International Journal of Electrical and Computer Engineering (IJECE)
Vol. 13, No. 6, December 2023, pp. 6216∼6227
ISSN: 2088-8708, DOI: 10.11591/ijece.v13i6.pp6216-6227 ❒ 6216

A novel visual tracking scheme for unstructured indoor environments

Fredy Martínez, Holman Montiel, Fernando Martínez
Facultad Tecnológica, Universidad Distrital Francisco José de Caldas, Bogotá D.C, Colombia

Article Info

Article history:

Received Mar 27, 2023
Revised Jun 8, 2023
Accepted Jun 22, 2023

Keywords:

Differential filter
Indoor environment
Interaction
K-means clustering
Real-time
Visual tracking

ABSTRACT

In the ever-expanding sphere of assistive robotics, the pressing need for advanced methods capable of accurately tracking individuals within unstructured indoor settings has been magnified. This research endeavours to devise a real-time visual tracking mechanism that encapsulates high performance attributes while maintaining minimal computational requirements. Inspired by the neural processes of the human brain's visual information handling, our innovative algorithm employs a pattern image, serving as an ephemeral memory, which facilitates the identification of motion within images. This tracking paradigm was subjected to rigorous testing on a Nao humanoid robot, demonstrating noteworthy outcomes in controlled laboratory conditions. The algorithm exhibited a remarkably low false detection rate, less than 4%, and target losses were recorded in merely 12% of instances, thus attesting to its successful operation. Moreover, the algorithm's capacity to accurately estimate the direct distance to the target further substantiated its high efficacy. These compelling findings serve as a substantial contribution to assistive robotics. The proficient visual tracking methodology proposed herein holds the potential to markedly amplify the competencies of robots operating in dynamic, unstructured indoor settings, and set the foundation for a higher degree of complex interactive tasks.

Corresponding Author:

Fredy Martínez
Facultad Tecnológica, Universidad Distrital Francisco José de Caldas
Carrera 7 No 40B-53, Bogotá D.C., Colombia
Email: fhmartinezs@udistrital.edu.co

1. INTRODUCTION

The integration of robots into everyday human life is a rapidly developing field that presents numerous research challenges and opportunities [1]. Assistive robotics, which is the use of robots in homes to provide care, entertainment, and company, has gained significant attention in recent years [2]. This type of application requires the development of advanced robotics technologies, such as control, learning, bipedal walking, fine manipulation, and human-robot interaction [3]. However, the use of robots in unstructured and dynamic environments, designed for human interaction, presents new and complex challenges. For assistive robots to be effective, they must be able to understand the intentions of human beings by sensing their behavior in the environment. This task requires the integration of multiple technologies and approaches, including computer vision, machine learning, and human-robot interaction. Despite the progress that has been made in these areas, many of the proposed solutions are still computationally demanding and require further refinement [4]. In order to overcome these challenges and fully realize the potential of assistive robotics, it is essential to conduct rigorous and interdisciplinary research that advances the state-of-the-art in control, learning, bipedal walking, fine manipulation, and human-robot interaction.

As the field of robotics continues to evolve, the integration of robots into daily human life has become a central focus for researchers. Assistive robots are designed to perform tasks such as surveillance, entertainment, companionship, and care, which require them to interact with humans in unstructured, dynamic environments [5]. These robots present new challenges, as they must recognize people, track them, and understand their state of mind and activity [6], [7]. The foundation of social interaction is visual information, which is why robots that interact with humans must process visual information in a way that imitates the human brain [8], [9]. The processing of visual information by the human brain is a complex process that requires the ability to recognize and track people in real-time [10]. The challenge of determining the position of a human body in a video stream is significant due to the many variations of the body throughout the sequence [11].

In recent literature, there have been numerous applications of tracking and control from camera images, including autonomous applications in real-time, where the robot extracts information from the image to make motion decisions [12]. These applications rely on specific hardware configurations and algorithms that have been developed to support them. However, the ability to process visual information in a way that mimics the human brain and operates with a low computational cost remains an open engineering problem [13], [14]. The robotics has made significant progress in recent years, with advancements in areas such as visual perception and tracking. For this purpose, various techniques have been proposed to capture the visual information and process it to identify and track people. One such technique is pattern recognition, which looks for specific shapes, colors, and movement characteristics in the images captured by the robot's cameras [15]. This process is complex, as the robot must identify the correct set of parameters to characterize a person, which can be challenging given the diversity of operating environments [16]. Moreover, some techniques, such as the use of histograms of oriented gradient (HOG) descriptors and support vector machine classifiers, are computationally expensive for real-time applications on small autonomous robots [17].

The robot must know the pattern beforehand to use pattern recognition effectively. However, this is not always possible, especially in a service robot, which must operate in a variety of environments and interact with a diverse set of people. In these cases, it is better for the robot to autonomously learn the pattern through its interactions with people [18]. For example, the robot could learn to recognize a person's silhouette through observing their movements. This approach is particularly important when the pattern (in this case, the silhouette of a person) is not rigid, but instead changes with their movements. Thus, learning the pattern should focus more on movement observation than on the detection of people in images.

The majority of visual tracking systems only employ monocular vision, which is prone to high false detection rates due to the absence of depth information in two-dimensional images [19]. To overcome this limitation, stereoscopic vision has been proposed as a solution, which utilizes two cameras to provide depth information, leading to a reduction in the number of false detections [20]. However, the use of stereoscopic vision comes with an increased computational cost, which may restrict its use in smaller platforms.

