A Review on Enhancing Image Annotation with Object Tracking

Hina Sharma, Department of Computer Science& Engineering, Laxmi Devi Institute of Engineering and Technology, Chikani, Alwar, 301001, Rajasthan, India hinasharma427@gmail.com,
Pratap Singh Patwal, Department of Computer Science& Engineering, Laxmi Devi Institute of Engineering and Technology, Chikani, Alwar, 301001, Rajasthan, India pratappatwal@gmail.com

Abstract

Object tracking is a core component of computer vision systems, enabling automated and consistent annotation of image and video data for machine learning applications. By accurately following objects across video frames, tracking techniques significantly reduce manual labelling efforts and improve annotation reliability. This paper presents a systematic review of object tracking and image acquisition methods, highlighting their role in enhancing annotation efficiency and accuracy. The review covers traditional and deep learning–based tracking approaches and examines their effectiveness in domains such as medical imaging, surveillance, and urban monitoring. Despite recent advancements, challenges related to robustness, scalability, and human–AI collaboration persist. This study provides insights into the current capabilities of object tracking technologies and outlines their potential in advancing automated annotation systems.

Keywords- Image annotation; object tracking; Artificial Intelligence; deep learning

Introduction

Image annotation plays a crucial role in a wide range of computer vision and artificial intelligence applications [1, 2]. With the rapid growth of visual data, there is an increasing demand for efficient and scalable approaches to annotate large image and video datasets [3]. Traditional manual annotation methods are time-consuming and often inconsistent, highlighting the need for automated or semi-automated solutions. In this context, object tracking has emerged as a promising technique for improving the efficiency and accuracy of image annotation.

Recent advances in artificial intelligence, particularly in machine learning and deep learning, have significantly transformed the way visual data is analyzed and interpreted [4]. Machine learning algorithms form the backbone of automatic annotation systems by learning meaningful patterns from data [1, 5, 6]. Among these, deep learning models—especially deep neural networks— have demonstrated remarkable performance in extracting complex visual features, enabling tasks such as object detection, recognition, and image classification [7].

Convolutional Neural Networks (CNNs) remain a fundamental component of image processing systems [8]. Inspired by the human visual cortex, CNNs process images through hierarchical layers that capture visual information at multiple levels of abstraction [9]. This makes them highly effective for image annotation, as they can accurately identify and label objects within images. More recently, Visual Transformers (ViTs) have gained attention as a robust alternative to CNN-based approaches [10]. By leveraging attention mechanisms, ViTs are capable of modeling global relationships across an image, allowing for better understanding of complex visual scenes and improving annotation quality [11].

Object tracking extends these capabilities by continuously identifying and following objects across consecutive frames in video sequences or image streams [12]. By tracking object trajectories over time, annotations can be automatically propagated across frames, significantly reducing manual effort and improving consistency [13]. This is particularly beneficial in dynamic environments where objects move, change appearance, or interact with one another.

This review focuses on the role of object tracking in enhancing image annotation processes. It systematically examines recent research, methodologies, and real-world applications that integrate tracking techniques with modern deep learning models. Additionally, the review discusses existing challenges, including robustness in complex scenes and adaptability across diverse domains. By analyzing current developments and practical use cases, this study

ISSN: 2393-8048
International Advance Journal of Engineering, Science and Management (IAJESM)
Multidisciplinary, Multilingual, Indexed, Double Blind, Open Access, Peer-Reviewed, Refereed-International Journal.
SJIF Impact Factor = 8.35, January-June 2026, Submitted in June 2026

provides a comprehensive overview of how object tracking can advance automated image annotation and outlines future directions for research in this area.

Related Work

The field of image annotation using deep learning has evolved rapidly, leading to several systematic reviews that examine different facets of this growing research area. Recent studies provide valuable insights into the techniques, applications, and challenges associated with automated image annotation.

Adnan et al. [14] present a comprehensive review of Automatic Image Annotation (AIA) methods, with a strong focus on deep learning–based approaches. Their study categorizes AIA techniques into five main groups, including CNN-, RNN-, DNN-, LSTM-, and SAE-based models. While the review highlights significant progress in these methods, it also identifies ongoing challenges, particularly the need for more accurate and computationally efficient annotation techniques.

Ojha et al. [15] concentrate on the role of convolutional neural networks in image annotation. Their work explores how ConvNets extract meaningful visual features for complex computer vision tasks, emphasizing their effectiveness in object detection and localization. The review underlines the importance of CNNs in addressing visual perception challenges within image annotation systems.

