

Table II. Steps of background modeling and motion extraction by PCA algorithm

Input: n video frames $f_{1,\dots,}f_n$ and the target bounding box (rectangle) in the first frame r_1

Output: the estimated bounding box $r_2, \dots, r_n$
For $i=2$ to k^{frm} do
  Compute the PCA background color model using frame $f_1, \dots, f_{i-1}$

  Compute the background confidence map using frame f_i, f_{i-1}, and the PCA model

  Estimate r_i according to the confidence map and motion model

End for
For $i=k^{frm}+1$ to n do
  Compute the PCA background color model using frame $f_{i-k^{frm}}, \dots, f_{i-1}$

  Compute the background confidence map using frame f_i, f_{i-1}, and the PCA model

  Estimate r_i according to the confidence map and motion model

End for

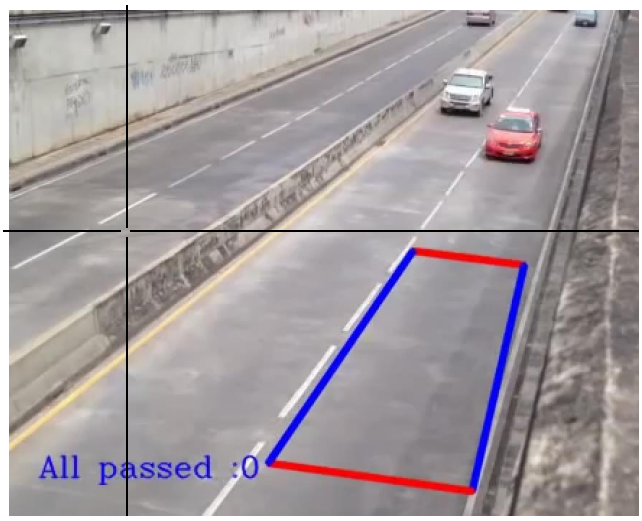

Figure 3. A new frame that uses only elements inside the box range

By applying the PCA algorithm in background modeling, motion information can be extracted properly. Figure 4 is a block diagram of the steps in Table 2.

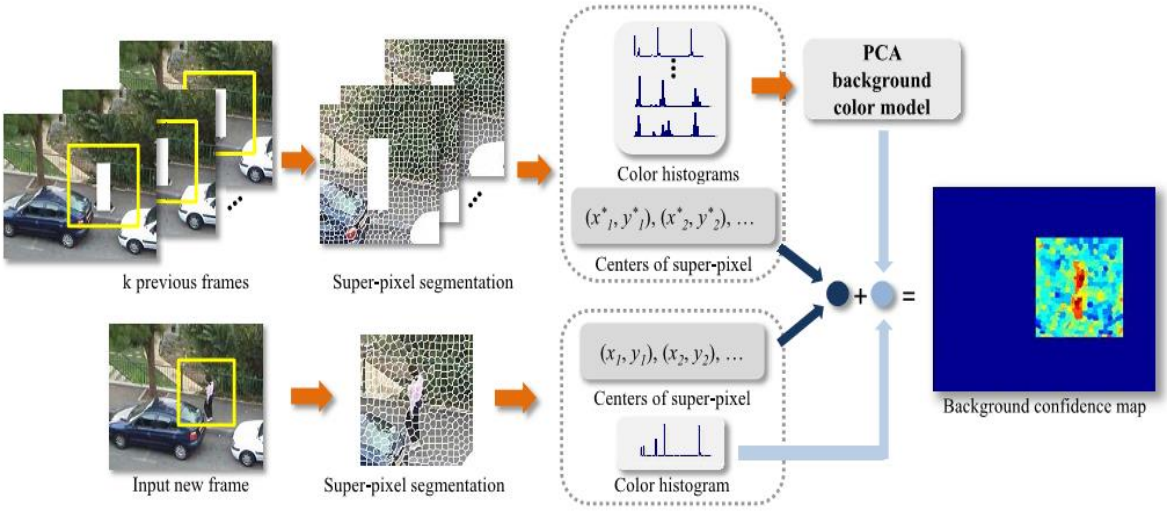

Figure 4. Block diagram of motion extraction model

Finally, by applying the algorithm (Figure 4), the results of Figure 5 are obtained.

Figure 5. An example of motion detection results

Obviously, moving objects (rectangular in shape) are properly extracted. This can include modeling despite extreme optical changes as well as shape changes in the environment.

