International Journal of Research and Development in Applied Science and Engineering (IJRDASE)
ISSN: 2454-6844

Using Kalman Filtering to Enable Object Detection in Video Frames

Km. Deepu Yadav, Shyam Shankar Dwivedi
Computer science and Engineering Department,
Rameshwaram Institute of Technology & Management, Lucknow
kmdeepu621@gmail.com

Abstract: Traditional object detection methods rely on basic features and handcrafted architectures, often resulting in suboptimal performance. To address these limitations, advanced machine learning techniques have been increasingly adopted. This work presents an effective object detection model for video frames, leveraging machine learning to enhance accuracy and efficiency.

Initially, data collection from video frames is performed, followed by background subtraction to extract moving objects. A smart estimation and prediction approach is employed, integrating object features for optimal performance. The study demonstrates that the proposed method outperforms state-of-the-art models, achieving high accuracy and effectiveness in object detection.

Keywords: Object Detection, Kalman filter, Data Mining, AI.

1. Introduction:

Many applications utilize video for various purposes such as detecting pedestrians and recognizing unusual behavior in parking lots. Retrieving moving objects and automating video analysis are becoming increasingly common. Video, a form of multimedia data, combines different types of information like text, images, metadata, visual, and audio elements. The primary aim of video data analysis is to recognize and track moving objects, such as people, across frames.

Video Data Mining is used in diverse fields such as security, surveillance, entertainment, medical and legal applications, medical education, and sports. It involves finding and analyzing patterns in massive amounts of video data, which consist of a series of images. Video material can be categorized into two types: low-level feature data, such as color, texture, and form, and high-level feature details, such as audio and video. Syntactic information in video material includes conspicuous objects, their spatial-temporal locations, and their spatial-temporal connections. Semantic data explains the events in the video, including spatial features offered by a video frame and the movement of characters on the screen.

The temporal aspects of video characterize a succession of frames, capturing actions and movements in sequence. Detecting moving objects in video streams is the first step in extracting information about the objects. This is crucial in many computer vision applications, including video surveillance and people annotation. Figures 1(a) and 1(b) illustrate the overall use of emerging technology in object detection and summarize key research works in the field

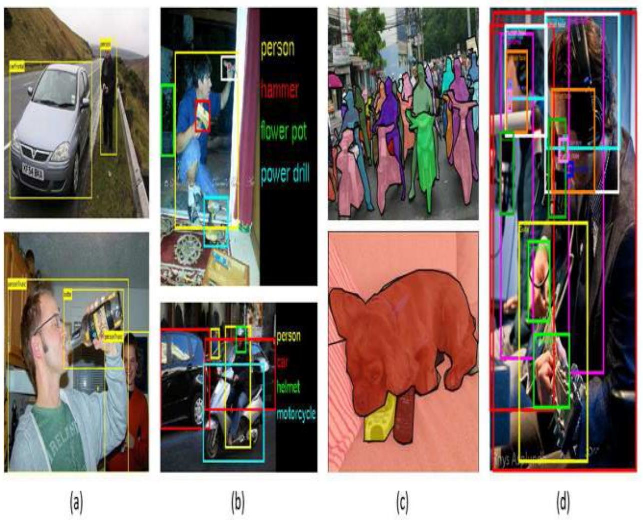

Fig. 1. Some example images and annotations

2. Related Work:

The journal by Schmidhuber (2015) elaborates on the workings of deep learning models and distinguishes between deep and shallow learners. Deep learning algorithms enable computational models with many layers to process requirements through multiple abstraction layers, effectively filtering necessary data. Deep learning models have state-of-the-art applications in object detection, voice processing, speech recognition, and various other scrutinizing applications (Lecun, Bengio, and Hinton, 2015). These models recognize complex structures in large datasets and learn their internal parameters, observed from the previous layer and used in the current layer, through the backpropagation method.

Convolutional neural networks (CNNs) have significantly advanced the processing of audio, video, speech, and images compared to recurrent neural networks. CNNs are iterative neural networks where convolution layers alternate with subsampling layers (Ciresan et al., 2003; Schmidhuber, 2015), resembling the structure of simple and complex cells in neurons. The training of neural nets and the identification of convolution and subsampling layers play crucial roles in this process. An optional image processing layer handles data preprocessing, and the convolutional layer is parameterized by the dimensions and the variety of maps, kernel sizes, skipping elements, and connection tables.