Pande et al. [16] offer a comparative study of image annotation tools designed for object detection. Their analysis evaluates these tools based on functionality, performance, and suitability for different application scenarios. The authors highlight that the choice of annotation tool plays a critical role, as annotation quality directly influences the performance of object detection models.

Collectively, these reviews provide a broad overview of existing methods, applications, and challenges in deep learning–based image annotation. However, they tend to examine individual techniques in isolation, which limits insight into how multiple approaches can be combined to address complex annotation challenges more effectively.

In contrast, our review adopts a more integrated perspective by jointly examining object tracking and image retrieval as complementary techniques for image annotation. By considering the combined strengths of these approaches, this work presents a more holistic view of the annotation process. Such integration has the potential to significantly improve both efficiency and accuracy, opening new directions for advancing automated image annotation systems.

Literature Review Methodology

A. Eligibility Criteria

To ensure a focused and unbiased selection of studies, clear eligibility criteria were defined prior to the review process. These criteria were designed to include only research that directly addresses the objectives of this study, while excluding work that falls outside the scope of object tracking for image and video annotation. By establishing these parameters in advance, the review concentrates on the most relevant literature and maintains alignment with the research questions. Applying the criteria before conducting the literature search also helps minimize selection bias and improves the objectivity and reliability of the review.

B. Identification Phase

The literature search was conducted using three widely recognized academic databases: IEEE Xplore, Scopus, and SpringerLink. The search focused on object tracking–related studies by applying combinations of keywords within article titles, abstracts, and author-defined keywords. The primary search query was:

("Object Tracking") AND ("Image" OR "Dataset" OR "Video") AND ("Annotation")

The search was performed and was limited to publications released between 2020 and 2025. This time frame was selected to capture recent advances in deep learning–based object tracking

ISSN: 2393-8048
International Advance Journal of Engineering, Science and Management (IAJESM)
Multidisciplinary, Multilingual, Indexed, Double Blind, Open Access, Peer-Reviewed, Refereed-International Journal.
SJIF Impact Factor = 8.35, January-June 2026, Submitted in June 2026

methods relevant to automated annotation of image and video datasets. The initial search returned 5,455 documents, which were subsequently considered for screening.

In addition to the database search, a small number of highly relevant studies were identified through reference analysis and expert judgment. Although these works did not strictly match the predefined search terms, they provided valuable insights into object tracking techniques for annotation. Such studies were included selectively to strengthen the depth and completeness of the review, based on their clear relevance to the research objectives.

Literature on Single Object Tracking

Tao Yu et al. [17] propose an efficient method for polyp detection and tracking in colonoscopy videos by integrating an Instance Tracking Head (ITH) into the Scaled-YOLOv4 framework. By sharing low-level features between detection and tracking tasks, the approach achieves approximately 30% faster performance while maintaining high accuracy, with a detection mAP of 91.70% and a MOTA of 92.50%.

Xiong et al. [18] present a Siamese network–based tracking approach using SiamFC++, where object tracking is formulated as a similarity learning problem. Combined with the Box2Segmentation module, the method enables accurate video object segmentation from an initial bounding box, forming the basis of the Td-VOS framework.

Hu et al. [19] introduce SiamMask, a real-time tracking and segmentation method that operates at 55 fps. Using offline-trained Siamese networks with integrated similarity, bounding box regression, and segmentation branches, SiamMask delivers efficient and accurate online tracking and is adaptable to multiple objects.

Schörkhuber et al. [20] address the challenge of annotating night-time driving videos through a semi-automatic tracking-based approach. Tested on the CVL dataset, their method improves recall by 23% while maintaining consistent precision, demonstrating its effectiveness in low-light and rural scenarios.

Henschel et al. [21] propose a hybrid tracking framework that combines video data with body-mounted IMUs to overcome appearance ambiguities. By jointly modeling visual and inertial information, the method achieves an IDF1 score of 91.2%, highlighting its robustness in challenging multi-person tracking environments.

Literature on Multiple Object Tracking and Segmentation

Xu et al. [22] propose PointTrackV2, a novel framework that represents images as unordered 2D point clouds to effectively separate foreground and background instances. By enriching point features with multiple data modalities, the method achieves strong tracking and segmentation performance at near real-time speed (20 FPS on a single GPU). The authors also introduce the APOLLOMOTS dataset, which is more challenging than existing benchmarks, and demonstrate the versatility of PointTrackV2 across tasks such as video segmentation, pose estimation, and fine-grained image classification.

