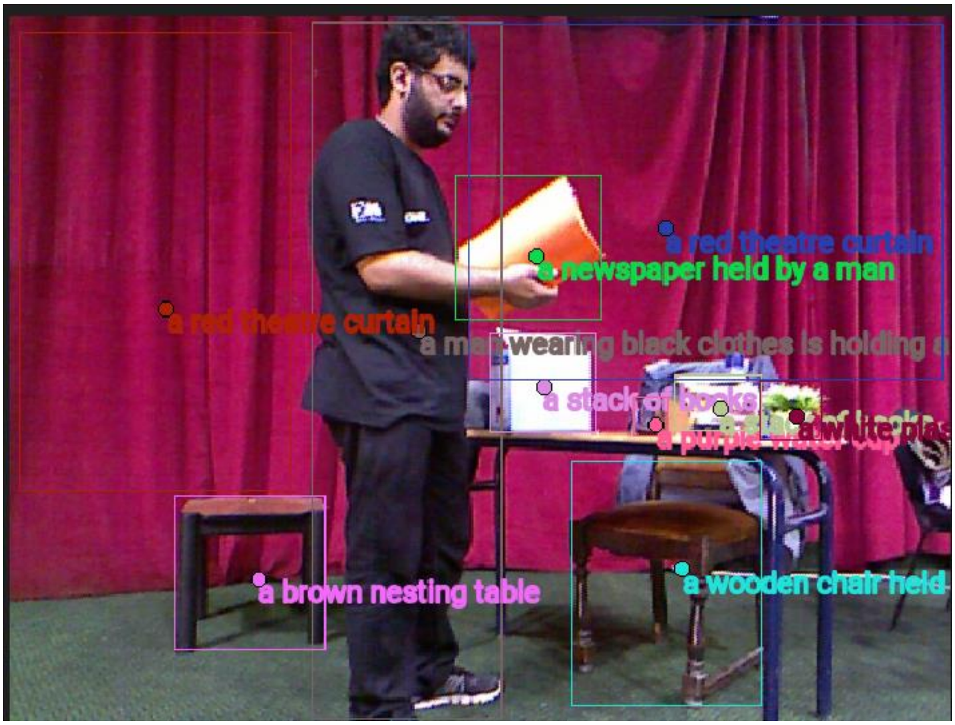
ISSN 2830-9820

VisualComputing MAGAZiNE

Vol. 1, Issue 4
2024

Visual Computing at the Computer Science Faculty of USTHB University



Towards a Smart Campus:
Embedded Cheating Detection with NVIDIA Jetson Nano

The Foreword

Dear Students, Dear Professors,

After the success of the first issues, we have the great pleasure of presenting to you issue 4 of the visual computing magazine, the fruit of the talent and ingenuity of our Master Visual Computing students as well as our doctoral students.
This magazine is much more than just a publication; it embodies the excellence and perseverance of our faculty's researchers for the creation of knowledge and its translation into the startups of the future. The work presented in these pages demonstrates that the specialty of visual computing is at the forefront of innovation and that it opens up great prospects in fields as varied as Computer Vision, Artificial Intelligence, Virtual Reality, Data Visualization and many more.

We would like to warmly congratulate all those who contributed to the creation of this journal, and invite students, doctoral students and teachers to follow this path of disseminating knowledge in order to raise the scientific level of our faculty and the vitality of our community academic.

Prof. Malika Ioualalen-Boukala
Dean of Computer Science Faculty
USTHB University
Algiers, Algeria

© USTHB University
Visual Computing: A quarterly magazine. Volume 1, Issue 4, 2024
***ISSN:* 2830-9820**
Chief Editor: Prof. Slimane LARABI

Skeleton-based Human Action Recognition System with GCN

L. Benhamida, Prof. S. Larabi, Computer Science Faculty, USTHB University

1. Introduction

The skeleton-based action recognition task has been addressed using Graph Convolution Networks (GCN) by treating the sequences of skeleton movements as spatio-temporal graphs (Figure 1), where the joints are treated as nodes and the links between different joints represent the edges linking the nodes. First GCN-based method was proposed by Yan et al [2] using a successive spatial graphs and one-dimensional temporal graph convolution blocks: ST-GCN. The adjacency matrix and the feature map of the spatio-temporal graph are injected into the model's input layer. When tested on benchmark datasets, this new approach achieved state-of-the-art performance. Thus, many ST-GCN variants have been developed in the last few years, each addressing a specific limitation in the original implementation.

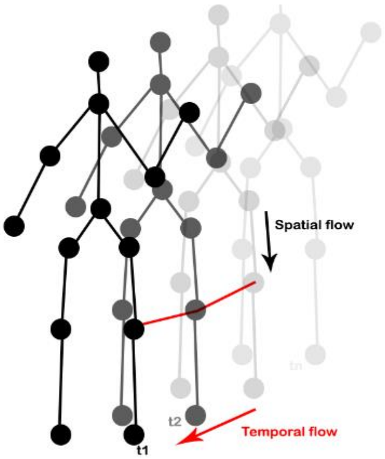


Figure1: Spatio-temporal skeleton representation: edges in black are spatial edges and red links are the temporal edges. [5]

However, the performance of these models remain unclear when applied to realistic applications with untrimmed videos for an online recognition. Most of the existing solutions for an online recognition use deep learning models in order to identify the starting point of each action in an untrimmed video. Hence, the flexibility provided by these methods comes at an enormous computational cost. As a result, the timely responses that are essential in some scenarios might not be provided. In addition, very few of the provided online HAR methods are based on skeleton data that are captured by the Kinect sensor. Most of them are based on RGB videos resulting in a relatively low performance due to the difficulty of human segmentation task.

