IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 7 May 2024, accepted 21 May 2024, date of publication 27 May 2024, date of current version 11 June 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3406018

SURVEY

Enhancing Image Annotation With Object Tracking and Image Retrieval: A Systematic Review

RODRIGO FERNANDES[1,2], ALEXANDRE PESSOA[3], MARTA SALGADO[4], ANSELMO DE PAIVA[3], ISHAK PACAL[5], AND ANTÓNIO CUNHA[1,2]

[1]Institute for Systems and Computer Engineering, Technology and Science–INESC TEC, 4200-465 Porto, Portugal
[2]School of Sciences and Technology, University of Trás-os-Montes and Alto Douro, 5000-801 Vila Real, Portugal
[3]Applied Computing Group, NCA-UFMA Federal University of Maranhão, São Luís 65080-805, Brazil
[4]University Hospital Centre of Santo António, 4099-001 Porto, Portugal
[5]Department of Computer Engineering, Faculty of Engineering, Igdir University, 76000 Igdir, Turkey

Corresponding author: Rodrigo Fernandes (rodrigo.c.fernandes@inesctec.pt)

This work was supported in part by the National Funds through the Portuguese Funding Agency, FCT—Fundacão para a Ciência e a Tecnologia, under Project PTDC/EEI-EEE/5557/2020; in part by the European Union under Grant 101095359; and in part by the U.K. Research and Innovation under Grant 10058099.

ABSTRACT Effective image and video annotation is a fundamental pillar in computer vision and artificial intelligence, crucial for the development of accurate machine learning models. Object tracking and image retrieval techniques are essential in this process, significantly improving the efficiency and accuracy of automatic annotation. This paper systematically investigates object tracking and image acquisition techniques. It explores how these technologies can collectively enhance the efficiency and accuracy of the annotation processes for image and video datasets. Object tracking is examined for its role in automating annotations by tracking objects across video sequences, while image retrieval is evaluated for its ability to suggest annotations for new images based on existing data. The review encompasses diverse methodologies, including advanced neural networks and machine learning techniques, highlighting their effectiveness in various contexts like medical analyses and urban monitoring. Despite notable advancements, challenges such as algorithm robustness and effective human-AI collaboration are identified. This review provides valuable insights into these technologies' current state and future potential in improving image annotation processes, even showing existing applications of these techniques and their full potential when combined.

INDEX TERMS Image annotation, object tracking, image retrieval, deep learning.

I. INTRODUCTION

Image annotation is essential in various computer vision and artificial intelligence applications. With the significant increase in the volume of image data available, efficient methods to annotate large data sets are needed. Object tracking and image retrieval techniques are relevant methods for facilitating and potentially automating image annotation in this context.

The associate editor coordinating the review of this manuscript and approving it for publication was Kathiravan Srinivasan.

Artificial intelligence, particularly machine learning and deep learning has revolutionized how we process and interpret large volumes of visual data [1]. Machine learning, which includes algorithms capable of learning and making predictions or decisions based on data, is the foundation for automatic image annotation. Within this domain, deep learning, especially using deep neural networks, has shown an extraordinary ability to extract complex features and patterns from images, facilitating tasks such as object recognition [2], [3], [4] and image classification [5], [6], [7].

Convolutional neural networks (CNNs) are a fundamental pillar in image processing [8]. They simulate how the human

visual cortex interprets images, using layers of neurons to process visual data in various abstractions [9]. This capability makes CNNs particularly suitable for computer vision tasks, including image annotation, where they can identify and label objects in images with high precision.

More recently, Visual Transformers (ViTs) [10], [11] have emerged as an alternative and robust approach. Inspired by the success of Transformers in natural language processing, ViTs apply attention mechanisms to capture global relationships between different parts of an image. This makes them particularly effective at understanding complex visual contexts, a valuable feature for automatic image annotation.

Accurate data annotation is fundamental in machine learning applications, directly impacting the effectiveness of trained models. Traditionally, annotation is carried out manually, a process that can be slow and subject to inconsistencies. Automating this process, partially or entirely, is a relevant objective for increasing the efficiency and consistency of annotations.

Object tracking involves identifying and following objects over time in videos or image sequences. This technique can be used for automatic annotation [12], [13], following the trajectory of moving objects and marking them in each frame. This approach can reduce the time needed for annotation and improve consistency, especially in contexts with dynamic objects.

On the other hand, image retrieval involves searching for and identifying similar images in large databases. Using algorithms to identify common patterns and characteristics, this technique can suggest annotations for new images [14], [15], based on previously annotated data, providing a starting point for annotation.

The joint application of object tracking and image retrieval to image annotation offers a promising approach to automation in computer vision. This systematic review aims to explore the current state of these techniques, assessing how they can be applied to optimize image annotation. The review focuses on analyzing recent studies and practical applications, aiming to provide a detailed overview of the benefits and challenges of these methodologies.

A. MOTIVATION

The growing demand for annotated image datasets in fields such as medicine, security and pattern recognition highlight the importance of efficient and accurate image annotation methods [16]. The motivation for this systematic review arises from the opportunity to explore how object tracking and image retrieval techniques can contribute to this process, offering solutions to existing challenges in manual annotation and providing a more automated and efficient approach [14], [17].

It is important to emphasize that, despite the growing research and development in image annotation, systematic reviews in this area are remarkably scarce in recent years. This gap in the literature highlights the critical need for a comprehensive review that synthesizes recent advances and contextualizes the current state of the art. Thus, this review stands out by compiling and presenting the most current studies, reflecting the significant advances in image annotation with emerging technologies as object tracking and image retrieval.

Manual image annotation, although traditional, presents significant challenges. These include high time demands, variability in the accuracy and consistency of annotations due to human intervention, and difficulty scaling to large data volumes. Automating image annotation, or at least offering automated assistance in this process, can speed up the work and increase its accuracy and consistency [18].

The application of techniques designed to maximize learning from limited data such as transfer learning [19], [20], [21], [22], [23], data augmentation [24], [25], [26], [27] and few-shot learning [28], [29], [30], [31], [32] complements this move towards automation. While these methods are valuable for training robust models with sparse annotated datasets, the ultimate goal remains to minimize their need by improving the automation of the annotation process itself. This approach not only addresses the immediate challenges of data scarcity, but also aligns with the long-term vision of creating self-sustaining deep learning ecosystems that can learn and adapt with minimal human oversight.

The emphasis on developing automated annotation systems is particularly pertinent given the exponential increase in digital data [18]. The ability to automatically annotate and categorize this data becomes not only beneficial, but also essential for its management and value extraction. Automated annotation systems fueled by object tracking and image retrieval therefore represent a significant advance in this regard, offering scalable, efficient and accurate solutions to meet the growing demands of various industries.

Additionally, human-AI collaboration in image annotation introduces unique challenges that deserve further exploration [33], [34]. While AI can significantly improve efficiency and accuracy in identifying and tracking objects in image sequences, properly integrating human judgement and expertise is crucial to ensuring the relevance and semantic accuracy of annotations. The interaction between human annotators and AI systems needs to be intuitive and flexible, allowing for easy corrections and adjustments, and ensuring that human knowledge is effectively incorporated into the annotation process.

Object tracking and image retrieval techniques have already demonstrated their effectiveness in several practical applications, suggesting significant potential for innovation in image annotation. Object tracking can automate the identification and tracking of objects in image sequences, reducing the human effort required to annotate each frame individually [35], [36], [37]. On the other hand, image retrieval can facilitate annotation by identifying similar images with existing annotations, providing a reliable starting point and speeding up the annotation process [33], [38], [39].

Therefore, this review seeks to evaluate and synthesize current knowledge on object tracking and image retrieval in image annotation, identifying potential advances, challenges and opportunities. The aim is to provide a comprehensive understanding of how these techniques can improve the efficiency and accuracy of image annotation in different domains, contributing to advancing research and practice in computer vision and related areas.

B. OBJECTIVE AND RESEARCH QUESTIONS

This systematic review aims to explore the use of object tracking and image retrieval techniques to automate or assist in image annotation. The focus is to investigate how the integration of these technologies can optimize the annotation process, enhancing its efficiency and accuracy. The review will be guided by the following research questions:

Q1: How are object tracking and image retrieval techniques being used to automate or assist in image annotation, and what are the current developments associated with these technologies?

Q2: How can the integration of object tracking, and image retrieval be optimized to improve the image annotation process? This question aims to discover innovative approaches to combining object tracking and image retrieval efficiently. It seeks to understand how the synergy between these two technologies can be maximized to speed up and improve image annotation.

Q3: What are the main challenges and limitations faced when applying these techniques to image annotation? Here, the focus is on identifying the technical and practical challenges and limitations that currently prevent the effective implementation of object tracking and image retrieval in image annotation. This issue also seeks to explore potential solutions or approaches to overcome these challenges.

An in-depth understanding of these issues will provide valuable insights into the opportunities, challenges and future directions for using advanced computer vision techniques in image annotation, boosting efficiency and accuracy in various fields of application, such as medical diagnosis, surveillance and large-scale pattern recognition.

II. RELATED WORK

In the dynamic field of image annotation with deep learning, several systematic reviews offer valuable insights and explore different aspects of this evolving domain. Recent studies in the dynamic field of image annotations can be summarized as follows.

Adnan et al. [40] devoted themselves to a comprehensive analysis of Automatic Image Annotation (AIA) methods, with a special emphasis on deep learning models. This review is significant in that it categorizes AIA methods into five distinct categories: CNN-based, RNN-based, DNN-based, LSTM-based and SAE-based. The study not only highlights recent advances in these techniques, but also points to persistent challenges, such as the need for more accurate and efficient techniques to improve automatic image annotation.

Ojha et al. [41] focused specifically on the use of convolutional neural networks (ConvNets) for image annotation. This review details how ConvNets are applied to image content annotation, exploring their ability to extract visual features for complex computer vision tasks. The review highlights the crucial role of ConvNets in object identification and localization, underlining their effectiveness in dealing with visual perception challenges in images.