Convolutional networks, a subset of deep learning models, are robust for feature extraction and visual model understanding. Krizhevsky, Sutskever, and Hinton successfully utilized CNNs for grasp detection as a classifier in the detection pipeline using the sliding window concept. While this approach addresses a problem similar to the one at hand, it differs in processing pipeline and network architecture, leading to increased accuracy at greater speeds. Contemporary work on grasp detection mainly focuses on detecting grasps solely from RGB-D data (Saxena, Driemeyer, and Ng, 2008). These algorithms rely on machine learning techniques to identify good grasps from the data. Grasp visual models are well-known for single object views, not a complete physical model.

A core problem in computer vision is object detection (Nowozin and Lampert, 2010). Detection pipelines typically start with the extraction of robust features from input images (such as SIFT, HOG, or convolutional features). Classifiers or localizers are then applied in the feature space to detect objects. Subsequently, the sliding window concept is implemented either across the entire image or a part of it with the help of classifiers or localizers.

3. Methodology:

Artificial intelligence (AI) has revolutionized countless industries and aspects of daily life, permeating everything from virtual assistants to advanced robotics. At its core, AI involves developing computer systems capable of performing tasks that typically require human intelligence, such as learning, problem-solving, perception, and language understanding. Through machine learning algorithms, AI systems can analyze vast amounts of data to identify patterns, make predictions, and continuously improve their performance over time. Natural language processing enables AI to comprehend and generate human language, facilitating communication between machines and humans.

From personalized recommendations on streaming platforms to autonomous vehicles navigating complex environments, AI is reshaping how we work, live, and interact with technology. Its potential for further innovation and societal impact continues to grow. However, ethical considerations and concerns about the societal implications of AI deployment remain significant, prompting ongoing discussions about responsible development, transparency, and accountability in the use of artificial intelligence.

Proposed Work:

A suggested architecture for object detection is shown in Figure 2, where a video is taken as input and a single frame is selected for object detection. Initially, the video is broken into individual image frames. During the first stage of processing, the images are analyzed for their characteristics. The structure and form of the pixels provide enough information to identify significant elements in an image, so not every pixel is sent to the neural network. The feature extraction procedure recognizes the edges of the picture. Subsequently, the features undergo dimensionality reduction, where SE (Squeeze-and-Excitation) blocks help select the most relevant features. Finally, these features are passed to a classification stage to obtain the required results.

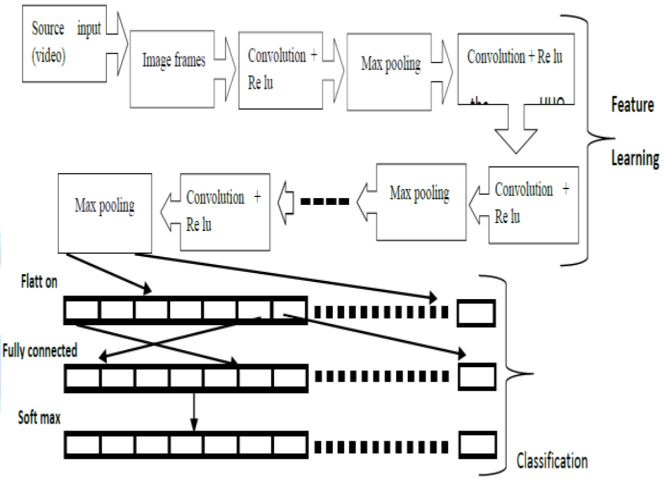

Fig 2: Flow of objective detection

4. Result and Discussion:

This work demonstrates the use of computer vision in object detection by configuring the Kalman Filter approach to track objects through image differencing. The results are discussed by outlining the main steps of the algorithm and applying it to a specific video of moving objects, highlighting the performance in terms of detected and actual tracks. The object detection algorithm has numerous applications, including control, navigation, computer vision, and time series econometrics. This section illustrates how to use the Kalman filter for tracking objects and focuses on three important features:

(a) Prediction of the object's future location.
(b) Reduction of noise introduced by inaccurate detections.
(c) Facilitating the process of associating multiple objects with their tracks.

The main challenges of object tracking arise from the movement of the background or the camera. To illustrate these challenges without the use of a Kalman filter, consider a video where a green ball moves from left to right on the floor. This example showcases the difficulties in tracking an object in a dynamic environment and underscores the importance of the Kalman filter in improving tracking accuracy.