In this work, we focus on exploiting one of the most powerful state-of-the-art Graph Convolution Network: Disentangled Unifying Multi-Scale GCN (MS-G3D)[3], using skeleton data provided by the Kinect sensor in order to develop an online human action recognition(HAR) system.

Skeleton-based Human Action Recognition System with GCN

L. Benhamida, Prof. S. Larabi, Computer Science Faculty, USTHB University

2. Method

Skeleton Data Captured by Kinect Sensor:

The Kinect sensor, developed by Microsoft, represents a groundbreaking advancement in the field of motion sensing technology. It utilizes an array of cameras and sensors to track the movements of users in three-dimensional space. Developers have leveraged the Kinect's capabilities to create interactive and immersive experiences, where users can control devices or interact with virtual environments using natural body movements. One of its key features is its ability to capture skeletal data, providing a highly detailed and accurate representation of the user's body movements. This skeletal data is provided as a set of 3D points representing different body joints as shown in the figures 2, 3.

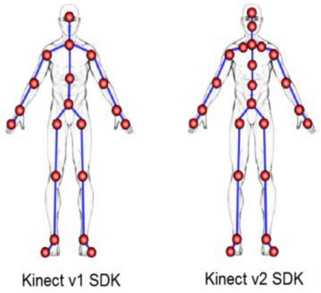


Figure 2: Skeleton joints provided by Kinect

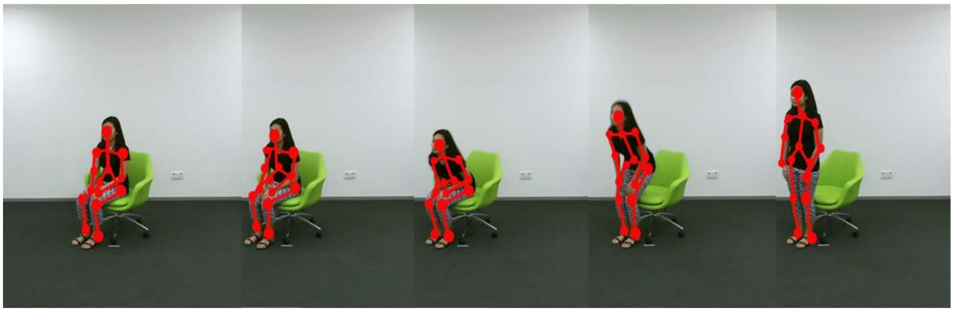


Figure3: Skeleton sequence of a person standing up.

Skeleton-based Human Action Recognition System with GCN

L. Benhamida, Prof. S. Larabi, Computer Science Faculty, USTHB University

MS-G3D

This GCN model uses a disentangled multi-scale aggregator that obtains direct information from farther nodes of the graph in input and removes redundant dependencies between node features. It also uses a unified spatial-temporal graph convolution operator to facilitate direct information flow across space and time. The combination of these two methods results in a powerful feature extraction across both spatial and temporal dimensions.

MS-G3D with Sliding Window Strategy

MS-G3D was designed to be used with trimmed videos, not for online recognition where the boundaries of each action are unknown. Some researchers handle the problem of untrimmed videos by using sliding window strategy. This strategy consists of giving the model a starting point from the untrimmed video and a size **n** to the frame-window to be classified, and then sliding the frame-window by a stride step **p** to classify the rest of the video (Figure 4).

The performance of this strategy relies on the values of **n** and **p** chosen for the sliding window and the stride step respectively. To fix the values of these two parameters that strike balance between the classification performance and the computation cost, we conducted a study on different action sequences. However, to our knowledge, no dataset is available on the internet that contains untrimmed RGB-D action sequences. Thus, we concatenated actions from the NTU RGB+D60 human action dataset resulting in ten untrimmed videos. We tested the sliding window strategy with n = [20, 25, 30, 35], and p = [3, 6, 10] for each value of n.

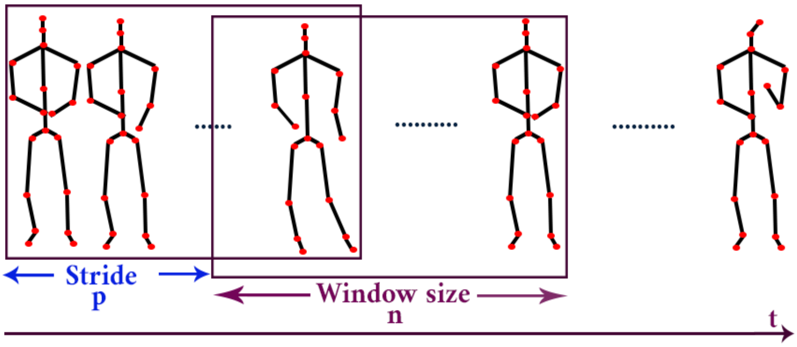


Figure 4: Sliding window strategy [1]

Skeleton-based Human Action Recognition System with GCN

L. Benhamida, Prof. S. Larabi, Computer Science Faculty, USTHB University

Action Transition Detection

We propose a method to locate actions by detecting the transition from one action to another using skeleton's joints distribution with SVM which is known for producing significant accuracy with less computation power. We use a sliding window W of n frames. Once a transition is detected in W, the recognition classifier (MS-G3D) is called to recognize the corresponding action (see figure 5).

