Computer vision solution applied to a fixed-station robot for target tracking

Gustavo Arthur Dutra
Universidade Federal de Santa Maria
Santa Maria - RS, BRASIL
email: gustavo.dutra@acad.ufsm.br

Claudenir Rocha Alves Filho
Universidade Federal de Santa Maria
Santa Maria - RS, BRASIL
email: claudenir.filho@acad.ufsm.br

Ricardo Dias Schirmer
Universidade Federal de Santa Maria
Santa Maria - RS, BRASIL
email: ricardo.schirmer@gmail.com

Anselmo R. Cukla
Universidade Federal de Santa Maria
Santa Maria - RS, BRASIL
email: anselmo.cukla@ufsm.br

Marcelo Serrano Zanetti
Universidade Federal de Santa Maria
Santa Maria - RS, BRASIL
email: marcelo.zanetti@ufsm.br

Gabriel Tarnowski
Universidad Nacional de Misiones
Obera, Misiones, Argentina
email: gabriel.tarnowski@fio.unam.edu.ar

***Abstract*—This article addresses the development and integration of artificial intelligence (AI) technologies, such as Tracking Learning Detection (TLD) and Google MediaPipe (MP), into a military robotic system known as Robotic Sniper (SR). The SR aims to replace the sniper and observer with cameras equipped with AI algorithms, reducing soldiers' exposure in surveillance missions and increasing operational efficiency. TLD is used for target detection and tracking, while MP is employed for recognizing keypoints of the target's face. The AIs control the robot's motors. The article details the methodology for developing the operator interface software, as well as the integration of TLD and MP technologies, and discusses the results obtained, the system's effectiveness in identifying and tracking targets in different military scenarios. Ethical concerns related to the increasing use of robotic weapons and the importance of maintaining human control over target selection and attack decisions are highlighted.**

I. INTRODUCTION

Technological advancement has significantly contributed to the evolution of conflicts throughout the centuries. With the rise of robotics and AI, the military landscape has witnessed an unprecedented transformation. Armed forces are facing the reality of accessible equipment, challenging and sometimes surpassing sophisticated equipment, as seen in current scenarios where low-cost drones can neutralize vehicles, aircraft, or ships whose costs are orders of magnitude higher. This contrast underscores the urgency of adaptation and the importance of quickly incorporating these technologies by the countries or groups involved, as is the case with the conflict between Russia and Ukraine.

Historically, the role of the sniper and observer has been decisive in military operations. As described in the United States Army Sniper Manual (1994) [1], the sniper's mission is to provide precise long-range fire support on selected targets, while the observer assists in locating targets on the ground and determining environmental conditions. However, both are exposed to the risk of being identified and targeted by enemy fire. The development of the Robotic Sniper (SR) can help minimize soldiers' exposure by replacing the sniper and observer with cameras equipped with AI-based algorithms. Snipers are also employed in hostage situations, and SRs should minimize the risk of injuring hostages by increasing the accuracy of shots through more robust weapon control via a more realistic modeling of local weather conditions.

In the Robotic Sniper (SR), the roles of the sniper and observer are replicated by two cameras integrated into a computer vision system that utilizes AI, such as MP and TLD, respectively, for target detection and tracking. The TLD will be used in the observer's camera to distinguish objects, while the MP will be used in the sniper's camera to accurately determine the center of the eyes for targeting the subject.

When it comes to studies on target tracking based on AI technologies, two notable ones are MP and TLD. According to [2] and [3], MP is a framework developed to enable users to quickly and easily utilize its resources through the Machine Learning (ML) process. This algorithm provides a structure for sensory data inference, performance evaluation tools, and reusable processing components called Calculators, which allow improvements in perception through their language configuration. MP is capable of recognizing various pieces of information, such as people's orientation and limb arrangement, and determines, through a convolutional network, which person deserves the detection layer, prioritizing the one more centrally located in the image.