Pande et al. [42] presented a comparative analysis of a variety of image annotation tools for object detection. This study is notable for its comprehensive approach, evaluating different annotation tools in terms of functionalities, effectiveness and applicability in varied object detection contexts. The review highlights the importance of the appropriate choice of annotation tool, emphasizing that the quality of the annotation has a direct and significant impact on the performance of object detection models.

Existing reviews in the field of image annotation with deep learning, including the work of Adnan et al. [40], Ojha et al. [41] and Pande et al. [42], offer a comprehensive overview of current methodologies and applications, focusing on different aspects of this evolving area. They illustrate the technological advances and challenges that still need to be overcome, providing an overview of the trends and future directions of this emerging field. However, a notable limitation of these reviews is their tendency to focus on specific types of techniques in isolation, which can limit understanding of the full capability of image annotation technologies. Especially when considering the synergistic potential of combining different approaches to tackle complex challenges, this perspective can prove restrictive.

In contrast, our review distinguishes itself by exploring not just one, but two complementary techniques: object tracking and image retrieval. By integrating these two approaches, we propose a more holistic view of image annotation, recognizing that combining these technologies can bring significant benefits to the efficiency and accuracy of the annotation process. This integration represents an evolution in the field of image annotation, leveraging the potential of each technique to complement and enrich the other, and opening up new possibilities for significant advances in the automation and accuracy of image annotation.

III. LITERATURE REVIEW METHODOLOGY

A. ELIGIBILITY CRITERIA

To guarantee a relevant and objective selection of studies, we established strict eligibility criteria, detailed in Table 1. The primary purpose of these criteria is to filter out research that effectively addresses the questions proposed by this study, excluding those that fall outside the scope of our research. These parameters were carefully formulated to capture the most pertinent literature and restrict our analysis to documents strictly related to the research questions at hand. Implementing these criteria before the literature search is

TABLE 1. Inclusion (IC) and exclusion (EC) criteria.

Criteria	Description
ICO	Published since 2020
IC1	The title, abstract, or keywords match the search query
IC2	Work published in refereed journals or conference
IC3	Direct or indirect applicability of object tracking techniques for image annotation
IC4	Direct or indirect applicability of image retrieval techniques for image annotation
EC0	Work not published in refereed journal or conference
EC1	Literature/Systematic Review
EC2	Full text is not available
EC3	The paper is not written in English
EC4	Does not consider the use of Object Tracking or Image Retrieval
EC5	Out of scope
EC6	Technique used cannot be leveraged for image annotation

a crucial strategy for reducing bias in the study selection process.

B. IDENTIFICATION PHASE

In the initial selection phase, three reference databases were chosen: IEEE Xplore, Scopus and SpringerLink. The search centred on combinations of keywords in the titles, abstracts and keywords of the articles searched, with the following query:

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

This process was carried out on 14 November 2023, considering publications from 2020 to 2023. This choice of time was strategic to ensure the inclusion of the most recent and relevant studies in deep learning applied to object tracking and image retrieval techniques for annotating image datasets, resulting in the identification of 5455 documents for the initial screening phase.

Additionally, during the review and selection of studies, we identified some highly relevant works that did not strictly fit the terms of the original search but offered valuable contributions to the topic under discussion. These studies were carefully included to enrich the analysis and discussion, considering their direct or indirect relevance and the potential to provide additional insights into object tracking and image retrieval techniques applied to image annotation.

IV. RESULTS

Based on the criteria established in Section III, this part of the article explores the results achieved. We focus our analysis on object tracking and image retrieval techniques, considering how these technologies can be adapted for image annotation. The research involves a careful analysis of the algorithmic approaches examined, focusing on the data sets used, the viability of the methods in various annotation contexts and the real-time execution capacity of the models. This review's main findings and observations are summarized and organized in Tables 2 and 3.

A. SCREENING PHASE AND ELIGIBILITY

After defining the eligibility criteria in Section III, we began the screening and eligibility determination phase, a crucial stage in the systematic review process to ensure the relevance and quality of the included studies. This phase involves a thorough evaluation of the documents retrieved from the selected databases, based on the criteria previously established. The aim is to refine the initial set of documents to include only those that strictly fulfil the eligibility criteria, thus guaranteeing the integrity and relevance of the subsequent analysis. The screening phase begins with a review of the titles and abstracts of the 5455 documents initially identified, as detailed in Section III, "Literature Review Methodology". This process allows us to identify and exclude studies that do not fall within the scope of our investigation, focussing on those papers that offer valuable insights into the use of object tracking and image retrieval techniques in the annotation of images and video datasets. Through individual reading of titles and abstracts, we identified that a substantial number of these documents did not meet our inclusion and exclusion criteria and were therefore excluded, leaving 95 unique works for the screening phase. This initial stage was crucial to ensure the relevance and uniqueness of the studies within our research scope. Subsequently, in the eligibility phase, we conducted a detailed evaluation of the full text of these 95 documents, strictly guided by the previously defined exclusion criteria.

This meticulous analysis resulted in the exclusion of a significant portion of the documents, based on various factors of incompatibility with the established criteria, such as non-compliance with the research objective. Of the evaluated articles, 15 were selected in the Object Tracking area and 17 in the "Image Retrieval" area, totaling 32 studies for data extraction and qualitative analysis. These studies were chosen not only for their direct relevance to the themes of interest but also for the quality of their methodologies, data sets used, and relevance to the proposed research questions. The selection of these studies reflects our commitment to covering a broad and in-depth spectrum of the applications of the techniques mentioned in image annotation.

Following Figure 1, the next image presents a bar chart outlining the number of articles selected per year from the initial set of studies. This visual representation allows for a clear and immediate understanding of the distribution and volume of relevant research within the specified time period.

The bar chart shown illustrates the annual distribution of the articles selected from 2020 to 2023. It is possible to observe a progressive increase in the number of articles, with the highest bar corresponding to the year 2023. This suggests growing interest and progress in the research fields of object tracking and image retrieval, as they become increasingly relevant to the development of more sophisticated image annotation techniques. The graph serves not only as a quantitative analysis of research output over the years but can also reflect the growing importance of these technologies in addressing complex challenges in computer vision and artificial intelligence.

TABLE 2. Selected object tracking reviewed articles with their respective authors, year of publication, methodology/algorithms, main area, application and real-time capacity.

Year Authors	Methodology/Algorithms	Main Area	Application in Image Annotation	Real-time Capacity
2022, Tao Yu et al. [43]	Instance Tracking Head (ITH), Scaled-YOLOv4, Similarity Metric based on Learning	Detecting and Tracking Polyps	Yes, for annotation and tracking polyps in colonoscopy videos	Yes
2020, Shaopan Xiong et al. [44]	Object segmentation approach in video based on tracking, combining Box2Segmentation with general object tracking	Object Segmentation in Videos	Yes, for segmenting individual objects in videos	Not specified
2023, Weiming Hu et al. [45]	SiamMask: object tracking and real-time video segmentation with Siamese networks trained offline.	Object Tracking and Segmentation in Videos	Yes, for real-time object tracking and segmentation in videos	Yes, it processes around 55 frames per second
2021, Dominik Schörkhuber et al. [46]	Semi-automatic video annotation method using object detection and tracking.	Semi-automatic video annotation, focus on night driving	Yes, for analysing night-time driving data and developing computer vision algorithms for autonomous driving.	Not specified
2021, Zhenbo Xu et al [48]	PointTrackV2, image conversion to 2D point clouds, SpatialEmbedding for instance segmentation	Multi-object tracking and segmentation (MOTS)	Yes, for automatic annotation in scenes with multiple moving objects	Yes, 20 FPS speed on 2080Ti GPU
2020, Trung-Nghia Le et al. [50]	Interactive Self-Annotation (ISA) framework based on recurrent self-supervised learning, with Automatic Recurrent Annotation (ARA) and Interactive Recurrent Annotation (IRA) processes, and Hierarchical Correction module.	Automatic bounding box annotation in videos	Video annotation for moving objects, especially in the context of autonomous driving and intelligent transport systems	Yes, focused on reducing annotation time and human effort
2023, Bhavani Sambaturu et al. [51]	Utilising latent feature perturbation in DNN for efficient interactive annotation. Integration with LabelMe software.	Interactive image annotation for semantic segmentation in urban scenes.	Efficient annotation of urban images, reducing time and human effort, with the capacity to correct multiple labels simultaneously.	Yes, with a focus on reducing annotation time.
2021, Jinsong Zhu et al [52]	Use of YOLO-v4 for vehicle detection and 3D bounding box reconstruction.	Computer vision for monitoring vehicle loads on bridges.	Vehicle detection and tracking.	Yes
2021, Quan Liu et al [53]	Uses CycleGAN for image-annotation synthesis and RSHN for pixel embedding. Includes annotation deformation strategies for HeLa cells.	Computer vision and deep learning in cell biology and microscopy.	Facilitates segmentation and tracking in microscopy videos without manual annotation.	Not specified.
2023, Fahad Lateef et al [54]	Presents a framework for object identification in autonomous vehicles using cameras. It uses image registration and optical flow for motion analysis. It combines moving object detection with semantic segmentation and encoder-decoder techniques, employing the Semi-Global Matching algorithm for depth estimation.	Computer vision in autonomous driving.	Identification and classification of moving objects in urban scenarios.	Not specified.
2020, Roberto Henschel et al. [47]	Use of a neural network to associate people detections with IMU orientations, formulation of a graph labelling problem, use of the perspective correction (PC) algorithm and integration of IMU signals for trajectory reconstruction.	Tracking multiple people in videos with wearable IMU sensors.	Identification and long-term tracking of people in videos, useful for behavioural and sports analysis.	Not specified
2023, Zeren Chen et al [55]	Use of a Siamese self-supervised pre-training approach for the Transformer architecture in DETR, with an emphasis on learning vision-invariant and detection-oriented representations. Implements two self-supervised pretraining tasks: Multi-Vision Region Detection and Multi-Vision Semantic Discrimination.	Self-supervised learning and object detection using Transformers.	Improved object detection and semantic discrimination in images, useful for object recognition and tracking tasks.	Not specified
2023, Liqi Yan et al. [49]	STC-Seg, video instance segmentation, unsupervised depth estimation, optical flow	Instance segmentation, weak supervised learning, spatio-temporal collaboration	Yes, for video instance segmentation	Not specified
2020, Thiago T. Santos et al. [56]	Use of CNNs, Mask R-CNN, YOLO, and three-dimensional association.	Object detection, instance segmentation, object tracking	Public dataset for detecting and segmenting grape bunches.	Not specified

B. OBJECT TRACKING

In this section, we will discuss in detail the algorithms that represent significant advances in object tracking. The approaches vary widely, from traditional machine learning techniques to advanced methods employing convolutional neural networks, each offering innovative solutions to

TABLE 3. Selected image retrieval reviewed articles with their respective authors, year of publication, methodology/algorithms, main area, application and datasets.