Results and discussion

In this paper, considering that the simulation database is a monitoring image of the outdoor environment, it is expected to have a good performance. Therefore, in this article, the combination of PCA + LDA is used. Figure 6 shows the motion detection flowchart in this paper.

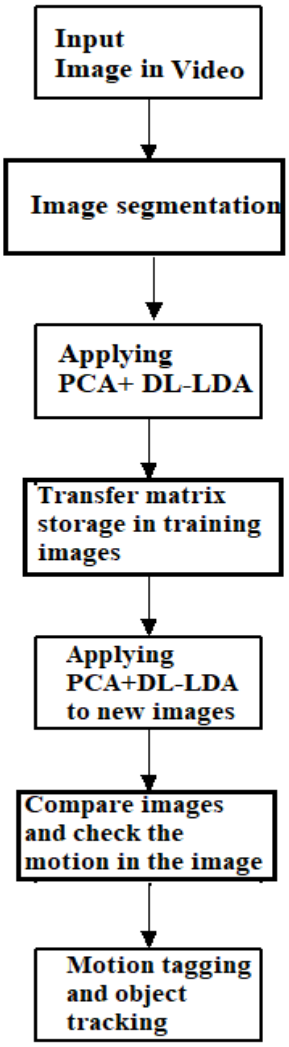

Figure 6. The main flowchart of this article combines local information

To remove the background information and extract the motion information using the block-based models, a window with dimensions of 32*32 is established in the picture, and PCA + DL-LDA is computed for each block. A total of 12 frames are required to calculate motion information. The motion in the image is estimated by using a similarity value of 600 between the new pixels and the background pixels as a threshold. The window size in this article was set to account for the various dimensions of items in the environment.

