Review

Machine Learning for Object and Action Recognition in Augmented and Mixed Reality: A Literature Review

Iolanda Chamusca [*] , Ingrid Winkler [*] , Tiago Pagano , Rafael Loureiro , Alex Santos , Thiago Murari

Posted Date: 4 October 2023

doi: 10.20944/preprints202310.0200.v1

Keywords: machine learning; augmented reality; mixed reality; object recognition; action recognition; context analysis

Review

Machine Learning for Object and Action Recognition in Augmented and Mixed Reality: A Literature Review

Iolanda L. Chamusca [1,*], Tiago P. Pagano [2], Rafael B. Loureiro [2], Erick G. S. Nascimento [2], Alex Á. B. Santos [2], Thiago B. Murari [3] and Ingrid Winkler [3,*]

1 Ford Motor Company, Camaçari 42831710, Brazil; ichamusc@ford.com (I.L.C.)
2 Department of Computational Modelling and Industrial Technology, SENAI CIMATEC University Center, Salvador 41650-010, Brazil; tiagopp@gmail.com (T.P.P); rafael.loureiro@fbter.org.br (R.B.L); alex.santos@fieb.org.br (A.A.B.S.); erick.sperandio@fieb.org.br (E.G.S.N)
3 Department of Management and Industrial Technology, SENAI CIMATEC University Center, Salvador 41650-010, Brazil; thiago.murari@fieb.org.br (T.B.M.); ingrid.winkler@doc.senaicimatec.edu.br (I.W.)
* Correspondence: ichamusc@ford.com (I.L.C.); ingrid.winkler@doc.senaicimatec.edu.br (I.W.)

Abstract: A major challenge of augmented and mixed reality applications is identifying the context and semantics of the real environment. Studies on object and action recognition were developed based on the improvement of machine learning techniques, allowing them to be annotated and recognized. This study aims to characterize current knowledge on the use of machine learning for recognizing objects and actions in augmented and mixed reality environments, increasing context awareness. Therefore, a systematic literature review of works related to these topics was made, using the Scopus and Web of Science knowledge bases. We searched articles and conference reviews or papers published between 2018 and 2022 and selected fifteen studies to be reviewed. The results indicate that there is a great demand for using machine learning to immersive technologies in factories, engineering, entertainment, education, health, among other application domains. However, these real-time interactive systems still have challenges and limitations to be solved, involving network communication, prediction time and the creation of a model that recognize objects and actions in broad contexts. Furthermore, additional research is needed to investigate how object and action recognition can increase context awareness in augmented reality applications.

Keywords: machine learning; augmented reality; mixed reality; object recognition; action recognition; context analysis

1. Introduction

A major challenge of augmented reality (AR) and mixed reality (MR) applications is identifying the context and semantics of the real environment. Contextual interfaces are an opportunity to make AR and MR interfaces responsive to both the user and the surrounding environment, making them more useful. AR and MR-based systems can be used in a variety of everyday applications, such as industrial maintenance or equipment assembly training, improving task execution, and the production performance analysis [1].

The sensors connected to an immersive device can help the user detect the context of the environment, such as who they are talking to, what they are looking at, and what they are doing. An interface with these characteristics can intelligently act as an assistant, enhancing productivity in support of current systems. Many AR/MR applications expand the world, but their connection is not meaningful. Augmented and mixed reality platforms add virtual content to an unknown real environment but do not recognize the semantics of related objects and actions, using the real world only as a backdrop [2].

Studies in the area of object and action recognition have evolved from the improvement of machine learning (ML) techniques, allowing these elements to be annotated and recognized. This

 doi:10.20944/preprints202310.0200.v1

opens up a field for studying the context of action in environments with greater immersion for the user, how objects may be used, and what action is being performed. It is important to consider that identifying objects can help with action recognition, because there is a direct connection between actions, objects, and context in the actual world, so action recognition is frequently based on object identification.

Augmented reality is defined as an immersive experience that overlays virtual 3D objects onto the user's direct view of the surrounding real environment, generating the illusion that these virtual objects exist in that space [3]. The increased perception of real-world information enables a more natural interface when working with computer-generated data and images [1]. Mixed reality, on the other hand, not only overlaps 3D elements with the real world but also combines them with real objects and actions in a not easily separable way [4]. Mixed reality devices provide users with a combined experience of both virtual reality (VR) and augmented reality, using sensors that can detect elements of the surroundings but providing greater immersion in the virtual world, not provided by AR devices [5].

Machine learning (ML) is characterized by inserting intelligence into systems and can be used as a means to improve the computer vision used in AR or MR. This technique has the ability to generalize problems such as identifying an object or an action; in addition, deep learning (DL) techniques are multilayer neural networks with high levels of flexibility, allowing efficient and effective classifications of various problems. The generalization of problems is possible due to the change of internal parameters in order to represent the structure of what is desired through different dataset training. Eventually, it is possible to automatically identify the optimal combinations of complex input data. This specific ability allows the development of autonomous systems for human-like decision-making. However, even with the growth of applications that use artificial intelligence, acquiring appropriate data for training a neural network is a great challenge in the field.

Previous studies have addressed various aspects of machine learning for object and action recognition [6] or applying deep learning to enhance the functionality of an augmented reality application [7], but they lack the connection between the three main elements addressed in this study: machine learning, object/action recognition, and augmented/mixed reality. One study focuses only on hand gesture recognition and does not explore the benefits of using this technique in immersive applications [6], while the other focuses on the aspects of implementing semantic webs and knowledge graphs in AR applications but does not explore object and action recognition [7]. Therefore, there is a knowledge gap when it comes to the connection between these elements and the associated challenges and opportunities.

Augmented and mixed reality applications does not necessarily recognize the semantic meaning of the real world, most systems use the real world only as a background for virtual elements [2]. Understanding the context in AR/MR environments through ML, with models that recognize objects and actions related to the activity, allows a greater immersion experience for the user, identified as one of the limitations in AR/MR environments. Given the above, this study aims to characterize current knowledge on the use of machine learning for recognizing objects and actions in augmented and mixed reality environments by providing context awareness.

This document is organized as follows: Section 2 describes the materials and methods utilized, Section 3 presents and analyzes the results, and Section 4 provides conclusions and suggestions for further research.

2. Materials and Methods

In order to identify the key issues in the field and to summarize the literature by highlighting these key issues, a qualitative approach was used in this systematic literature review. Because the study is exploratory, there hasn't been much research on the application of machine learning to the recognition of objects and actions in augmented and mixed reality. This idea needs to be investigated

 doi:10.20944/preprints202310.0200.v1

and understood, and qualitative research is especially helpful when the researcher is unsure of the key variables to look at [8].

This review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines, which was designed to "help systematic reviewers transparently report why the review was done, what the authors did, and what they found" [9]. Additionally, it was followed a process comprising the following seven steps: planning, defining the scope, searching the published research, assessing the evidence base, synthesizing, analyzing, and writing [10]. This study is registered on open science framework, number https://osf.io/q3h2a.

To assess the risk of bias in the included studies, as per PRISMA item five establishes [9], the preliminary search strategy was designed by two machine learning model researchers. Then, the candidate strategy was peer-reviewed by two senior ML researchers and by an expert in extended reality-based design testing of product development. Qualitative research is interpretative research [8]; therefore, it is relevant to the outcomes of this study that the authors have a strong background and experience with machine learning, augmented/mixed reality, and object and action recognition, among others.