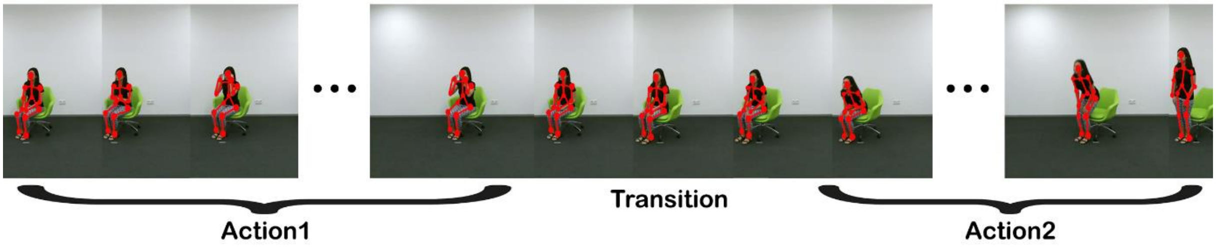


Figure5. Transition between two actions in a video sequence

$$W = F_1\, F_2\ \dots F_i \dots F_n$$

F_i is a matrix of 25×4 where 25 is the number of joints, and 4 refers to the joint's coordinates and time (x, y, z, t).

$$F_i \begin{bmatrix} j_{1,1} & \cdots & j_{1,4} \\ \vdots & \ddots & \vdots \\ j_{25,1} & \cdots & j_{25,4} \end{bmatrix}$$

First, we transform each $F_i \in W$ to a vector f_i of 25 components by reducing the dimensionality of skeleton coordinate system using Principal Component Analysis (PCA). As a result, W is transformed to a matrix of 25×n.

$$W = f_1\, f_2\ \dots f_i \dots f_n$$

Second, we apply PCA dimensionality reduction again on W in order to obtain a vector V of n points.

Finally, an SVM is used to learn the distribution of the points of different vectors and find a hyper-plane that distinctly classifies the vectors.

To train the SVM model, we generated two classes: **same-action** class and **transition** class, using skeleton sequences from NTU–RGBD60 dataset. **same-action** class was generated by calculating vectors from each NTU-RGBD action sequence, and **transition** class was generated by calculating vectors of pairwise combinations of different action classes.

Skeleton-based Human Action Recognition System with GCN

L. Benhamida, Prof. S. Larabi, Computer Science Faculty, USTHB University

MS-G3D with the Action Transition

By combining the obtained results from both experiments on MS-G3D with a simple sliding window method and the action transition detection method, we implemented the following system for online HAR: Once a transition is determined using a 20-frame window using the proposed method, the next 35-frame window is fed to the MS-G3D model to recognize the following action, and then we slide the window by 10 frames (Figure6).

3. Results

After analyzing the findings of the conducted statistics, we found that a sliding window of n = 35 with a stride step of 10 frames is the most effective so far to guarantee better classification with the least computation time. However, even with the chosen parameter values, the computation time is still relatively high, due to the use of the MS-G3D model every 10 frames.

To locate actions by detecting the transition from one action to another, we trained the SVM with different values of n, the results showed that SVM obtained the best accuracy with 20-frame window. We conclude that the best size for best action transition detection is n=20.

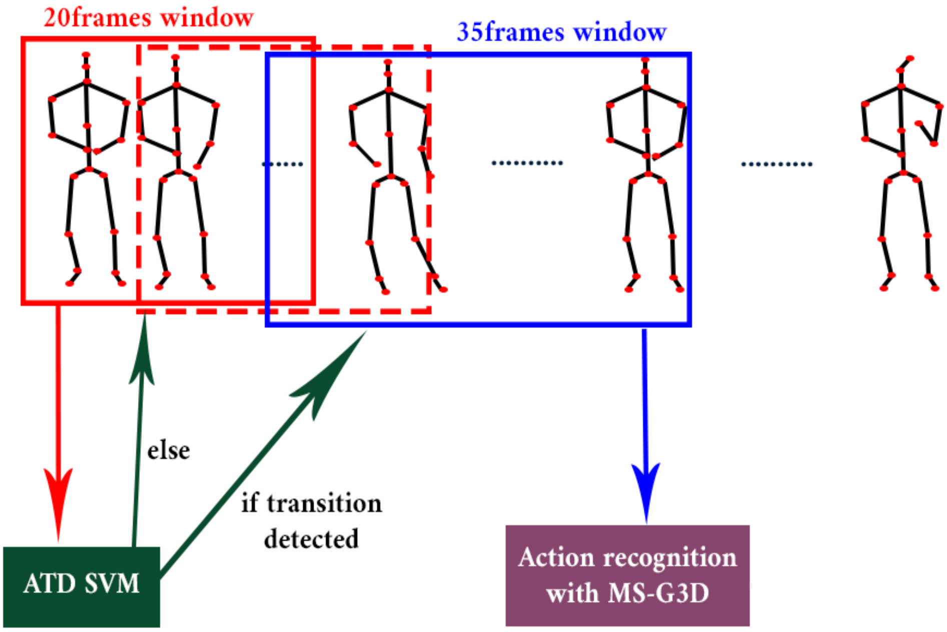


Figure 6: Action Transition method with MS-G3D for an online HAR [1]

Skeleton-based Human Action Recognition System with GCN

L. Benhamida, Prof. S. Larabi, Computer Science Faculty, USTHB University

After a comparison study between the proposed online HAR system and the simple sliding window, we observed a huge difference in computation time with quite similar and sometimes better recognition performance. The minimum time cost with a simple sliding window was 7 seconds, whereas the maximum time cost using the proposed system is 6 seconds. This proves that our system is effective at reducing computation time, which is crucial when using a deep learning model for real-time HAR applications, while preserving the same recognition performance of the model.