However, in cases of multiple people on the screen or a cluttered center where facial perception is ambiguous, MP may encounter difficulties in identifying the correct target, which is resolved by using TLD. According to studies [4] and [5], the Region of Interest (ROI) process is used with real-time object tracking, combining technologies such as the CamShift algorithm, OpenCV, and ConvNet. This allows for precise identification of individuals, even in complex situations such as kidnappings, human shields, protests, among others. ROI technologies have a limitation related to defining a generic rectangle that does not specify what is contained within it. This can result in the exclusion of other objects that are not entirely within the area delimited by the rectangle. For example, if a human is within the ROI, the center of the

rectangle will be close to their navel. However, if the person raises an arm, the rectangle will expand, and the center will no longer be aligned with the body. Thus, the combination of these two technologies provides a robust and effective system for detecting and tracking targets in various defense or public security contexts.

Previous studies have explored the use of cameras for gesture recognition purposes to transmit information to control systems, such as prototypes of robotic arms. These investigations aim to achieve more intuitive and efficient control of physical devices, leveraging the keypoint recognition capabilities of MP. The articles [6] and [7] present an approach congruent with the project discussed in this study. Both works focus on identifying single-hand gestures and subsequently transmitting this information to control systems. Although the mentioned studies demonstrate the feasibility of this approach, it is important to note that comprehensive tests have not yet been conducted, especially in scenarios involving additional challenges, such as simultaneous detection of gestures from multiple hands or gestures in unconventional positions.

The technology of TLD can be observed in [8], where a mobile robot is tasked with recognizing and tracking a human. This study utilizes not only TLD but also MedianFlow, another native AI algorithm of OpenCV, and a generic facial recognition algorithm. Here, there was also a concern with scattering components such as the tracked human being in motion or obstructed. The approach of this study is related to the creation of a skeleton of 20 joints. The system appears to resemble the keypoints of MP. Tests demonstrate the technology's ability to continue recognizing the target even after it disappears from the screen and reappears. The study also addresses the high rate of false positives of TLD. The developed technology, called SIFace-TLD, can successfully track the target even when the target is facing away from the camera. Furthermore, the Monte Carlo Particle Filter was adopted to reduce the influence of measurement noise and obtain an accurate estimate of the state for robot motion control. Lastly, a simple and effective controller for human tracking was designed. In the case of SR, the project uses MP to fulfill this need.

On the other hand, robotic systems utilizing the concept of sentinels already exist in the military domain, such as the SGR-A1 developed by Samsung Heavy Industries [9]. It is a robot with autonomous shooting capability, allowing it to identify and shoot targets without human intervention. This robot has various levels of human interaction, operating in both semi-autonomous mode, where the final decision on shooting is made by a human, and fully autonomous mode, where the robot makes independent decisions on when and where to shoot. Its operating interface consists of two main screens, a camera with night vision, thermal imaging, and zoom capabilities.

Additionally, Dodaam Systems Ltd., a company recognized for its leadership in manufacturing military robots, also invests in research, as evidenced in [10], which presents a surveillance system equipped with a target tracking algorithm. This system utilizes a direction vector calculated from the position error between the center of the viewport and the center of the object in the image captured by the camera. The company is not limited to a static robot but has a range of lethal and non-lethal robots, both static and mobile, with or without humans. It has a control center and a variety of complementary technologies.

In this In this paper, we cannot fail to mention the manifesto by Human Rights Watch [11], in partnership with the International Committee of the Red Cross [12], regarding their concerns about the increasing proliferation of robotic weapons that have the potential to replace humans in combat situations. The SR robot aims to position itself at the safest level of human interaction, known as "Human-in-the-Loop," where robots require human authorization to select targets and inflict damage, following established international norms.