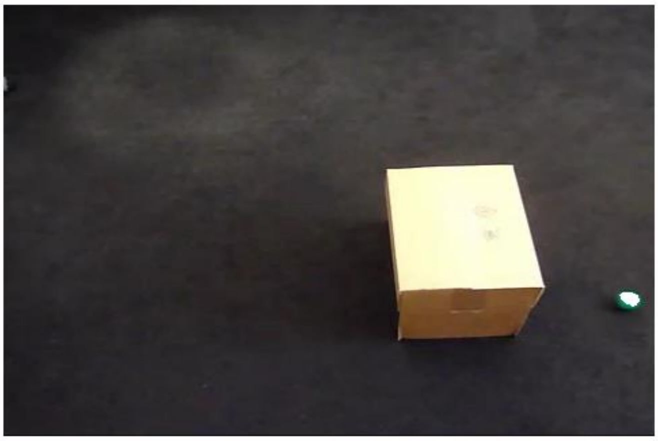

Fig 3: Detection of moving ball by simple image difference method.

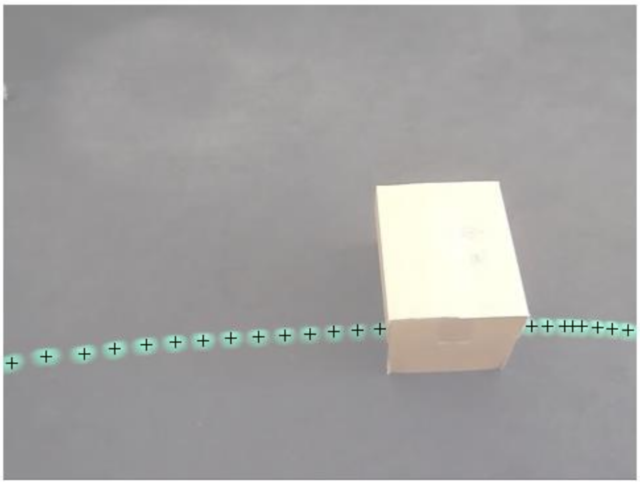

Fig 4: Overlaid view of image frames of video to show the detected object

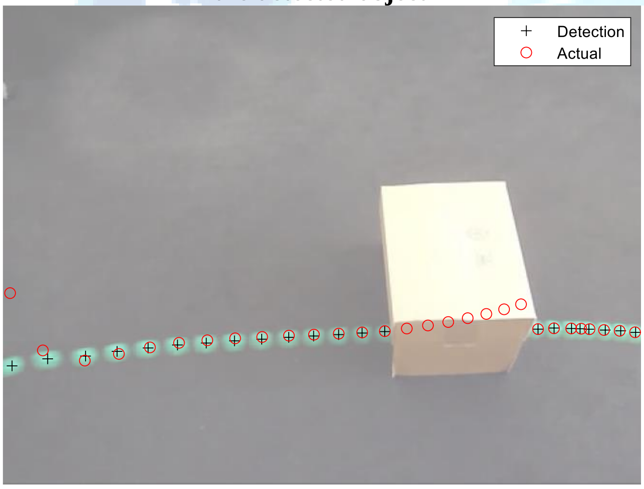

Fig 5: Object actual and detected location without kalman filter

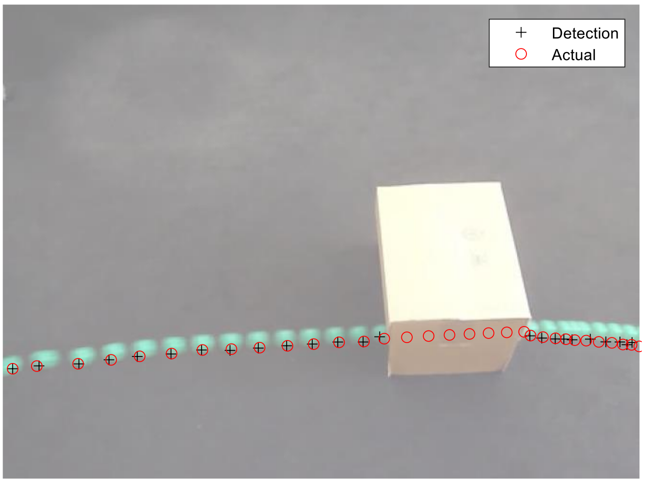

Fig 6: Object actual and detected location with kalman filter

5. Conclusion:

In this work, a visual surveillance system with moving object detection and tracking capabilities has been presented. Object tracking, both for single and multiple moving objects, has been successfully implemented on standard datasets using the Kalman filter with image differencing. The system works on videos captured in both indoor and outdoor environments using a static camera under moderate to complex background conditions. This implemented module can be applied to any computer vision application for moving object detection and tracking.

