BULETINUL INSTITUTULUI POLITEHNIC DIN IAŞI
Publicat de
Universitatea Tehnică „Gheorghe Asachi” din Iaşi
Volumul 69 (73), Numărul 3, 2023
Secţia
ELECTROTEHNICĂ. ENERGETICĂ. ELECTRONICĂ
DOI:10.2478/bipie-2023-0018

A SHORT REVIEW OF DEEP LEARNING METHODS IN VISUAL SERVOING SYSTEMS

BY

ADRIAN-PAUL BOTEZATU* and ADRIAN BURLACU

“Gheorghe Asachi” Technical University of Iași,
Faculty of Automatic Control and Computer Engineering, Iași, Romania

Received: July 31, 2024
Accepted for publication: October 14, 2024

Abstract. This survey explores the evolution and applications of Visual Servoing Systems in robotics, emphasizing the transition from traditional image processing techniques to the incorporation of neural networks for feature extraction and control. Robotic systems, integral to manufacturing, surveillance, and healthcare, increasingly rely on Visual Servoing for enhanced interaction within various work environments. Initially focused on auxiliary sensors like visual sensors for robustness and accuracy, recent advances have seen a shift towards integrating deep learning methods for direct control and feature extraction. The survey covers the differences that emerge from classical Visual Servoing architectures to novel methods involving Deep Learning, highlighting their respective advantages and limitations regarding stability, precision, and real-time applicability. Innovative approaches, such as Direct Visual Servoing and the use of siamese networks for camera position estimation, demonstrate significant progress in overcoming the challenges of traditional Visual Servoing. Through detailed examination of leading research, the survey highlights the potential of neural networks to revolutionize this domain by enhancing feature extraction,

*Corresponding author; *e-mail*: adrian-paul.botezatu@academic.tuiasi.ro

reducing reliance on precise calibration, and improving control laws for complex robotic tasks.

Keywords: visual feedback control, neural networks, robotic systems.

1. Introduction

Robotic systems are playing an increasingly important role in fields that involve the manufacturing of parts, surveillance systems, health, and others. These areas involve robots operating in various work environments with limited information. Recent research in the field of robotics aims to use auxiliary sensors that can improve the system's robustness and accuracy. One of the most used sensors in this case is the visual sensor, which, through the complete information it provides, gives the robot the ability to interact with the work environment without coming into contact with it. The concept of extracting visual features from images acquired with the help of the visual sensor, viable for controlling the robot's movements, is known as Visual Servoing System.

The term "visual servoing" was introduced in (Hill and Park, 1979), where the authors wanted to distinguish between their theoretical approaches and the then-current literature, which tried various experiments with a system that grasped and moved objects. The concept of Visual Servoing System has been attributed to those cases where a robot, in the control loop, uses information from a visual sensor. In recent years, the fusion of several research areas such as image processing, kinematics, dynamics, control theory, artificial intelligence, and real-time programming has been used in this new field of interest. Thus, the purpose of a Visual Servoing System has been defined as the ability of a robot to solve specific tasks in its environment, using information extracted from an image (Hancock and Langer, 1998). However, the fundamental operating principle of a Visual Servoing System should not be to interpret the scene and then model it entirely, but to focus its attention on certain areas of interest relevant to the task it has to perform.

Therefore, the performance of visual feedback starts with the selection of visual features that lead to properties such as stability, robustness, and accuracy. In (Mahony *et al*., 2002; Chaumette, 2004), and (Marchand and Chaumette, 2005), these types of features are introduced. By selecting high-quality features, such as points of interest (centroids, corners, edges) or image moments, it becomes possible to develop control laws that achieve the desired properties of stability, robustness and accuracy.

A Visual Servoing System can have two distinct camera configurations: one where the camera is mounted in a fixed position in the work space, eye-to-hand, and the other where the camera is mounted on the last joint of the robotic manipulator (end-effector), eye-in-hand. Visual Servoing Systems are divided into three types, depending on the architecture they use: position-based visual

servoing (PBVS - Position Based Visual Servoing), image-based visual servoing (IBVS - Image Based Visual Servoing), and the hybrid approach of the first two (Chang, 2007).

A PBVS system uses 3D information about the work scene. A coordinate system is attached to the camera, and thus the position and direction of an object can be estimated relative to this coordinate system. In the Cartesian space of the robot, the error is defined as the difference between the current position and the desired one. Figure 1 represents a block diagram of the architecture of a position-based servoing system, where X_c^* represents the desired posture, X_c is the current posture, and the error e, the difference between the two, is used in the position-based controller to calculate the speed of the robot.

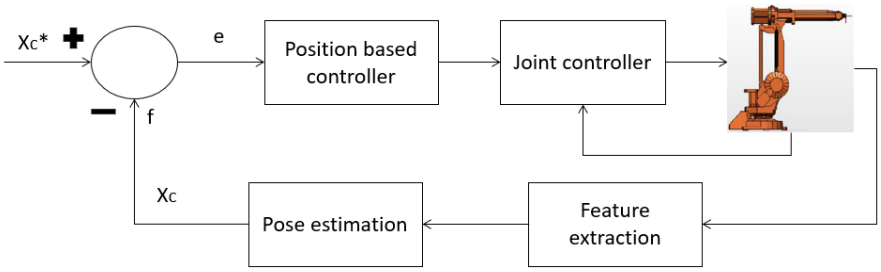

Fig. 1 – Diagram for PBVS control.

Starting from the premise that a task of a Visual Servoing System can be to achieve and maintain a constant posture between a certain object and the robot's end-effector, Hutchinson defines in (Hutchinson *et al*., 1996) this concept as being an EOL (end point open loop) system in which only the target of the object is observable. From this arises a principal advantage of a PBVS structure, namely the control that can be made in the Cartesian coordinates of the robot's trajectory and the camera.

Although it seems that trajectory planning could be more easily achieved, according to (Chaumette, 2002), in the case of an eye-in-hand configuration (camera positioned on the manipulator's effector), it can happen that some image features for an estimated position are outside of the image. Because of this, it's possible that the current position and the desired position of the camera might not be accurately estimated, which can lead to poor performance in terms of precision. The solution presented in (Hutchinson *et al*., 1996) is to regard the servoing system as an ECL (end point closed loop) system, where both targets can be observed in a task of the system. Due to these characteristics that influence the system's stability, the architecture of position-based visual servoing has proven to be limited regarding robustness and the mathematical modeling for real-time implementations. Other disadvantages of a PBVS architecture include the need to know both the object's position and the need for precise camera calibration.

IBVS uses features extracted from 2D images acquired to directly model the control law for commanding the robot. The aspect that is pursued in using this architecture is the minimization of the error between the desired image features, as compared to the current image features (Chaumette and Hutchinson, 2006). Fig. 2 illustrates the block diagram of the architecture of an image-based servoing system, where the difference between f^* (the desired features obtained from an image corresponding to a certain position of the effector relative to the object) and f (the feature vector) represents the error which is used in the image-based controller to calculate the robot's speed.

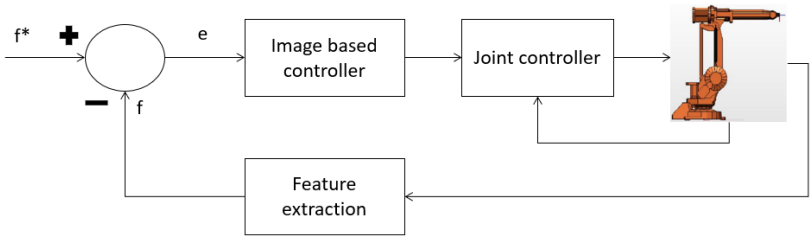

Fig. 2 – Diagram for IBVS control.