Yan et al. [23] introduce STC-Seg, a weakly supervised video instance segmentation framework that relies on box-level annotations. The method uses unsupervised depth estimation and optical flow to generate reliable pseudo-labels, supported by a novel *puzzle loss* for end-to-end training. An advanced tracking module improves robustness to appearance changes. STC-Seg outperforms fully supervised methods on KITTI MOTS and YT-VIS, highlighting the potential of weak supervision in video segmentation.

Several studies focus on improving annotation efficiency. Le et al. [24] present an interactive, self-supervised annotation framework combining Automatic Recurrent Annotation (ARA) and Interactive Recurrent Annotation (IRA). By iteratively refining detectors using unlabeled videos and limited human corrections, the approach significantly reduces annotation time while maintaining high accuracy. Similarly, Sambaturu et al. [25] propose ScribbleNet, an interactive segmentation tool that uses sparse scribble inputs to rapidly refine annotations, achieving substantial time savings over manual and existing interactive methods.

Zhu et al. [26] apply multiple object tracking to reconstruct 3D vehicle bounding boxes for bridge load monitoring. Their system integrates deep learning–based vehicle detection, camera calibration, and 3D reconstruction to extract accurate spatial–temporal vehicle information, demonstrating reliable performance in real-world bridge environments.

Liu et al. [27] introduce ASIST, a fully automated method for instance segmentation and tracking in microscopy videos without manual annotations. Using adversarial simulations and pixel embedding learning, ASIST achieves superior or comparable performance to supervised approaches, offering an efficient solution for analyzing cellular dynamics.

Lateef et al. [28] propose an object identification framework for autonomous driving that relies solely on camera data. By combining optical flow, image registration, and deep segmentation networks, the system effectively detects and semantically classifies moving objects in complex urban scenes.

Finally, Chen et al. [29] present Siamese DETR, a self-supervised extension of the DETR framework that integrates Siamese learning with transformer-based cross-attention. The method achieves state-of-the-art transfer performance on standard detection benchmarks, improving localization accuracy and training efficiency, while highlighting future directions toward unified CNN–Transformer training.

Conclusion

This systematic review examined the role of object tracking in automating and supporting image and video annotation. Based on the analyzed studies, object tracking has emerged as a key technology for improving both the efficiency and accuracy of annotation processes across a wide range of application domains.

Automatic image annotation continues to advance rapidly due to recent developments in object tracking techniques, particularly those driven by deep learning. Object tracking enables the continuous identification and localization of objects across image sequences and videos, allowing annotations to be propagated over time with minimal human intervention. This capability significantly reduces manual annotation effort while improving consistency and reliability, which is critical in applications such as medical imaging, urban surveillance, autonomous systems, and large-scale dataset creation.

The progress of object tracking has been strongly supported by advances in deep neural networks and modern machine learning frameworks. These approaches enhance the robustness of tracking under challenging conditions, such as occlusions, appearance changes, and crowded scenes, thereby enabling accurate and continuous annotation of objects in dynamic environments.

A notable example is the work by Yu et al. [17], who introduced the Instance Tracking Head (ITH) integrated into the Scaled-YOLOv4 detector. This approach improves object detection and tracking in medical videos, such as colonoscopy footage, allowing polyps to be automatically and consistently annotated over time. Such automation reduces reliance on manual labeling, which is often labor-intensive and prone to errors.

Similarly, Hu et al. [19] proposed SiamMask, a real-time framework that combines object tracking with segmentation. Operating at 55 frames per second, SiamMask enables continuous and automated annotation of fast-moving objects, making it well suited for real-time applications such as traffic monitoring and urban surveillance, where precise object localization is essential.

Henschel et al. [21] further extended object tracking capabilities by integrating video data with inertial measurement units (IMUs) for multi-person tracking. This approach is particularly effective in scenarios where visual appearance alone is insufficient, such as environments with frequent appearance changes. By maintaining identity consistency, the method enables accurate annotation of individual movements even under challenging conditions.

In addition, Xu et al. [22] introduced PointTrackV2, which represents images as 2D point clouds to improve instance segmentation and multi-object tracking in dense and dynamic scenes. This technique is especially effective in crowded urban environments or public events, where accurate tracking and annotation of multiple objects are critical for downstream analysis and decision-making.

Overall, the studies reviewed demonstrate that object tracking plays a transformative role in image annotation by enabling automation, reducing manual workload, and improving annotation quality. Its successful application across diverse fields—from healthcare to public safety—highlights its strong potential for future research. Continued advancements in tracking robustness, scalability, and integration with human-in-the-loop systems are expected to further enhance automated image annotation and support more reliable analysis of visual data.