To further improve the performance of visual tracking systems, researchers have proposed combining the camera system with other sensors, such as acoustic sensors or laser range finders (LRF). This strategy allows for the integration of additional information, increasing the robustness of the system. However, this approach also increases the computational load and the complexity of the algorithms involved, making it challenging to implement in small platforms [15], [21]. Additionally, the use of people as tracking elements may not be feasible in service robots, which need to operate in a variety of environments without specific cues.

In this research, we propose a novel identification and tracking scheme for human beings based on their movement in video frames. Our scheme operates under the assumption that the movement in the frames is caused by a human. The scheme involves three main steps. First, the movement is detected by a differential filter, which has a memory effect on the robot and detects combined patterns of moving objects and colors in the frames. Second, the information gathered from the differential filter is then classified using a k-means clustering algorithm. The k-means clustering algorithm is used to determine the presence and position of the human, information that is then used by the robot to react and track the person. The algorithm has been implemented directly in Python, without relying on image processing libraries such as OpenCV, to reduce its computational requirement and allow for fast execution in parallel with other tasks. Finally, the algorithm has been implemented on the Nao robot developed by Aldebaran Robotics (SoftBank Group), however, its design is scalable and can be implemented on small robots as well [22]. The scheme can also be augmented with auditory information through acoustic tracking [23], either as parallel schemes or by integrating the two sensors into a single algorithm [24]. The integration of auditory information with visual information can provide a more robust tracking solution, which is especially useful in noisy or cluttered environments.

2. BACKGROUND

In recent years, there has been a growing interest in the development of assistive navigation systems and mobile robots to aid blind and visually impaired individuals with indoor travel and to support human-robot interaction experiments. A number of studies have explored different approaches to solving the challenges associated with these systems. For instance, [25] presented a novel vision-based mobile assistive navigation system that helps blind and visually impaired individuals to travel independently indoors. The system integrates various sensors and algorithms to provide a holistic solution that enables these individuals to navigate and avoid obstacles in their environment. In [26] introduced the Georgia Tech Miniature Autonomous Blimp (GT-MAB), which is designed to support human-robot interaction experiments in indoor spaces. The GT-MAB is equipped with sensors and algorithms that allow it to interact with humans and respond to their movements. This system provides a unique platform for studying human-robot interaction and exploring new approaches to assistive navigation. In [27] focused on an alternative solution to existing filtering techniques by introducing the belief condensation filter (BCF) for localization via Bluetooth low energy (BLE)-enabled beacons. The BCF is designed to provide a more efficient and effective way to determine the location of a mobile robot within an environment. By incorporating the BCF, the researchers were able to demonstrate improved accuracy and robustness in the localization of mobile robots.

In another study, Feng *et al.* [28] proposed an integrated indoor positioning system (IPS) that combines the use of inertial measurement units (IMU) and ultra-wideband (UWB) technology. By integrating these technologies, the researchers were able to improve the robustness and accuracy of the IPS by using the extended Kalman filter (EKF) and unscented Kalman filter (UKF). The IPS provides a comprehensive solution for determining the location of a mobile robot within an indoor environment. Al Khatib *et al.* [29] presented a low-cost approach for solving the navigation problem of wheeled mobile robots in indoor and outdoor environments. The researchers proposed an efficient geometric outlier detection method that uses dynamic information from previous frames and a novel probability model to judge moving objects, with the help of geometric constraints and human detection [30]. Liu and Miura [31] proposed a fuzzy detection strategy to prejudge the tracking result and to improve the accuracy of the navigation system. Additionally, they proposed an efficient geometric outlier detection method that uses dynamic information from previous frames and a novel probability model to judge moving objects with the help of geometric constraints and human detection. Xue *et al.* [32] explored the tracking problem of cluster targets and proposed new solutions to improve the accuracy of target tracking.

The recent trend towards the development of internet of things (IoT) architectures has led to the transformation of standard camera networks into smart multi-device systems that are capable of acquiring, elaborating, and exchanging data, and adapting to the environment dynamically. Giordano *et al.* [33] proposed a novel distributed solution that guarantees real-time monitoring of 3D indoor structured areas and tracking of multiple targets. The solution employs a heterogeneous visual sensor network composed of both fixed and pan-tilt-zoom (PTZ) cameras. Finally, Shi *et al.* [34] proposed a novel hybrid method that combines visual and probabilistic localization results to improve the accuracy of indoor positioning systems. By combining these two techniques, the researchers were able to demonstrate improved accuracy in the determination of the location of a mobile robot within an indoor environment.

In conclusion, the field of robotics has seen significant advancements in the development of solutions to support the navigation and localization needs of mobile robots operating in indoor environments. A range of approaches have been proposed, including vision-based systems, integration of IMU and UWB, low-cost approaches, and the incorporation of probabilistic methods. The increasing trend towards the development of IoT architectures has led to the transformation of standard camera networks into smart multi-device systems capable of acquiring, elaborating and exchanging data, and adapting dynamically to the environment. In this line, novel solutions have been proposed that guarantee real-time monitoring and tracking of multiple targets using heterogeneous visual sensor networks. These advancements represent significant steps forward in the development of robust and accurate navigation systems for indoor environments and provide promising avenues for future research and development.

3. METHOD

The visual tracking scheme presented in this study is designed to achieve high performance while being being simple and computationally efficient. The scheme aims to minimize false detections and maintain stability in tracking the target object. The approach utilizes monocular vision and only processes images captured by the robot's camera when it is not undergoing any movement. The images captured by the robot's camera undergo a processing stage, where the existence of movement in the environment is determined. This is achieved through the use of image analysis algorithms and algorithms that detect changes in position and movement patterns. The proposed scheme does not rely on additional sensors, making it a computationally efficient solution for human tracking in robotics.