This architecture involves the calculation of a Jacobian matrix or interaction matrix (Chaumette and Hutchinson, 2008; Chaumette *et al.*, 1991), (Mahony *et al.*, 2002), which will mathematically express the relationship between the variations of the features in the image caused by the movement of the camera. Thus, the construction of the servoing system will be based solely on image information and should be robust to calibration errors since there is no longer a need to approximate the location of an object of interest in a 3D space. This idea is supported by Fig. 3 (Copot, 2012), which presents an example of the image-based architecture.

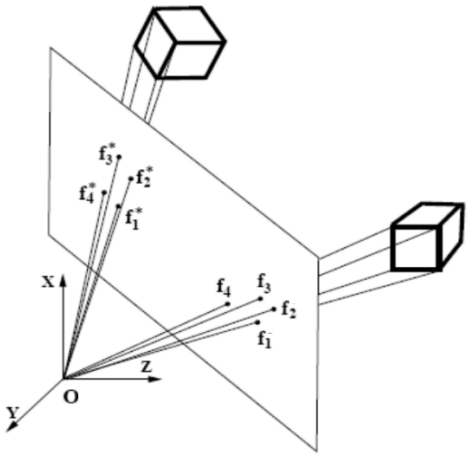

Fig. 3 – Example of IBVS architecture.

In Fig. 3, an eye-to-hand configuration is used, where it is assumed that the object is moving while the camera is fixed. The current feature vector f is generated using a number of object features. The error function is defined as the difference between the current and desired features. This information, calculated together with the interaction matrix in each acquired frame, is further used in the robot's control law.

Compared to the position-based architecture, IBVS offers several advantages (Chaumette, 1998) due to the fact that it is not as sensitive to camera calibrations. In the case of IBVS, miscalibration can be corrected by minimizing the squared error, meaning that in the event of a possible miscalibration, the robot can reach the desired position. Thus, a more pronounced dependence on the precision with which the camera's relative position is determined is observed in position-based servoing systems. However, both architectures present disadvantages due to convergence and stability (Chaumette, 1998; Malis, 1999), such as the case of the PBVS architecture which requires geometric information about the location of the object of interest and the fact that it is necessary for the extracted points to be outside the visual field of the next frame. On the other hand, in the eye-to-hand architecture, if objects of interest are placed in certain static zones, the IBVS architecture presents disadvantages in minimizing the error (Chesi *et al*., 2004). Although presented as an advantage in comparison with PBVS, the interaction matrix brings certain disadvantages due to the need for a priori knowledge of the camera's intrinsic parameters.

The third architecture mentioned, the hybrid one, has tried to reduce the disadvantages mentioned so far because it does not require camera calibrations nor knowledge of the 3D model of the object and the work scene. In (Malis *et al*., 1999), it is shown that this has better stability properties and does not have the same convergence and stability issues as those presented by Chaumette in (Chaumette, 1998). However, a disadvantage of using the hybrid architecture comes from sensitivity to various noises that can disturb the image (e.g., low brightness). In (Collewet and Chaumette, 2002), the authors used the concept of hybrid architecture by dividing the task the servoing system had to perform into two tasks, a primary and a secondary one. The primary task was to keep the image of interest centered to also maintain the features of interest in the camera's visual field. The secondary task was to find different positions while keeping the object centered in the image so as to bring the camera to the desired position.

Both in the case of IBVS and in the case of PBVS (and implicitly in the case of hybrid architectures), visual features extracted from images of interest are used to characterize the objects in the image plane in a certain form. In the case of the position-based architecture, these properties are used with a view to correlating with the 3D space of the plane, and in the case of the image-based architecture, visual features are used in the construction of the interaction matrix. Over years of research, special attention has been paid to the area of image processing algorithms through which these visual features have been extracted.

The disadvantage arises in attempts to generalize the finding of features of interest, regardless of the shape or texture of the object or the lighting conditions that could disturb the images from which these features are extracted.

All the mentioned disadvantages can be countered by using the concept of a neural network. In this case, a neural network will learn various patterns and features of objects during the training process and will be able to apply these learned features to recognize new, unseen objects.

2. From Classical methods to Deep Learning methods

A disadvantage of the image-based visual servoing system architecture was that features from images had to be extracted through classical image processing methods. With the emergence of neural networks, these began to be used as feature extraction structures in control architectures. Thus, in Fig. 2, the feature extraction block will be replaced with a neural network architecture whose role will consist in extracting features.

In (Ahlin *et al*., 2016), the authors use AlexNet to detect and track damaged leaves, and then use the SURF (Speeded-Up Robust Features) descriptor for feature extraction. The points of interest are further given as input to an IBVS controller that performs the robot control. An IBVS architecture is also used in (Cheng *et al*., 2018) where the authors train and test Faster R-CNN for both object detection and feature determination. Based on these, the error is calculated as the difference between the desired features and those obtained, so that this information is later used in the control law.

A different architecture is proposed in (Liu and Li, 2019) where, from an architectural point of view, a convolutional neural network is used on two streams, as shown in Fig. 4. The network receives as input the image from the desired configuration and the image from the current configuration, and the output is given by the control law for a robot with 4 degrees of freedom. The estimated position and orientation parameters are further used as input by an IBVS-based controller.

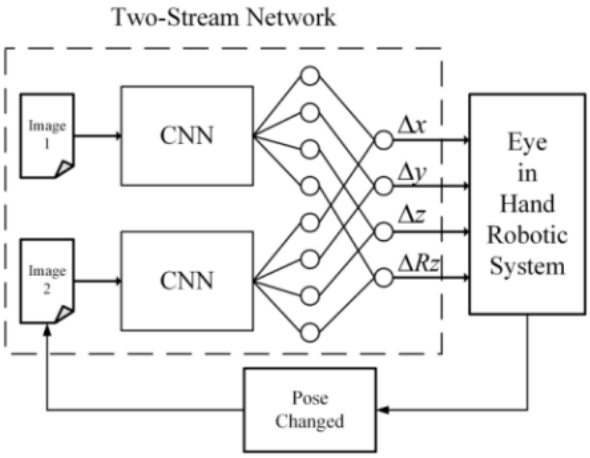

Fig. 4 – Two stage CNN architecture proposed in (Liu and Li, 2019).

Another approach is described in (Harish, 2020), where the authors use FlowNet, a network with very good results obtained in feature determination. The network receives as input the image from the desired configuration and the image from the initial configuration and directly predicts the error as the difference between the desired and initial features. This information, together with the depth information, is further used by an IBVS controller. The architecture proposed by them is represented in Fig. 5. A dashed line describes a version of the architecture in which the depth information is calculated from disparity maps.

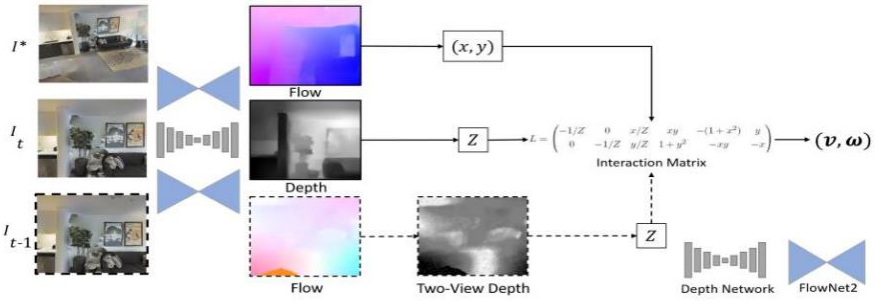

Fig. 5 – FlowNet architecture from (Harrish *et al.*, 2020).

Another analyzed paper where a neural network is used solely as a feature extractor is presented in (Nicholas *et al*., 2022) where the authors employs KeyPointNet (Tang *et al.*, 2020), an architecture built for extracting features. Pairs of images representing the desired and initial configurations are input into the neural network, and the pipeline includes feature detection, matching features in the two images, and eliminating features that did not match. Using depth

information as well, the information is further processed by an IBVS controller through linear and angular velocities. The architecture is schematically represented in Fig. 6.