Thus, the study was initiated in 2022, resulting in an SR, as shown in Figure 1, with two axes, equipped with a Taurus 9mm pistol and two cameras, one for terrain vision and another equipped with a scope. The robot has compact dimensions and a modular structure consisting of two distinct parts: a base and a weapon-carrying module. It features two Dynamixel Robotis MX-106R motors, each responsible for a specific axis. In the SR, the functions of the shooter and observer are replicated by two cameras integrated into a computer vision system that utilizes AI, such as MP and TLD, respectively, for target detection and tracking. TLD will be used in the observer's camera to distinguish objects, and MP in the shooter's camera to accurately determine the center of the eyes for targeting the subject. The cameras used in this study are *Webcam Full HD 1080p*, which have autofocus and good contrast, adapting to the lighting conditions.

Fig. 1. Profile image of the SR prototype.

The proposed article aims to explain the operation of the operator interface software, as well as the development and integration of TLD and MP technologies, computer vision algorithms based on artificial intelligence, which control the movements of the robot's motors presented in [13]. The main objective is to explain the algorithms that enable the operation of the interface, where the user can indicate a target, and the robot identifies and tracks it, keeping it centered on the

screen. Specific objectives include how AIs interact with the SR, including the implementation of real-time ROI in the TLD algorithm, integration between the various AI tools employed, the use of keypoints from the MP algorithm for robot motor control, and the development of an operational interface for the user.

This article is structured into four distinct sections. In addition to the introduction, which presents relevant concepts and technologies related to the topic, it includes a detailed methodology on the development and construction of the software. The third section addresses the obtained results and the discussions arising from these results. Finally, the fourth section encompasses the conclusions, highlighting potential future applications related to facial differentiation.

II. Methodology

This study provides a detailed analysis of the operation of the software and user interface created for the SR, as well as describing the process by which the robot is created, the exchange of Artificial Intelligences (AIs), and finally, the execution of the shot. The entire procedure is described sequentially, as illustrated in the flowchart shown in Figure2.

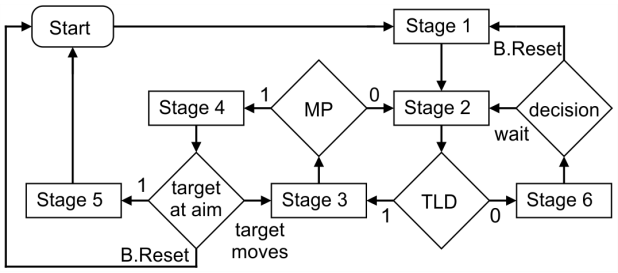

Fig. 2. Steps flowchart.

In the operation of the SR, initially when the system is initialized (Start), the robot is in a surveillance state and responds to the operator's commands, which must be provided via the keyboard and can move it in any direction or activate the ROI algorithm that leads to the first stage. At this moment (Stage 1), on the screen where the user views the actions of the two cameras, the creation of a small green rectangle, controlled by the mouse, around the desired target selected by the mouse is made available, which can be moved or resized to select the target quickly and accurately. After the rectangle is created by the operator and clicked on the target, the TLD comes into action. For better visualization of the beginning of this phase, the rectangle changes from green to blue. The TLD has two internal cycles, search or track. The search is when the target is not detected, and the track is when the target is detected.

The TLD algorithm uses pixel values to transform this information into motor speed, and for this conversion, a mapping function is used, which we call here the MAP function. The pseudocode for this function is presented in Figure 3.

As mentioned, the MAP function is responsible for transforming input values from a pixel displacement range to an output range of velocity displacement, which translates

```
function MAP(refer){
    A ← minimum camera position
    B ← maximum camera position
    C ← minimum engine speed
    D ← maximum engine speed

    vel ← ((refer − A) × (D − C)) / (B − A) + C
    return vel
}
```

Fig. 3. Map function.

the information coming from the camera into commands understandable for the motors and was adapted from the C language library of Arduino to Python language. The algorithm uses the reference of the distance between the central point of the target, given by the TLD or MP algorithms, and the central point of the camera, and uses this distance to calculate the velocity.