Conclusion

This work provides a method that can be employed with any offline skeleton-based HAR deep learning model for real-time applications. The method is based on the concept of sliding window, but rather than calling the model to classify the window at each stride step, we do so only when a transition between two actions is detected. We were able to determine the ideal values for both parameters: window size n and stride step n, which ensure the best recognition performance of MS-G3D with the lowest computation cost. This helps to reduce the computation time of the recognition system while maintaining the model's performance.

References

[1] Benhamida, Leyla, and Slimane Larabi. "Human Action Recognition and Coding based on Skeleton Data for Visually Impaired and Blind People Aid System." 2022 First International Conference on Computer Communications and Intelligent Systems (I3CIS). IEEE, 2022.
[2] Yan, Sijie, Yuanjun Xiong, and Dahua Lin. "Spatial temporal graph convolutional networks for skeleton-based action recognition." Proceedings of the AAAI conference on artificial intelligence. Vol. 32. No. 1. 2018.
[3] Liu, Ziyu, et al. "Disentangling and unifying graph convolutions for skeleton-based action recognition." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.
[4] Shahroudy, Amir, et al. "Ntu rgb+ d: A large scale dataset for 3d human activity analysis." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
[5] Benhamida, Leyla, and Slimane Larabi. "Theater Aid System for the Visually Impaired Through Transfer Learning of Spatio-Temporal Graph Convolution Networks." arXiv preprint arXiv:2306.16357 (2023).

Visual Computing Magazine

Design and development of VR-based exergames for functional hand rehabilitation after stroke

A. Bouatrous(1), N. Zenati(2), A. Meziane(3), C. Hamitouche(4), (1) Computer Science Faculty, USTHB University, (2)CDTA, (3)CERIST, (4) IMT Atlantique Bretagne, France.

1. Introduction

Stroke is one of the world's leading causes of disability, particularly when it comes to the motor skills needed to carry out daily tasks and professional commitments. A large proportion of stroke victims suffer from impaired motor function in their upper limbs. Even after recovery, many stroke survivors encounter difficulties in their daily activities due to persistent problems with upper limb functionality, necessitating medical intervention through rehabilitation. The field of functional rehabilitation in rehabilitation centers is in need of significant enhancements in terms of rehabilitation tools. Conventional clinical protocols for functional rehabilitation often fall short due to issues such as boredom, affecting patients' motivation and subsequently hindering the recovery of their motor functions. In response to these challenges, innovative therapies like virtual rehabilitation have emerged as promising alternatives. Virtual rehabilitation involves the use of virtual reality (VR) therapy, utilizing software systems that simulate real-world tasks to enhance rehabilitation for stroke patients (Figure 1).

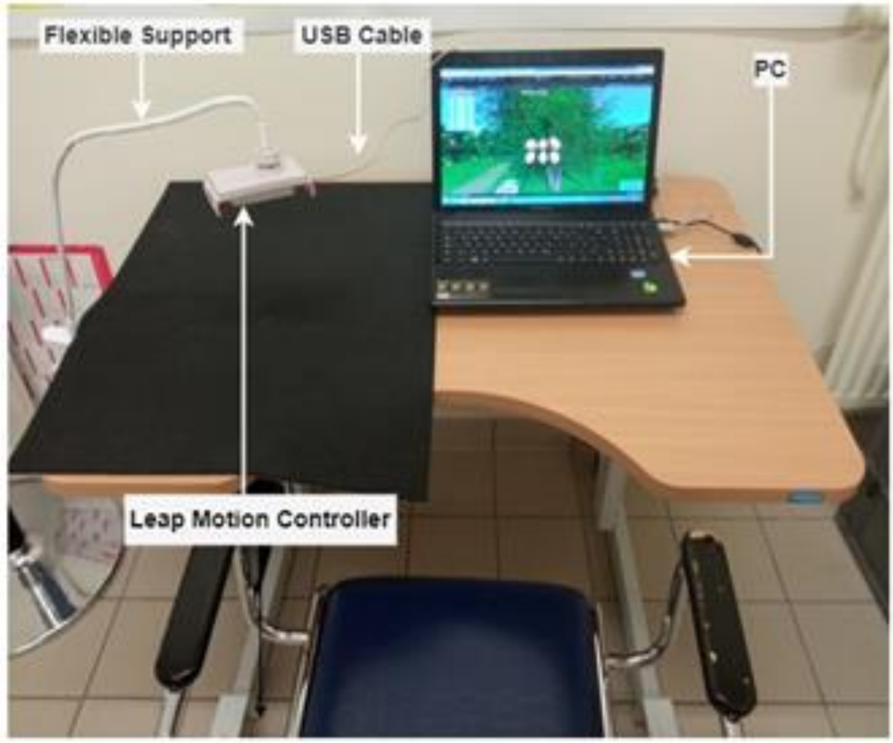


Figure 1. System hardware components [5].

Design and development of VR-based exergames for functional hand rehabilitation after stroke

A. Bouatrous[1], N. Zenati[2], A. Meziane[3], C. Hamitouche[4], (1) Computer Science Faculty, USTHB University, (2)CDTA, (3)CERIST, (4) MT Atlantique Bretagne, France.