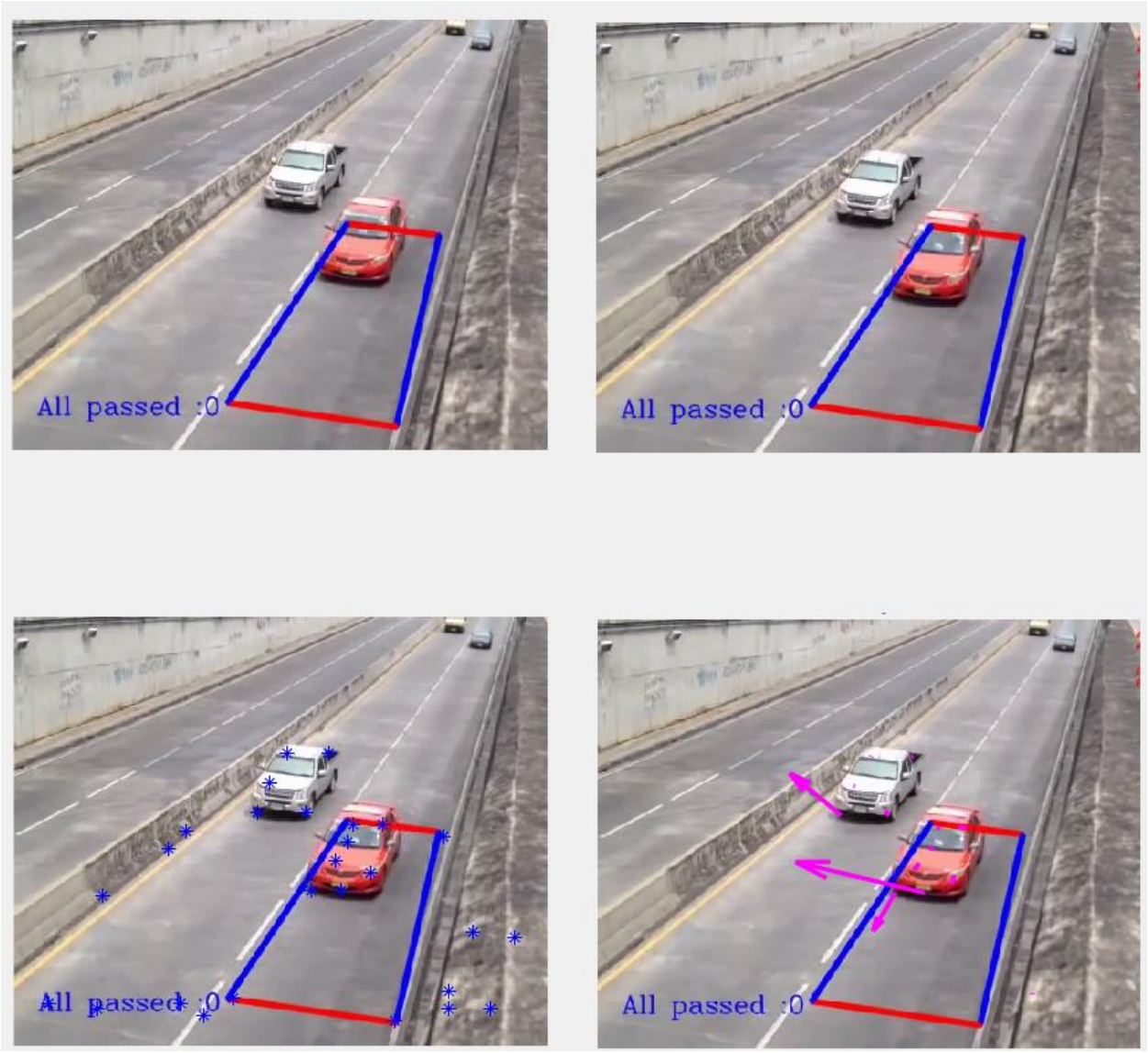

Figure 7. Several sample frames of the test video

The scenario begins with 45 automobiles entering near the frame, followed by 80 traffic jams moving around the frame, and then 140 people entering near the frame. Figure 8 depicts the outcomes of using the proposed technique in Figure 6. In this figure, the simulator places a bounding box around objects in the main frames, which is moved by the movement of objects.

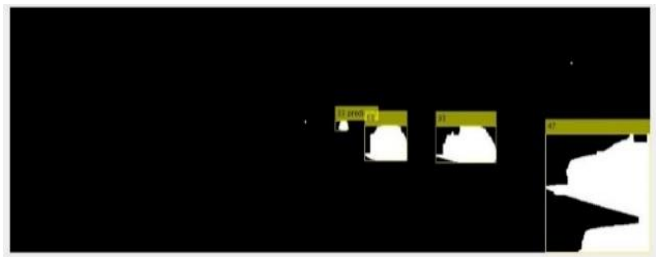

Figure 8. Results of the proposed algorithm generated by block-based PCA + DL-LDA method

Qualitative comparison of the proposed approach

In this paper, a new model is presented which is based on image blocks and PCA + DL-LDA. This method is used in background modeling and motion detection. One of the important applications of this approach is proper accuracy in motion detection and the use of advanced models. The results of this article can also be used to identify targets, track objects, and analyze scenes. To evaluate the behavior of the 3 algorithms, according to the qualitative issues presented in the previous section, they are given the range of "low" to "high" (Table III). This table compares a selected method of pixel-based algorithms and one of the block-based methods. The results show that PCA + DL-LDA method, using the proposed algorithms and after processing, is more resistant to noise.

Table III. Qualitative evaluation. 6 parameters have been selected to compare the proposed algorithms. For each parameter, a vote from "low" to "high" is considered

Detection of half shadow and indirect shadow	**Computational load**	**Scene independence**	**Independence of the object**	**Flexibility to shade**	**Noise resistance**	**Approaches**
low	High	low	low	low	low	FDM
medium	Low	medium	medium	medium	medium	PCA
High	medium	High	High	High	High	PCA+DL-LDA

The capacity to deal with shadow size and power in the pixel-based approach (FDM) is weak and the capacity of the PCA block-based approach in this area is average. However, higher flexibility is obtained by the PCA+DL-LDA algorithm which is able to detect even half of the shadows in an effective method. Background hypotheses make the PCA, LDA, and specifically PCA + DL-LDA approaches more scene-dependent than the other two algorithms. Although we can not claim that these algorithms are implemented in a more efficient way, it seems that PCA + DL-LDA takes more time than PCA and LDA algorithms due to the amount of processing required, but it is faster than the FDM algorithm.

The capacity to deal with shadow size and power in the pixel-based approach (FDM) is weak and the capacity of the PCA block-based approach in this area is average. However, higher flexibility is obtained by the PCA + DL-LDA algorithm which is able to detect even half shadows in an effective method. Background hypotheses make the PCA, LDA, and specifically PCA + DL-LDA approaches more scene-dependent than the other two algorithms. Although we can not claim that these algorithms are implemented in a more efficient way, it seems that PCA + DL-LDA takes more time than PCA and LDA algorithms due to the amount of processing required, but it is faster than the FDM algorithm.