Year, Authors	Methodology/Algorithms	Main Area	Application in Image Annotation	Dataset
2022, J Faritha Banu et al. [57]	Image segmentation and feature extraction using grid-based colour histogram and texture techniques	Content-Based Image Retrieval (CBIR)	Yes, for image annotation and retrieval	WANG
2020, Yi-Hui Chen et al [55]	Automatic semantic annotation of images, Natural language analysis, Candidate phrase extraction, RDF (Resource Description Framework), SPARQL, LSI (Latent Semantic Indexing)	Social Image Retrieval	Yes, for automatic semantic annotation and identification of semantic intentions in social images	NBA Blogs (January 2015 to November 2015) - Manual annotations with RDF
2020, Binqiang Wang et al [61]	Recurrent Topic Memory Network (RTRMN), Recurrent Neural Network, Memory Network, Convolutional-MaxPooling	Remote Sensing Image Processing, Legend Generation	Yes, to generate automatic semantic descriptions of remote sensing images	UCM-CaptioRSICD (Remote Sensing Image Captioning Dataset)ns:and
2023, Myasar Mundher Adnan et al. [40]	ResNet50-SLT, word2vec, principal component analysis (PCA), t-SNE	Automatic Image Annotation, Deep Learning	Yes, by improving accuracy in image annotation.	Corel-5K, ESP-Game and Flickr8k
2021, Mona Zamiri et al [66]	Multi-View Robust Spectral Clustering (MVRSC), Maximum Correntropy Criterion, Half-Quadratic optimisation framework	Image Annotation, Semantic Retrieval	Model for image annotation based on multi-view fusion.	Flickr, 500PX and Corel-5K
2022, Jordão Bragantini et al [67]	Interactive image segmentation annotation guided by feature space projection, metric learning, dimensionality reduction	Interactive Image Segmentation, Interactive Machine Learning	Method for mass annotation of images through projection in feature space.	iCoSeg, DAVIS, Rooftop and Cityscapes
2023, Ikhlaq Ahmed et al. [59]	Use of RESNET-50 and BERT for image and text feature extraction, with inductive learning for feature fusion.	Content-Based Image Retrieval (CBIR)	Recovery of modified images on e-commerce platforms, using deep learning.	Fashion-200K and MIT-States
2022, Umer Ali Khan et al. [62]	Use of local tetra angle patterns (LTAP) and colour moment features to improve image retrieval accuracy, optimised with genetic algorithm.	Content-Based Image Retrieval (CBIR)	Efficient image retrieval on social media platforms, using advanced colour and texture features	Corel-1K,Oxford Flower and CIFAR-10
2020, Yikun Yang et al. [63]	Use of DNN and CNN for saliency prediction and acquisition of deep image representations.	Content-Based Image Retrieval (CBIR)	Retrieval of quality images from large databases using deep learning.	ImageNet, Caltech256 and CIFAR-10
2021, Jhilik Bhattacharya et al. [68]	Use of Capsule Networks and decision fusion with W-DCT and RBC for classification and retrieval of medical images.	Content-Based Medical Image Retrieval (CBIR)	Use of capsule architecture for accurate retrieval and classification of medical images in large databases.	IRMA (Image Retrieval in Medical Applications) and ImageCLEFMed-2009
2021, Dhupam Bhanu Mahesh et al. [69]	Development of an OLWGP descriptor for data retrieval and classification, using a heuristic J-BMO algorithm for optimal feature point selection and an optimised CNN for medical data classification.	Content-Based Image Retrieval (CBIR) and Medical Image Classification	Use of optimised OLWGP and CNN descriptors for accurate retrieval and classification of medical images in large databases.	Kaggle datasets named: CT (computerised tomography), CT head, Fundus Iris (DIARETDB1), Mammogram breast (MIAS), MRI brain, US (ultrasound), X-ray bone, X-ray chest and X-ray dental
2022, Zafran Khan et al. [64]	Use of DenseNet to generate visual characteristics of images and BERT for text embeddings. Deep learning for joint image and text representation.	Content-Based Image Retrieval (CBIR) and Medical Image Classification	Multi-modal CBIR that processes image and text queries to retrieve images from a substantial database, adjusting to the wishes expressed in the query.	Fashion200k, MIT States and FashionIQ
2022, Anna Guan et al. [65]	Use of the DenseNet-121 model pre-trained with the C2L method; introduction of interpretable saliency maps; fusion of global and local features; definition of three loss functions to optimise hash codes.	Hash-based medical image retrieval, with a focus on chest X-rays.	Improving accuracy in medical image retrieval, with special attention to injured areas in chest X-rays.	Chest X-ray8
2021, P. Das et al. [60]	Method based on robust descriptors using Zernike Moments, curvalet features and gradient orientation for biomedical image retrieval.	Biomedical Image Retrieval	Use of robust descriptors for effective retrieval of biomedical images from large databases	HRCT dataset, Emphysema CT database, OASIS MRI database and NEMA MRI database
2023, Felipe Cadar et al. [70]	Learned keypoint detection method for non-rigid image matching, using an end-to-end convolutional neural network (CNN).	Keypoint Detection, Non-rigid Image Matching	Improvement in deformable object matching and object retrieval through learnt keypoint detection, improving accuracy in non-rigid images.	HRCT dataset, Emphysema CT database, OASIS MRI database and NEMA MRI database
2022, Seyed Mahdi Roostaiyan et al. [71]	Coupled dictionary learning with marginalised loss function and L1 regularisation.	Machine Learning and Image Processing	Improvement of image annotation through coupled dictionary learning, addressing label imbalance.	IAPRTC-12, FLICKR-60K and FLICKR-125K

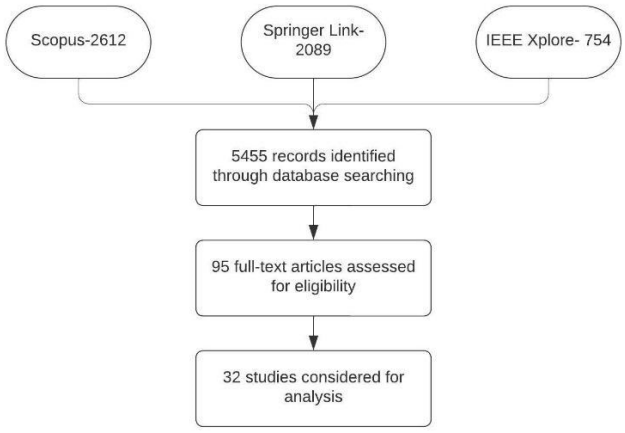

FIGURE 1. Study selection flow diagram.

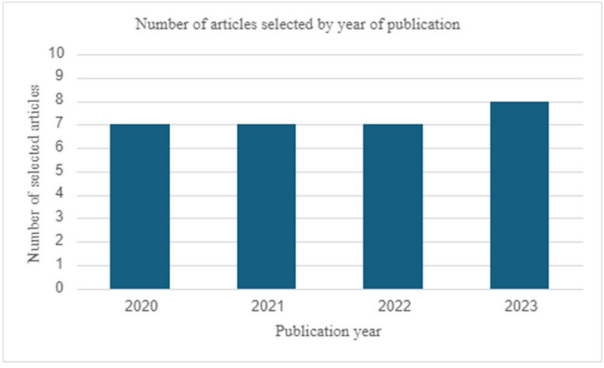

FIGURE 2. Studys selected separated by years.

specific challenges within diverse application contexts, these algorithms are presented in Table 2 with their respective research articles.

1) SINGLE OBJECT TRACKING

Tao Yu et al. [43] present a revolutionary method that integrates the ''Instance Tracking Head'' (ITH) module into object detection frameworks to detect and track polyps in colonoscopy videos. This method, aligned with the Scaled-YOLOv4 detector, allows sharing of low-level feature extraction and progressive specialization in detection and tracking. The approach stands out for its speed, being around 30 % faster than conventional methods, while maintaining exceptional detection accuracy (mAP of 91.70 %) and tracking accuracy (MOTA of 92.50 %, Rank-1 Acc of 88.31 %).vai

Xiong et al. [44] develop a solution based on a tracking module that uses Siamese networks, specifically SiamFC++, to accurately localize objects. The innovation here lies in modelling visual tracking as a similarity learning problem, complemented by the Box2Segmentation module, which efficiently transforms bounding boxes into segmentation masks, trained on the COCO dataset. This method forms the basis of Td-VOS, allowing precise segmentation of objects in videos from the initialization of a bounding box in the first frame.

Hu et al. [45] (SiamMask) uniquely combine object tracking with real-time video segmentation. Using convolutional Siamese networks fully trained offline with an additional binary segmentation task, SiamMask operates online with the initialization of a bounding box, processing video at 55 fps. This innovative approach employs two- and three-branch variants, integrating similarity, bounding box regression and binary segmentation tasks, with mask refinement to improve accuracy. SiamMask is adaptable to multiple objects and stands out for its efficiency and speed.