The strategy resulting from this validation process is described in the sections that follow.

2.1. Planning

The knowledge bases that will be investigated are determined during the planning step [10]. The investigation was carried out using the scientific databases Scopus and Web of Science. These databases were chosen because they are reliable, multi-disciplinary scientific databases of international scope with comprehensive coverage of citation indexing, providing the best data from scientific publications. Scopus now includes 87 million curated documents [11], whereas the Web of Science covers more than 82 million entries [12].

2.2. Defining the scope

The scope definition step ensures that questions relevant to the research are considered before the actual literature review is carried out [10]. A brainstorming session was held with an interdisciplinary group composed of eleven experts on machine learning models, which selected two pertinent research questions to this systematic review address, namely:

Q1: What is the state of the art on using machine learning for the object and action recognition in an augmented and mixed reality environment? Q2: What are the challenges and opportunities for using machine learning for object and action recognition in an augmented and mixed reality environment?

2.3. Literature search

In the literature search step, a particular string is used to search the database set up in the planning step based on the research questions asked in the defining the scope step [10]. This string takes into account the search for works on *augmented* and *mixed reality* supported by *machine learning* models; it also considers the terms *algorithm*, *neural network*, *dataset* and *virtual assistant* as synonyms or directly related to the ML techniques. In the third part of the search string, the terms *action recognition*, *object recognition*, *emotion recognition*, *human-computer interaction*, and *real-time interaction* were used to specify the expected results of applying ML to the immersive application context.

Thus was formed the final search phrase: TITLE-ABS-KEY ((*mixed reality* OR *augmented reality*) AND (*machine learning* OR *algorithm* OR *neural network* OR *dataset* OR *virtual assistant*) AND (*action recognition* OR *object recognition* OR *emotion recognition* OR *human-computer interaction* OR *real-time interaction*)).

 doi:10.20944/preprints202310.0200.v1

2.4. Assessing the evidence base

The assessing step uses inclusion and exclusion criteria filters to reduce the number of documents found in the searching the literature step—selecting those that are relevant to the research questions [10]. These criteria were applied to the researched articles in three phases, as follows:

Phase 1: exclusions through filter options provided by the database used on the research.

- E1.1.: The entry title, abstract or keywords did not have one or more of the terms described on the search phrase;
- E1.2.: Published before 2018;
- E1.3.: Entry not written in English language;
- E1.4.: Duplicate entry.

Phase 2: exclusions through screening of the abstract of publications.

- E2.1.: Entry is not an article, conference paper or conference review;
- E2.2.: Entry is focused on medicine/health area, 3D modelling/prototyping, learning or teaching processes, metaverse, big data or robotics development;
- E2.3.: Entry focus on virtual reality, instead of augmented or mixed reality.

Phase 3: exclusion through screening of the entire article.

- E3.1.: Entry with less than 5 pages;
- E3.2.: Entry does not describe or use machine learning technologies to provide object, action or emotion recognition;
- E3.3.: Entry not apply the object, action or emotion recognition to an augmented or mixed reality context, even mentioning a possible application.

2.5. Synthesizing and Analyzing

Figure 1 depicts the flow of the systematic review from searching the published research to synthesizing processes.

Textual information is analyzed in qualitative research designs, allowing researchers to interpret themes or patterns that emerge from the data [8]. The procedures for data analysis aim to extract meaning from text; they entail segmenting, deconstructing, and reconstructing the data [8]. In terms of data analysis steps and procedures used for interpreting and validating collected data, we employed text non-numerical analysis, with the type of interpretation consisting of themes and patterns to identify challenges and opportunities of using machine learning in AR/MR, and peer experts debriefing strategy for validating findings.

Regarding the non-numeric analysis and interpretation of themes and patterns, we began by reading the retrieved articles and identifying the *artifact* developed in the study, which could be a review, a new ML model, an application of a previous model to a new interface, or a framework, among others. We also highlighted the *keywords* used in the articles to understand the main technical terms used in the field, which was used afterwards as a base for a bibliographic analysis.

The characteristics of the reviewed works (3.1) were also addressed through a word graph and an association network, generated by a bibliometric analysis, to show the words most related to the evaluated works and the strength between them. Then, different aspects of the studies were qualitatively discussed in six results subsections, each one evaluating a part of the recognition: *Object interaction recognition* (3.2), *Action recognition by frames* (3.3), *Human movements recognition* (3.4), *Real-time interactive systems processing* (3.5), *Other aspects* (3.6) and *Challenges and opportunities analysis* (3.7).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 October 2023 doi:10.20944/preprints202310.0200.v1

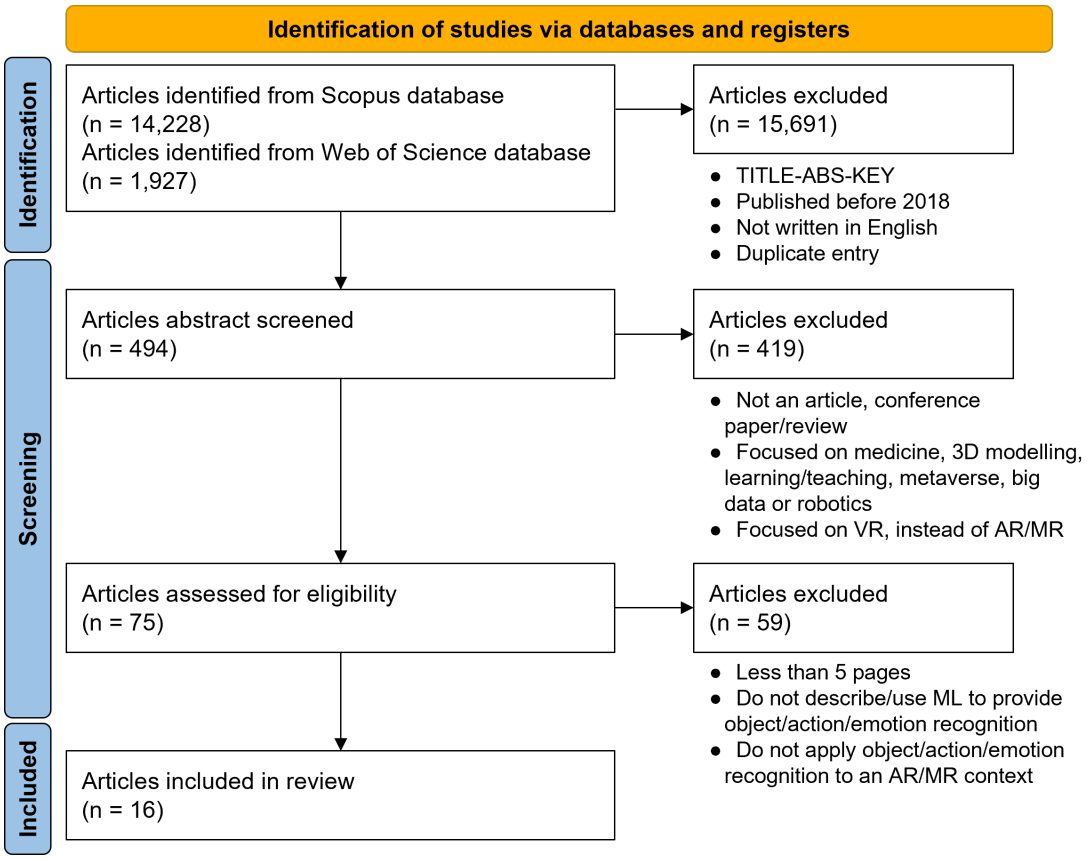

Figure 1. Systematic review flow diagram, adapted from PRISMA [9].