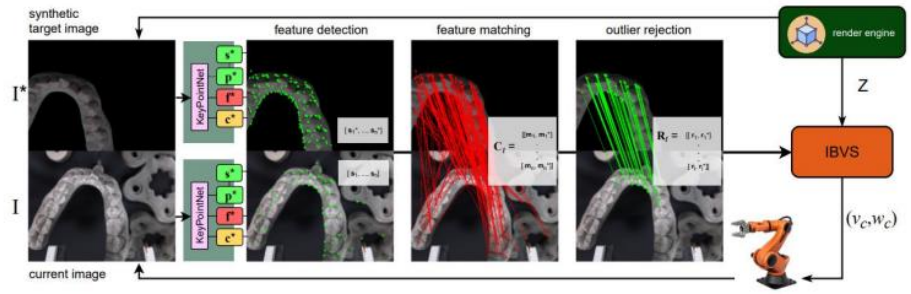

Fig. 6 – KeyPointNet architecture from (Nicholas *et al*, 2022).

3. From Direct Visual Servoing to Deep Learning methods

As shown so far, remarkable progress has been made in how features are extracted from images, from classical image processing methods to the use of neural networks. Another way in which the disadvantages of IBVS and PBVS have been diminished in applications was the emergence of Direct Visual Servoing systems, which use the entire image without the need for feature extraction. This method is based on finding the differences between the image in the initial configuration and the image in the desired configuration. One of the most appreciated works that introduce Direct Visual Servoing is by Bateux (Bateux, 2018), in which the calculation of histograms as a similarity function, by analyzing the differences between the histogram of the initial image and the histogram of the desired image, is proposed. With the help of a set of virtually obtained image data, it tries to minimize the error represented by the similarity function. Based on the obtained squared error, a CNN architecture based on VGG16 is used to learn the control law of a six-degree-of-freedom robot.

Real progress has been made in the field of Direct Visual Servoing even without using CNNs. For example, in (Marchand, 2019), where two control laws are proposed that are based on the principle of decomposing the main components from the image, or in (Marchand,2020), where an interesting idea is proposed by not considering the image as a whole in its space but transforming it into the frequency domain through the Discrete Cosine Transform method. This method involves representing the image as a sum of cosine functions of different frequencies. The coefficients of this transform are subsequently used in the control law. However, as shown in (Bateux, 2018) and (Marchand, 2019), a disadvantage of this method is the nonlinearities that can be induced in minimizing the cost function due to the small convergence space (compared to IBVS or PBVS). For this reason, special attention has been paid to the use of

neural structures not just as feature extractors, but also in the control part. The most popular and cited works developed are presented and commented on next.

The first significant work in this regard is from (Saxena *et al*., 2017) where the authors use FlowNet to develop a visual servoing system that performs in different working environments, without knowing its geometry. With pairs of images of the initial and desired configurations as input, the network predicts the necessary homogeneous transformations to reach a desired configuration from an initial one. The tests developed by the authors were done on a quadcopter, with promising results both in indoor and outdoor working environments.

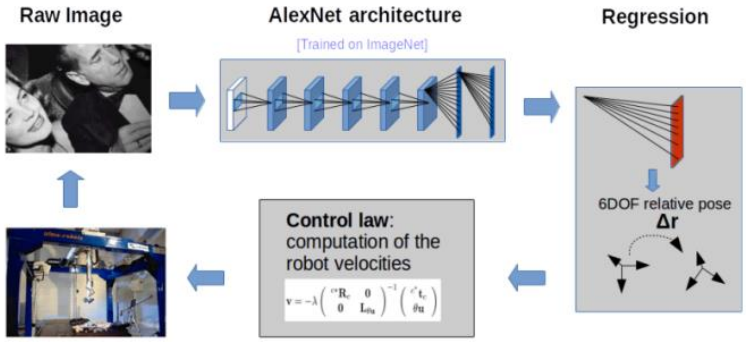

Fig. 7 – Architecture of a Visual Servoing System based on CNN from (Bateux *et al*., 2018).

Another example of a visual servoing system built through knowledge transfer was developed in (Bateux *et al*., 2017) and (Bateux *et al*., 2018) where the authors use a neural network architecture configured based on AlexNet to predict the relative position of the camera on the effector of a six-degree-of-freedom robot. The network is used to move the robot from an initial relative position to a desired one. The innovation was that the neural network receives as input both the images of the mentioned configurations and the relative position between the current image and the reference image. The output of the neural network is represented by the transformation matrix of the current camera position relative to the initial position. To determine the relative position of the desired configuration, the network estimates the relative position between the desired image and the reference image, and between the current image and the reference image. From these two transformation matrices, the relative position between the current image and the reference image is subsequently determined. This last information is used as a cost function during training through AlexNet, an innovative approach at the time of publication. Another innovation of the work was the synthetically generated dataset used for training. The control loop architecture is represented in Fig. 7. A disadvantage of this approach is that with every newly emerged reference position, the network needs to be trained again, which could be problematic in a real-time application.

An interesting approach is proposed in (Yu *et al*., 2019), where the authors configure an architecture based on siamese networks (Bromley *et al.*, 1994) to estimate the camera position for an eye-in-hand manipulator. As can be seen from Fig. 8, the network receives as input images of the initial and desired configurations, which are processed through two different branches of convolutional layers that have the same structure and weights. Used as a regression network, the network outputs the camera position. The obtained results are good, with an error of 0.6 mm for translation and 0.4 degrees for rotation.

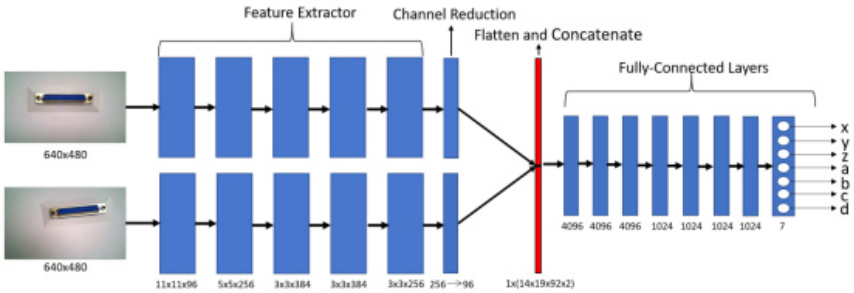

Fig. 8 – Architecture of a Visual Servoing System based on Syamese neural networks, from (Yu *et al.*, 2019).

Another solution based on siamese networks is presented in (Tokuda *et al*., 2021), where the authors configure DEFINet (Difference of Encoded Features driven Interaction Matrix Network), a convolutional neural network for an eye-to-hand servoing system. DEFINet consists of two parts, a feature extractor inspired by (Yu *et al*., 2019) and a regression block. The innovation in the feature extraction block model comes from the fact that the difference between the features extracted from the initial and desired images is directly parsed into the regression block, resulting in the relative position of the camera in the desired configuration. The proposed architecture is presented in Fig. 9.

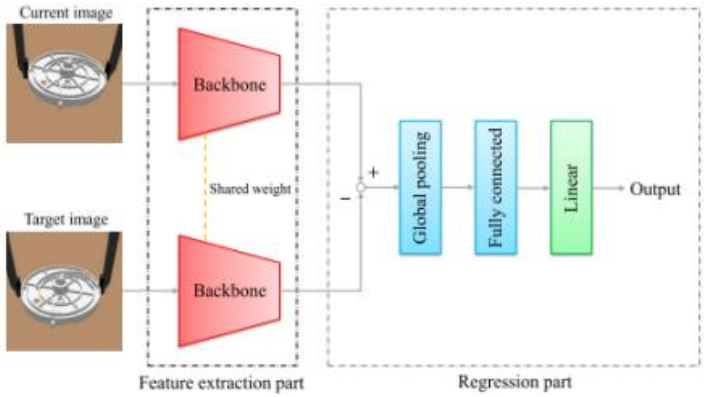

Fig. 9 – DEFINet architecture from (Tokuda *et al*., 2021).