One notable aspect of virtual rehabilitation is the incorporation of serious games, a form of gamification, into motor rehabilitation after stroke. Exergames, which inject an element of enjoyment into rehabilitation exercises. Gamification transforms rehabilitation into an enjoyable experience, requiring participants to engage in physical activities that involve specific body movements. The primary goal of stroke rehabilitation is to promote cerebral plasticity and functional compensations, requiring a personalized approach tailored to individual patient needs based on factors such as stroke severity, age, and pre-stroke occupation , and VR technology offers customizable and engaging environments for rehabilitation . The ideal VR platform for functional rehabilitation, should not only motivate patients but also offer personalization and adaptability in the form of progressively challenging exercises [1]. In recent years, artificial intelligence (AI) has played a pivotal role in advancing human welfare, particularly in the realm of rehabilitation methods. In this context, our paper presents a functional rehabilitation system based on VR and integrating serious games, aimed at enhancing patients' motivation, commitment and therapeutic adherence.

The system has been designed and developed, taking into account various criteria associated with the clinical specifications of the medical context concerned. The design process involved close collaboration with therapists and patients, and validation was carried out with post-stroke patients. Simulated clinical exercises were chosen to ensure that the system was perfectly aligned with the clinical specifications. These exercises were tested, evaluated and validated beforehand, demonstrating their therapeutic effectiveness. Particular attention was paid to the scenario, customization and adaptive difficulty during the design and development phases. Subjective evaluation of the system was carried out using standardized questionnaires.

2. Method

The activity scenario

Gamified exercises help the user train complex skills required for everyday activities that involve the same movements. Thus, an activity is linked to the values and culture of the person and the occupational therapist must ensure that the proposed activity is meaningful for the person who is sufficiently committed to relearn how to practice the activity.

Consequently, as part of the VR-based serious games we designed, an environment representing a virtual farm was chosen, which constitutes a user-friendly environment. Great care was taken in designing this environment.

The environment consists of a set of fruit trees. In which the user is encouraged to perform a fruit harvesting activity, in order to carry out the predefined movements.

Design and development of VR-based exergames for functional hand rehabilitation after stroke

A. Bouatrous[1], N. Zenati[2], A. Meziane[3], C. Hamitouche[4], (1) Computer Science Faculty, USTHB University, (2)CDTA, (3)CERIST, (4) MT Atlantique Bretagne, France.

Hardware and software system

The hardware system used in this work consists of a computer, connected to a hand motion sensor which is Leap Motion controller (LMC). The LMC is a non-contact motion capture device that allows the user to interact in a virtual environment. This sensor has two infrared cameras in a stereo vision configuration [2]. It provides data on the position of the user's finger phalanges [3]. It tracks in real time the 3D positions of each finger bone in a field of view of 135 to 120 frames per second [2] . The LMC SDK uses an internal model of the human hand to provide predictive tracking, even when parts of the hand are not visible. This model always provides the positions of the five fingers, but tracking is optimal when the outline of the hand and all its fingers are clearly visible. The software uses the visible parts of the hand, its internal model, and past observations to calculate the most likely positions of the parts that are not currently visible. Thus, given that a stroke patient with moderate to severe symptoms may only be able to move their arm to a small extent [4]. Therefore, to allow patients with low mobility/strength to be able to place their paretic limb on the table so that they do not have to fight against gravity, we placed the Leap Motion© sensor upside down over the patient's arm. Thus, the Leap Motion© sensor was placed above the table on a universal flexible support for smartphone. The complete hardware architecture of the system is presented in Fig. 1. The exergames was developed using Unity 3D and the scripts were realized by the C# programming language. Thus, to obtain the 3D vectors of the palm positions and finger joints, the LMC Software Development Kit (SDK) was used. To implement the exergame adaptation system, we used the Python programming language.

Exergames

The system is designed primarily for functional rehabilitation of the hand. Two activities have been selected, which can be performed in a seated position, with the aim of recovering motor skills in the hand. One consists of training a coarse hand movement called ”palmar grip”, and the other a fine hand movement called ”bidigital grip”. Indeed, we agreed to simulate two clinical exercises, in an attractive virtual environment with intuitively simple virtual activity scenarios.

1) Palmar grip exergame: Coarse dexterity, essential for handling large objects, is crucial in a variety of everyday and professional activities. To improve this motor skill, we created an Exergame simulating two clinically proven exercises focused on strengthening palmar prehension. The first exercise involves manipulating a ball to develop finger flexion and hand-finger coordination. The second, uses modelling clay to strengthen the muscles associated with finger flexion and improve movement control.
2) Bidigital grip exergame: Fine dexterity is crucial for handling small objects in everyday and professional life. To improve this motor skill, we have created a game simulating two clinically proven exercises to train bidigital grasping. These exercises consist of a terminal opposition pinch, for gripping very fine objects, and a sub-terminal opposition pinch, ideal for holding objects such as a pencil or sheet of paper.

Design and development of VR-based exergames for functional hand rehabilitation after stroke

A. Boudtrous[(1)], N. Zenati[(2)], A. Meziane[(3)], C. Hamitouche[(4)], (1) Computer Science Faculty, USTHB University, (2)CDTA, (3)CERIST, (4) MT Atlantique Bretagne, France.

3) Virtual environment: In the VR games, the patient has a semi-immersive experience on a virtual farm with fruit trees. The palmar grip exergame involves medium-sized fruits, such as oranges and apples, corresponding to the dimensions of a human hand and resembling the balls used in clinical exercises, while the bidigit grip exergame includes relatively small-sized fruits, such as cherries, in the virtual environment. Here, patients are asked to perform finger flexion movements resembling those used in simulated clinical exercises to pick the virtual fruit.