Qualitative validity means that the researcher checks for the accuracy of the findings by employing certain procedures; is based on determining whether the findings are accurate from the standpoint of the researcher or the readers of an account [8].

This process led to the results presented in the following section.

3. Results

In the following sections, the research questions Q1 and Q2 are addressed.

3.1. Characteristics of the reviewed works

The fifteen studies were reviewed to analyze the research question *Q1: What is the state of the art on using machine learning for the object and action recognition in an augmented and mixed reality environment?*, and the results are presented between the Sections 3.1 and 3.6.

Table 1 lists the selected studies, showing the *artifact* developed and the *keywords* associated. Related to that, Figure 2 shows, using bibliometric analysis techniques, a graph of the words most cited in the studies that will be evaluated in the review. *Augmented reality*, *mixed reality*, *machine learning* and *action recognition* can be seen as the top four most cited words in the reviewed articles. In Figure 3, a word association network presents the connection strength between the most cited words according to the frequency at which they appear together. Analyzing Figure 3, we notice that the top four terms shown in Figure 2 are positioned in the center and are more strongly connected, which can be compared with their positions in the previous Figure. It is also important to highlight the proximity of *action recognition* to *augmented* and *mixed reality*, as well as the proximity of the terms *physical* and *virtual objects* to *machine learning*.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 October 2023 doi:10.20944/preprints202310.0200.v1

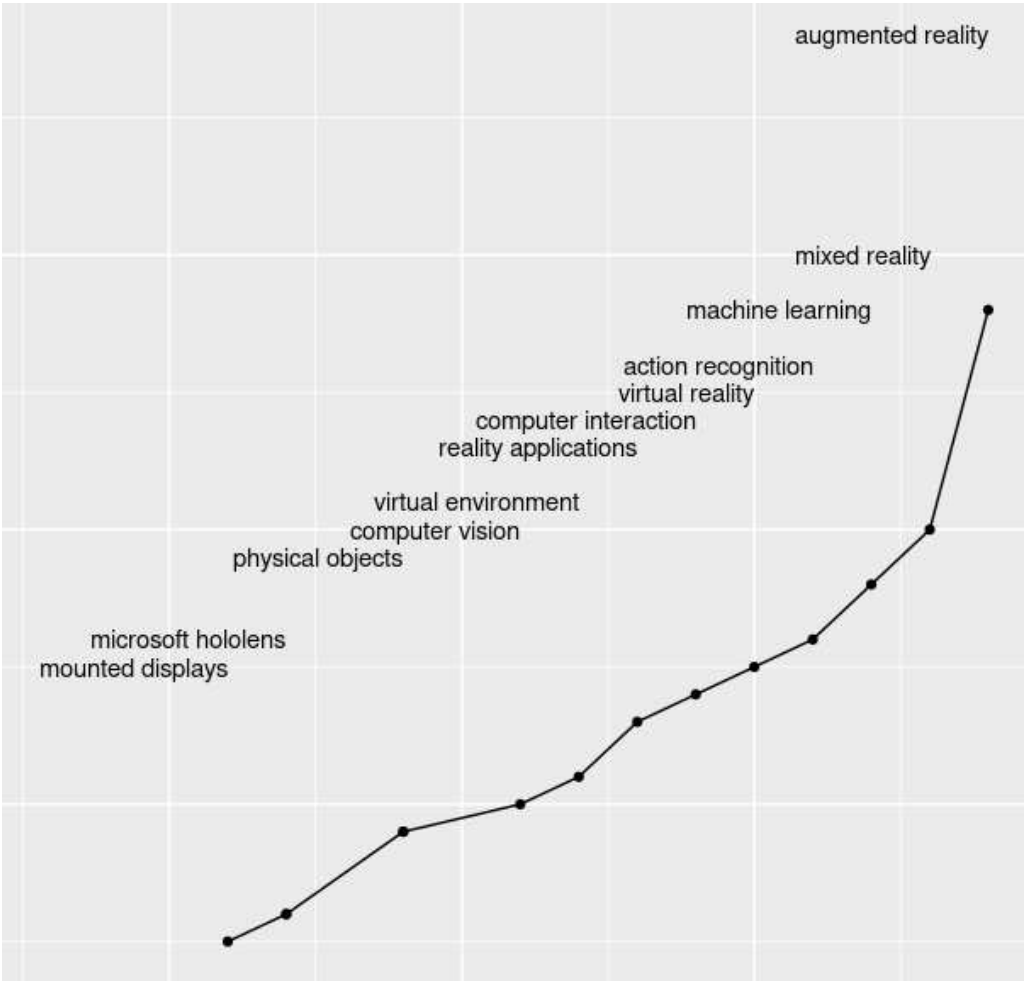

Figure 2. Graph of the most cited words in the reviewed articles

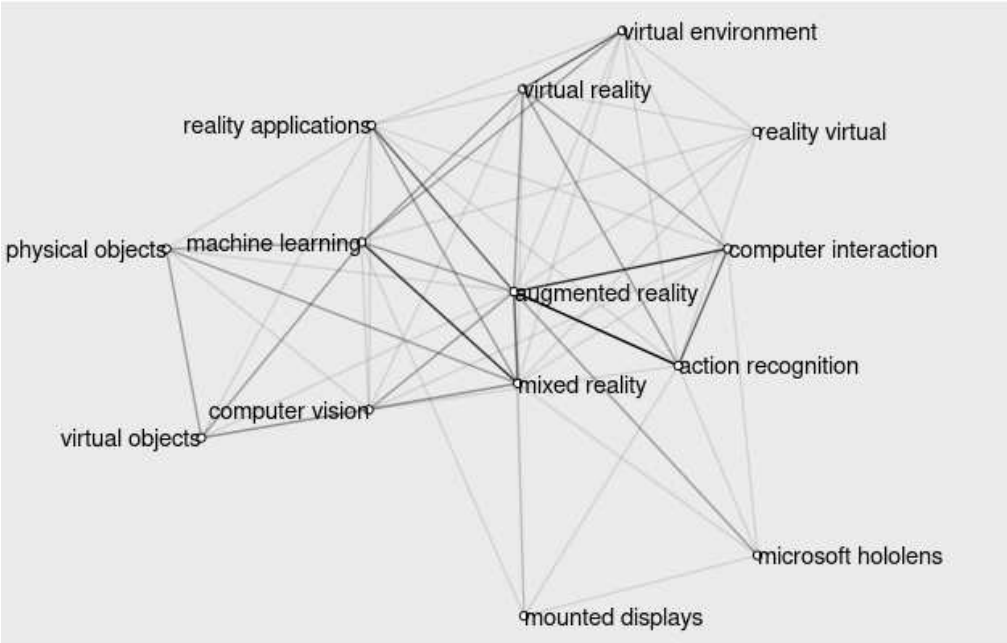

Figure 3. Word association network presenting the connection strength between the most cited words

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 October 2023 doi:10.20944/preprints202310.0200.v1