An innovation appears in (Ribeiro *et al*., 2021), where the authors concatenate the images of the initial and desired configurations. The authors develop three different types of neural networks through which they estimate the linear and angular velocities that should be transmitted to the effector to move from the initial configuration to the desired one. The first architecture (Fig. 10 – Direct regression) uses a single branch, receiving as input the two images concatenated in depth and outputting the six velocities.

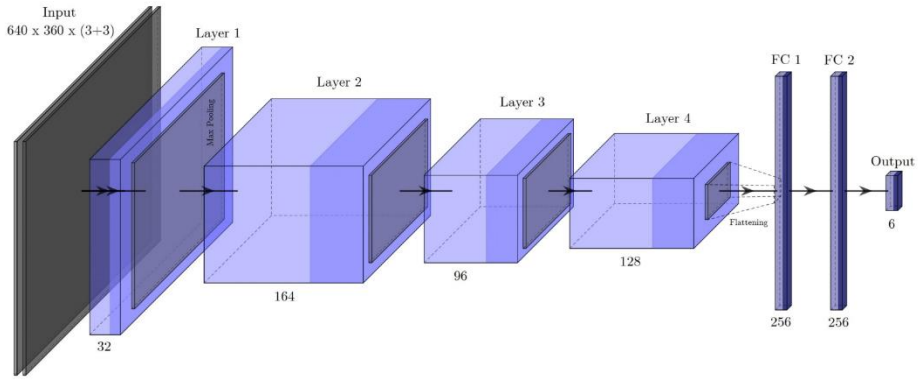

Fig. 10 – Task-specific architecture from (Ribeiro *et al.*, 2021).

The second model (Fig 11 – Direct regression from associated features) uses the same information as input, but has two different output branches, one for linear velocities and one for angular velocities.

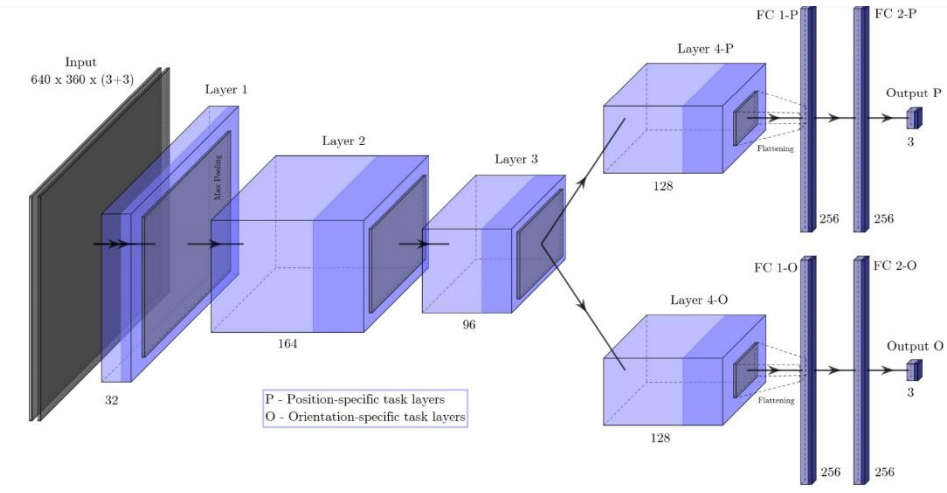

Fig. 11 – Direct regression from associated features architecture from (Ribeiro *et al*., 2021).

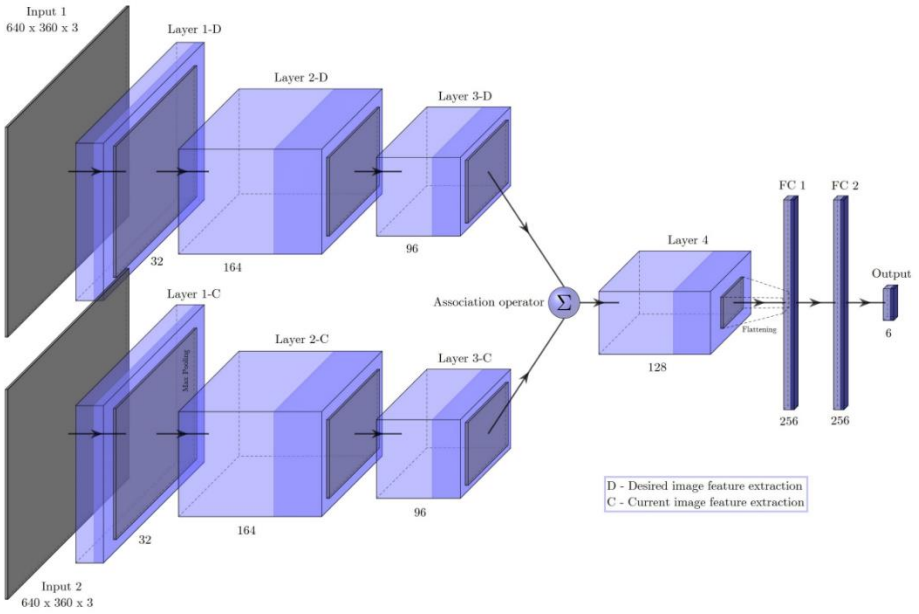

Fig. 12 – Direct regression architecture from (Ribeiro *et al*., 2021).

The last architecture (Fig. 12 – Task-specific architecture) has a feature extractor for each image, concatenating the extracted features and using a single regression block. The best results were obtained by the authors with the first model. Along with these architectures, the authors construct a dataset of input pairs of images of the initial and desired configurations, as well as the velocities that should be given to the effector to achieve control. These velocities were measured manually by moving the effector in multiple configurations.

An interesting approach is presented in (Katara *et al*., 2021) where DeepMPCVS is introduced, a novel deep learning- based Model Predictive Control (MPC) framework for VS. It innovatively combines classical and deep learning techniques, leveraging dense optical flow predictions as intermediate representations to guide control actions in 6-DoF (degrees of freedom). The core of the approach lies in an unsupervised recurrent neural network (LSTM-based) that optimizes control commands in real-time, minimizing the photometric error between the current and target image. The architecture is presented in Fig. 13. The system improves upon existing methods by predicting future states over a time horizon and adapting to dynamic environments through online optimization, without the need for retraining. Results from simulations on photorealistic indoor environments demonstrate superior performance in terms of convergence speed, trajectory length, and reduced computational cost compared to state-of-the-art visual servoing methods.

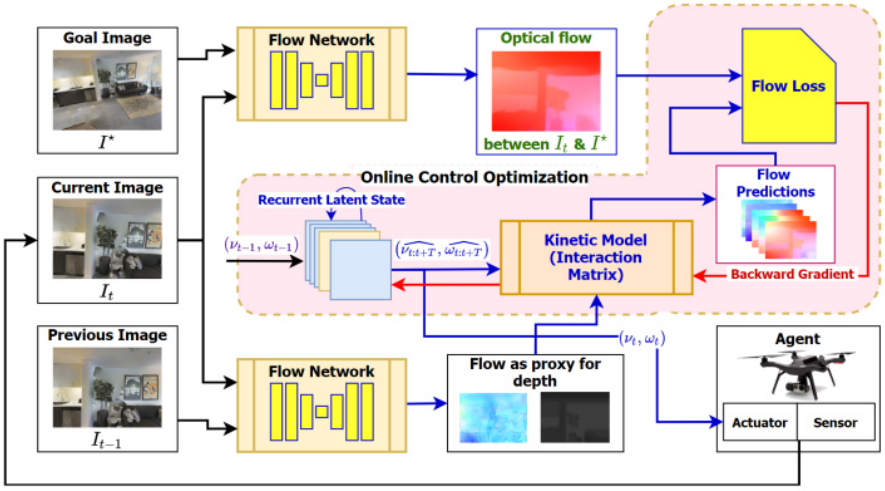

Fig. 13 – DeepMPCVS architecture from (Katara *et al*., 2021).