Tasks

In order to increase the realism of this activity, we added to the set of movements evoked by the simulated exercises a set of sensory-motor tasks necessary for the natural performance of the activity. In both exergames, to succeed in the game, the patient must perform the following tasks:

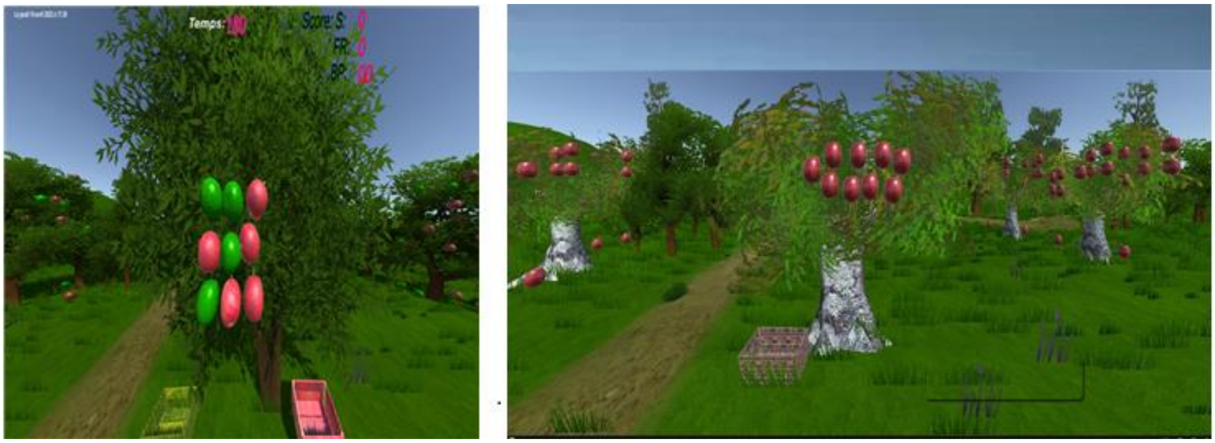


Figure 2. The scenes representing the virtual environments where the games occurs .

- A pointing task: In this task, participants perform the palmar grasping exercise by moving their hand towards the fruit and using their palm to indicate it. In the bidigital grasping exercise, the index finger is used to point at the fruit, enabling the patient to pick it up more easily afterwards.
- A grasping task: In the palmar grip exercise, patients have to flex their fingers to bring them closer to the palm. In the bidigital grip exercise, the task is to bring the thumb and index finger as close together as possible, enabling patients to effectively grasp the fruit they have pointed to.
- A deposit task: In the palmar grip exercise, participants extend their fingers away from the palm to open the hand. In the bidigital grip exercise, they separate the thumb and forefinger to free the fruit placed between them, set it down, then place the freed fruit in a basket.

Design and development of VR-based exergames for functional hand rehabilitation after stroke

A. Bouatrous(1), N. Zenati(2), A. Meziane(3), C. Hamitouche(4), (1) Computer Science Faculty, USTHB University, (2)CDTA, (3)CERIST, (4) MT Atlantique Bretagne, France.

User representation

As mentioned earlier, our system is particularly interested in the functional rehabilitation of the hand. The hand is therefore the limb that will be in direct interaction with the entities of the virtual environment. As patients' hands are often deficient and retracted, we have chosen to display a virtual hand instead of the patient's real hand (Figure 3). This is supposed to give a more positive perception to the patients. The movements of the virtual hand are synchronized with those of the real hand.

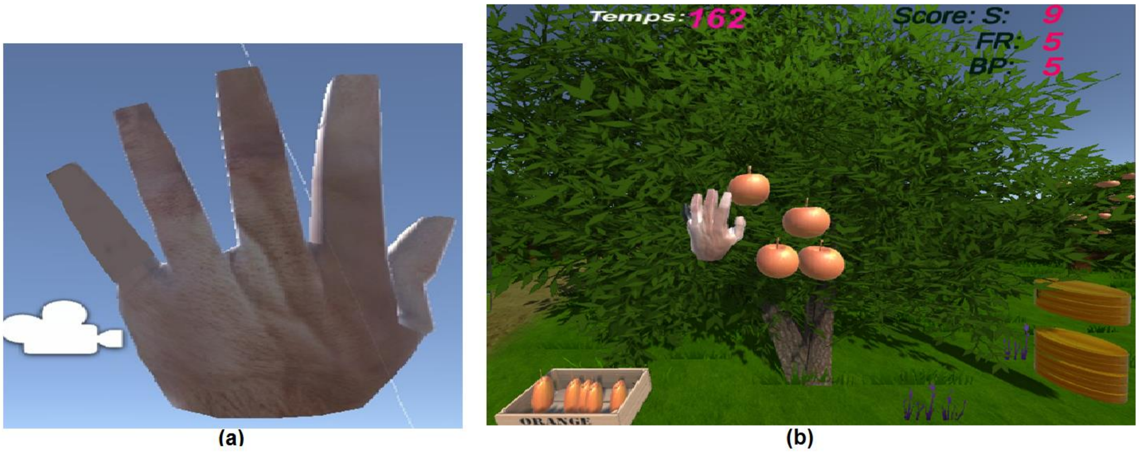


Figure 3. The virtual hand used in the exergame. (a) The hand's avatar . (b) Representation of the hand's avatar in the virtual environment [5].

Measurement of patient motion

The games on offer incorporate feedback from therapists to determine the amplitude of hand movements. The palmar grip exercise is characterized by the ability to open and close the hand, while the bidigital grip exercise involves moving the thumb away from the index finger. Range of motion is defined as the ability of the hand to open and close. This measure is determined by the distance between the fingertips (excluding the thumb) and the palm in the first exercise [5], and by the distance between the tip of the thumb and the index finger in the second exercise.