Table 1. Characteristics of the reviewed articles

Ref	Artifact	Keywords	Year
[5]	Study of the various approaches and the techniques of emotion recognition	Augmented reality; Machine learning; Deep learning, Semantic web; Knowledge graph; Human computer interaction	2018
[13]	AR and VR interaction interface based on Human Action Recognition (HAR) with a binary motion descriptor that describe and recognize actions in videos	Augmented Reality; Virtual Reality; Interaction; Human Action Recognition; Binary motion descriptor Proximity patches; Real-time	2018
[14]	Virtual assistant to help mild cognitive impaired subjects to carry out elementary food preparation	Intelligent Assistive Technologies; Augmented Reality; Action Recognition; Object Localization; Sensor Fusion	2018
[15]	Interface using input from voice, hand gestures, and eye gaze to interact with information in a virtual environment	Multimodal interface; Gesture recognition; Virtual environment	2019
[7]	Describe the concept of augmented reality and mixed reality and present deep learning, semantic web and knowledge graphs technologies	Augmented reality; Machine learning; Deep learning, Semantic web; Knowledge graph; Human computer interaction	2020
[16]	Novel approach and the corresponding framework by introducing a hardware level separation between the two tasks of object recognition and rendering virtual components	Mixed/Augmented reality; Human computer interaction (HCI); Machine learning; Real-time systems; Object recognition; Information visualization	2020
[4]	Model order reduction methods to estimate the physical behavior of deformable objects in real-time	Model order reduction; nonlinear materials; real-time interaction; solids contact	2020
[17]	Study the problem of Paired Egocentric Interaction Recognition (PEIR) that recognizes the interactions between face-to-face AR users	Paired egocentric interaction recognition; bilinear pooling; action recognition	2020
[18]	Named Data Networking (NDN) based framework that address networking challenges by offering a hybrid edge-cloud model for the execution of AR/VR computational tasks	Cloud computing; Quality of service; Bandwidth; Virtual reality; Head-mounted displays; Delays; Software	2020
[19]	Mixed reality glasses that supports the work of the operator integrated via a cloud with a technology line	Machine learning; Industry 4.0; Augmented Reality; Mixed Reality	2020
[20]	Studies survey in which Convolutional Neural Networks (CNNs) were used to estimate the 3D hand pose from data obtained from cameras	3D Hand Pose Estimation; 3D Hand Skeleton Estimation; Convolutional Neural Network	2020
[21]	Demonstrate descriptive power of 2D skeleton modality by achieving accuracy on daily action recognition competitive to 3D skeleton data	Augmented reality; Economic and social effects; Human computer interaction; Medical computing; Musculoskeletal system	2021
[22]	Semantic gaze analysis of eye-tracking data from interactive 3D scenes by training a CNN on synthetic datasets derived from virtual models using image augmentation techniques	Volumes of Interest (VOI), Semantic Gaze Analysis, Synthetic Training Data, Neural Network, User Centered Dynamic Recordings	2021
[23]	Explore object detection and classification technologies leveraging superresolution (SR), that can be integrated into small, mobile and low-power AR/VR devices	Object Detection and Classification, Super-Resolution, AR/VR applications	2022
[24]	Mobility-aware, lightweight, and hybrid 3D object detection framework for improving the user experience of AR/MR on mobile headsets	3D Object Detection, Hybrid Mobile Vision, Augmented and Mixed Reality, Mobile Headsets	2022

The discussions are organized in following sections.

 doi:10.20944/preprints202310.0200.v1

3.2. Object interaction recognition

In this section, works related to object interaction recognition are analysed. In the work of [19] and [4] the importance of interactive systems is clear, where processing for decision-making can be a limitation of their solutions.

The work of [19] highlights the importance of supporting employees in their duties, proposing the implementation of a robot control system using the Microsoft HoloLens (MHL) device. Eight robots and a cooperating operator, equipped with an AR device, were put in place. Then, the images captured by the AR device were processed for analysis, avoiding collisions between robots when placing an object on a tray in a free slot. This solution deploys ML models in an AR environment, where the AR device maps the room and displays three objects scaled in space. Projections are then arranged virtually, and the device has the ability to exclude all objects in the augmented environment [19].

For this project, a demonstration program was developed as a proof of concept, along with a library that allowed communication between the AR device and the industrial controllers. The basis for the ML training resource were the photos taken with the AR device from different angles and different lighting intensities, while the training and prediction processing were done on a server (cloud). Therefore, the application rendering and the predictive processing happened on different devices [19].

In [19], eighteen neural network models were trained, with three different optimizers and six different activation functions. The optimizers used were stochastic gradient descent (SGD), ADAM, and RMSprop, while the activation functions used were ReLU, Leaky ReLU, ELU, and Tanh. Each model was trained five times to reduce the randomness of the RNA. To compare the models, the average accuracy was used. Models with sigmoid functions in conjunction with the SGD optimizer obtained the worst results with 71% accuracy for 1000 training epochs. The other models had accuracy close to 95%, with the hyperbolic tangent function reaching the value more quickly, while the optimizer with the best result was ADAN. The RNA was also implemented to control the actual industrial process by affecting robot movement.

Meanwhile, in [4], virtual objects are manipulated in AR environments in order to simulate the laws of physics, increasing the degree of realism perceived by the user. According to the authors, this type of interaction had not yet been explored until their research. The interaction takes into account not only the deformation capacity of inserted virtual objects but also that of real objects available in the real environment. For that, techniques of machine learning, computer graphics, and computer vision were used.

The study addressed some related works in the area of solid deformation and sought to define a machine learning model that could learn the behavior of deformable objects, a field of computational mechanics. Previous studies that dealt with the issue of user interaction were also presented, highlighting that [4] work aims for the user to also interact with real objects through collision estimation. The method used by [4] was demonstrated with examples of contact between two virtual objects and a real object and the interaction of virtual objects with the real environment. In the end, it is mentioned that the importance of this study is not only in entertainment but also in medicine and industry in general.

Developed by [24], DeepMix is a hybrid technique that combines mature deep neural networks (DNN)-based 2D object identification with lightweight on-device depth data processing. This results in minimal end-to-end latency and greatly improved detection accuracy in mobile circumstances. DeepMix, in particular, offloads just 2D RGB pictures to the edge for object recognition and uses the returned 2D bounding box to dramatically minimize the quantity of to-be-processed 3D data. The accuracy of typical methods is assessed to understand the influence of input-data quality on 3D object recognition; they employ point clouds with varied densities generated from depth pictures with different resolutions. Further, the inference time of classic 2D object recognition models like YOLOv4 is compared with the aforementioned typical 3D object detection models and a DeepMix prototype is evaluated on Microsoft HoloLens.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 October 2023 doi:10.20944/preprints202310.0200.v1

The typical 3D object detection models serched in [24] are separated into: *point-cloud-based 3D object detection* (VoteNet , COG, and MLCVNet), which can directly take raw point cloud as input and achieve high detection accuracy; *image-based 3D object detection*, that utilize 2D detectors to achieve 3D object detection, for example D4LCN, which estimates the depth information from monocular images and generate 3D bounding boxes; and *3D object detection with RGB-D input*, that utilizes both RGB images and depth data for 3D object detection, such as F-PointNet. Different from this models, DeepMix benefits from 2D object detection models that have low computation latency and utilizing real-time depth information from sensors, it can achieve high 3D object detection accuracy with low end-to-end latency. The end-to-end latency of DeepMix is only 34 ms, much lower than that of existing DNN-based models (ranging from 91 to 311 ms).

The approach used by [24], [19] and [4] to process the interaction of objects is a relevant factor when it comes to recognizing actions, since objects need to be mapped before the actions involved can be recognized. Thus, the aspects of action recognition are discussed in the next section.