The architecture of the visual tracking scheme is illustrated in Figure 1. The figure provides a high-level overview of the components and steps involved in the tracking algorithm. The algorithm is designed to operate in real-time, enabling the robot to quickly respond to changes in its environment. The approach taken in the design of the algorithm balances computational efficiency with performance, making it a valuable contribution.

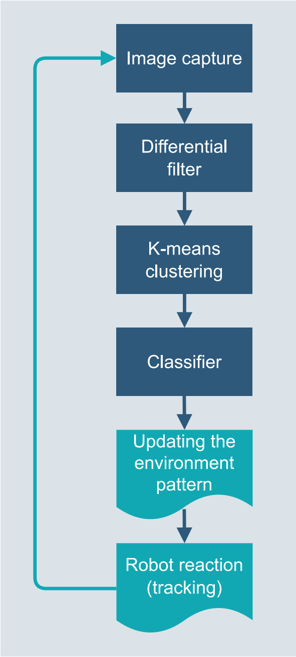

Figure 1. Architecture of the proposed visual tracking scheme

3.1. Differential filter

Our differential filter operates based on the principles of how the human brain processes visual information from the eyes. Just like the retina in the eye, which is located at the back and receives light through specialized cells (rods and cones), our filter differentiates between peripheral and central information. The rods in the eye, responsible for peripheral vision, have high sensitivity to light but are unable to detect details, while the cones, located mostly in the center of the retina, are capable of detecting details and colors. This means that despite having a wide peripheral vision, the eye only processes detailed information in the center of vision where the gaze is focused [35]. As the center of focus moves away, the level of detail decreases.

Our differential filter mimics this biological strategy by focusing its attention on the object of interest, in this case a moving object (human body), and disregarding irrelevant information [36]. The filter detects the object by comparing the current image with a pattern image stored in memory that is constructed from the changes in the pixel values of previous images. This process is a representation of the brain's autonomous mechanism for processing unknown information by relating it to patterns stored in memory based on individual experience, reducing the amount of information to be processed and the total processing time.

The human brain's ability to recognize patterns and relate new information to previously stored information is a crucial aspect of visual processing. This process of relating new information to stored patterns is a powerful strategy for reducing the amount of information that needs to be processed and stored, thus optimizing the processing time and storage costs. The mechanism behind this is autonomous and enables individuals to work with unknown information by relating it to the stored memories. This biological strategy is what makes people identify figures in clouds or faces on toast. The brain continually tries to complete the visual information received from the eyes with previous information stored in memory. By comparing the incoming information with patterns already stored, the brain can quickly make connections and categorize the information, thereby reducing the amount of time and resources required for processing.

Moreover, the brain performs a series of information selection processes to discard irrelevant information, such as information that is not directly related to the task at hand. This selection process further optimizes the visual processing system by reducing the amount of data that needs to be processed, increasing the speed and accuracy of the process. In this way, the functioning of the human brain when processing visual information from the eyes can be seen as a highly optimized system that balances the need for detail and speed, and ensures the most efficient use of resources. This mechanism of processing visual information serves as a model for our differential filter, which mimics the functioning of the human brain to optimize the processing of visual information in our robotic system.

The operating model of our system incorporates the essential features of the human brain's visual information processing mechanism. Detection and tracking of objects in motion is achieved through the identification of movements, with the moving object, in this case, the human body, serving as the center of attention. The rest of the visual information is categorized as background and is considered irrelevant. In this approach, the visual information is differentiated into two main components: the object and the background. The object is the focus of attention, the item in motion that is being monitored, while the background encompasses all the surrounding elements. This differentiation between the object and background is crucial in enabling the system to efficiently process and analyze the visual information, allowing it to selectively focus its attention on the moving object.

By replicating the human brain's visual information processing mechanism, our system is able to achieve a high level of accuracy and efficiency in detecting and tracking objects in motion, making it a valuable tool in various fields, including but not limited to, robotics, surveillance, and human-computer interaction. The object (human body) is detected by comparing the current image *CI*(*t*) with another pattern image stored in memory *BI*(*t*) (background image). This pattern image is constructed from the pixel changes of the previous images. The value of each pixel is updated with each image captured according to (1):

$$g_i = \beta_i g_i + (1 - \beta_i) \times p\left(CI\left(t\right)\right)_i \quad \diagup i = 1,\, 2,\, \ldots,\, l \tag{1}$$

where:

- l is the number of pixels in the image, i.e. $l = m \times n$ for an image of $m \times n$ pixels.
- p_i is the grayscale color (amount of light in the pixel) of the *i*-th pixel in the image.
- b_i is the distance between the pixels *i* of the images *BI*(*t*) and *BI*(*t-1*) calculated as (2):

$$\beta_i = 1 - \frac{\left|p\left(BI\left(t\right)\right)_i - p\left(BI\left(t-1\right)\right)_i\right|}{255} \tag{2}$$

This β inherently exhibits an inverse relationship with pixel distances, ranging from 0 to 1, allocating high values when pixel distances are compact, and conversely, low values when these distances are expansive. One of the defining features of our filter is its ability to create and systematically update a pattern image, denoted by *BI* (*t*), which is temporarily archived in memory. This pattern image formation revolves around the vigilant tracking of fluctuations in pixel values within a sequence of captured images over a specified timeframe, as mathematically represented in (1).