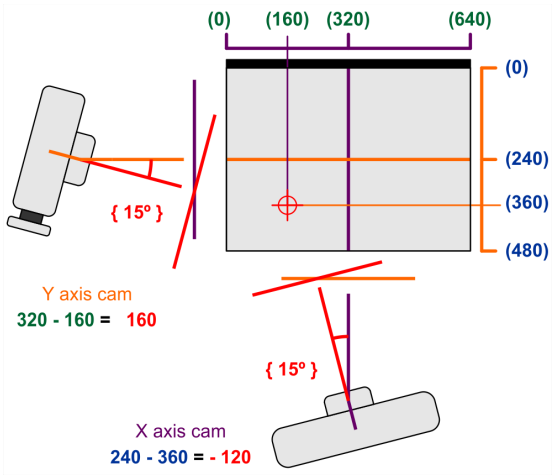

Fig. 4. Capture of Pixels that will be converted by the Map function.

As illustrated in Figure 4, each motor has a different speed limit because their behaviors vary according to the central points of the cameras. The cameras have a resolution of 640x480 pixels; therefore, the central point of the camera for the x-axis is 320, and for the y-axis, it is 240 pixels. Another variant is the x-axis motor, with a movement ratio of 1:1, while for the y-axis motor, there is a movement ratio of 20:1, due to a worm gear. This variation in robot movement is addressed with the Map function and different speed limits on the Dynamixel motors, being for the x and y axes, respectively, 60 and 1000.

Continuing with the flowchart of the software operation methodology of this work, we observe that the second stage (Stage 2) is based on target selection and confirmation. The central x and y coordinates of the rectangle, that is, of the target, are identified. With the use of the Map function, it becomes possible to calculate the motor speeds. The robot will move autonomously and continuously (TLD) until it reaches the central point of the target, adjusting the motor

speeds according to the verification of the distance between the desired central point and the camera's central point, which occurs with the receipt of each new frame. An error rate can be defined as the minimum difference between the central point of the target and the camera's central point. The sixth stage (Stage 6) occurs if the operator decides to turn off the TLD algorithm. This situation may occur if the target is outside the camera's coverage area or if it is erroneously selected by the operator, not corresponding to the desired one.

Subsequently, in the third stage (Stage 3), with the robot centered on the rectangular target of the ROI, the operator has the option to switch the identification algorithm via the keyboard. Thus, the robot will perform a more accurate identification using the MP, which results in the identification of key points on the target's face and body, differentiating yellow points for the left side and green for the right side. In this phase, the displacement of the target in relation to a reference point between the eyes is calculated, allowing the "sniper" camera to control the robot's movement assertively (MP). If disabled or fails to find the target within a timeout of 2 seconds, it will return to Stage 2.

In the fourth stage (Stage 4), the target is in sight, and the robot precisely awaits the operator's decision to interrupt and return to the initial stage or proceed to the final stage of the mission. If the target moves, the MP automatically centers the target in real-time, using the Map function to pass data to the motors. Finally, in the fifth stage (Stage 5), the operator fires, and the robot returns to the initial surveillance position, awaiting new instructions. Thus, completing the cycle that the SR goes through from visual identification to the actual shot, carried out under operator supervision.

The supervision and control of the system are carried out by a designated operator, whose operations are conducted through an interface, as shown in Figure 5. In the image, two main screens stand out: the primary screen for the observer and a secondary screen for the shooter. Given the academic context of this study, the interface is equipped with a series of interactions, including buttons to activate AIs, reset and emergency buttons, firing buttons, as well as a text block for code monitoring, among other essential functionalities.

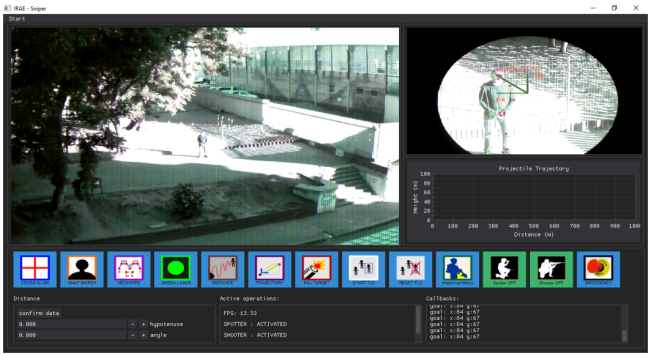

Fig. 5. Interface between operator and SR.