Visual Computing Magazine

Design and development of VR-based exergames for functional hand rehabilitation after stroke

A. Bouatrous(1), N. Zenati(2), A. Meziane(3), C. Hamitouche(4), (1) Computer Science Faculty, USTHB University, (2)CDTA, (3)CERIST, (4) MT Atlantique Bretagne, France.

Exergames difficulty

To enable actions in the games and determine their level of difficulty, we have established specific thresholds for the patient's finger movements. These thresholds serve as benchmarks and enable us to assess the patient's ability to perform specific tasks [5]. The patient's ability to perform these tasks successfully becomes the main criterion for defining the complexity of a game. By adjusting the difficulty of the game according to the patient's performance, we aim to provide a personalized and stimulating rehabilitation experience that matches the patient's skill level and promotes continuous improvement.

Exergames personalization

Game calibration is crucial to tailoring the difficulty level to the patient's current abilities. Before starting the game, the patient performs the specific finger movements required to complete the game tasks. These movements include actions such as opening and closing the hand in the palmar grip exercise, and separating and bringing together the thumb and index finger in the bidigital grip exercise. This initial phase encourages the patient to perform the fundamental movements essential for play.

To reinforce autonomy in the game, the system offers on-screen video tutorials illustrating the required movements. The aim of these tutorials is to provide visual and practical guidance, making it easier for the patient to understand and perform the required movements precisely [6].

The system assesses the patient's motor skills by measuring the amplitude of finger movements, in particular the maximum and minimum amplitudes achieved. Based on these measurements, the system establishes personalized thresholds corresponding to the patient's abilities. These thresholds are used as a reference to determine the appropriate level of difficulty for the game, ensuring that it is adapted to the patient's individual abilities.

Adaptive difficulty

In upper limb motor rehabilitation, preventing patient fatigue and maintaining motivation is crucial [7]. To achieve this, a dynamic approach is implemented that adjusts the difficulty level of the exergames based on the patient's motor skills. The system analyzes previous performances and utilizes unsupervised machine learning (clustering) to categorize patients into low, medium, and high-performance clusters. The K-means algorithm is employed, and the patient's performance is assigned to the most relevant cluster. By monitoring parameters such as attempts to grasp an object, the system detects the risk of fatigue and updates the game's difficulty accordingly. This personalized approach aims to strike a balance, ensuring the game is challenging enough to promote intensity but not too difficult to discourage engagement. The update involves adjusting thresholds with the average values of the cluster representing the patient's current motor abilities. Figure 4 shows a diagram of the dynamic difficulty adaptation approach.

Design and development of VR-based exergames for functional hand rehabilitation after stroke

A. Bouatrous(1), N. Zenati(2), A. Meziane(3), C. Hamitouche(4), (1) Computer Science Faculty, USTHB University, (2)CDTA, (3)CERIST, (4) MT Atlantique Bretagne, France.

3. Results

Preliminary experimental tests

In order to evaluate and validate our system, we conducted a subjective evaluation of the acceptability and usability of the system by patients, using questionnaires to assess patient satisfaction with our system. In this context, we chose the standardized System Usability Scale (SUS) questionnaire [8] to assess the usability level of our system, and the Intrinsic Motivation Inventory (IMI) [9] to assess the subjective experience of the participants who used the system.

1) Experimental setup: Each participant was asked to sit on a chair facing a table with a matte infra-absorbent surface on it, and for visual feedback, a standard laptop screen was used (see Figure 5). The Leap Motion© sensor was placed above the table on a universal flexible support for smartphone, its rotation along the vertical axis will have been 0, and its position on the table is defined relative to the paretic limb, and it was placed at a height of 23 cm above the table. The sensor was calibrated (before the experiments) according to the method recommended by Leap Motion© until it reached at least the recommended score of 90. In order to ensure a better reading quality during the experiments, some precautions were taken, such as reducing interference from outside light, closing the windows so as not to be influenced by sunlight.

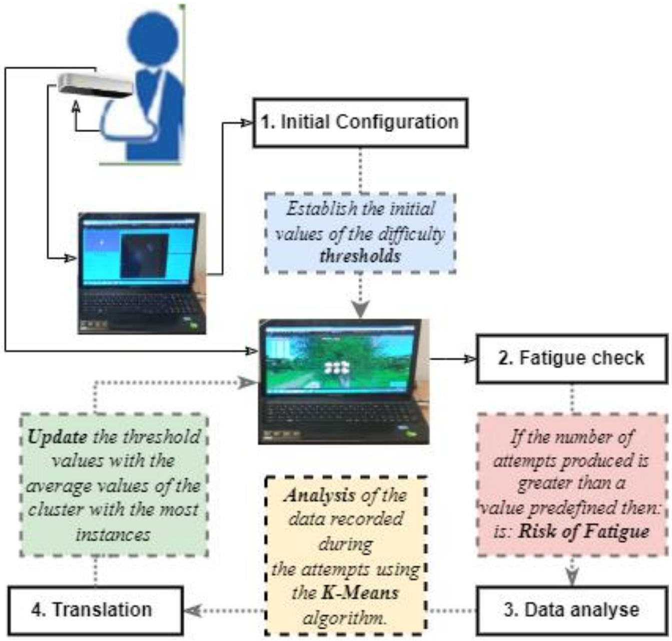

1. Initial Configuration
Establish the initial values of the difficulty thresholds
2. Fatigue check
If the number of attempts produced is greater than a value predefined then: is: Risk of Fatigue
3. Data analyse
Analysis of the data recorded during the attempts using the K-Means algorithm.
4. Translation
Update the threshold values with the average values of the cluster with the most instances

Figure 4. Dynamic difficulty adaptation diagram [5].