3.3. Action recognition by frames

Here, we discuss works that address action recognition for AR/MR systems directly, considering object recognition as a support element for activities and actions. Among those that address the problem of action recognition, [13] proposes a non-intrusive and fast approach to support an MR interface. To achieve this, scene comprehension is used as an element of system interaction with the user. Because it is applied in real time, the model is not enough to be accurate; it also needs to be fast and use little memory so that it can be used on devices such as smartphones.

Therefore, a new standard for motion description is proposed, called Proximity Patch (PP), which uses a structure to ensure that all pixels around an object are used to calculate the descriptor. It was also presented as a new algorithm through PP called Binary Proximity Patches Ensemble Motion (BPPEM) to calculate the texture change. Finally, an extended version of BPPEM, the eBPPEM, was implemented as a small and fast descriptor for movement using three consecutive frames, which obtained competitive results for the Weizmann and KTH datasets. In future work, the use of classifiers that are more suitable for multi-class problems, such as Randon Forest, will be explored, as will the implementation of an AR application that uses the method as an interaction interface [13].

In parallel, the study [17] aims to recognize human-machine interaction through an egocentric vision between two users. Paired Egocentric Interaction Recognition (PEIR) is the task of recognizing the interaction of two people collaboratively through videos by correlating paired images. Using pairs provides more accurate recognition than a single view, helping to provide more accurate assistance for everyday situations. The experiments portray seven categories, including: Pointing, Attention, Positive, Negative, Passion, Receiving and Gesture. All results using two views together outperform models using only one view. It used bilinear pooling to capture all information in pairs in a consistent way, achieving success in the correlation between paired images with cutting-edge performance for the PEV dataset. The work did not describe limitations or suggest future work [17].

The studies [13] and [17] presented an action recognition method through image frames, because they developed models to map movements previously registered in the Weizmann, KTH and PEV datasets. This is a different method from the one used in [19] and [4], which needed to recognize the object to indirectly indicate the action. It is important to notice that studies such as [14] merge the object recognition proposed by [19] and [4] with the action recognition proposed by [13].

Accordingly, [14] seeks to recognize actions for application in an Intelligent Assistive Technologies (IAT) capable of supporting people with mild cognitive disabilities in cooking. To achieve this, the IAT works on two simultaneous fronts: one to understand the environment by capturing the position of the user and objects, or recognizing the user's actions to estimate the progress degree; and the other to provide interaction with the user through an interface to explain tasks, detect the progress of operations, provide the next task, and control the objects or the user from their location. To control the tasks, a finite state machine was used, which was activated based on inputs from the location of

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 October 2023 doi:10.20944/preprints202310.0200.v1

objects obtained from a Kinect ToF camera. The system achieved an accuracy of 85% for the action recognition of two objects (pot and cup) manipulated in a virtual kitchen with five different actions (Reach, Move, Mix, Hold, and Tilt) using a Random Forest classifier. Future activities aim to expand support for other everyday activities.

Even related to action recognition, the identification of human movements opens the way to another discussion that can be observed in the studies [15], [20], and [21], presented in the next section.

3.4. Human movements recognition

Assimilating the existence of the human body into AR/MR applications is one of the factors that brings immersion to the virtual world, which is why XR devices have evolved into adding features to follow the human movement in the virtual world, such as eye-tracking, hand-tracking, pupillometry, heart rate, and face cams, among others [25].

The work [21] proposes a compact model for recognizing human movements using 2D skeleton sequences. This sequence represents the configuration of the body's joints in relation to time and space. The counterpart of the technique is the lack of information in relation to depth, which is present in 3D techniques, which are more complex. However, the advantage of using 2D skeletons over HD video images can be summarized by the size of the information per second, where 2D skeletons occupy approximately three orders of magnitude less. To implement the model, three steps were used: skeleton capture, pre-processing, and action identification.

To capture the skeletons, two techniques considered state-of-the-art for detecting people [26] and estimating pose [27] were used. The pre-processing methods sum up to normalization to reduce the impacts of variance in position, direction, and height of the skeletons. It also reduced the variance of the *joint coordinate space scaling* by placing the values in a shorter range and of the *joint coordinate space quantization* by making the joint coordinates a discrete value, reducing the number of domains and thus reducing the error in the estimated positions [21].

To recognize actions, the Bi-LSTM network was used, which is a lightweight network for action recognition using 2D skeletons. The experiments showed that the technique has competitive results in relation to 3D skeleton techniques, such as dataset PKU-MMD, while also remaining competitive against other techniques that use a greater data intensity (2D skeleton, RGB, and heat map), such as dataset Action PENN. Thus showing the superiority of the 2D skeleton modality for task recognition, as it remains competitive, using less complexity and with easier training than the alternatives [21].

An important subcategory for action recognition of human movements in the AR/MR environment are techniques for estimating hand position. The work [20] reviews the techniques focused on estimating 3D hand poses using CNNs. The study aims to estimate a hand in 3D using tracking, parsing, contour and segmentation of the hand, fingertip detection, and recognizing gestures, among others. The general estimation difficulties are: low resolution, self-similarity, occlusion, incomplete data, annotation difficulties, hand segmentation, and real-time performance. Sixty studies were reviewed between 2015 and 2019, divided into categories based on the type of CNN, input data, and method.

The types of CNNs analyzed were 2D, 3D, or *no CNNs*. The input data was divided into the following categories: depth map, color image (RGB), stereo, RGB-D, point cloud data, and scaled gray image. The methods were divided into model-based, appearance-based, and hybrid methods. Model-based methods compare the estimated hypothetical hand and the real hand obtained from sensor data to measure the discrepancy between them. Appearance-based ones are based on learning features from a discrete set of annotated poses. The hybrid method uses a mixture of the two. The works are compared in three datasets: the ICVL, the NYU, and the MSRA. In addition, the performance between the techniques is compared with a frames per second (FPS) test [20].

The study concludes that when comparing 3D CNN with 2D CNN, the average error of techniques that use 3D CNN is smaller than that of 2D CNN, thus being superior in terms of accuracy. On the other hand, because the 3D CNN input data has a greater number of dimensionalities, the computational

 doi:10.20944/preprints202310.0200.v1

time is also greater. It is worth noting that most techniques do not produce results based on *datasets* from an egocentric view; these results tend to have a greater error due to the fact that occlusions occur more frequently. For comparison, the average error in *dataset* EgoDexter (with egocentric view) is 32.6mm, while the other *datasets* have a varying average error: ICVL from 6.28 to 10.4mm, NYU from 8.42 to 20.7mm, and MSRA from 7.49 to 13.1m [20].

Also in the field of hand-tracking, the work [15] uses action detection to recognize hand gestures, aiming to bring more commands for interaction in the virtual environment. For that, it is used the *inertial an inertial measurement unit* (IMU) system, which consists of several sensors placed on the user. Each IMU consists of a gyroscope, magnetometer, and accelerometer responsible for translating the user's movements. To perform the prediction, a CNN was trained to recognize a set of 22 gestures. Training data was comprised of 3D finger joint rotation data recorded from IMU gloves. To recognize actions in real time, the complexity of the network was reduced by decreasing the number of feature layers and weight parameters and by making the network find the archetypal characteristics of each gesture. Thus, maintaining high prediction accuracy.