Another novely approach at the neural level is developed in (Gubbi and Bell, 2021) where a novel DL-based photoacoustic VS system is introduced that directly processes raw sensor data to guide robotic tool tip tracking, bypassing traditional image formation and segmentation. The core innovation lies in using raw photoacoustic sensor data as input to a deep learning model, eliminating the need for beamforming and segmentation, which often produce unreliable results in challenging acoustic environments. The deep learning model, implemented with the ResNet-101 architecture, allows for more robust and accurate detection of needle tips compared to conventional image-based methods. The VS sytem is presented in Fig. 14.

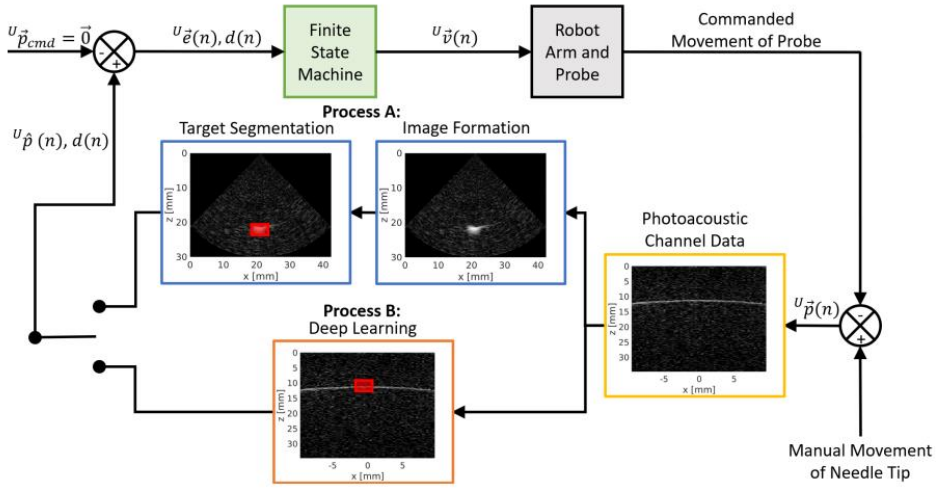

Fig. 14 – VS system workflow from (Gubbi and Bell, 2021).

Experiments using a plastisol phantom and ex vivo tissue demonstrated significant improvements in tracking accuracy, with error reductions of 67.7% and 55.3%, respectively, over segmentation-based approaches. The system also operates at real-time speeds, maintaining the 10 Hz frame rate imposed by the laser pulse frequency, highlighting its suitability for practical robotic surgery applications.

An interesting approach based on a Bayesian Deep Neural Network is presented in (Shi *et al*., 2021). The authors focus on collision avoidance with human hands by predicting a "repulsive pose" that moves the robot away from potential contact. By incorporating Monte Carlo (MC) dropout to transform a DNN into a Bayesian DNN, the system can predict both the position and uncertainty of the robot's movements, enhancing safety by avoiding undesired robot actions, especially in unseen spaces. The Bayesian DNN outperforms standard DNNs by generalizing better to unseen data and ensuring more robust collision avoidance. The experimental results demonstrate that the predictive accuracy of the Bayesian DNN is high, with predictive interval coverage probabilities (PICP) of 0.84, 0.94, and 0.95 along the x, y, and z axes, respectively. Additionally, the approach improves the convergence speed of IBVS by utilizing the predicted repulsive pose, significantly reducing the iterations needed compared to opposite repulsive poses. By validating their system on a real-world UR10 robot, the authors show that their method effectively prevents the robot from moving to unsafe positions while maintaining task efficiency. The diagram of their system is presented in Fig. 15.

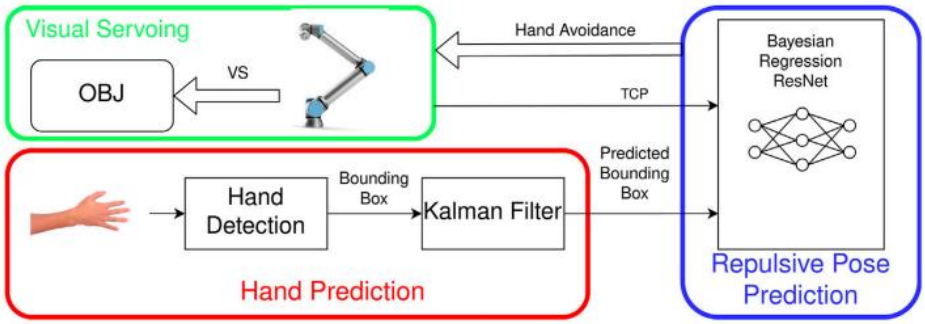

Fig. 15 – Diagram of the system for safe interaction with the robot from (Shi *et al*., 2021).

In the same manner of working with a CNN, in (Lazo *et al*., 2022) a novel approach for autonomous intraluminal navigation using a 3D-printed soft endoscopic robot integrated with a model-less visual servoing system is developed. The control architecture is presented in Fig. 16. The CNN, trained with phantom and in-vivo data, is used to segment the lumen and guide the robot's movement without the need for predefined models, addressing challenges such

as varying image conditions and the need for precise control in narrow luminal structures. The system autonomously detects the center of the lumen and navigates through it, avoiding collisions with the inner walls. The methodology includes testing the robot in various phantom scenarios with different path configurations, where the robot's performance was analyzed using metrics like task completion time, smoothness, and steady-state error.

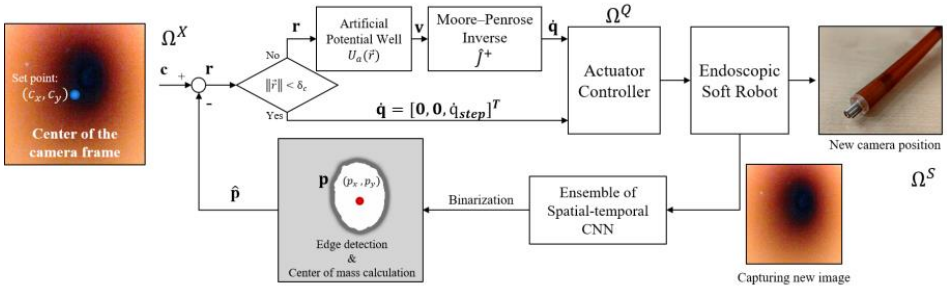

Fig. 16 – Model-less VS system proposed in (Lazo *et al*., 2022).

Based on image similarity, in (He *et al*., 2022) is presented a deep learning-based pose prediction method for VS of robotic manipulators using a novel joint training strategy that combines an L1 loss function with a mean similarity image measurements (MSIM) loss. The MSIM loss evaluates image similarity in terms of brightness, contrast, and structure, improving pose prediction accuracy by guiding the network to better align predicted and target images. The approach is trained using a data generator based on spherical projection and applied to position-based visual servoing (PBVS) using monocular images. Tested in both simulation and real-world scenarios with a UR3 robotic manipulator, the method demonstrated superior accuracy and robustness to occlusions compared to traditional L1 and L2 loss functions, showing promise for more precise and reliable visual servoing tasks in robotics. The algorithm framework is presented in Fig. 17.

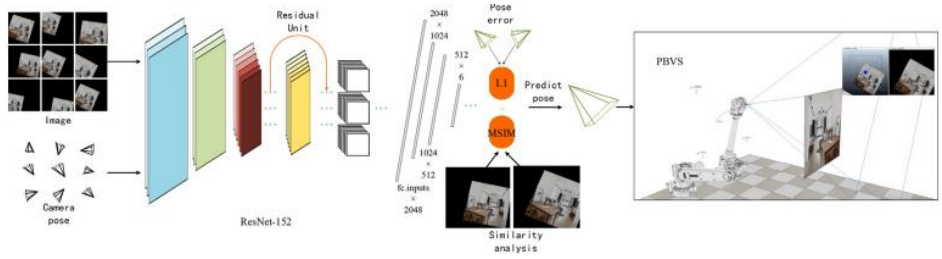

Fig. 17 – Algorithm framework from (He *et al*., 2022).