Schörkhuber et al. [46] introduce a technique for semi-automatic annotation of night-time driving videos. The method includes generating trajectory proposals by tracking, extending and verifying these trajectories with single object tracking and semi-automatic annotation of bounding boxes. Tested on the CVL dataset, focused on European rural roads at night, the method demonstrated a 23 % increase in recall with near-constant precision, outperforming traditional detection and tracking approaches. This work addresses the gap of rural and night scenes in driving datasets, proposing significant improvements for efficient annotation in challenging autonomous driving contexts.

Henschel et al. [47] propose an advanced method for multi-person tracking, combining video and body Inertial Measurement Units (IMUs). This method stands out by addressing the challenge of tracking people in situations where appearance is not discriminating or changes over time, such as changes in clothing. Using a neural network to relate person detections to IMU orientations and a graph labelling problem for global consistency between video and inertial data, the method overcomes the limitations of video-only approaches. With a challenging new dataset that includes both video and IMU recordings, the method achieved an impressive average IDF1 score of 91.2 % demonstrating its effectiveness in situations where it is feasible to equip people with inertial sensors.

2) MULTIPLE OBJECT TRACKING AND SEGMENTATION

Xu et al. [48] (PointTrackV2) stand out with an innovative method that converts compact image representations into disordered 2D point clouds, facilitating the rigorous separation of foreground and background areas from instance segments. This process is enriched by a variety of data modalities to enhance point features. PointTrackV2 surpasses existing methods in efficiency and effectiveness, achieving speeds close to real time (20 FPS) on a single 2080Ti GPU. In addition, the study introduces the APOLLO MOTS dataset, more challenging than KITTI MOTS, with a higher density of instances. Extensive evaluations demonstrate the superior performance of PointTrackV2 on various datasets, also discussing the applicability of this method in areas beyond tracking, such as detailed image classification, 2D pose estimation and object segmentation in videos.

Yan et al. [49] (STC-Seg) present a novel framework for instance segmentation in videos under a weakly supervised approach. Using unsupervised depth estimation and optical flow, STC-Seg generates efficient pseudo-labels to train deep networks, focusing on the accurate generation

of instance masks. One of the main contributions is 'puzzle loss', which allows end-to-end training using box-level annotations. In addition, STC-Seg incorporates an advanced tracking module that utilizes diagonal points and spatio-temporal discrepancy, increasing robustness against changes in object appearance. This method demonstrates exceptional performance, outperforming supervised alternatives on the KITTI MOTS and YT-VIS datasets, evidencing the effectiveness of weakly supervised learning in segmenting instances in videos.

3) IMPROVEMENTS IN ANNOTATION AND EFFICIENCY

Le et al. [50] propose an interactive and self-supervised annotation framework that significantly improves the efficiency of creating object bounding boxes in videos. Based on two main networks, Automatic Recurrent Annotation (ARA) and Interactive Recurrent Annotation (IRA), the method iterates over the improvement of a pre-existing detector by exposing it to unlabeled videos, generating better ground pseudo-truths for self-training. IRA integrates human corrections to guide the detection network, using a Hierarchical Correction module that progressively reduces the distance between annotated frames with each iteration. This innovative system has proven capable of generating accurate, high-quality annotations for objects in videos, substantially reducing annotation time and costs.

Sambaturu et al. [51] (ScribbleNet) present an interactive annotation method called ScribbleNet, designed to improve the annotation of complex urban images for semantic segmentation, crucial in autonomous navigation systems. This technique offers a pre-segmented image, which iteratively improves segmentation using scribbles as input. Based on conditional inference and exploiting correlations learnt in deep neural networks, ScribbleNet significantly reduces annotation time - up to 14.7 times faster than manual annotation and 5.4 times faster than current interactive methods. In addition, it integrates with the LabelMe image annotation tool and will be made available as open-source software, notable for its ability to work with scenes in unknown environments, annotate new classes and correct multiple labels simultaneously.

Zhu et al. [52] present an accurate method for reconstructing 3D bounding boxes of vehicles in order to obtain detailed spatial-temporal information about vehicle loads on bridges. The study uses a deep convolutional neural network (DCNN) and the You Only Look Once (YOLO) detector to detect vehicles and obtain 2D bounding boxes. A model for reconstructing the 3D bounding box is proposed, making it possible to determine the sizes and positions of vehicles. Spatial-temporal information on vehicle loads is obtained using multiple object tracking (MOT). The system developed, Bridge Vehicle Load Identification System (BVLIS), was tested on an operating cable-stayed bridge, demonstrating the accuracy and reliability of the method. This approach is innovative in that it combines deep learning-based vehicle detection, camera calibration and 3D bounding box reconstruction, providing an effective alternative to conventional methods for assessing the condition of bridges and their behavior under vehicle loads.

Liu et al. [53] propose a novel technique for segmenting and tracking instances in microscopy videos without the need for manual annotation. Using adversarial simulations and pixel embedding-based learning, the ASIST method is able to simulate variations in the shape of cellular and subcellular objects, overcoming the challenge of consistent annotations required by traditional methods. The study demonstrates that ASIST achieves a significant improvement over supervised approaches, showing superior performance in segmentation, detection and tracking of microvilli and comparable performance in videos of HeLa cells. This method represents a breakthrough in the quantitative analysis of microscopy videos, offering an efficient and automated solution for the quantification of cellular and subcellular dynamics without the labor-intensive manual annotation.

Lateef et al. [54] propose an innovative object identification framework (FOI) for autonomous vehicles, focusing exclusively on camera data to detect and analyze objects in urban driving scenarios. This framework uses image registration algorithms and optical flow estimation to compensate for self-motion and extract accurate motion information from moving objects from a mobile camera. At the heart of this system is a moving object detection (MOD) model, which combines an encoder-decoder network with a semantic segmentation network to perform two crucial tasks: the semantic segmentation of objects into specific classes and the binary classification of pixels to determine addition, the article presents a unique dataset for detecting moving objects, covering a variety of dynamic objects. The experiments demonstrate the effectiveness of the proposed framework in providing detailed semantic information about objects in urban driving environments.

Chen et al. [55] (Siamese DETR) present ''Siamese DETR'', a new method for self-supervised training of DETR (DEtection TRansformer) transformers, introduced at the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). This study proposes combining the Siamese network with DETR's cross-attention mechanism, focusing on learning vision-invariant and detection-oriented representations. The method achieved state-of-the-art transfer performance on the COCO and PASCAL VOC detection benchmarks. The team highlights the effectiveness and versatility of Siamese DETR, demonstrating significant improvements in localization accuracy and acceleration in convergence. However, Siamese DETR relies on a pre-trained CNN, such as SwAV, and future work aims to integrate the CNN and Transformer into a unified training paradigm.

C. IMAGE RETRIEVAL

In this section, significant advances in the image retrieval field will be discussed with the intent of presenting a wide range of different algorithms and approaches to this technique. The approaches vary from methods to solve problems

in the field of image retrieval to new approaches to this technique that show potential to innovate this field, each offering solutions to specific challenges within diverse application contexts, these algorithms are presented in Table 3 with their respective research articles.

1) ADVANCES IN CBIR AND OVERCOMING THE SEMANTIC GAP

Faritha Banu et al. [57]: This paper proposes an innovative CBIR system that incorporates both model and content annotations within an ontology framework. Using distinct visual features such as color and texture, the system applies advanced image segmentation techniques and extracts features using grid-based color histograms and texture analysis. This multi-faceted approach not only significantly improves the accuracy and speed of image retrieval, but also addresses the semantic gap, especially useful in medical image search and retrieval contexts.

Chen et al. [58]: In this study, the authors present an innovative solution to the challenge of the semantic gap in social image retrieval. Combining multiple visual features and textual matching, the model employs ontologies and linked open data to perform automatic, semantic annotation of images. Using the Stanford Parser, the study extracts candidate phrases related to images and maps them to entities in DBpedia, facilitating the generation of accurate RDF annotations. The method was validated on NBA blogs, demonstrating significant improvements in the accuracy and relevance of search results.

Ahmed et al. [59]: The article describes the development of the Deep-view Linguistic and Inductive Learning (DvLIL) framework, which stands out for combining visual and textual modalities to improve image retrieval. Using state-of-the-art techniques such as ResNet-50 and BERT, the system extracts detailed visual features and generates semantic and contextual representations of the text. The fusion of these features, realized through a sequence of multilayer perceptrons based on inductive learning, demonstrates remarkable effectiveness on challenging datasets, offering a more adaptable and robust solution for image retrieval compared to traditional CBIR systems.

Das and Neelima [60]: This paper introduces a robust methodology for biomedical image retrieval, centered on the use of a feature vector that combines Zernike moments, curvlet features and gradient orientation. This holistic approach not only captures texture and shape information effectively, but also minimizes redundant data. The methodology has been validated on four biomedical databases, demonstrating a superior retrieval rate, which represents a significant advance in overcoming the semantic gap in medical images.

2) IMPROVED ANNOTATION AND AUTOMATIC CAPTION GENERATION

Wang et al. [61]: The study presents the Recurrent Topic Retrieval Memory Network (RTRMN), an innovative approach for generating captions in remotely sensed images (RSI). By analyzing five annotated sentences per image and identifying common keywords between them, the RTRMN employs a memory network where these keywords serve as guides for the formation of more precise and determined captions. This method offers a solution to the ambiguity often found in captions generated by previous techniques, improving both the accuracy and contextual relevance of captions for remotely sensed images.

Adnan et al. [18]: This paper proposes an advanced automatic image annotation system that combines the ResNet50-Slantlet transform with technologies such as word2vec, principal component analysis (PCA) and t-SNE for effective image characterization. The system is able to extract high and low-level visual features, including shape, texture and color, outperforming traditional methods in terms of precision, recall and F-measure across multiple datasets. This approach not only improves automatic image annotation, but also provides valuable insights for semantic gap reduction.