The functional essence of this pattern image resides in its role as a point of reference for identifying moving entities in the present image, termed as *CI* (*t*). This is achieved through a comparative analysis between the pattern image and the current image. Given the continuous updating mechanism of the pattern image, it possesses the capacity to rapidly register alterations and discern between elements that exhibit motion and stationary background elements. Interestingly, background entities that cease motion are eventually disregarded in this process. Figure 2 provides a graphical illustration encapsulating the operational behavior of our differential filter. Furthermore, the practical application and overall functionality of our differential filter is vividly demonstrated in Figure 3.

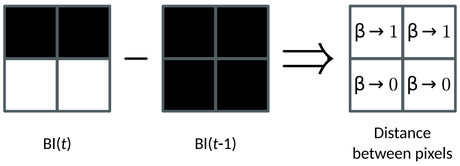

Figure 2. Pattern image construction from in-memory images

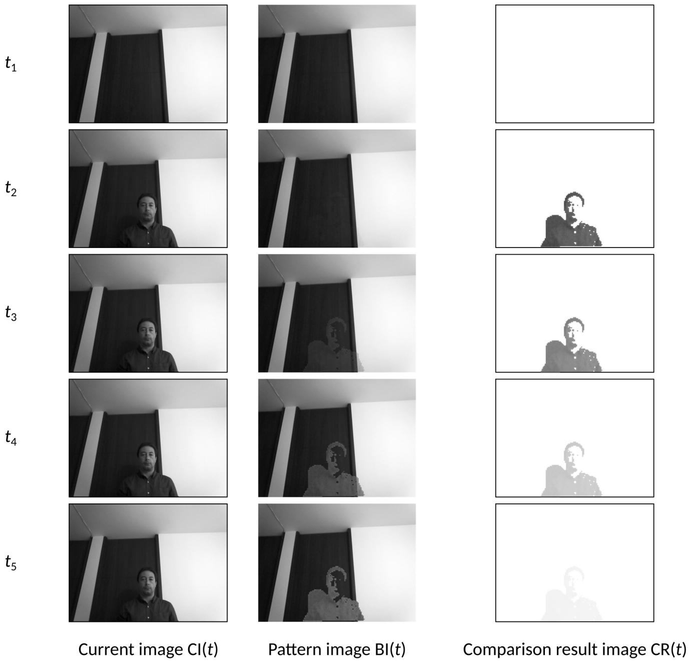

Figure 3. Differential filter operation when a person appears in front of the robot

3.2. Position identifier

The k-means clustering algorithm is a widely used unsupervised machine learning technique that aims to partition a given dataset into a predefined number of clusters, *k*, based on the similarity of their features. The algorithm operates by iteratively updating the centroids of the clusters until convergence, where the data points are assigned to the nearest centroid. One of the key strengths of k-means is its simplicity and computational efficiency, making it a popular choice for many real-world applications such as image segmentation, text clustering, and customer segmentation. Despite its simplicity, k-means is a powerful tool that can effectively identify meaningful patterns and structures in large and complex datasets.

To effectively apply k-means, it is essential to choose an appropriate value for *k*. In general, a larger value of *k* can lead to finer grained clusters, but may also result in over-fitting the data. On the other hand, a smaller value of *k* may result in coarser clusters that do not accurately reflect the underlying structure of the data. A popular heuristic for choosing *k* is the elbow method, which involves analyzing the relationship between the number of clusters and the within-cluster sum of squares.

Once *k* has been selected, k-means operates by first initializing the centroids randomly, and then iteratively updating the cluster assignments and centroids. The algorithm terminates when the cluster assignments no longer change, or when a maximum number of iterations has been reached. The final result of the k-means algorithm is a partition of the data into *k* clusters, with each cluster represented by its centroid. In our study, we use a dataset consisting of pairs of differential filter results, $CR(t)$, and corresponding current images, $CI\ (t)$, in RGB color format. To effectively incorporate both spatial and color information in our analysis, we represent each pixel as a vector that consists of both its position (x, y) and its color information.

However, it is important to note that the color information is only used for the pixels that are identified in the comparison result $CR(t)$. For the background pixels in the image, which are not identified by the differential filter, we assign a white color value of (R = 255, G = 255, B = 255). This allows us to effectively incorporate both the result of the differential filter comparison and the color information of the current image in our analysis. By defining the dataset in this manner, we are able to effectively capture both the spatial and color information of the pixels in the image, which will be useful for subsequent processing and analysis. This is a crucial step in our study, as it provides the necessary foundation for our analysis and allows us to effectively utilize both the information provided by the differential filter and the color information of the current image. Vectors are defined as (3).

$$\boldsymbol{q}_i = (q_{R.i},\, q_{G,i},\, q_{B,i},\, q_{x,i},\, q_{y,i}) \quad \diagup i = 1,\, 2,\, \ldots,\, l \tag{3}$$

In our study, we have chosen to define a total of *k*=3 clusters to be identified in each dataset using the k-means algorithm. The purpose of this is to gain a deeper understanding of the data and to be able to accurately estimate various attributes related to the presence of a person in the dataset. The k-means algorithm provides the coordinates (*x*, *y*) of three points that, when combined, form a triangular area. This triangular area is easily analyzed to estimate if the object within the area is indeed a person. Furthermore, we are able to use the information provided by the k-means algorithm to estimate the approximate distance from the person to the robot, as well as the relative position of the person's head.

Additionally, the triangular area and the information provided by the k-means algorithm can also be used to estimate whether the person is standing or not. This is a crucial step in our study, as it provides valuable information that can be used to inform the robot's actions and interactions with the person. The objective of the robot is to attentively monitor the movements of human beings. To achieve this, we use the information obtained from the k-means algorithm, which provides valuable insights into the presence and attributes of a person in the dataset.