The contribution of [22] aims to answer the challenge of semantic gaze analysis in 3D interactive scenes. It is demonstrated how new training datasets for the annotation of volumes of interface (VOIs) may be created using image augmentations with Cycle-GAN (Generative Adversarial Network) and accurately reflect virtual and even real environments. The machine learning approach is utilized to annotate VOIs at the feature level and only with synthetically generated training data, achieving state-of-the-art accuracy.

To assess the performance share of the picture augmentations, a ResNet50v2 architecture was trained only on the unaugmented simulation dataset, i.e. on the underlying 100,000 simulation images. Research demonstrated that sophisticated, extensively trained neural networks may not necessarily produce the greatest results. This was notably true for Cycle-GAN, where superior results were obtained utilizing 50 epochs rather than 200. The authors think that this is due to overfitting effects generated by homogeneous training data. When utilizing sophisticated architectures like ResNet50v2 or Cycle-GAN, it is recommended that the models be trained on a small number of epochs initially to determine whether a larger number of epochs is necessary. Further research is expected to enhance prediction accuracy and image augmentation approaches [22].

Another aspect of object, action, and human movement recognition is being able to process them in real-time interactive systems, which is addressed in the next session.

3.5. Processing in real-time interactive systems

To recognize actions and objects, ML techniques require high speed for predictions as they are a real-time system. This concern was addressed by [16], aiming for real-time object recognition for AR environments through a device that adds contextual awareness to the user by detecting and classifying the object with ML. Here, the term *context awareness* was used in place of *action recognition*, since it allows the system to provide an action based on the environment related to the task in progress.

It is worth noting that the hardware capabilities of MR devices are generally used to process renderings of virtual objects in the augmented environment and that these devices are not capable of detecting objects and rendering them at the same time, as in [13]. Therefore, they proposed a separation at the hardware level between the rendering of virtual components and object detection with ML, which sends labels that will be rendered on the MR device. Objects are detected by cameras with a 360-degree view of the user's environment, and the images are processed on an NVIDIA Jetson Nano that sends the processing results to the Microsoft HoloLens device. Note that the proposed solution separates rendering and prediction processing on different devices.

As a proof of concept, [13] implemented the new approach by assembling a Christmas tree. In this use case, the system identifies which ornament the user picked up and indicates the position that should be placed on the tree in the environment through a virtual projection, so that assembly is carried

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 October 2023 doi:10.20944/preprints202310.0200.v1

out more efficiently. The development started in three stages: object recognition, communication, and RM.

For object recognition, it was used YOLOv3 and GoogLeNet, aiming to compare them, which were processed on the Jetson GPU using NVIDIA's TensorRT. It was noticed that YOLOv3 is more robust than GoogLeNet, but only the objects desired in the application needed to be recognized, so GoogLeNet ended up being used since it was faster on the NVIDIA Jetson Nano. In terms of performance, YOLOv3, which identifies a greater number of objects, spent an average of 0.347 seconds to process each iteration, while GoogLeNet, focused on identifying only the object of interest, spent 0.054 seconds, 6.4 times faster [13]. Therefore, it can be seen that a solution that identifies objects in a broader way may suffer limitations when considering the real-time requirement.

Regarding communication between devices, the time spent was an average of 0.057 seconds for transmitting 1000 messages under the MQTT protocol. In the last stage, where the object label was printed in the projection of the virtual Christmas tree, the rendering time was considered insignificant. The MQTT protocol proved to be important for real-time communication between the headset and the NVIDIA Jetson Nano. In the future, they intend to evaluate the usability of the proposed approach with potential users [13].

3.6. Other aspects to recognize

Another studied aspect of action recognition in augmented and mixed reality is the recognition of emotions. The study [5] discusses different techniques and datasets for emotion recognition and carries out an experiment using Microsoft HoloLens for emotion recognition in MR. The MHL was chosen over its competitors, such as *Google Glass* and *Meta Quest 2*, for being wireless and containing important features for the experiment. Features such as environmental recognition, human interaction with holograms, light sensors, and a depth camera, among others, make the MHL headset robust to various environmental conditions.

In the datasets used for the training of the recognition models, different types of images were observed: 2D, 3D, or thermal, while the capture could be classified as artificial or spontaneous. From the study of techniques and datasets, it is concluded that 2D RGB datasets do not have intensity labels, making them less convenient for the experiment and compromising efficiency. Meanwhile, thermal datasets do not work well with variations in human poses, temperature, and age because they have low resolution. The 3D datasets are not available in sufficient quantity to carry out experiments, despite providing the best accuracy [5].

What concerns the classification of the dataset capture is that the first *artificial* ones express more extreme emotions, but they are easier to create. On the other hand, the *spontaneous* ones are more natural; however, annotation is costly, which is why a hybrid dataset is the ideal one. The best results in terms of accuracy involve using CNN, reaching up to 97.6% of accuracy, SVM with *clustering*, with 94.34% of accuracy, and Pyramid Histogram of Oriented Histogram (PHOG), with 96.33% of accuracy. The models trained with MHL images were superior compared to the ones trained with webcams in similar scenarios, thus demonstrating that sensors are extremely important for recognizing emotions, both their quantity and quality [5].

At the same time, [7] stands that the combination of AR/MR and deep learning technologies can provide an important contribution to interactive, user-centered real-time systems. The study highlights limitations in AR/MR related to handling camera location, object and user tracking, lighting estimation of a scene, registration of complex scenes with image classification, context-based image segmentation, and text processing, pointing to deep learning techniques as a solution to these problems. Therefore, the use of these technologies guarantees an improved user experience in dynamic, adaptive applications with customizable digital content in both virtual and real environments. As a suggestion for future works, [7] suggests the development of a smart application that combines the technologies and recognizes objects in various conditions, providing relevant information about them in an interactive way.

 doi:10.20944/preprints202310.0200.v1

Another interesting aspect to mention is the investigation of new object identification and classification methods in [23], which take use of super-resolution (SR). A low-footprint Generative Adversarial Network (GAN)-based system capable of taking low-resolution input and producing an SR-supported recognition model is provided. High-resolution pictures can provide excellent reconstruction quality for imaging data, which can be used in a variety of real-world settings, including satellite and medical imaging, facial recognition, security, and others. A considerable amount of previous research has developed CNN designs and functions that calculate loss between the reconstructed and actual picture, specifically Mean Square Error (MSE), to improve the performance of these networks, as well as GANs to successfully apply SR. The Enhanced Super-Resolution Generative Adversarial Networks (ESRGAN) model is utilized in citeli2022super to handle the Super Resolution issue.

It is worth mentioning that SR processing can increase object identification and classification performance in a variety of use-cases from medium to distant ranges. FRRCNN, Retina, and YOLOv3 object identification algorithms that have demonstrated better performance in terms of accuracy and inference time on well-known datasets in recent years were employed in [23]. While the Retina and YOLOv3 models improved significantly in terms of accuracy in the last experiment after being retrained, the FRRCNN did not generalize as well as the others. YOLOv3, on the other hand, received the greatest ratings of the three models and should be researched further. The YOLOv3 model was better at identifying and categorizing small items, but struggled with bigger ones.

3.7. Challenges and opportunities analysis

This section addresses the research question *Q2: What are the challenges and opportunities for using machine learning for object and action recognition in an augmented and mixed reality environment?*.