Quantitative comparison of results

This section presents the methodology used to compare the two approaches. In order to systematically evaluate the detection of different shadows and to identify, two scales are important in measuring quality: good detection (low probability of incorrectly classifying a shadow spot) and good differentiation (probability of classifying non-shadow spots as shadow should be low, In other words, the false alarm ratio is low). Here, two criteria for evaluating object motion detection are proposed: error detection rate (FRR) and false alarm rate (FAR). Assuming FD is the number of incorrectly detected pixels (in other words, shadow points that are not correctly identified) and FM is the number of incorrect pixels of moving objects that are lost, these two criteria are defined as follows:

$$) 7 \qquad (FRR\% = FM/(FM + TD)\times 100$$
$$FAR\% = FD/(FD + TD) \times 100 \qquad (8)$$

Another evaluation criterion used in this report is the accuracy of recognizing objects as a percentage, known as Accuracy. The relationship of this criterion is defined as follows:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (9)$$

In this respect, TP (True Positive) refers to the number of correctly detected and correctly classified objects. TN (True Negative) refers to the number of correctly detected background or non-object regions. FP (False Positive) refers to the number of objects that were incorrectly detected or mistakenly classified as a specific class., and FN (False Negative) refers to the

number of objects that were not detected or mistakenly classified as background or another class. Another criterion used is the recall rate, the relationship of which is as follows:

$$\text{Recall} = \frac{TP}{TP+FN} \quad (10)$$

Table IV. Quantitative comparison table

%	PCA	[27]	PCA+DL-LDA
FRR	10.2	-	7.5
FAR	0.2	-	0.05
Accuracy	0.67	0.986	0.988
Recall	0.69	0.97	0.98

Table IV shows a small comparison between PCA and PCA + DL-LDA methods according to the above two criteria. As can be seen, the proposed approach is more efficient in all features.

Conclusions

In this research, the suggested model in motion detection utilizing block-based models was studied. A novel technique based on picture blocks and PCA + LDA was also developed. Hierarchical convolutional features were suggested as DL-LDA in this work. Important uses of this method include appropriate precision in motion detection and target identification, object tracking, and scene analysis. A novel approach for tracking a moving target may be described by replacing similarity with a moving area extraction criteria that applies to both static and dynamic backdrop models since the number of classes in the scene can be specified in the LDA model. In existing approaches, the interaction between the number of items in the scene and the backdrop model is restricted. However, putting the two together into a single model may fix the difficulties. According to the acquired findings, motion detection and removal of fixed pieces are done more precisely. The findings also suggest that the PCA + DL-LDA technique, employing the provided methods and after processing, is more resistant to noise.

Acknowledgment

I would like to express my sincerest gratitude to all those who have contributed to the successful completion of this paper.

First and foremost, I would like to thank my supervisor, Dr. Mohammadi, for his invaluable guidance, support, and expertise throughout the entire process.

I would like to extend my thanks to the members of committee, for their time and effort in reviewing and evaluating this paper.

I cannot overlook the contributions of my family and friends who have offered their support, advice, and encouragement during this journey.

References

[1] A. Yilmaz, O. Javed and M. Shah, “Object tracking: A survey,” ACM Computer Survey, vol. 38, no. 4, p. 13, 2006.

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

[3] G. Deng and K. Guo, “Self-adaptive background modeling research based on change detection and area training,” IEEE Workshop on Electronics, Computer and Applications, pp. 59-62, 8-9 May 2014.

[4] N. Papenberg, A. Bkuhn and T. Brox, “Highly accurate optic flow computation with theoretically justified warping”, [J].International Journal of ComputerVision,” vol. 67, pp. 141-158, 2006.

[5] F. Li, C. Tian, W. Zuo, L. Zhang and M.-H. Yang, “Learning spatial-temporal regularized correlation filters for visual tracking,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4904–4913, Salt Lake City, UT, USA, June 2018.

[6] S. Li, Z. Qin and H. Song, “A temporal-spatial method for group detection, locating and tracking,” IEEE Access, vol. 4, pp. 4484–4494, 2016.