The robot has been programmed to respond to human movements, with the head being the only response element in our laboratory tests. By utilizing the information provided by the k-means algorithm, we are able to coordinate the movements of the robot's head in a way that ensures that the robot always tracks the head of the person. Based on the size and shape of the triangular area, we estimate the distance between the robot and the person. Additionally, we are able to determine the position of the person's head, which is critical information for the robot to effectively track the person. In order to ensure that the robot is always attentively monitoring the person, it is crucial that the face of the robot is always directed towards the person. This is a critical step in our study, as it allows us to ensure that the robot is effectively able to monitor and respond to the movements of human beings.

4. RESULT AND DISCUSSION

We utilize two distinct robotic platforms to address the challenges of human-robot interaction and indoor navigation. The humanoid Nao robot from SoftBank Group serves as the primary interface for human interaction, providing a natural and intuitive means for communication and collaboration with the environment [14]. Equipped with advanced sensors and actuators, the Nao robot allows for real-time monitoring of human behavior and the ability to respond dynamically to changing scenarios. On the other hand, the ARMOS TurtleBot 1 robot from the Arquitecturas Modernas para Sistemas de Alimentación (ARMOS) research group was employed for indoor navigation, leveraging its superior mobility and navigation capabilities. The TurtleBot robot was integrated with the Nao robot through a Wi-Fi connection, providing seamless communication between the two platforms as seen in Figure 4.

To implement our proposed solution, we developed the algorithm in Python, which was programmed onto the Nao robot. The use of Python allowed us to effectively harness the full power of the Nao robot's hardware and software capabilities, while also providing a highly accessible and scalable platform for future developments and advancements. The combination of these two robotic platforms, along with the implementation of our novel algorithm, has the potential to deliver significant improvements in the field of assistive robotics, opening up new avenues for innovative applications and research [29].

We utilized high resolution images of size 1280×720 pixels to capture the movements of objects in the environment. In order to ensure that the movements recorded in the images were solely the result of objects external to the robot, a decision was made to only capture a single image when the robot was not making any movements. This approach was based on the biological model of visual perception, which suggests that in the absence of any movement, the image captured at a given time remains static. To further enhance the accuracy of our results, the images were captured at an interval of 100 milliseconds when no movement was detected. This allowed us to accurately record and process the movements of external objects in real-time, enabling the robot to respond appropriately to its environment.

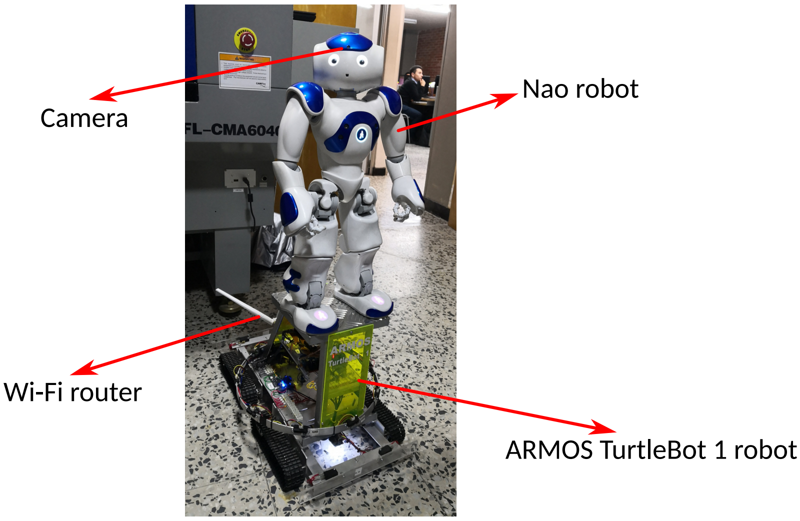

Figure 4. Experimental setup for the human-tracking robot. It is composed of a humanoid Nao robot from SoftBank Group at the top and an ARMOS TurtleBot 1 tank robot from the ARMOS research group at the bottom

In our experiments, we aimed to thoroughly test the capabilities of our assistive robot under various conditions. To achieve this, we conducted a series of experiments with varying lighting conditions, environments, and distances between the robot and the people involved. This was to ensure that our robot could effectively perform its intended tasks in various scenarios, and provide meaningful assistance to the people it interacts with. To further validate our approach, we also performed tests with multiple people present in the robot's field of vision. These tests allowed us to evaluate the robot's ability to distinguish between multiple individuals, and track their movements accurately. This information is crucial in ensuring that our robot can provide personalized assistance to multiple individuals simultaneously.

We documented our experiments using video footage, which can be seen in the accompanying video link [37]. This footage provides visual evidence of the robustness and effectiveness of our assistive robot, and highlights its ability to perform its intended tasks under different conditions. The video footage is also accompanied by a still image see Figure 5, which provides a visual representation of one of our experiments.

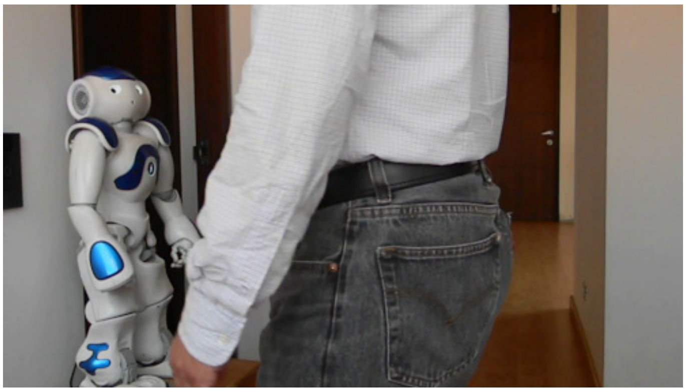

Figure 5. Development of one of the visual tracking tests

The robot has the capability to detect movement in its environment with high accuracy. The algorithm used to achieve this demonstrated remarkable performance in laboratory tests, surpassing the expectations of the researchers. False detections occurred infrequently, only in 4% of the cases, and were primarily due to the movement of the ARMOS TurtleBot 1 platform, which currently does not have direct communication with the Nao robot. A small percentage (12%) of target losses were observed, typically when the object was moving at high speeds and exceeded the robot's image capture rate.