3) FEATURE FUSION WITH DEEP LEARNING

Khan and Javed [62]: In this paper, the authors develop a hybrid CBIR system that integrates Local Tetra Angle Patterns (LTAPs) and color moment features for more efficient image retrieval. The combination of these textural and chromatic features, together with the application of a genetic algorithm for attribute selection, results in a hybrid feature vector that significantly improves image retrieval performance. This system effectively addresses the semantic gap and offers a robust solution for image retrieval in large databases.

Yang et al. [63]: The paper proposes an advanced image retrieval algorithm that utilizes a deep content-based quality model, integrating DNN-based saliency prediction and image quality assessment (IQA). This method selects high-quality salient regions and concatenates them in a way that mimics human visual perception, improving image retrieval in large datasets. The study demonstrates that this approach outperforms several state-of-the-art algorithms, offering an effective solution to the semantic gap in large-scale image retrieval.

"DenseBert4Ret" by Khan et al. [64]: This study develops an image retrieval system based on multimodal content, using the integration of DenseNet for visual feature extraction and BERT for textual analysis. This bi-modal system simultaneously processes images and text as queries, improving accuracy in retrieving images that match the combination of users' visual and textual desires. The approach demonstrates superiority on real-world datasets, highlighting the potential of deep learning in creating joint image and text representations.

Guan et al. [65]: The article presents an advanced method for retrieving medical images, focusing on feature fusion and information interpretability. Using the DenseNet-121 model to learn relevant medical features without the need for manual annotation, the method applies interpretable saliency maps and integrates global and local networks to extract complete

information, resulting in a significant improvement in the accuracy of retrieval results. These advances promise valuable applications in computer-aided diagnosis systems.

4) SPECIFIC APPLICATIONS AND INNOVATIONS IN IMAGE ANNOTATION

Zamiri and Sadoghi Yazdi [66]: This study introduces the Multi-View Robust Spectral Clustering (MVRSC) method for image annotation, modelling the relationship between semantic and multi-features of training images. Using the Maximum Correntropy Criterion and semi-quadratic optimization, the method suggests tags based on a new fusion distance at the decision level. Experimental results on real datasets demonstrate the method's effectiveness in generating accurate and meaningful annotations, integrating geographic and visual information.

Bragantini et al. [67]: In this article, the authors propose an innovative approach to interactive image annotation, allowing the simultaneous annotation of segments of multiple images through projection onto the feature space. This technique results in a faster process and avoids redundancies by annotating similar components in different images. The results show a significant improvement in the efficiency of image annotation, suggesting possibilities for integration with other existing image segmentation methodologies.

Bhattacharya et al. [68]: The paper proposes an advanced approach to medical image search, utilizing capsule architecture and decision fusion to address challenges such as data imbalance, insufficient labels and obscured images. Tested on the IRMA dataset, the method demonstrates superior efficiency, significantly improving diagnostic efficiency by grouping similar images for automatic retrieval and annotation.

Bhanu Mahesh et al. [69]: This paper presents a medical image retrieval and classification model based on the Optimized Local Weber Gradient Pattern (OLWGP), using a new heuristic algorithm to improve image retrieval. The study also employs an optimized CNN model for image classification, demonstrating superior performance on several public databases and offering significant advances in medical image retrieval and classification.

Cadar et al. [70]: The study presents a novel technique for keypoint detection in non-rigid images using a CNN trained with true correspondences. This method not only improves the accuracy of the matches, but also the efficiency of object retrieval, representing a significant advance in the detection of keypoints in non-rigid images and in improving the matching performance of existing descriptors.

Roostaiyan et al. [71]: This paper introduces Marginalised Coupled Dictionary Learning (MCDL) as a new approach for real-time image annotation. Focusing on learning a limited number of visual prototypes and their associated semantics, the method overcomes common challenges in image annotation by offering an efficient and fast solution with a publicly available implementation.

V. DISCUSSION

This discussion section aims to further analyze the advances and implications of object tracking and image retrieval technologies, with a special focus on their practical applications in various domains and the significant real-world impact that these techniques have demonstrated. Given the insights provided by the studies analyzed, we have undertaken a comparative assessment of the techniques studied, highlighting their strengths, weaknesses and suitability for different application scenarios. This analysis not only illuminates the unique contributions of each method, model or algorithm, but also sheds light on the synergies and challenges that arise when integrating these technologies to solve complex real-world problems.

We recognize the complexity and depth of the topics covered by this review and have therefore expanded our discussion to provide a more nuanced view of the limitations and challenges faced by current methodologies. In addition, we will discuss the potential interdisciplinary applications of these technologies in more detail, highlighting areas that go beyond those primarily considered in the review. This includes exploring how object tracking and image retrieval can be innovatively applied in fields such as health, public safety and environmental conservation, where they have the potential to promote significant advances.

A. OBJECT TRACKING

The field of object tracking has been marked by significant innovations, especially with the application of advanced neural networks and deep learning techniques. The introduction of the Instance Tracking Head (ITH) by Yu et al. [43] exemplifies this evolution, offering a notable improvement in tracking accuracy in medical contexts, such as colonoscopy videos. This innovation underlines the ability of these new technologies to adapt to specialized applications where precision is crucial.

Advancing the complexity of applications, Xiong et al. [44] explored the use of Siamese networks to improve object segmentation in videos. This approach not only strengthens tracking accuracy, but also highlights the versatility of modern techniques in dealing with dynamic and complex scenes. The convergence of these technologies' points to a horizon where object tracking can be adapted to a wider range of scenarios, from controlled environments to busy urban contexts.

The accuracy and adaptability of object tracking in adverse conditions represent ongoing challenges, as demonstrated by Schörkhuber et al. [46] in their studies of night-time driving videos. This work illustrates the importance of developing systems that can operate efficiently under variations in visibility, a critical factor for security and monitoring applications.

Collaboration between humans and artificial intelligence has emerged as a recurring theme, with studies such as those by Le et al. [50] and Sambaturu et al. [51] highlighting

collaborative approaches to image annotation and analysis. This human-machine interaction suggests a future in which the precision and efficiency of AI can be combined with human sensitivity and discernment to create more robust and accurate solutions in a variety of applications.

The expansion to multiple objects tracking and segmentation, as demonstrated by Xu et al. [48] and Yan et al. [49], opens up new possibilities for real-time monitoring and analysis of complex scenes. These techniques, which transform images into more malleable representations such as point clouds, highlight the potential of deep learning to extract and analyze information in an efficient and innovative way.

The challenge of detecting and analyzing movements in dynamic scenarios is addressed by Lateef et al. [54] and Henschel et al. [47], who apply object tracking to contexts of urban mobility and human interactions, respectively. These studies illustrate how technology can be adapted to improve public safety and understand complex behavior in crowded environments.

Finally, the diversity of applications and continuous innovation in object tracking, as reflected by the works of Chen et al. [55] and. Santos et al. [56], highlight the importance of ethical approaches, especially in public contexts where privacy and consent are paramount concerns. The evolution of this technology not only promises improvements in a variety of fields, but also imposes the need for careful reflection on its responsible use.

B. IMAGE RETRIEVAL

The evolution of Content-Based Image Retrieval (CBIR) has been driven by significant technological advances, as demonstrated by Faritha Banu et al. [57], who developed an innovative CBIR system using ontologies to integrate model and content annotations. This system not only improves the accuracy and speed of image retrieval, but also paves the way for practical applications, especially in medical fields where precision in image search and retrieval is vital.

This quest for accuracy and efficiency is complemented by the efforts of Chen et al. [58] to address the ''semantic gap'' in social image retrieval by combining multiple visual features and textual matching. This breakthrough highlights a growing trend in CBIR: the integration of multiple data modalities to enrich the retrieval process, making the results more aligned with the users' intentions.

In this context of multimodal enrichment, Wang et al. [61] made progress with the Recurrent Topic Retrieval Memory Network (RTRMN), which generates accurate captions for remotely sensed images. This development highlights the importance of contextualization and detail in the generation of captions, which are crucial aspects for the interpretation and use of images in areas such as environmental and geographical research.

The integration of multimodalities and technological innovation, as seen in the work of Ahmed et al. [59] and Khan et al. [64], exemplify how the combination of in-depth visual features and semantic textual representations can refine image retrieval. This approach not only improves accuracy, but also personalizes the image retrieval experience, adapting to the specific needs of users in a variety of contexts, from e-commerce to multimedia.

As we explore specific applications and advances in segmentation and classification, studies such as those by Zamiri et al. [66] and Bhattacharya et al. [68] bring to light innovative methods that improve image annotation and retrieval in urban and medical contexts. They use advanced clustering techniques and capsule networks to model semantic relationships and multiple features, demonstrating the adaptability of these technologies to specific annotation and retrieval needs.

However, beyond specific applications, CBIR faces the ongoing challenge of detecting and analyzing complex patterns in images. Guan et al. [65], for example, focus on medical image retrieval, using hashing techniques based on feature fusion and interpretability to better represent injured areas on X-rays. This approach not only advances computer-aided diagnosis, but also emphasizes the importance of interpretable and transparent systems.

Looking to the future, CBIR should continue to explore data fusion and deep contextualization. Deep learning, exemplified by Khan et al. [62] and Yang et al. [63], promises to transform image retrieval by dynamically adapting to a variety of contexts and user requirements. Furthermore, the emphasis on interpretation and user interaction, as evidenced by Roostaiyan et al. [71], highlights the need for methods that can effectively deal with unbalanced labels and maintain data sparsity.

Thus, the trajectory of CBIR is marked by an intersection of technological innovation, practical applicability and integration challenges. Roostaiyan et al.'s approach [71], which introduces marginalized coupled dictionary learning for real-time image annotation, illustrates the need for adaptive approaches capable of dealing with the diversity and complexity of image datasets, while maintaining computational efficiency and the relevance of retrieval results.