When analyzing the selected works, one can notice common problems with possible common solutions or limitations. Some of the studies showed the importance of analyzing the topic of ML in AR/MR application in the industry field ([4] [7] [19]) and its application in equipment maintenance ([16]). Others highlight the context awareness provided by intelligent systems [7] [16] [19] [22]. The user interaction with the virtual environment was also observed as an element of study in [4] and [7] and is an important element in improving the experience in the virtual environment in [7] and [18].

Analyzing the challenges of the works presented, it is possible to recognize four main technical limitations for applying machine learning in the context of augmented and mixed reality, namely: *context limitations*, *real-time processing*, *communication protocols*, and *environmental conditions*. In this section, we will discuss these limitations as well as their possible solutions.

Context limitations are noticed in techniques that perform well in a specific and limited context. This problem may occur due to limitations in the network, as in [13], [14], [20] and [22], where the techniques perform well in their limited universe of context, but there are difficulties in expanding them for more general uses. In this case, the solution would be to change the ML model in order to implement a more robust one. The second cause of context limitation is limited datasets. The works [14], [13], and [19] reported action recognition methods, but only in datasets for specific purposes. One of the most impacted areas by limited datasets is object and action recognition in an egocentric view, as in [17] and [20].

The unaugmented dataset in [22] does not appear to sufficiently represent the experimental data, particularly in the real world. On the other hand, the suggested image augmentation utilizing Cycle-GAN appears to be capable of compensating for these reality deficits to a significant degree. Manually generating large and high-quality datasets for training CNNs is a time-consuming and economically inefficient operation. As can be observed, network trained on synthetic datasets substantially greater accuracy with less time input, making it more economically viable. The study should ideally be replicated with various use cases other than the coffee machine. Nevertheless, this requires the collection of a database with annotated ground truth labeling, which is a time-consuming manual process.

 doi:10.20944/preprints202310.0200.v1

Machine learning models must be able to serve real-time interactive applications. To achieve this, the models must consider some requirements related to communication, data complexity, processing, and application scope. In [18], the challenges of computer networks for AR services are discussed, as is whether they meet the requirements that are demanded by the next generation of applications. With the development of new hardware for AR/MR applications supported by machine learning, a new era of applications is beginning that allows better user interaction between the physical world and the virtual world.

According to [24], mobile headsets should be capable of recognizing 3D physical settings in order to provide a genuinely immersive experience for augmented/ mixed reality (AR/MR). Nevertheless, their compact form-factor and restricted computer resources make real-time 3D vision algorithms, which are known to be more compute-intensive than their 2D equivalents, exceedingly difficult to implement. The large calculation overhead frequently leads in considerable data processing delay. Furthermore, the quality of input data has a significant impact on the performance of 3D vision algorithms (e.g., point cloud density or depth image resolution). As a result, present AR/MR systems are mostly focused on 2D object detection.

The objective of [18] is to evaluate networks, carrying out quantitative and qualitative analysis, in order to identify obstacles, such as latency problems and a lack of quality in cloud services. It was verified that the cloud latency for Microsoft HoloLens was low, which highlighted other network challenges for AR applications in addition to bandwidth and latency. The results showed that new network protocols for cutting-edge AR applications are needed, as latency makes image processing and computer vision in the cloud unfeasible. Thus, a Named Data Networking (NDN) was suggested as a network solution, the Low Latency Infrastructure for Augmented Reality Interactive Systems (LLRIS), which offers support for service discovery, task offloading, computing reuse, and caching, promoting a hybrid model for cloud computing.

According to [18], the concept of offloading compute-heavy operations to cloud/edge servers is also a potential solution for speeding up 3D object identification. Their DeepMix technique is reported to be the first to reach 30 FPS (i.e., an end-to-end latency substantially lower than the 100 ms demanding criterion of interactive AR/MR). DeepMix sends just 2D RGB photos to the edge for object recognition and uses the returning 2D bounding box to dramatically minimize the quantity of to-be-processed 3D data, which is done similarly by [21].

In the work [5], the existing real-time processing limitations are caused by the use of complex data, centralized processing, and heavy networks. A closer look at this limitation shows, as seen in [16], that AR devices are not capable of processing rendering and ML predictions simultaneously. To solve this problem, two approaches were observed: one that tries to separate processing on different devices, as in [16], and a second approach that aims to create ML models with faster predictions, as in [14] and [17]. Another approach that seeks to solve this limitation is presented in [21] and involves replacing the use of images with the representation of 2D skeletons, reducing the complexity of the data. Another study, [16], reinforces the need to reduce inputs to enable processing on the AR device.

These solutions have their own limitations. For example, the attempts to make a ML model that provides general object recognition, as seen in [16], cause limitations related to the real-time processing. To overcome this problem, a trade-off is suggested, since processing a model built to recognize only specific objects for a desired application would be faster and lighter. Meanwhile, separating rendering and prediction processing on different devices, whether on a dedicated and embedded device or in the cloud, needs to establish proper communication between all the parts. The communication protocol between these devices must be carefully selected, as it requires low latency in addition to affecting the pre-processing of large amounts of data to extract skeletons [16]. To overcome the communication problem between devices, a particular strength of [22] approach is, that in comparison to other methods for semantic gaze analysis, neither markers nor motion tracking systems are required, which minimizes the risk of errors through latencies and registration failures.

 doi:10.20944/preprints202310.0200.v1

The limitations regarding environmental conditions can be seen in [22], [20] and [7]. Limitations related to occlusions are highlighted in [4], which can be solved through object segmentation. Other errors appeared when measuring the depth of objects, caused by areas not visible by cameras due to lighting problems, which could be solved through texture and perspective corrections. In [22], when made of translucent or reflective materials, VOIs may be hidden behind other portions and only partially visible, which is addressed by combining simulation with the image enhancement technique.

In [20], techniques are used to limit the problem of environmental conditions such as lighting and position variation; however, there is no solution for occlusion, which is present mainly in egocentric vision datasets. The approach of using AR through an egocentric view of the user in relation to the scene, as seen in [17] and [20], was little observed, and it is considered an important limitation.

The opportunities identified in the reviewed works can be summarized as an official deployment of the applications developed, expansion of the use cases and exploration of the use of techniques on AR/MR multiplatforms/devices. The deployment is primarily limited by real-time processing and is identified in [13], [16], and [7], but could bring a lot of feedback from users that could result in improvements for the applications. Expanding the use cases is an opportunity brought by [14], [13], and [21]. However, it requires changes in the network model used or the evolution of the datasets for training, identified mainly in the egocentric vision area.

The opportunity of exploring the ML techniques to multiple AR/MR platforms and devices is mentioned by [22], [23] and [24]. The techniques shown could be integrated into small, mobile and low-power AR/VR devices, mobile eye tracking used in combination with Powerwalls, CAVEs or real prototypes and HMD integrated eye tracking, among others. The current generation of AR/MR headsets is in charge of performing computation-intensive activities locally. According to [24] in the future, with new network technologies like as 5G and beyond, the majority of jobs requiring intensive computing will be offloaded to faraway cloud/edge servers, fully using the mobility of headsets.

4. Conclusions

In this study, fifteen articles were reviewed to understand how machine learning is being used in augmented and mixed reality to improve the recognition of objects and actions. Our results indicate that there is great demand for ML and AR/MR technologies to be explored in daily activities, industry, medicine, entertainment, personal assistance, and operational training, bringing a more immersive experience to the user.