We aimed to ascertain the direct, real-world distance separating the robot and a human entity within the same environmental context [8]. To achieve this, we employed a method centered around geometric analysis of a triangular region distinctly defined by three distinct data points. These specific points were procured by leveraging the statistical of the k-means clustering algorithm. The triangular area constituted by these three points was then meticulously scrutinized. Our method extrapolated the spatial relation of the points within the triangle, using their relative positioning and the magnitude of the area they enclosed, to draw inferences about the actual physical distance between the robot and the individual.

Despite the simplicity of the algorithm, the results obtained were quite remarkable. The estimated distance obtained from the geometric parameters was compared with the real distance of the object, and the results showed a high degree of accuracy as seen in Figure 6. This suggests that the distance estimation scheme derived from this simple algorithm is a valuable tool for robotic applications where distance measurement is important, such as navigation and human-robot interaction.

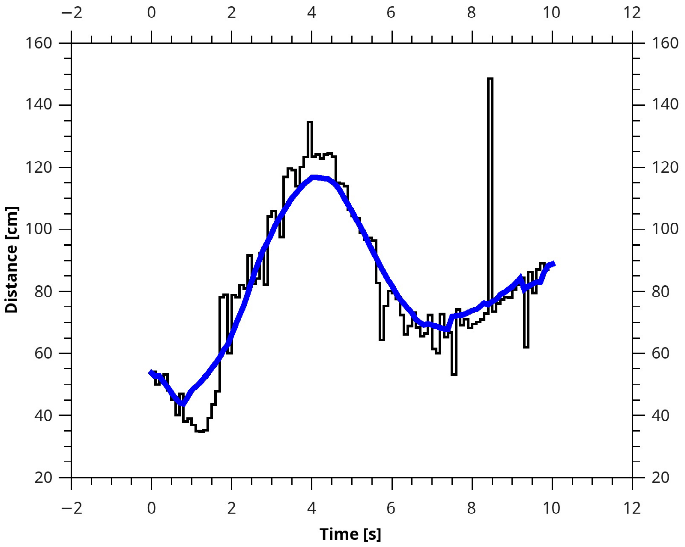

Figure 6. Direct distance from the robot to the person during a movement estimated from the tracking strategy. The black curve represents the raw data obtained from each image analyzed, and the blue curve corresponds to the data smoothed by means of a moving average of 20 points

The success of this method can be attributed to it is reliance on the fundamental principle of geometry, which states that the larger the area of the triangle, the farther the distance between its vertices. By utilizing this principle, our algorithm was able to provide robust and reliable estimates of the distance between the robot and the object in its environment. These results open up new avenues for further research in the field of robotics, and demonstrate the potential for using simple geometric principles to improve the performance of robotic systems.

5. CONCLUSION

The people tracking strategy presented in this paper is a novel contribution to the field of small assistive robots. The strategy's unique combination of high performance and low computational and memory storage cost makes it an attractive solution for the development of small assistive robots. The strategy mimics the human brain's behavior when processing visual information and utilizes a differential filter and a k-means clustering algorithm to track objects in unstructured indoor environments. The results of the performance evaluation showed that the strategy has a low percentage of false detections and accurately tracks the motion parameters of the robot head. The strategy's ability to store short-term memory of previous observations and to construct a pattern image to detect movement is a testament to the robustness of the approach. The strategy's low computational cost makes it an attractive solution for small assistive robots, which are often limited by the available processing power. Additionally, the strategy's low memory storage cost makes it suitable for small robots with limited memory capacity.

In future work, we plan to extend the scope of the strategy's evaluation to include different scenarios with different levels of complexity, such as tracking multiple objects simultaneously or tracking objects in outdoor environments. We also plan to integrate the people tracking strategy into the control architecture of a small assistive robot, to demonstrate its practical applicability. In conclusion, the people tracking strategy presented in this paper represents a significant step forward in the development of small assistive robots. The high performance and low computational and memory storage cost make it a promising solution for the development of future small assistive robots.

ACKNOWLEDGEMENT

This work was supported by the Universidad Distrital Francisco José de Caldas, in part through CIDC, and partly by the Facultad Tecnológica. The views expressed in this paper are not necessarily endorsed by Universidad Distrital. The authors thank the research group ARMOS for the evaluation carried out on prototypes of ideas and strategies.

REFERENCES

[1] O. A. Wudarczyk *et al.*, "Robots facilitate human language production," *Scientific Reports*, vol. 11, no. 1, Aug. 2021, doi: 10.1038/s41598-021-95645-9.

[2] N. Fields, L. Xu, J. Greer, and E. Murphy, "Shall I compare thee… to a robot? An exploratory pilot study using participatory arts and social robotics to improve psychological well-being in later life," *Aging and Mental Health*, vol. 25, no. 3, pp. 575–584, Mar. 2021, doi: 10.1080/13607863.2019.1699016.

[3] A. E. Eiben, E. Hart, J. Timmis, A. M. Tyrrell, and A. F. Winfield, "Towards autonomous robot evolution," in *Software Engineering for Robotics*, Cham: Springer International Publishing, 2021, pp. 29–51.