Continued innovation in CBIR, particularly in the integration of advanced deep learning techniques and multimodal analysis, as demonstrated by Cadar et al. [70] in their research on keypoint detection in non-rigid images, highlights the potential for significant advances in image retrieval accuracy and capacity. The application of these techniques in a variety of contexts, from medical analyses to pattern recognition in remote sensing images, suggests a broad spectrum of possibilities for improving both the granularity and applicability of CBIR.

However, as technologies advance and their applications expand, ethical and privacy considerations emerge, especially in contexts involving sensitive or identifiable data. The need for responsible and transparent approaches to image retrieval is becoming increasingly pressing, emphasizing the importance of incorporating ethical principles into the development and deployment of CBIR systems.

C. POTENTIAL OF COMBINING TECHNIQUES FOR IMAGE ANNOTATION

The fusion of object tracking and image retrieval techniques promises to revolutionize the field of image annotation, offering more sophisticated and efficient methods for identifying and cataloguing visual content. This convergence has the potential to significantly automate the annotation process, improving accuracy and reducing the manual effort required, particularly in large datasets.

In the context of object tracking, the ability to continuously follow an entity through a sequence of images or videos provides a solid basis for dynamic and contextually rich annotations. When integrated with image retrieval systems, this continuous tracking can be enriched with historical or semantic information extracted from extensive databases, allowing for annotations that capture not only the identity of the object, but also its behavior, interactions and evolution over time.

For example, in surveillance video analysis, the combination of these techniques can automate the annotation of activities, identifying and cataloguing specific actions by individuals or vehicles. This not only saves time manually reviewing hours of footage, but also improves search and retrieval capabilities, allowing users to quickly find moments or events of interest based on detailed annotations.

In scientific and environmental research, the combined application of these technologies can facilitate the cataloguing of species or natural phenomena by integrating movement information captured by object tracking with taxonomic or behavioral knowledge derived from image retrieval systems. This can significantly speed up the annotation of large sets of images captured in field studies, allowing researchers to concentrate their efforts on analyzing and interpreting the data.

In the medical field, this integrated approach could transform the way diagnostic images are annotated and stored. By combining the precise tracking of injuries or medical conditions in sequential images with the ability to link these observations to similar or relevant cases in medical literature, annotation systems can provide a wealth of clinical context, potentially revealing previously hidden patterns or correlations.

However, the successful implementation of this integrated approach requires overcoming significant challenges, including managing large volumes of data, the need for high-performance processing algorithms and ensuring accuracy and relevance in the annotations generated. In addition, ethical and privacy issues remain paramount, especially in sensitive applications such as surveillance and medicine.

VI. CONCLUSION

This systematic review investigated the current use of object tracking and image retrieval techniques in automating or assisting image annotation. From the studies analyzed, the answers to the proposed questions are as follows:

Q1: How are object tracking and image retrieval techniques being used to automate or assist in image annotation, and what are the current developments associated with these technologies?

A: Automatic image annotation is an area that continues to evolve with the development of object tracking and image retrieval techniques. These techniques are essential for improving the accuracy and efficiency of annotation, which is important in various applications such as medical diagnosis, urban surveillance and the management of large image databases.

Object tracking and image retrieval techniques, if used for this purpose, can play key roles in automating and assisting image annotation, greatly helping annotators who would otherwise have to do it manually. These technologies are very useful not only for improving the efficiency of annotation processes, but also for increasing the accuracy of the annotations generated, which is essential in fields such as medical diagnosis, urban monitoring and automatic multimedia content management.

Object tracking has benefited from the advancement of deep neural networks and sophisticated machine learning methods, which have contributed to automation and accuracy in image annotation. The integration of advanced technologies not only improves the identification and tracking of objects in image sequences, but also facilitates the automatic and continuous annotation of these objects.

An example of this evolution is the work of Yu et al. [43] who developed the ''Instance Tracking Head'' (ITH), integrated into the Scaled-YOLOv4 detector. This innovation offers improvements in the detection and tracking of objects in medical videos, such as colonoscopies. Improved detection accuracy and continuous tracking enable automatic and accurate annotation of polyps over time, facilitating medical monitoring and analysis by reducing the need for manual annotation, which is often prone to errors and inconsistencies.

Another significant development is SiamMask, created by Hu et al. [45], which combines object tracking and segmentation in real time. This tool processes video at a rate of 55 frames per second, enabling continuous and automated annotation of fast-moving objects. SiamMask is particularly useful in scenarios that require real-time responses, such as urban surveillance and traffic monitoring, where the precise identification and tracking of objects is essential for security and incident management.

The study by Henschel et al. offers an advanced method for tracking multiple people using both video and inertial measurement units (IMUs). This method is especially effective in environments where the appearance of individuals changes frequently, such as at sporting events or concerts, allowing accurate annotation of movements and positions without loss of subject identity, even in challenging conditions.

Additionally, Xu et al. [48] with PointTrackV2 transform images into 2D point clouds, which facilitates the segmentation of instances and the tracking of multiple objects in crowded and dynamic environments. This technique

enables effective annotation in congested urban areas or at public events, where accurate tracking and annotation of multiple moving objects is crucial for subsequent analyses and decision-making.

These are just a few examples of the advances that demonstrate how object tracking can transform the task of image annotation, making it more efficient and reducing the manual workload. The application of these technologies in a variety of fields, from medicine to public safety, highlights the significant potential for future innovations that can lead to an even deeper understanding and better practices in analyzing visual data.

Continuing the discussion on the automation of image annotation, we have also seen significant advances in image retrieval that complement the improvements brought about by object tracking. Image retrieval techniques have benefited greatly from deep learning and semantic analysis, which increase the accuracy and speed of retrieval and enrich the quality of automatic annotations.

For example, Faritha Banu et al. have developed a system that employs grid-based color histograms and texture analysis within an ontological framework, which not only speeds up the retrieval of medical images but also improves the accuracy of automatic annotations. This advance is quite important for clinical applications, where accurate annotations can mean better diagnosis and treatment.

Chen et al. implemented a method that combines visual and textual analysis for semantic retrieval of social images. This method not only improves the relevance and accuracy of annotations, but also facilitates the categorization and retrieval of social content based on clear semantic intent, thus improving the management of large image databases.

Wang et al. advanced the automatic generation of captions for remote sensing images through recurrent memory networks, which use common keywords to generate accurate and contextual descriptions. This process is vital for the effective interpretation and use of images in environmental monitoring and urban planning.

Furthermore, Ali Khan and Javed have created a hybrid CBIR system that combines local tetra angle patterns with color moment features to improve image retrieval. This system addresses the challenge of the ''semantic gap'' found in large image databases by enriching the automatic annotation process with more detailed and accurate features, which is essential for better categorization and use of the retrieved images.

These developments in image retrieval, together with advances in object tracking, are broadening the possibilities for using images in a variety of practical applications, ensuring that visual information is maximized to its full potential. With these advanced technologies, it is possible to automate the annotation of large image datasets, reducing manual labor and increasing the reliability of the information.

Q2: How can the integration of object tracking and image retrieval be optimized to improve the image annotation process?

A: The efficient integration of object tracking and image retrieval techniques represents a significant advance in the field of automatic image annotation. Both techniques have complementary capabilities which, when aligned correctly, can substantially improve the accuracy, efficiency and applicability of image annotation in a variety of contexts. Existing object tracking techniques usually have difficulties in some areas, namely in situations of low light or visibility, in situations where you want to follow an individual in a crowd, situations where the scenery changes abruptly or even when there is occlusion of objects due to overlapping objects momentarily blocking the object being followed, although image retrieval cannot solve these issues it can offer support for annotating these complicated situations using previously studied information providing the annotator not only with different perspectives but also with comparisons that would not have been observed previously, a paper that experimented with this approach was Wei and Huang [72], this paper shows how the combination of both techniques used for the purpose of autonomous driving, their approach shows also that it could also be used to improve image annotation, this being a step towards an interdisciplinary collaboration for image annotation. In the following, we present a detailed approach to how this integration can be optimized, exploring connections between the techniques discussed in the selected articles.

As previously mentioned, object tracking often faces challenges in conditions of poor lighting and visibility. Image retrieval techniques, such as the one presented by Faritha Banu et al, which employ ontologies to improve the accuracy and speed of image retrieval, can be used to complement and enrich the training datasets of tracking models. In addition, the integration of advanced image attributes and semantic annotations extracted through retrieval methods can help tracking models to better adapt to varying conditions, using historical or similar data to adjust their predictions in real time and improve accuracy in challenging environments.

Multiple object tracking, as explored in works such as that by Xu et al. [48] (PointTrackV2), can benefit significantly from image retrieval. For example, techniques that use textual and visual analysis for semantic annotation of images (Chen et al.) can be integrated to provide additional context that makes it easier to distinguish between similar objects in crowded scenes. This approach can enable tracking systems to assign more accurate identities and maintain tracking consistency over time, even when objects interact or hide from each other.

Even well-defined tracking systems, such as Hu et al.'s SiamMask [45], can be improved with image retrieval techniques that process contextual and appearance variations. Using advanced image retrieval algorithms that integrate deep and semantic features (such as the DvLIL system by Ahmed et al. [59]), it is possible to develop an adaptive layer that adjusts the parameters of the tracking model in real time, based on features previously observed in similar situations. This not only improves the robustness of the tracking, but also

reduces errors caused by abrupt changes in the scenario or the appearance of objects.

To maximize the benefits of this integration, it is crucial to implement effective synchronization between object tracking and image retrieval systems. This can be achieved by developing integrated frameworks that combine real-time data streams with dynamic access to annotated image databases, allowing for fluid and complementary interaction between the tracking and retrieval processes.

Q3: What are the main challenges and limitations faced when applying these techniques to image annotation?