[7] S. Hare, S. Golodetz and Ph. S. H. Torr, “Struck: structured output tracking with kernels,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 38, no. 10, pp. 2096–2109, 2016.

[8] D. Bolme, J. Beveridge, B. Draper and Y. Lui, “Visual object tracking using adaptive correlation filters,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2544–2550, San Francisco, CA, USA, June 2010.

[9] J. Henriques, R. Caseiror, P. Martins and P. Jorge, “Exploiting the circulant structure of tracking-by-detection with kernels,” in Proceedings of the 12th European Conference on Computer Vision, pp. 702–715, Florence, Italy, October 2012.

[10] J. Henriques, J. Carreira, C. Rui and B. Jorge, “Beyond hard negative mining: efficient detector learning via block-circulant decomposition,” in Proceedings of the IEEE International Conference on Computer Vision, pp. 2760–2767, Sydney, Australia, April 2013.

[11] J. Henriques, R. Caseiro, P. Martins and J. Batista, "High-speed tracking with kernelized correlation filters," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, no. 3, pp. 583–596, 2015.

[12] H. Galoogahi, A. Fagg and S. Lucey, "Learning background-aware correlation filters for visual tracking," in Proceedings of the IEEE International Conference on Computer Vision, pp. 1369–1378, Venice, Italy, October 2017.

[13] T. Liu, G. Wang and Q. Yang, "Real-time part-based visual tracking via adaptive correlation filters," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4902–4912, Boston, MA, USA, June 2015.

[14] M. Danelljan, F. S. Khan and M. Felsberg, "Adaptive color attributes for real-time visual tracking," in Proceedings of the Computer Vision and Pattern Recognition (CVPR), pp. 1090–1097, IEEE, Columbus, OH, USA, June 2014.

[15] G. H. Zhao, S. Zhuo and X. L. Xu, "Multi-object tracking algorithm based on kalman filter," Computer Science, vol. 45, no. 8, pp. 253–257, 2018.

[16] Z. L. Zhang and Y. X. Wang, "SiamRPN target tracking method based on kalman filter," Intelligent Computer and Applications, vol. 10, no. 3, pp. 44–50, 2020.

[17] J. Zhang, H. M. Huang and J. M. Wang, "An improved TLD real-time target tracking algorithm based on CN algorithm," Computer Engineering & Science, vol. 42, no. 7, pp. 1215–1225, 2020.

[18] D. Yuan, N. Fan and Z. He, "Learning target-focusing convolutional regression model for visual object tracking," *Knowledge-Based System*, vol. 194, p. 105526, 2020.

[19] C. Ma, X. K. Yang, C. Y. Zhang and M. H. Yang, "Long-term correlation tracking," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, June 2015.

[20] T. Y. Xu, Research on Correlation Filter Based Visual Object Tracking. Jiangnan University, Wuxi, China, 2019.

[21] D. Yuan, X. Zhang, J. Liu and D. Li, "A multiple feature fused model for visual object tracking via correlation filters," Multimedia Tools and Applications, vol. 78, no. 19, pp. 27271–27290, 2019.

[22] Y. Li and J. Zhu, "A scale adaptive kernel correlation filter tracker with feature integration," in Proceedings of the European Conference On Computer Vision, pp. 254–265, Zurich, Switzerland, September 2014.

[23] P. Zhao, Y. N. Zhang, T. Yang, X. W. Zhang and Y. H. Yang, "A novel multi-object detection method in complex scene using synthetic aperture imaging," Pattern Recognition, vol. 45, no. 4, pp. 1637–1658, 2012.

[24] C. Ma, J.-B. Huang, X. Yang and M.-H. Yang, "Robust visual tracking via hierarchical convolutional features," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 11, pp. 2709–2723, 2019.

[25] Y. Qi, S. Zhang, L. Qin, H. Yao, Q. Huang, J. Lim and M.-H. Yang, "Hedged deep tracking," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4303–4311, Las Vegas, NV, USA, June 2016.

[26] J. Valmadre, L. Bertinetto, J. Henriques, A. Vedaldi and P. H. S. Torr, "Endto-end representation learning for correlation filter based tracking," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2085–2813, Honolulu, HI, USA, July 2017.

[27] S. Zhang, J. Wang, Z. Wang, Y. Gong and Y. Liu, "Multi-target tracking by learning local-to-global trajectory models," Pattern Recognit, vol. 48, no. 2, 2015.