This interface allows observations of the behavior of the AIs and the response of the motors. Analyzing the interactions between the operator and the system can provide valuable insights to optimize operational efficiency and ensure the safety and reliability of operations.

Fig. 6. Result of Stage 1.

III. Results and Discussion

As described in the methodology presented in Figure 2, the process begins with the selection of a target by the operator. Then, it proceeds to Stage 1, as illustrated in Figure 6.

The camera shown in Figure 6 has a wide capture area to cover a larger number of objects and people that can be selected as targets. Selecting an area of interest allows the Sniper's camera to provide greater detail of the target. It is noticeable that the image does not have high quality, as the camera was configured for a resolution of 640x480 pixels. Additionally, the image used corresponds to the observer, which does not have an appropriate zoom system.

In Stage 2, it is important to highlight that the area of interest can be created in any area of the image, making it possible to select a non-human object and pursue it. This could be advantageous from the perspective of a broader scope target search. However, the technology developed is aimed at human targets. Therefore, it may be necessary to repeat the target selection stage to proceed with activating the MP for motor control and target pursuit. The result of this stage is illustrated in Figure 7.

Thus, with the use of MP in Stage 3, illustrated in Figure 8, in conjunction with TLD, it is possible to work with great precision. However, the system does not have effective and specific techniques for real distinction between people. This ability could be acquired with the use of other tools that work in conjunction with the OpenCV library, using the calculation of the distance between specific points on the face for differentiation.

Finally, after the transition from target tracking to firing in Stage 4, a laser pointer was used in place of a weapon. When the firing was activated, the laser pointer would activate as well, in order to demonstrate the tests conducted in Stage 5.

Fig. 7. Result of Stage 2.

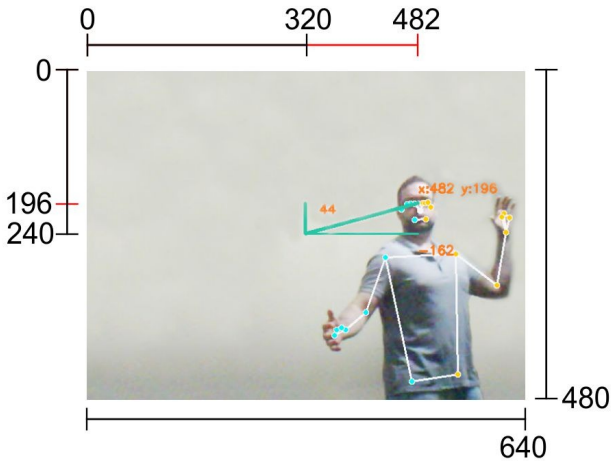

Fig. 8. Result of Stage 3.

Figure 9 demonstrates the test conducted during the laser activation. In the wider image on the left, the observer's view is shown, with the target already selected and centered. On the right is the sniper's view, providing greater detail of the target and the exact position of the sight.

Regardless of its movement, the sight tends to track the target, which can also be used as a powerful tool for intimidation and confusion of the target due to the difficulty in opening the eyes considering the location of the light beam.

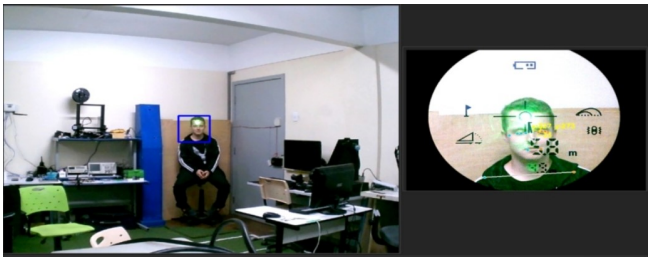

Fig. 9. Result of Stage 5.