[4] D. Reverter Valeiras, X. Lagorce, X. Clady, C. Bartolozzi, S.-H. Ieng, and R. Benosman, "An asynchronous neuromorphic event-driven visual part-based shape tracking," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 12, pp. 3045–3059, Dec. 2015, doi: 10.1109/TNNLS.2015.2401834.

[5] W. Takano, "Annotation generation from IMU-based human whole-body motions in daily life behavior," *IEEE Transactions on Human-Machine Systems*, vol. 50, no. 1, pp. 13–21, Feb. 2020, doi: 10.1109/THMS.2019.2960630.

[6] D. Das, M. G. Rashed, Y. Kobayashi, and Y. Kuno, "Supporting human-robot interaction based on the level of visual focus of attention," *IEEE Transactions on Human-Machine Systems*, vol. 45, no. 6, pp. 664–675, 2015, doi: 10.1109/THMS.2015.2445856.

[7] N. F. Duarte, M. Rakovic, J. Tasevski, M. I. Coco, A. Billard, and J. Santos-Victor, "Action anticipation: reading the intentions of humans and robots," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 4132–4139, Oct. 2018, doi: 10.1109/LRA.2018.2861569.

[8] D. Fernandez-Chaves, J.-R. Ruiz-Sarmiento, A. Jaenal, N. Petkov, and J. Gonzalez-Jimenez, "Robot@virtualhome, an ecosystem of virtual environments and tools for realistic indoor robotic simulation," *Expert Systems with Applications*, vol. 208, Dec. 2022, doi: 10.1016/j.eswa.2022.117970.

[9] D. F. P. C, "Performance evaluation of ROS on the Raspberry Pi platform as OS for small robots," *Tekhnê*, vol. 14, no. 1, pp. 61–72, 2017.

[10] J. Castañeda and Y. Salguero, "Adjustment of visual identification algorithm for use in stand-alone robot navigation applications," *Tekhnê*, vol. 14, no. 1, pp. 73–86, 2017.

[11] P. Cigliano, V. Lippiello, F. Ruggiero, and B. Siciliano, "Robotic ball catching with an eye-in-hand single-camera system," *IEEE Transactions on Control Systems Technology*, vol. 23, no. 5, pp. 1657–1671, 2015, doi: 10.1109/TCST.2014.2380175.

[12] K. Zhang, J. Chen, Y. Li, and Y. Gao, "Unified visual servoing tracking and regulation of wheeled mobile robots with an uncalibrated camera," *IEEE/ASME Transactions on Mechatronics*, vol. 23, no. 4, pp. 1728–1739, Aug. 2018, doi: 10.1109/TMECH.2018.2836394.

[13] L. H. Juang and J. Sen Zhang, "Visual tracking control of humanoid robot," *IEEE Access*, vol. 7, no. 1, pp. 29213–29222, 2019, doi: 10.1109/ACCESS.2019.2901009.

[14] K.-H. Lee, J.-N. Hwang, G. Okopal, and J. Pitton, "Ground-moving-platform-based human tracking using visual SLAM and constrained multiple kernels," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 12, pp. 3602–3612, Dec. 2016, doi: 10.1109/TITS.2016.2557763.

[15] E. Petrović, A. Leu, D. Ristić-Durrant, and V. Nikolić, "Stereo vision-based human tracking for robotic follower," *International Journal of Advanced Robotic Systems*, vol. 10, no. 5, May 2013, doi: 10.5772/56124.

[16] T. Sonoura, T. Yoshimi, M. Nishiyama, H. Nakamoto, S. Tokura, and N. Matsuhir, "Person following robot with vision-based and sensor fusion tracking algorithm," in *Computer Vision*, InTech, 2008.

[17] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 2005, vol. 1, pp. 886–893, doi: 10.1109/CVPR.2005.177.

[18] M. Gupta, S. Kumar, L. Behera, and V. K. Subramanian, "A novel vision-based tracking algorithm for a human-following mobile robot," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 47, no. 7, pp. 1415–1427, Jul. 2017, doi: 10.1109/TSMC.2016.2616343.

[19] K. Wang, Y. Liu, and L. Li, "Vision-based tracking control of underactuated water surface robots without direct position measurement," *IEEE Transactions on Control Systems Technology*, vol. 23, no. 6, pp. 2391–2399, 2015, doi: 10.1109/TCST.2015.2403471.

[20] J. P. C. de Souza, A. M. Amorim, L. F. Rocha, V. H. Pinto, and A. P. Moreira, "Industrial robot programming by demonstration using stereoscopic vision and inertial sensing," *Industrial Robot*, vol. 49, no. 1, pp. 96–107, 2022, doi: 10.1108/IR-02-2021-0043.

[21] W. Ye, Z. Li, C. Yang, J. Sun, C. Y. Su, and R. Lu, "Vision-based human tracking control of a wheeled inverted pendulum robot," *IEEE Transactions on Cybernetics*, vol. 46, no. 10, pp. 2423–2434, 2015, doi: 10.1109/TCYB.2015.2478154.

[22] F. Hernán, M. Sarmiento, J. Gómez, D. Alexander, and Z. Díaz, "Concepto de robot humanoide antropométrico para investigación en control," *Tecnura*, vol. 19, no. SPE, pp. 55–65, 2015, doi: 10.14483/22487638.10372.

[23] E. D. Lasso, A. Patarroyo Sánchez, and F. H. Martinez, "Acoustic tracking system for autonomous robots based on TDE and signal intensity," *Tecciencia*, vol. 10, no. 19, pp. 43–48, 2015, doi: 10.18180/tecciencia.2015.19.7.