An interesting approach is described in (Cheng *et al*., 2022) by a vision-based robotic grasping system designed for grasping various objects using a simple parallel gripper. The system employs a deep learning-based approach with a densely connected Feature Pyramid Network (FPN) to enhance grasp pose estimation by fusing multi-layer features and refining predictions through a two-stage detection process. The first stage generates horizontal grasp areas, while the second stage refines these into rotated grasp poses, improving the system's ability to handle objects of diverse shapes and sizes. The architecture is presented in Fig. 18.

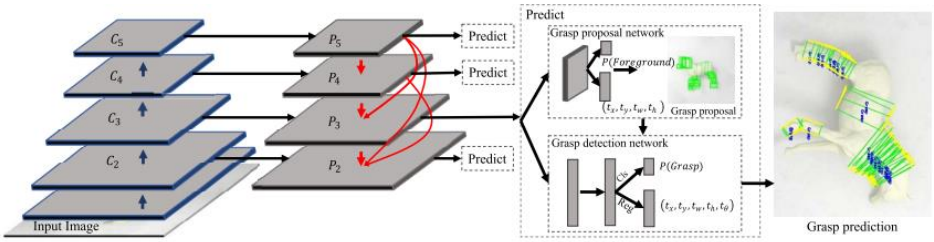

Fig. 18 – Model structure from (Cheng *et al*., 2022).

The up-sampling part (C2, C3, C4, C5) consists of each stage of ResNet. The down-sampling part (P5, P4, P3, P2) is constructed progressively by fusing the corresponding layers in the up-sampling part and all the generated higher layers. A grasp prediction unit is attached to each fused layer. The grasp proposal network first generates the horizontal grasp candidates and crops the corresponding features from the fused feature maps. The second stage grasp detection network refines those proposals and predicts the grasp angles. The output of each predict unit is the dense grasp pose predictions and its probabilities. The final grasp pose is given by the grasp pose with the maximum probability. Evaluated on the Cornell and Jacquard datasets, the system achieves high accuracy (93.3% and 89.6%) and performs well in real-world experiments, including grasping unseen objects and densely overlapped items, demonstrating a 90.8% success rate in handling 51 diverse objects.

Another novelty is presented in (Hao and Xu, 2023) where a robotic system is used for grasping and assembling screws using VS. The system consists of two stages: grasping and installation, facilitated by a binocular vision setup. A convolutional neural network (CNN)-based feature extraction algorithm is employed, shown in Fig. 19, which segments objects and extracts point features for alignment. Image-based visual servoing (IBVS) is used for precise control of the robot's end-effector, achieving sub-millimeter accuracy in grasping and successful assembly of M6-sized screws in experiments. The paper demonstrates the effectiveness of deep learning in feature extraction and the integration of IBVS in handling small components with high precision. The proposed method shows improved alignment and assembly accuracy compared to prior methods,

particularly under natural lighting and complex backgrounds, highlighting its potential for industrial automation.

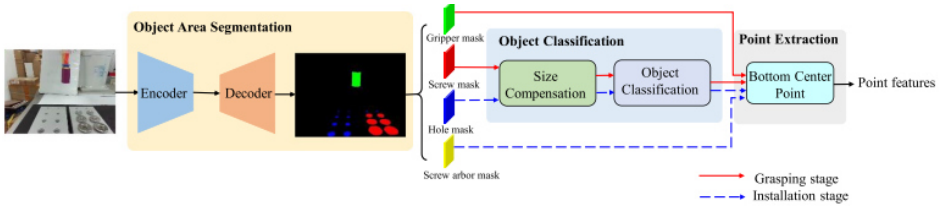

Fig. 19 – Workflow of feature extraction from (Hao and Xu, 2023).

In comparison, (Guo *et al*., 2023) presents a novel vision-based control method for robots with an eye-in-hand configuration, integrating deep learning-based CNN object detectors into traditional robot control frameworks. The proposed system utilizes real-time CNN-based object detection, such as YOLO, to generate task variables from the bounding box parameters of detected objects, and employs LSTM to smooth chattering in bounding box outputs. A vision-based controller is designed using task-space motion control principles to keep objects of unknown and variable aspect ratios centered in the camera's field of view. The stability of the control system is rigorously analyzed using a Lyapunov-like approach, ensuring predictable behavior. Experimental results demonstrate the method's effectiveness in tracking and controlling the positioning of objects, with applications ranging from human detection to crack inspection. The proposed architecture is presented in Fig. 20.

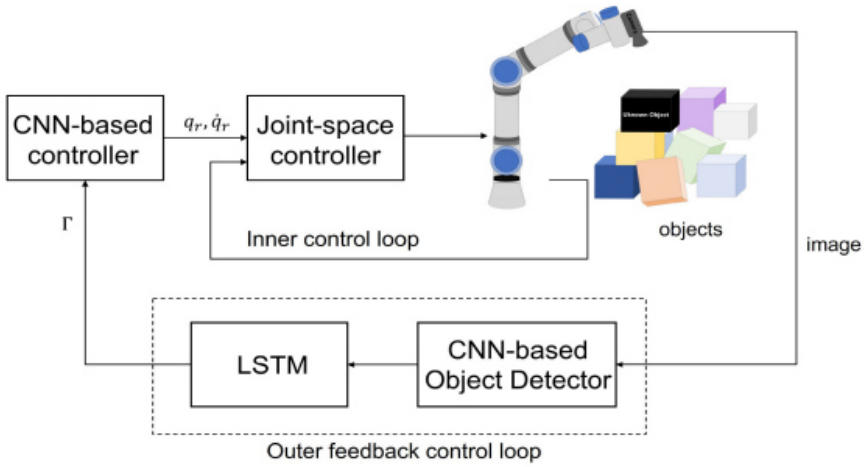

Fig. 20 – Overall block diagram of VS system from (Guo *et al*., 2023).

Alongside CNNs, a novel topic in VS represents the usage of Reinforcement Learning (RL). In (Gao *et al*., 2023) is presented such an

approach, aimed at improving robotic control tasks such as target tracking by addressing key challenges in continuous state and action spaces. To enhance exploration and training efficiency, the authors introduce two innovative components: a novelty measurement method and a hybrid probability sampling strategy. The novelty measure, based on centralized features extracted by a neural network, encourages exploration by quantifying the uniqueness of visited states. The hybrid sampling method integrates intrinsic and external rewards with temporal-difference (TD) error, optimizing the sampling process in RL by balancing the novelty of transitions and their training relevance. The proposed method is implemented using the soft actor-critic (SAC) RL framework, and its performance is tested through simulated experiments in CoppeliaSim, where it achieves improved reward values and completion rates compared to baseline methods. Additionally, real-world experiments with a quadrotor further validate the system's ability to handle both static and dynamic target tracking scenarios. Overall, this work demonstrates how combining intrinsic rewards and optimized sampling can significantly boost RL-based VS performance, with potential applications in more complex robotic tasks.

4. Discussions

In this chapter, we analyze the different visual servoing methodologies reviewed in this survey, focusing on their applications and advantages in various robotic tasks. The transition from classical image-based and position-based visual servoing systems to advanced techniques leveraging deep learning approaches is explored in detail. We highlight how the integration of neural networks and reinforcement learning into visual servoing architectures has enhanced feature extraction, improved control accuracy, and expanded the range of applications for these systems. A summary table is presented, mapping the methodologies to their corresponding applications for a clear comparison of the strengths and limitations of each approach.