IV. Conclusions

The study highlights that the combination of AI and computer vision technologies in the SR represents a promising solution for military operations. The specific results emphasize the real-time use of a ROI converted to the TLD algorithm, as well as motor control and alignment with the target through AI. The operator interface harmonizes with combat robotic equipment, ensuring ethical care in a topic still under discussion.

Future studies focused on optimizing AI algorithms and developing a proprietary combat system are necessary for future advancements in this area. These new studies open the possibility of facial recognition through the replacement or blending of real-time facial recognition technologies with those currently used. Articles such as [2] and [7] demonstrate that better results are obtained with combined AI algorithms.

References

[1] Department of the Army. *Sniper Training*. Washington: U.S. Government Printing Office: 1994—528-027/80156, 1994, p. 329.

[2] Camillo Lugaresi et al. "MediaPipe: A Framework for Building Perception Pipelines". In: (June 2019).

[3] Valentin Bazarevsky et al. "BlazePose: On-device Real-time Body Pose tracking". In: (June 2020).

[4] Zdenek Kalal, Krystian Mikolajczyk, and Jiri Matas. "Tracking-Learning-Detection". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34.7 (2012), pp. 1409–1422. DOI: 10.1109/TPAMI.2011.239.

[5] Nadja Dardagan et al. "Multiple Object Trackers in OpenCV: A Benchmark". In: Oct. 2021. DOI: 10.1109/ISIE45552.2021.9576367.

[6] P D Rathika et al. "Gesture Based Robot Arm Control". In: *Nat. Volatiles & Essent. Oil* (2021). DOI: 8(5):3133-3143.

[7] Muneera Altayeb. "Hand Gestures Replicating Robot Arm based on MediaPipe". In: *Indonesian Journal of Electrical Engineering and Informatics (IJEEI)* 11 (Sept. 2023), pp. 727–737. DOI: 10.52549/ijeei.v11i3.4491.

[8] Jing Yuan et al. "Fusing Skeleton Recognition With Face-TLD for Human Following of Mobile Service Robots". In: *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 51.5 (2021), pp. 2963–2979. DOI: 10.1109/TSMC.2019.2921974.

[9] Mark Prigg. *Who goes there? Samsung unveils robot sentry that can kill from two miles away. SGR-1 has heat and motion detectors to identify potential targets more than 2 miles away Being used in the Demilitarised Zone*. Acessado em 01 jan. 2023. 2014. URL: https://www.dailymail.co.uk/sciencetech/article-2756847/Who-goes-Samsung-reveals-robot-sentry-set-eye-North-Korea.html.

[10] Bong-Cheol Seo, Sung-Soo Kim, and Dong-Youm Lee. “Target-Tracking System for Mobile Surveillance Robot Using CAMShift Image Processing Technique”. In: *Transactions of the Korean Society of Mechanical Engineers A* (2014). DOI: 10.3795/KSME-A.2014.38.2.129.

[11] Bonnie Docherty. *Losing Humanity. The Case against Killer Robots*. Acessado em 01 jan. 2023. 2012. URL: https://www.hrw.org/report/2012/11/19/losing-humanity/case-against-killer-robots.

[12] CICV. *Armas autônomas: os Estados devem discutir os grandes desafios éticos e humanitários.)* Acessado em 01 jan. 2023. 2013. URL: https://www.icrc.org/pt/content/armas-autonomas-os-estados-devem-discutir-os-grandes-desafios-eticos-e-humanitarios.

[13] Ricardo Dias Schirmer et al. “Project of a Sentinel Robot Controlled with a Tracking Algorithm”. In: *2022 Latin American Robotics Symposium (LARS), 2022 Brazilian Symposium on Robotics (SBR), and 2022 Workshop on Robotics in Education (WRE)* (2022), pp. 241–246. DOI: 10.1109/LARS/SBR/WRE56824.2022.9995943.