References

1. L'Heureux, K. Grolinger, H. F. Elyamany, and M. A. M. Capretz, ‘‘Machine learning with big data: Challenges and approaches,’’ IEEE Access, vol. 5, pp. 7776–7797, 2017, doi:10.1109/ACCESS.2017.2696365.
2. L. Ren, J. Lu, J. Feng, and J. Zhou, ‘‘Uniform and variational deep learning for RGB-D object recognition and person re-identification,’’ IEEE Trans. Image Process., vol. 28, no. 10, pp. 4970–4983, Oct. 2019, doi:10.1109/TIP.2019.2915655.
3. J. Seo and H. Park, ‘‘Object recognition in very low resolution images using deep collaborative learning,’’ IEEE Access, vol. 7, pp. 134071–134082, 2019, doi: 10.1109/ACCESS.2019.2941005.
4. S. H. Kasaei, ‘‘OrthographicNet: A deep transfer learning approach for 3-D object recognition in open-ended domains,’’ IEEE/ASME Trans. Mechatronics, vol. 26, no. 6, pp. 2910–2921, Dec. 2021, doi:10.1109/TMECH.2020.3048433.
5. S.-J. Liu, H. Luo, and Q. Shi, ‘‘Active ensemble deep learning for polarimetric synthetic aperture radar image classification,’’ IEEE Geosci. Remote Sens. Lett., vol. 18, no. 9, pp. 1580–1584, Sep. 2021, doi:10.1109/LGRS.2020.3005076.
6. Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, ‘‘A survey of convolutional neural networks: Analysis, applications, and prospects,’’ IEEE Trans. Neural Netw. Learn. Syst., vol. 33, no. 12, pp. 6999–7019, Dec. 2022, doi: 10.1109/TNNLS.2021.3084827.
7. R. Chauhan, K. K. Ghanshala, and R. C. Joshi, ‘‘Convolutional neural network (CNN) for image detection and recognition,’’ in Proc. 1st Int. Conf. Secure Cyber Comput. Commun. (ICSCCC), Dec. 2018, pp. 278–282, doi: 10.1109/ICSCCC.2018.8703316.
8. Y. Liu, Y. Zhang, Y. Wang, F. Hou, J. Yuan, J. Tian, Y. Zhang, Z. Shi, J. Fan, and Z. He, ‘‘A survey of visual transformers,’’ IEEE Trans. Neural Netw. Learn. Syst., early access, Mar. 30, 2023, doi: 10.1109/TNNLS.2022.3227717.
9. K. G. Ince, A. Koksal, A. Fazla, and A. A. Alatan, ‘‘Semi-automatic annotation for visual object tracking,’’ in Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW), Oct. 2021, pp. 1233–1239. doi: 10.1109/ICCVW54120.2021.00143.
10. L. Porzi, M. Hofinger, I. Ruiz, J. Serrat, S. R. Bulo, and P. Kontschieder, ‘‘Learning multi-object tracking and segmentation from automatic annotations,’’ in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 6845–6854.
11. X. Li, L. Chen, L. Zhang, F. Lin, and W.-Y. Ma, ‘‘Image annotation by large-scale content-based image retrieval,’’ in Proc. 14th ACM Int. Conf. Multimedia. New York, NY, USA: Association for Computing Machinery, Oct. 2006, pp. 607–610, doi: 10.1145/1180639.1180764.
12. O. Pelka, F. Nensa, and C. M. Friedrich, ‘‘Annotation of enhanced radiographs for medical image retrieval with deep convolutional neural networks,’’ PLoS ONE, vol. 13, no. 11, Nov. 2018, Art. no. e0206229, doi:10.1371/journal.pone.0206229.
13. L. Porzi, M. Hofinger, I. Ruiz, J. Serrat, S. R. Bulo, and P. Kontschieder, ‘‘Learning multi-object tracking and segmentation from automatic annotations,’’ 2019, arXiv:1912.02096.
14. M. M. Adnan, M. S. M. Rahim, A. Rehman, Z. Mehmood, T. Saba, and R. A. Naqvi, ‘‘Automatic image annotation based on deep learning models: A systematic review and future challenges,’’ IEEE Access, vol. 9, p. 50253–50264, 2021, doi: 10.1109/ACCESS.2021.3068897.