The table provides a comprehensive overview of visual servoing methodologies and their specific applications, emphasizing the shift from traditional IBVS and PBVS techniques to modern deep learning-based approaches. Classical methods, such as IBVS and PBVS, rely on visual feature extraction and 3D information for robotic control, with a focus on stability and error minimization. In contrast, newer methods incorporating neural networks, such as CNNs, FlowNet, and Siamese networks, offer more robust and flexible feature extraction, improving system performance in dynamic environments like object grasping, navigation, and human-robot interaction. The inclusion of Direct Visual Servoing (DVS) and reinforcement learning further showcases the growing trend towards more autonomous, data-driven control systems that adapt to changing scenarios in real-time. These advancements demonstrate the

increasing versatility and capability of visual servoing systems across diverse applications.

Table 1
Overview of analyzed papers

Methodology	**Paper**	**Approach Type**	**Appplication**
DL-based Feature Extraction	Ahlin *et al*., 2016	DL + IBVS	Leaf picking, object tracking with CNN
Direct VS with CNN	Bateux, 2018	DL + DVS	Camera pose estiamation
Direct VS with CNN	Bateux *et al*., 2017	DL + DVS	Camera pose estimation
Direct VS with CNN	Bateux *et al*., 2018	DL + DVS	Camera pose estimation
Siamese CNN for Camera Pose Estimation	Bromley *et al*., 1994	DL + PBVS	Eye-in-hand manipulator control and camera pose estimation
PBVS	Chang, 2007	PBVS	Precise positioning of binocular eye-to-hand robotic manipulators.
PBVS	Chaumette, 2002	PBVS	Challenges of PBVS in eye-in-hand configurations.
IBVS	Chaumette, 2004	Classical	Robot control using visual features for movement
IBVS, PBVS	Chaumette and Hutchinson, 2008	IBVS, PBVS	Overview of visual servoing techniques in the Handbook of Robotics.
IBVS, PBVS	Chaumette and Hutchinson, 2006	IBVS, PBVS	Comprehensive review of basic visual servo control approaches
PBVS	Chaumette *et al*., 1991	Classical	Cartesian control of robot trajectories
IBVS	Chaumette, 1998	IBVS	Stability and convergence issues in IBVS.
DL for Grasp Pose Estimation	Cheng *et al*., 2018	DL + PBVS	Robotic grasping of objects using FPN for feature extraction
Hybrid IBVS/PBVS	Chesi *et al*., 2004	Hybrid VS	Switching approach for keeping features in the field of view in eye-in-hand visual servoing.
IBVS	Copot, 2012	IBVS	Control techniques for image-based visual

			servoing using image moments.
Hybrid VS	Collewet and Chaumette, 2002	Classical	Improved stability and control without full calibration
RL-based VS	Gao *et al*., 2023	RL + VS	RL-based target tracking for dynamic environments
CNN-Based Visual Servoing for Object Detection	Guo *et al*., 2023	DL + IBVS	Real-time object detection and tracking for industrial robots
Photoacoustic Visual Servoing	Gubbi *et al*., 2021	DL + VS	Robotic surgery and tool tip tracking
Robotic Screw Grasping & Assembly	Hao *et al*., 2023	DL + IBVS	Visual servoing for robotic assembly of screws
FlowNet for VS	Harish *et al*., 2020	DL + IBVS	Quadcopter navigation, object tracking
DL for Pose Prediction	He *et al*., 2022	DL + PBVS	DL-based pose prediction in VS for robotics
Early Visual Servoing	Hill and Park, 1979	VS	Real-time control of a robot with a mobile camera (early work in visual servoing).
Visual Servoing with Laser Radar	Hancock and Langer, 1998	Active Laser Radar-Based Visual Servoing	High-performance measurements using active laser radar.
PBVS	Hutchinson *et al*., 1996	Classical	Cartesian control of robot trajectories
DL + IBVS	Liu and Li, 2019	DL + IBVS	Image-based visual servoing with deep learning for robotic manipulation
DeepMPCVS	Katara *et al*., 2021	DL + PBVS + MPC	Model predictive control for visual servoing
IBVS	Mahony *et al*., 2002	IBVS	Feature selection for depth-axis control in IBVS.
Hybrid VS	Malis *et al*., 1999	Classical	Improved stability and control without full calibration
IBVS	Marchand and Chaumette, 2005	IBVS	Feature tracking for visual servoing.
Direct VS	Marchand, 2019	Direct VS	Decomposition of image components for control.

DVS in Frequency Domain	Marchand, 2020	Direct VS	Image similarity in frequency domain for control
DFBVS: Deep Feature-Based Visual Servo	Nicholas *et al*., 2022	DL + IBVS	Feature-based visual servo control using keypoint detection
Real-time DL for Grasping & Detection	Ribeiro *et al*., 2021	DL + IBVS	Grasp detection, real-time autonomous robotic control
FlowNet for VS	Saxena *et al*., 2017	DL + IBVS	Quadcopter navigation, object tracking
Bayesian DNN for Collision Avoidance	Shi *et al*., 2021	DL + IBVS	Collision avoidance in human-robot interaction
Neural Outlier Rejection for Keypoint Learning	Tang *et al*., 2020	DL + Keypoint extraction	Outlier rejection for improved keypoint learning in self-supervised systems
Siamese networks for Camera Pose	Tokuda *et al*., 2021	DL + IBVS	Camera pose estimation, object tracking
Siamese networks for Camera Pose	Yu *et al*., 2019	DL + IBVS	Camera pose estimation, object tracking

5. Conclusions

This paper proposed to underscore a significant paradigm shift from classical image processing to the adoption of deep learning techniques, marking a pivotal advancement in robotic control and interaction capabilities. The integration of neural networks has not only enhanced feature extraction and control accuracy but also mitigated the limitations associated with traditional Visual Servoing architectures, such as sensitivity to calibration errors and the complexity of modeling dynamic environments.

Position-based, image-based, and hybrid servoing systems each present distinct approaches to integrating visual information for robotic control, with their respective strengths and challenges in terms of stability, accuracy, and computational demands. However, the emergence of Visual Servoing and sophisticated neural network models, such as siamese networks and convolutional neural networks has introduced innovative solutions that promise greater flexibility, robustness, and efficiency.

Despite the advances, the application of Deep Learning in Visual Servoing faces challenges, including the need for extensive datasets for training and the computational complexity of neural network models, which may impact real-time performance. Future directions should focus on optimizing network architectures for faster processing, exploring unsupervised and reinforcement learning approaches to reduce dependency on labeled data, and developing more

generalized models capable of adapting to a wider range of tasks and environments.

Overall, the convergence of robotics and artificial intelligence through visual servoing systems heralds a new era of robotic capabilities, offering the potential to significantly expand the scope of robotic applications in industry, healthcare, and beyond.

REFERENCES

Ahlin Konrad *et al*., *Autonomous leaf picking using deep learning and visual-servoing*, IFAC PapersOnLine 49.16, (2016), 177-183.

Bateux Q., *Going further with direct visual servoing*, Ph.D. Thesis, Universite, Rennes 1, (2018).

Bateux Q., Marchand E., Leitner J., Chaumette F., Corke P., *Visual servoing from deep neural networks*, arXiv:1705.08940, (2017).

Bateux Q., Marchand E., Leitner J., Chaumette F., Corke P., *Training deep neural networks for visual servoing*, In: 2018 IEEE International Conference on Robotics and Automation (ICRA), pp. 1-8. IEEE, (2018).

Bromley J., Guyon I., LeCun Y., Säckinger E., Shah R., *Signature verification using a 'Siamese' time delay neural network*, In: Proc. Adv. Neural Inf. Process. Syst., 1994, pp. 737-744.

Chang W.C., *Precise positioning of binocular eye-to-hand robotic manipulators*, Journal of Intelligent Robot System, 49(1):219-236, (2007).

Chaumette F., *A first step toward visual servoing using image moments*, Proc. of IEEE / RSJ IROS, 378-438, (2002).

Chaumette F., *Image moments: a general and useful set of features for visual servoing*, IEEE Trans. on Robotics, 20(4), 713-723, (2004).