References:

[1] Yiwei Wang, John F. Doherty and Robert E. Van Dyck, "Moving Object Tracking in Video", in proceedings of 29th applied imagery pattern recognition workshop, ISBN 0-7695-0978-9, page 95,2000.

[2] Bhavana C. Bendale, Prof. Anil R. Karwankar, "Moving Object Tracking in Video Using MATLAB", International Journal of Electronics, Communication and Soft Computing Science and Engineering ISSN: 2277-9477, Volume 2, Issue 1.

[3] Marcus A. Brubaker, Leonid Sigal and David J. Fleet, "Video-Based People Tracking", hand book of ambient intelligence under smart environments 2010, pp 57-87.

[4] Emilio Maggio and Andrea Cavallaro, "Video Tracking: Theory and Practice", _rst edition 2011, John Wiley and Sons, Ltd.

[5] Y.Alper, J.Omar, and S.Mubarak. "Object Tracking: A Survey" ACM Computing Surveys, vol. 38, no. 4, Article 13, December 2006.

[6] B. Triggs, P.F. McLauchlan, R.I. Hartley and A.W. "Fitzgibbon. Bundle adjustment – a modern synthesis". In Proceedings of the International Conference on Computer Vision, London, UK, 1999, 298?372.

[7] G.C. Holst and T.S. Lomheim. "CMOS/CCD Sensors and Camera Systems". Bellingham, WA, SPIE Society of Photo-Optical Instrumentation Engineering, 2007.

[8] E. Maggio, M. Taj and A. Cavallaro. "E_cient multi-target visual tracking using random finite sets". IEEE Transactions on Circuits Systems and Video Technology, 18(8), 1016?1027,
2008.

[9] G. David Lowe. Object recognition from local scale-invariant features. Proceedings of the
International Conference on Computer Vision. 2. pp. 1150?1157,1997.

[10] G. David Lowe. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 60(2), pp, 91-110, 2004.

[11] Bruce D. Lucas and Takeo Kanade. An Iterative Image Registration Technique with an Application to Stereo Vision. International Joint Conference on Arti_cial Intelligence, pages 674-679, 1981.

[12] Carlo Tomasi and Takeo Kanade. Detection and Tracking of Point Features. Carnegie Mellon University Technical Report CMU-CS-91-132, April 1991.

[13] Jianbo Shi and Carlo Tomasi. Good Features to Track. IEEE Conference on Computer Vision and Pattern Recognition, pages 593-600, 1994.

[14] Stan Birch_eld. Derivation of Kanade-Lucas-Tomasi Tracking Equation. Unpublished, January 1997.

[15] Y. Cui, S. Samarasekera, Q. Huang. Indoor Monitoring Via the Collaboration Betweena Peripheral Senson and a

Foveal Sensor, IEEE Work-shop on Visual Surveillance, Bomba y, India, 2-9, 1998.

[16] G. R. Bradski, Computer Vision Face T racking as a Component of a Perceptual User Interface, IEEE Work. on Applic. Comp. Vis., Princeton, 214-219, 1998.

[17] S.S. Intille, J.W. Davis, A.F. Bobick, Real-Time Closed-World T racking. IEEE Conf. on Comp. Vis. and Pat. Rec., Puerto Rico, 697-703, 1997.

[18] C. Wren, A. Azarbayejani, T. Darrell, A. Pentland, P_nder: Real-Time Tracking of the Human Body, IEEE Trans. Pattern Analysis Machine Intell, 19:780-785, 1997.

[19] A. Eleftheriadis, A. Jacquin. Automatic Face Location Detection and Tracking for Model-Assisted Coding of Video Teleconference Sequences at Low Bit Rates, Signal Processing- Image Communication, 7(3): 231-248, 1995.

[20] D. Fuiorea, V. Gui, D. Pescaru, and C. Toma. Comparative study on RANSAC and Mean shift algorithm, International Symposium on Electronics and Telecommunications Edition 8.
vol. 53(67) Sept. 2008, pp. 80-85.