15. U. Ojha, U. Adhikari, and D. K. Singh, ''Image annotation using deep learning: A review,'' in Proc. Int. Conf. Intell. Comput. Control (I2C2), Jun. 2017, pp. 1–5, doi: 10.1109/I2C2.2017.8321819.
16. Pande, K. Padamwar, S. Bhattacharya, S. Roshan, and M. Bhamare, ''A review of image annotation tools for object detection,'' in Proc. Int. Conf. Appl. Artif. Intell. Comput. (ICAAIC), May 2022, pp. 976–982, doi: 10.1109/ICAAIC53929.2022.9792665.
17. T. Yu, N. Lin, X. Zhang, Y. Pan, H. Hu, W. Zheng, J. Liu, W. Hu, H. Duan, and J. Si, ''An end-to-end tracking method for polyp detectors in colonoscopy videos,'' Artif. Intell. Med., vol. 131, Sep. 2022, Art. no. 102363, doi: 10.1016/j.artmed.2022.102363.
18. S. Xiong, S. Li, L. Kou, W. Guo, Z. Zhou, and Z. Zhao, ''Td-VOS: Tracking-driven single-object video object segmentation,'' in Proc. IEEE 5th Int. Conf. Image, Vis. Comput. (ICIVC), Jul. 2020, pp. 102–107, doi:10.1109/ICIVC50857.2020.9177471.
19. W. Hu, Q. Wang, L. Zhang, L. Bertinetto, and P. H. S. Torr, ''SiamMask: A framework for fast online object tracking and segmentation,'' IEEE Trans. Pattern Anal. Mach. Intell., vol. 45, no. 3, pp. 3072–3089, Mar. 2023.
20. Schörkhuber, F. Groh, and M. Gelautz, ''Bounding box propagation for semi-automatic video annotation of nighttime driving scenes,'' in Proc. 12th Int. Symp. Image Signal Process. Anal. (ISPA), Sep. 2021, pp. 131–137, doi: 10.1109/ISPA52656.2021.9552141.
21. R. Henschel, T. Von Marcard, and B. Rosenhahn, ''Accurate longterm multiple people tracking using video and body-worn IMUs,'' IEEE Trans. Image Process., vol. 29, pp. 8476–8489, 2020, doi: 10.1109/TIP.2020.3013801.
22. Z. Xu, W. Yang, W. Zhang, X. Tan, H. Huang, and L. Huang, ''Segment as points for efficient and effective online multi-object tracking and segmentation,'' IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 10, pp. 6424–6437, Oct. 2022, doi: 10.1109/TPAMI.2021.3087898.
23. L. Yan, Q. Wang, S. Ma, J. Wang, and C. Yu, ''Solve the puzzle of instance segmentation in videos: A weakly supervised framework with spatio-temporal collaboration,'' IEEE Trans. Circuits Syst. for Video Technol., vol. 33, no. 1, pp. 393–406, Jan. 2023, doi: 10.1109/TCSVT.2022.3202574.
24. T.-N. Le, S. Akihiro, S. Ono, and H. Kawasaki, ''Toward interactive self-annotation for video object bounding box: Recurrent self-learning and hierarchical annotation based framework,'' in Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV), Mar. 2020, pp. 3220–3229, doi:10.1109/WACV45572.2020.9093398.
25. B. Sambaturu, A. Gupta, C. V. Jawahar, and C. Arora, ''ScribbleNet: Efficient interactive annotation of urban city scenes for semantic segmentation,''Pattern Recognit., vol. 133, Jan. 2023, Art. no. 109011, doi:10.1016/j.patcog.2022.109011.
26. J. Zhu, X. Li, C. Zhang, and T. Shi, ''An accurate approach for obtaining spatiotemporal information of vehicle loads on bridges based on 3D bounding box reconstruction with computer vision,'' Measurement, vol. 181, Aug. 2021, Art. no. 109657, doi: 10.1016/j.measurement.2021.109657.
27. Q. Liu, I. M. Gaeta, M. Zhao, R. Deng, A. Jha, B. A. Millis, A. Mahadevan-Jansen, M. J. Tyska, and Y. Huo, ''ASIST: Annotation-free synthetic instance segmentation and tracking by adversarial simulations,'' Comput. Biol. Med., vol. 134, Jul. 2021, Art. no. 104501, doi:10.1016/j.compbiomed.2021.104501.
28. F. Lateef, M. Kas, and Y. Ruichek, ''Motion and geometry-related information fusion through a framework for object identification from a moving camera in urban driving scenarios,'' Transp. Res. C, Emerg. Technol., vol. 155, Oct. 2023, Art. no. 104271, doi: 10.1016/j.trc.2023.104271.
29. Z. Chen, G. Huang, W. Li, J. Teng, K. Wang, J. Shao, C. C. Loy, and L. Sheng, ''Siamese DETR,'' in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2023, pp. 15722–15731, doi:10.1109/cvpr52729.2023.01509.