[24] M. Wang *et al.*, "Accurate and real-time 3-D tracking for the following robots by fusing vision and ultrasonar information," *IEEE/ASME Transactions on Mechatronics*, vol. 23, no. 3, pp. 997–1006, 2018, doi: 10.1109/TMECH.2018.2820172.

[25] B. Li *et al.*, "Vision-based mobile indoor assistive navigation aid for blind people," *IEEE Transactions on Mobile Computing*, vol. 18, no. 3, pp. 702–714, 2019, doi: 10.1109/TMC.2018.2842751.

[26] N. shi Yao *et al.*, "Autonomous flying blimp interaction with human in an indoor space," *Frontiers of Information Technology and Electronic Engineering*, vol. 20, no. 1, pp. 45–59, 2019, doi: 10.1631/FITEE.1800587.

[27] S. Mehryar, P. Malekzadeh, S. Mazuelas, P. Spachos, K. N. Plataniotis, and A. Mohammadi, "Belief condensation filtering for RSSI-based state estimation in indoor localization," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 8385–8389, doi: 10.1109/ICASSP.2019.8683560.

[28] D. Feng, C. Wang, C. He, Y. Zhuang, and X. G. Xia, "Kalman-filter-based integration of IMU and UWB for high-accuracy indoor positioning and navigation," *IEEE Internet of Things Journal*, vol. 7, no. 4, pp. 3133–3146, 2020, doi: 10.1109/JIOT.2020.2965115.

[29] E. I. Al Khatib, M. A. K. Jaradat, and M. F. Abdel-Hafez, "Low-cost reduced navigation system for mobile robot in indoor/outdoor environments," *IEEE Access*, vol. 8, no. 1, pp. 25014–25026, 2020, doi: 10.1109/ACCESS.2020.2971169.

[30] Y. Liu and J. Miura, "KMOP-vSLAM: dynamic visual SLAM for RGB-D cameras using k-means and OpenPose," in *2021 IEEE/SICE International Symposium on System Integration (SII)*, Jan. 2021, pp. 415–420, doi: 10.1109/IEEECONF49454.2021.9382724.

[31] S. Liu, S. Wang, X. Liu, C.-T. Lin, and Z. Lv, "Fuzzy detection aided real-time and robust visual tracking under complex environments," *IEEE Transactions on Fuzzy Systems*, vol. 29, no. 1, pp. 90–102, Jan. 2021, doi: 10.1109/TFUZZ.2020.3006520.

[32] X. Xue, S. Huang, N. Li, and W. Zhong, "Resolvable cluster target tracking based on wavelet coefficients and JPDA," in *2021 International Symposium on Computer Technology and Information Science (ISCTIS)*, Jun. 2021, pp. 330–336, doi: 10.1109/ISCTIS51085.2021.00074.

[33] J. Giordano, M. Lazzaretto, G. Michieletto, and A. Cenedese, "Visual sensor networks for indoor real-time surveillance and tracking of multiple targets," *Sensors*, vol. 22, no. 7, 2022, doi: 10.3390/s22072661.

[34] H. Shi, J. Yang, J. Shi, L. Zhu, and G. Wang, "Vision-sensor-assisted probabilistic localization method for indoor environment," *Sensors*, vol. 22, no. 19, 2022, doi: 10.3390/s22197114.

[35] A. Marie, H. G. A. Burton, and P. F. ois Loos, "Perturbation theory in the complex plane: Exceptional points and where to find them," *Journal of Physics Condensed Matter*, vol. 33, no. 28, 2021, doi: 10.1088/1361-648X/abe795.

[36] K. Kollom *et al.*, "A four-country cross-case analysis of academic staff expectations about learning analytics in higher education," *Internet and Higher Education*, vol. 49, no. 2020, 2021, doi: 10.1016/j.iheduc.2020.100788.

[37] F. Mart´ınez, Colombia. Nao robot tracking people. (Apr. 20, 2019). Accessed: Apr. 10, 2023. [Online Video]. Available: https://youtu.be/Gnz0aRBoOsA.

BIOGRAPHIES OF AUTHORS

Fredy Martínez is a professor of control, intelligent systems, and robotics at the Universidad Distrital Francisco José de Caldas (Colombia) and director of the ARMOS research group (Modern Architectures for Power Systems). His research interests are control schemes for autonomous robots, mathematical modeling, electronic instrumentation, pattern recognition, and multi-agent systems. Martínez holds a Ph.D. in Computer and Systems Engineering from the Universidad Nacional de Colombia. He can be contacted at email: fhmartinezs@udistrital.edu.co.

Holman Montiel is a professor of algorithms, embedded systems, instrumentation, telecommunications, and computer security at the Universidad Distrital Francisco José de Caldas (Colombia) and a researcher in the ARMOS research group (Modern Architectures for Power Systems). His research interests are encryption schemes, embedded systems, electronic instrumentation, and telecommunications. Montiel holds a master's degree in computer security. He can be contacted at email: hmontiela@udistrital.edu.co.

Fernando Martínez is a doctoral researcher at the Universidad Distrital Francisco José de Caldas focusing on the development of navigation strategies for autonomous vehicles using hierarchical control schemes. In 2009 he completed his M.Sc. degree in Computer and Electronics Engineering at Universidad de Los Andes, Colombia. He is a researcher of the ARMOS research group (Modern Architectures for Power Systems) supporting the lines of electronic instrumentation, control and robotics. He can be contacted at email: fmartinezs@udistrital.edu.co.