In this scenario, two research questions were satisfactorily addressed. To answer *Q1: What is the state of the art on using machine learning for object and action recognition in an augmented and mixed reality environment?*, techniques were mapped and qualitatively discussed, aiming for *Object interaction recognition* (3.2), *Action recognition by frames* (3.3), *Human movements recognition* (3.4), *Real-time interactive systems processing* (3.5) and *Other aspects* (3.6).

To answer *Q2: What are the challenges and opportunities for using machine learning for object and action recognition in an augmented and mixed reality environment?*, four main technical limitations were discussed, namely *context limitations*, *real-time processing*, *communication protocols*, and *environmental conditions*, while three main opportunities, official deployment of the applications developed, expanding the use cases and exploration of the use of techniques on AR/MR multiplatforms/devices, were highlighted (3.7).

Concerning future research, there are still unresolved problems when it comes to real-time interactive systems, as it involves low prediction time in ML models, low-latency network communication, and limitations in object and action recognition models for a broad context. It is suggested that dedicated processing hardware be added next to the capture device. Therefore, more research is needed to investigate how AR can work together with ML to meet real-time interactive demands.

Author Contributions: Conceptualization, T.P.P, R.B.L. and E.G.S.N.; methodology, I.L.C. and I.W.; validation, I.L.C., I.W., A.A.B.S. and T.B.M.; formal analysis, I.W. and I.L.C.; investigation, T.P.P, R.B.L. and I.L.C.; resources,

I.W., A.A.B.S. and T.B.M; data curation, T.P.P, R.B.L. and I.L.C.; writing—original draft preparation, T.P.P and R.B.L.; writing—review and editing, I.L.C., I.W., A.A.B.S. and T.B.M.; visualization, T.P.P, R.B.L., E.G.S.N. and I.L.C.; supervision, I.W., A.A.B.S. and T.B.M.; project administration,I.W., A.A.B.S. and T.B.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Acknowledgments: The authors would like to thank the financial support from the National Council for Scientific and Technological Development (CNPq). Ingrid Winkler is a CNPq technological development fellow (Proc. 308783/2020-4).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Tori, R.; Hounsell, M.d.S. Introduçao a realidade virtual e aumentada. *Porto Alegre: Editora SBC* **2018**.
2. Azuma, R.T. The road to ubiquitous consumer augmented reality systems. *Human Behavior and Emerging Technologies* **2019**, *1*, 26–32.
3. Azuma, R.T. A survey of augmented reality. *Presence: Teleoperators & Virtual Environments* **1997**, *6*, 355–385.
4. Badías, A.; González, D.; Alfaro, I.; Chinesta, F.; Cueto, E. Real-time interaction of virtual and physical objects in mixed reality applications. *International Journal for Numerical Methods in Engineering* **2020**, *121*, 3849–3868.
5. Mehta, D.; Siddiqui, M.F.H.; Javaid, A.Y. Facial emotion recognition: A survey and real-world user experiences in mixed reality. *Sensors* **2018**, *18*, 416.
6. Nogales, R.E.; Benalcázar, M.E. Hand gesture recognition using machine learning and infrared information: a systematic literature review. *International Journal of Machine Learning and Cybernetics* **2021**, *12*, 2859–2886.
7. Lampropoulos, G.; Keramopoulos, E.; Diamantaras, K. Enhancing the functionality of augmented reality using deep learning, semantic web and knowledge graphs: A review. *Visual Informatics* **2020**, *4*, 32–42.
8. Creswell, J. W.; Creswell, J.D. *Research design: Qualitative, quantitative, and mixed methods approaches*; Sage publications, 2017.
9. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; others. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Systematic reviews* **2021**, *10*, 1–11.
10. Booth, A.; Sutton, A.; Clowes, M.; Martyn-St James, M. Systematic approaches to a successful literature review **2021**.
11. How Scopus Works. Available online: https://www.elsevier.com/solutions/scopus/how-scopus-works/content. (Accessed on 7 December 2022).
12. Matthews, T. LibGuides: Resources for Librarians: Web of Science Coverage Details. Available online: https://clarivate.libguides.com/librarianresources/coverage. (Accessed on 7 December 2022).
13. Fangbemi, A.S.; Liu, B.; Yu, N.H.; Zhang, Y. Efficient human action recognition interface for augmented and virtual reality applications based on binary descriptor. International Conference on Augmented Reality, Virtual Reality and Computer Graphics. Springer, 2018, pp. 252–260.
14. D'Agostini, J.; Bonetti, L.; Salem, A.; Passerini, L.; Fiacco, G.; Lavanda, P.; Motti, E.; Stocco, M.; Gashay, K.; Abebe, E.; others. An augmented reality virtual assistant to help mild cognitive impaired users in cooking a system able to recognize the user status and personalize the support. 2018 Workshop on Metrology for Industry 4.0 and IoT. IEEE, 2018, pp. 12–17.
15. Hansberger, J.T.; Peng, C.; Blakely, V.; Meacham, S.; Cao, L.; Diliberti, N. A multimodal interface for virtual information environments. International Conference on Human-Computer Interaction. Springer, 2019, pp. 59–70.
16. Dasgupta, A.; Manuel, M.; Mansur, R.S.; Nowak, N.; Gračanin, D. Towards real time object recognition for context awareness in mixed reality: a machine learning approach. 2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW). IEEE, 2020, pp. 262–268.

17. Li, Z.; Lyu, F.; Feng, W.; Wang, S. Modeling Cross-View Interaction Consistency for Paired Egocentric Interaction Recognition. 2020 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2020, pp. 1–6.
18. Shannigrahi, S.; Mastorakis, S.; Ortega, F.R. Next-generation networking and edge computing for mixed reality real-time interactive systems. 2020 IEEE International Conference on Communications Workshops (ICC Workshops). IEEE, 2020, pp. 1–6.
19. Kozek, M. Transfer Learning algorithm in image analysis with Augmented Reality headset for Industry 4.0 technology. 2020 International Conference Mechatronic Systems and Materials (MSM). IEEE, 2020, pp. 1–5.
20. Le, V.H.; Nguyen, H.C. A survey on 3D hand skeleton and pose estimation by convolutional neural network. *Advances in Science, Technology and Engineering Systems* **2020**, *5*, 144–159. doi:10.25046/aj050418.
21. Elias, P.; Sedmidubsky, J.; Zezula, P. Understanding the limits of 2D skeletons for action recognition. *Multimedia Systems* **2021**, pp. 1–15.
22. Stubbemann, L.; Dürrschnabel, D.; Refflinghaus, R. Neural Networks for Semantic Gaze Analysis in XR Settings. ACM Symposium on Eye Tracking Research and Applications, 2021, pp. 1–11.
23. Li, V.; Amponis, G.; Nebel, J.C.; Argyriou, V.; Lagkas, T.; Ouzounidis, S.; Sarigiannidis, P. Super resolution for augmented reality applications. IEEE INFOCOM 2022-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS). IEEE, 2022, pp. 1–6.
24. Guan, Y.; Hou, X.; Wu, N.; Han, B.; Han, T. DeepMix: mobility-aware, lightweight, and hybrid 3D object detection for headsets. Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services, 2022, pp. 28–41.
25. Nair, V.; Rosenberg, L.; O'Brien, J.F.; Song, D. Truth in Motion: The Unprecedented Risks and Opportunities of Extended Reality Motion Data. *arXiv preprint arXiv:2306.06459* **2023**.
26. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6517–6525. doi:10.1109/CVPR.2017.690.
27. Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep High-Resolution Representation Learning for Human Pose Estimation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.