Chaumette F., Hutchinson S., *Handbook of Robotics*, Springer, (2008).

Chaumette F., Hutchinson S., *Visual Servo Control Part I: Basic Approaches*, IEEE Robotics & Automation Magazine, 13(4), 82-90, (2006).

Chaumette F., Rives P., Espiau B., *Positioning a robot with respect to an object, tracking it and estimating its velocity by visual servoing*, Proc. of the IEEE International Conference on Robotics and Automation, 2248-2253, (1991).

Chaumette F., *Potential problems of stability and convergence in image-based and position-based visual servoing*, In the Conference of Vision and Control, Series 140 Lecture Notes in Control and Information Science, vol. 237, pp. 66-78, Verlag, New York, (1998).

Cheng H. *et al*., *Deep learning for manipulator visual positioning*, 2018 IEEE 8th Annual International Conference on CYBER Technology in Automation, Control, and Intelligent Systems (CYBER), IEEE, 2018.

Cheng H., Wang Y., Meng M.Q.-H., *A Vision-Based Robot Grasping System*, in *IEEE Sensors Journal*, Vol. 22, No. 10, pp. 9610-9620, 15 May15, 2022.

Chesi G., Hashimoto K., Prattichizzo D., Vicino A., *Keeping features in the field of view in eye-in-hand visual servoing: a switching approach*, IEEE Trans. Robot, 20(5), 908-914, (2004).

Copoț C., *Tehnici de control pentru sistemele servoing vizuale*, PhD Thesis, "Gheorghe Asachi" Technical University of Iași, (2012).

Collewet C., Chaumette F., *Positioning a camera with respect to planar objects of unknown shape by coupling 2-D visual servoing and 3-D estimations*, IEEE Trans. Robot. Autom. 18(3), 322-333, (2002).

Gao J., He Y., Chen Y., Li Y., *Learning end-to-end visual servoing using an improved soft actor-critic approach with centralized novelty measurement*, IEEE Transactions on Instrumentation and Measurement, 72, 1-12, (2023).

Guo J., Nguyen H.T., Liu C., Cheah C.C., *Convolutional neural network-based robot control for an eye-in-hand camera*, IEEE Transactions on Systems, Man, and Cybernetics: Systems, 53(8), 4764-4775, (2023).

Gubbi M.R., Bell M.A.L., *Deep learning-based photoacoustic visual servoing: Using outputs from raw sensor data as inputs to a robot controller*, In 2021 IEEE International Conference on Robotics and Automation (ICRA), pp. 14261-14267, (2021).

Hao T., Xu D., *Robotic grasping and assembly of screws based on visual servoing using point features*, The International Journal of Advanced Manufacturing Technology, 129(9), 3979-3991, (2023).

Harish Y.V.S., *DFVS: Deep flow guided scene agnostic image based visual servoing*. 2020 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2020.

He Y., Gao J., Chen Y., *Deep learning-based pose prediction for visual servoing of robotic manipulators using image similarity*, Neurocomputing, 491, 343-352, (2022).

Hill J., Park W.T., *Real-time control of a robot with mobile-camera*, 9th International Symposium on Industrial Robots, pp. 233-246, March 1979.

Hancock J., Langer D., *Active laser radar for high-performance measurements*, In Proc. of IEEE International Conference on Robotics and Automation (ICRA), vol. 2, pp. 1465-1470, 1998.

Hutchinson S., Hager G., Corke P., *A tutorial on visual servo control*, IEEE Transactions on Robotics and Automation, 12(5), (1996), 651-670.

Katara P., Harish Y.V.S., Pandya H., Gupta A., Sanchawala A., Kumar G., Krishna M., *Deepmpcvs: Deep model predictive control for visual servoing*, In Conference on Robot Learning, pp. 2006-2015, (2021).

Lazo Jorge F. *et al*., *Autonomous intraluminal navigation of a soft robot using deep-learning-based visual servoing*, *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2022.

Liu J., Li Y., *An Image Based Visual Servo Approach with Deep Learning for Robotic Manipulation*, arXiv preprint arXiv:1909.07727, (2019).

Mahony R., Corke P., Chaumette F., *Choice of image features for depth-axis control in image based visual servo control*, Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems, Lausanne, Switzerland, (2002), 390-395.

Malis E., Chaumette F., Boudet S., *2 1/2 d visual servoing*, IEEE Trans. Robot. Autom. 15(2), 238-250, (1999).

Marchand E., Chaumette F., *Feature tracking for visual purposes*, In Robotics and Systems, 52(1), 53-70, (2005).

Marchand E., *Subspace-based direct visual servoing*, IEEE Robot. Autom. Lett. 4(3), 2699-2706, (2019).

Marchand E., *Direct visual servoing in the frequency domain*, IEEE Robot. Autom. Lett. 5(2), 620-627, (2020).

Nicholas A., Van-Thach D., Quang-Cuong P., *DFBVS: Deep Feature-Based Visual Servo*. arXiv preprint arXiv:2201.08046, (2022).

Ribeiro E.G., Mendes R-Q., Grassi V.Jr., *Real-time deep learning approach to visual servo control and grasp detection for autonomous robotic manipulation*. Robotics and Autonomous Systems 139, (2021).

Saxena A., Pandya H., Kumar G., Gaud A., *Exploring convolutional networks for end-to-end visual servoing*, In: 2017 IEEE International Conference on Robotics and Automation, ICRA, IEEE, Marina Bay Sands, Singapore, 2017, pp. 3817-3823.

Shi L., Copot C., Vanlanduit S., *A bayesian deep neural network for safe visual servoing in human–robot interaction*, Frontiers in Robotics and AI, 8, 687031, (2021).

Tang J., Kim H., Guizilini V., Pillai S., Ambrus R., *Neural outlier rejection for self-supervised keypoint learning*, In International Conference on Learning Representations, 2020.

Tokuda F., Shogso A., Kosuge K., *Convolutional neural network-based visual servoing for eye-to-hand manipulator*, IEEE Access 9 (2021): 91820-91835.

Yu C., Cai Z., Pham H., Pham Q.-C., *Siamese convolutional neural network for sub-millimeter accurate camera pose estimation and visual servoing*, In Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS), Nov. 2019, pp. 935-941.

O SCURTĂ ANALIZĂ A METODELOR DE ÎNVĂȚARE AUTOMATĂ ÎN SISTEMELE SERVOING VIZUALE

(Rezumat)

Această lucrare analizează evoluția și aplicațiile sistemelor servoing vizuale în robotică, accentuând trecerea de la tehnicile tradiționale de procesare a imaginilor la integrarea rețelelor neuronale pentru extragerea trăsăturilor și control. Sistemele robotice, esențiale în domenii precum producția autonomă de piese, supraveghere video sau sănătatea se bazează tot mai mult pe controlul vizual pentru a îmbunătăți interacțiunea în diverse medii de lucru. Inițial concentrat pe senzori auxiliari, cum ar fi senzorii vizuali, pentru a crește robustețea și precizia, progresele recente au adus o schimbare spre integrarea metodelor de învățare profundă pentru control direct și extracția caracteristicilor. Acest studiu se concentrează pe diferențele dintre arhitecturile clasice de servoing vizual și metodele noi care implică învățarea profundă, subliniind avantajele și limitările acestora în ceea ce privește stabilitatea, precizia și aplicabilitatea în timp real. Abordări inovatoare, cum ar fi servoingul vizual direct și utilizarea rețelelor neuronale pentru estimarea poziției camerei, demonstrează progrese semnificative în depășirea provocărilor controlului vizual tradițional. Prin examinarea detaliată a cercetărilor de referință, lucrarea evidențiază potențialul rețelelor neuronale de a revoluționa acest domeniu prin îmbunătățirea extracției caracteristicilor, reducerea dependenței de calibrarea precisă și optimizarea legilor de control pentru sarcini robotice complexe.