A: Despite advances in automatic image annotation through object tracking and image retrieval techniques, significant challenges still persist in both areas, impacting the effectiveness of these technologies.

In object tracking, one of the main challenges faced is the management of occlusions, where objects of interest are temporarily blocked by other elements in the scene, complicating their detection and continuous tracking. In addition, rapid variations in the scene, such as sudden changes in lighting or rapid movements of objects, can challenge current algorithms, reducing tracking accuracy. The need for accurate tracking in low-visibility conditions also remains a technical obstacle, especially in applications such as night surveillance or in adverse weather conditions.

In image retrieval, the ''semantic gap'' - the discrepancy between the visual attributes of retrieved images and the semantic meaning that users attribute to those images - remains a prominent challenge. This gap often results in annotations that don't match user expectations or specific application needs, limiting the practical usefulness of image retrieval systems. Finding more robust and adaptive methods to fill this semantic gap is crucial to improving the relevance and accuracy of automatically generated annotations.

These challenges highlight the continued need for research and development in the areas of object tracking and image retrieval. Innovative solutions are needed to address these limitations, potentially through the development of more sophisticated algorithms that can better cope with adverse conditions and complex contexts, and more effective semantic processing techniques that better align image retrieval results with user needs. Improvements in these areas will not only advance the state of the art in automatic image annotation, but also expand its practical applications in fields such as medical diagnosis, urban monitoring and multimedia content management.

VII. FUTURE RESEARCH

The evolution of image annotation through object tracking and image retrieval technologies, as systematically analyzed, shows a promising trajectory towards more efficient and accurate machine learning models. Despite notable advances, the field faces challenges that require a research agenda geared towards promoting innovation and addressing the complexities of real-world applications.

A critical area for future exploration lies in improving algorithmic robustness and generalization. Current methodologies demonstrate varying degrees of effectiveness in different datasets and conditions, often struggling with low-visibility scenarios and rapid object movements. Solving these problems requires a concerted effort to develop algorithms that are not only adaptable to diverse environmental conditions, but also capable of learning from limited and unstructured data, a good example of this effort is shown by Schörkhuber et al. [46]. The integration of unsupervised and semi-supervised learning paradigms could offer a way to reduce dependence on extensively annotated data sets, thus expanding the applicability of these technologies in domains where such data is scarce or difficult to obtain in order to increase the amount of annotation data, creating a cycle where this kind of algorithms are continually less needed to fight scarce annotation.

At the same time, the synergy between human expertise and automated systems represents fertile ground for research. The current landscape of image annotation tools reflects a growing recognition of the invaluable role of human intuition and understanding in improving AI-generated annotations. Future research should endeavor to improve this symbiosis by developing more intuitive interfaces and feedback mechanisms, which are open source and easy to use. These systems should not only facilitate the incorporation of human corrections, but also learn from these interactions, thus continuously improving the accuracy and relevance of the annotations.

In addition, the imperative need for real-time annotation capabilities cannot be overemphasized, especially in domains that require instant decision-making, such as surveillance and live medical diagnosis. The search for real-time processing solutions requires innovations in terms of computational efficiency and algorithmic speed. This may involve taking advantage of the growing dominance of peripheral computing and developing models adapted for use in resource-limited environments, ensuring that the benefits of automated annotation can be realized across a broader spectrum of applications.

The ethical considerations and privacy concerns surrounding the use of these technologies, particularly in sensitive areas such as personal surveillance and healthcare, require rigorous attention. Future research should prioritize the development of ethical frameworks and privacy-preserving mechanisms. This includes exploring advanced data anonymization techniques and secure data sharing protocols to protect individual privacy, while enabling the beneficial applications of image annotation technologies.

ACKNOWLEDGMENT

This work was supported in part by the National Funds through the Portuguese Funding Agency, FCT—Fundacão para a Ciência e a Tecnologia, under Project PTDC/EEI-EEE/5557/2020; in part by the European Union under Grant 101095359; and in part by the U.K. Research and Innovation under Grant 10058099.

REFERENCES

[1] A. 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] K. Muhammad, S. Khan, J. D. Ser, and V. H. C. de Albuquerque, "Deep learning for multigrade brain tumor classification in smart healthcare systems: A prospective survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 2, pp. 507–522, Feb. 2021, doi: 10.1109/TNNLS.2020.2995800.

[7] W. Teng, N. Wang, H. Shi, Y. Liu, and J. Wang, "Classifier-constrained deep adversarial domain adaptation for cross-domain semisupervised classification in remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 5, pp. 789–793, May 2020, doi: 10.1109/LGRS.2019.2931305.

[8] 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.

[9] 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.

[10] Y. Xu, H. Wei, M. Lin, Y. Deng, K. Sheng, M. Zhang, F. Tang, W. Dong, F. Huang, and C. Xu, "Transformers in computational visual media: A survey," *Comput. Vis. Media*, vol. 8, no. 1, pp. 33–62, Mar. 2022, doi: 10.1007/s41095-021-0247-3.

[11] 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.

[12] 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. Accessed: Jan. 31, 2024, doi: 10.1109/ICCVW54120.2021.00143.

[13] 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. Accessed: Jan. 31, 2024.

[14] 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.

[15] D. D. Burdescu, C. G. Mihai, L. Stanescu, and M. Brezovan, "Automatic image annotation and semantic based image retrieval for medical domain," *Neurocomputing*, vol. 109, pp. 33–48, Jun. 2013, doi: 10.1016/j.neucom.2012.07.030.

[16] 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.

[17] 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*.

[18] M. M. Adnan, M. S. M. Rahim, A. R. Khan, A. Alkhayyat, F. S. Alamri, T. Saba, and S. A. Bahaj, "Automated image annotation with novel features based on deep ResNet50-SLT," *IEEE Access*, vol. 11, pp. 40258–40277, 2023, doi: 10.1109/ACCESS.2023.3266296.

[19] S. A. H. Minoofam, A. Bastanfard, and M. R. Keyvanpour, "TRCLA: A transfer learning approach to reduce negative transfer for cellular learning automata," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 5, pp. 2480–2489, May 2023, doi: 10.1109/TNNLS.2021.3106705.

[20] Z. Zhu, K. Lin, A. K. Jain, and J. Zhou, "Transfer learning in deep reinforcement learning: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 11, pp. 13344–13362, Nov. 2023, doi: 10.1109/TPAMI.2023.3292075.

[21] H. Han, H. Liu, C. Yang, and J. Qiao, "Transfer learning algorithm with knowledge division level," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 11, pp. 8602–8616, Nov. 2023, doi: 10.1109/TNNLS.2022.3151646.

[22] Z. Fan, L. Shi, Q. Liu, Z. Li, and Z. Zhang, "Discriminative Fisher embedding dictionary transfer learning for object recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 1, pp. 64–78, Jan. 2023, doi: 10.1109/TNNLS.2021.3089566.

[23] H. Shi, J. Li, J. Mao, and K.-S. Hwang, "Lateral transfer learning for multiagent reinforcement learning," *IEEE Trans. Cybern.*, vol. 53, no. 3, pp. 1699–1711, Mar. 2023, doi: 10.1109/TCYB.2021.3108237.

[24] W. Zhang, Y. Zhang, and L. Zhang, "Multiplanar data augmentation and lightweight skip connection design for deep-learning-based abdominal CT image segmentation," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–11, 2023, doi: 10.1109/TIM.2023.3328707.

[25] Y. Ma, M. Liu, Y. Tang, X. Wang, and Y. Wang, "Image-level automatic data augmentation for pedestrian detection," *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1–12, 2024, doi: 10.1109/TIM.2023.3336760.

[26] J. Cao, M. Luo, J. Yu, M.-H. Yang, and R. He, "ScoreMix: A scalable augmentation strategy for training GANs with limited data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 7, pp. 8920–8935, Jul. 2023, doi: 10.1109/TPAMI.2022.3231649.

[27] L. Zhang and K. Ma, "A good data augmentation policy is not all you need: A multi-task learning perspective," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 5, pp. 2190–2201, May 2023, doi: 10.1109/TCSVT.2022.3219339.

[28] X. Wang, X. Wang, B. Jiang, and B. Luo, "Few-shot learning meets transformer: Unified query-support transformers for few-shot classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 12, pp. 7789–7802, Dec. 2023, doi: 10.1109/TCSVT.2023.3282777.

[29] P. Tian, H. Yu, and S. Xie, "An adversarial meta-training framework for cross-domain few-shot learning," *IEEE Trans. Multimedia*, vol. 25, pp. 6881–6891, 2023, doi: 10.1109/TMM.2022.3215310.

[30] Y. Cui, W. Deng, X. Xu, Z. Liu, Z. Liu, M. Pietikäinen, and L. Liu, "Uncertainty-guided semi-supervised few-shot class-incremental learning with knowledge distillation," *IEEE Trans. Multimedia*, vol. 25, pp. 6422–6435, 2023, doi: 10.1109/TMM.2022.3208743.

[31] J. Li, M. Gong, H. Liu, Y. Zhang, M. Zhang, and Y. Wu, "Multiform ensemble self-supervised learning for few-shot remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4500416, doi: 10.1109/TGRS.2023.3234252.

[32] H.-J. Ye, L. Han, and D.-C. Zhan, "Revisiting unsupervised meta-learning via the characteristics of few-shot tasks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3721–3737, Mar. 2023, doi: 10.1109/TPAMI.2022.3179368.

[33] G. Carneiro, A. B. Chan, P. J. Moreno, and N. Vasconcelos, "Supervised learning of semantic classes for image annotation and retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 3, pp. 394–410, Mar. 2007, doi: 10.1109/TPAMI.2007.61.

[34] A. Ulges, M. Worring, and T. Breuel, "Learning visual contexts for image annotation from Flickr groups," *IEEE Trans. Multimedia*, vol. 13, no. 2, pp. 330–341, Apr. 2011, doi: 10.1109/TMM.2010.2101051.

[35] S. Li, Z. Zhou, M. Zhao, J. Yang, W. Guo, Y. Lv, L. Kou, H. Wang, and Y. Gu, "A multitask benchmark dataset for satellite video: Object detection, tracking, and segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, doi: 10.1109/TGRS.2023.3278075.

[36] P. Zhu, J. Zheng, D. Du, L. Wen, Y. Sun, and Q. Hu, "Multi-Drone-Based single object tracking with agent sharing network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 10, pp. 4058–4070, Oct. 2021, doi: 10.1109/TCSVT.2020.3045747.

[37] X. Zheng, H. Cui, and X. Lu, "Multiple source domain adaptation for multiple object tracking in satellite video," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5626911, doi: 10.1109/TGRS.2023.3336665.

[38] S. Zhang, J. Huang, H. Li, and D. N. Metaxas, "Automatic image annotation and retrieval using group sparsity," *IEEE Trans. Syst., Man, Cybern., B, Cybern.*, vol. 42, no. 3, pp. 838–849, Jun. 2012, doi: 10.1109/TSMCB.2011.2179533.

[39] D. Wang, S. C. H. Hoi, Y. He, J. Zhu, T. Mei, and J. Luo, "Retrieval-based face annotation by weak label regularized local coordinate coding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 3, pp. 550–563, Mar. 2014, doi: 10.1109/TPAMI.2013.145.

[40] 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, pp. 50253–50264, 2021, doi: 10.1109/ACCESS.2021.3068897.

[41] 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.

[42] B. 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.

[43] 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.

[44] 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.

[45] 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.

[46] D. 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.

[47] R. Henschel, T. Von Marcard, and B. Rosenhahn, "Accurate long-term 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.

[48] 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.

[49] 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.

[50] 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.

[51] 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.

[52] 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.

[53] 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.

[54] 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.

[55] 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.

[56] T. T. Santos, L. L. de Souza, A. A. dos Santos, and S. Avila, "Grape detection, segmentation, and tracking using deep neural networks and three-dimensional association," *Comput. Electron. Agricult.*, vol. 170, Mar. 2020, Art. no. 105247, doi: 10.1016/j.compag.2020.105247.

[57] J. Faritha Banu, P. Muneeshwari, K. Raja, S. Suresh, T. P. Latchoumi, and S. Deepan, "Ontology based image retrieval by utilizing model annotations and content," in *Proc. 12th Int. Conf. Cloud Comput., Data Sci. Eng. (Confluence)*, Jan. 2022, pp. 300–305, doi: 10.1109/Confluence52989.2022.9734194.

[58] Y.-H. Chen, E. J. Lu, and S.-C. Lin, "Ontology-based dynamic semantic annotation for social image retrieval," in *Proc. 21st IEEE Int. Conf. Mobile Data Manage. (MDM)*, Jun. 2020, pp. 337–341, doi: 10.1109/MDM48529.2020.00074.

[59] I. Ahmed, N. Iltaf, Z. Khan, and U. Zia, "Deep-view linguistic and inductive learning (DvLIL) based framework for image retrieval," *Inf. Sci.*, vol. 649, Nov. 2023, Art. no. 119641, doi: 10.1016/j.ins.2023.119641.

[60] P. Das and A. Neelima, "A robust feature descriptor for biomedical image retrieval," *IRBM*, vol. 42, no. 4, pp. 245–257, Aug. 2021, doi: 10.1016/j.irbm.2020.06.007.

[61] B. Wang, X. Zheng, B. Qu, and X. Lu, "Retrieval topic recurrent memory network for remote sensing image captioning," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 256–270, 2020, doi: 10.1109/JSTARS.2019.2959208.

[62] U. A. Khan and A. Javed, "A hybrid CBIR system using novel local tetra angle patterns and color moment features," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 34, no. 10, pp. 7856–7873, Nov. 2022, doi: 10.1016/j.jksuci.2022.07.005.

[63] Y. Yang, S. Jiao, J. He, B. Xia, J. Li, and R. Xiao, "Image retrieval via learning content-based deep quality model towards big data," *Future Gener. Comput. Syst.*, vol. 112, pp. 243–249, Nov. 2020, doi: 10.1016/j.future.2020.05.016.

[64] Z. Khan, B. Latif, J. Kim, H. K. Kim, and M. Jeon, "DenseBert4Ret: Deep bi-modal for image retrieval," *Inf. Sci.*, vol. 612, pp. 1171–1186, Oct. 2022, doi: 10.1016/j.ins.2022.08.119.

[65] A. Guan, L. Liu, X. Fu, and L. Liu, "Precision medical image hash retrieval by interpretability and feature fusion," *Comput. Methods Programs Biomed.*, vol. 222, Jul. 2022, Art. no. 106945, doi: 10.1016/j.cmpb.2022.106945.

[66] M. Zamiri and H. S. Yazdi, "Image annotation based on multi-view robust spectral clustering," *J. Vis. Commun. Image Represent.*, vol. 74, Jan. 2021, Art. no. 103003, doi: 10.1016/j.jvcir.2020.103003.

[67] J. Bragantini, A. X. Falcão, and L. Najman, "Rethinking interactive image segmentation: Feature space annotation," *Pattern Recognit.*, vol. 131, Nov. 2022, Art. no. 108882, doi: 10.1016/j.patcog.2022.108882.

[68] J. Bhattacharya, T. Bhatia, and H. S. Pannu, "Improved search space shrinking for medical image retrieval using capsule architecture and decision fusion," *Expert Syst. Appl.*, vol. 171, Jun. 2021, Art. no. 114543, doi: 10.1016/j.eswa.2020.114543.

[69] D. B. Mahesh, G. S. Murty, and D. R. Lakshmi, "Optimized local weber and gradient pattern-based medical image retrieval and optimized convolutional neural network-based classification," *Biomed. Signal Process. Control*, vol. 70, Sep. 2021, Art. no. 102971, doi: 10.1016/j.bspc.2021.102971.

[70] F. Cadar, W. Melo, V. Kanagasabapathi, G. Potje, R. Martins, and E. R. Nascimento, "Improving the matching of deformable objects by learning to detect keypoints," *Pattern Recognit. Lett.*, vol. 175, pp. 83–89, Nov. 2023, doi: 10.1016/j.patrec.2023.08.012.

[71] S. M. Roostaiyan, M. M. Hosseini, M. M. Kashani, and S. H. Amiri, "Toward real-time image annotation using marginalized coupled dictionary learning," *J. Real-Time Image Process.*, vol. 19, no. 3, pp. 623–638, Jun. 2022, doi: 10.1007/s11554-022-01210-6.

[72] H. Wei and Y. Huang, "Online multiple object tracking using spatial pyramid pooling hashing and image retrieval for autonomous driving," *Machines*, vol. 10, no. 8, p. 668, Aug. 2022, doi: 10.3390/machines10080668.

RODRIGO FERNANDES received the bachelor's degree in biomedical engineering from the University of Trás-os-Montes and Alto Douro, where he is currently pursuing the master's degree in biomedical engineering. He is a Researcher of computer assisted annotation methods for capsule endoscopy datasets, a project with INESC TEC and the University of Trás-os-Montes and Alto Douro. His work focuses mainly on object tracking and image retrieval deep learning methods for detecting and annotating gastric lesions.

ALEXANDRE PESSOA received the degree in computer science from the Federal University of Maranhão (UFMA), and the master's degree in computer science from the Institute of Mathematics and Statistics, University of São Paulo (IME-USP), in the area of artificial intelligence. He is currently pursuing the Ph.D. degree in computer science with the UFMA/UFPI Association Doctoral Program in Computer Science. He has experience in digital image processing, computer vision, and convolutional neural networks. He is interested in artificial intelligence, machine learning, computer theory, digital image processing, and computer vision.

MARTA SALGADO received the degree in medicine from the University of Porto, in 1997, and the Specialty degree in gastroenterology, in 2005. She is a Graduate Hospital Assistant with the Gastroenterology Department, University Hospital Centre of Porto. She is also a Guest Lecturer on the master's degree in medicine with the Abel Salazar Biomedical Sciences Institute. She is the author of dozens of articles presented at scientific meetings and published in scientific journals.

ANSELMO DE PAIVA received the degree in civil engineering from the State University of Maranhão, in 1990, and the master's degree in civil engineering–structures and the Ph.D. degree in informatics from the Pontifical Catholic University of Rio de Janeiro, in 1993 and 2001, respectively. He is currently a Full Professor with the Federal University of Maranhão. He is the Coordinator of the NCA-UFMA Applied Computing Centre. He has experience in computer science, with an emphasis on graphics processing, working mainly on the following subjects: virtual and augmented reality, computer graphics, GIS, medical image processing and volumetric visualization. He is a member of the Brazilian Computer Society (SBC) and the Association for Computing Machinery (ACM).

ISHAK PACAL received the bachelor's degree in computer engineering from Harran University, the master's degree in electronic communications and computer engineering from the University of Nottingham, and the Ph.D. degree in real-time polyp detection using deep learning from Erciyes University, in 2022. Currently, he is an Assistant Professor with Igdir University. With over 15 publications in SCI-indexed journals, his research interests span medical image processing, artificial intelligence in healthcare, and artificial intelligence in agriculture.

ANTÓNIO CUNHA is the Ph.D. Senior Researcher and an Auxiliary Professor with the Engineering Department, University of Trás-os-Montes and Alto Douro (UTAD). He has participated as a member in seven funded research projects. His research interests include medical image analysis, bio-image analysis, computer vision, machine learning, and artificial intelligence, particularly in computer-aided diagnosis applied in several imaging modalities, e.g., computed tomography of the lung and endoscopic videos. He is part of the organization committee HCIST–International Conference on Health and Social Care Information Systems and Technologies (2013–2015 and 2020–2023), and the organization chair (2012) and advisory board (2016–2023).