Design and development of VR-based exergames for functional hand rehabilitation after stroke

A. Bouatrous[1], N. Zenati[2], A. Meziane[3], C. Hamitouche[4], (1) Computer Science Faculty, USTHB University, (2)CDTA, (3)CERIST, (4) MT Atlantique Bretagne, France.

2) Participants: A total of 11 participants were recruited (9 male, 2 female). These patients were not the same as those involved in the design process. All were stroke patients and with unilateral affection. The age of the patients ranged from 43 to 81 years, with a mean age of 66.2 (±10.31) years.

Results of the standardized SUS questionnaire

The results we obtained are shown in (Table I). The total is calculated according to the method given by the SUS questionnaire and goes from 0 to 100. The higher the score, the more usable the device. Figure 6 shows the SUS score of each participant.

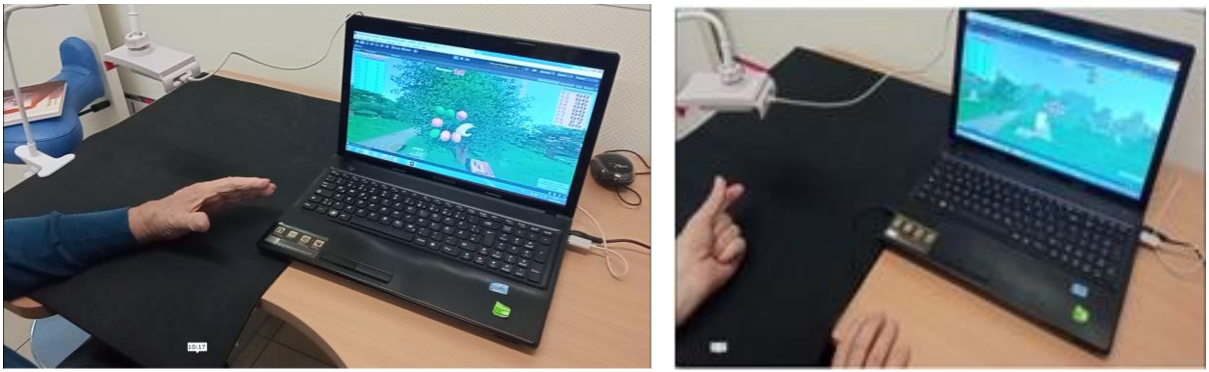
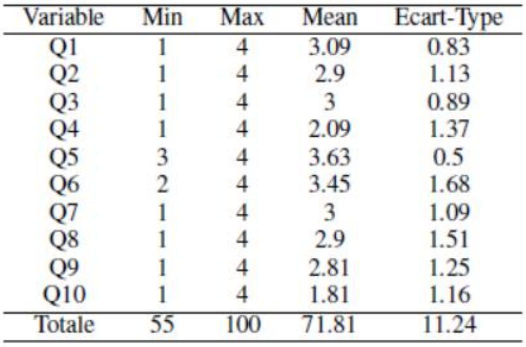


Figure 5. Patients in the process of using the system. (Left) Clinical test, (Right) Game test



Variable	Min	Max	Mean	Ecart-Type
Q1	1	4	3.09	0.83
Q2	1	4	2.9	1.13
Q3	1	4	3	0.89
Q4	1	4	2.09	1.37
Q5	3	4	3.63	0.5
Q6	2	4	3.45	1.68
Q7	1	4	3	1.09
Q8	1	4	2.9	1.51
Q9	1	4	2.81	1.25
Q10	1	4	1.81	1.16
Totale	55	100	71.81	11.24

TABLE I: Results SUS of Patients (n=11) [5].

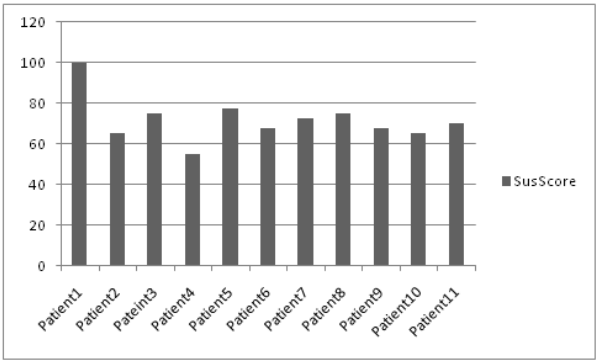


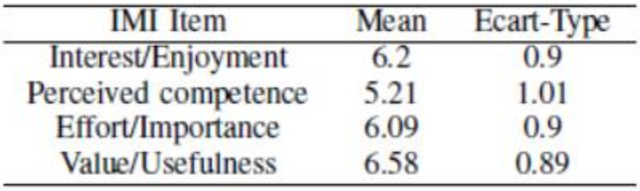
Figure 6. SUS Score [5].

Design and development of VR-based exergames for functional hand rehabilitation after stroke

A. Bouatrous(1), N. Zenati(2), A. Meziane(3), C. Hamitouche(4), (1) Computer Science Faculty, USTHB University, (2)CDTA, (3)CERIST, (4) IMT Atlantique Bretagne, France.

Results of the Intrinsic Motivation Inventory (IMI)

The Intrinsic Motivation Inventory (IMI) assesses participants' subjective experience with a target activity [9]. This 7-item Likert scale instrument was administered to assess participants' interest/pleasure; perceived competence; effort/ importance; and value/utility. Thus, we have calculated scores for each item: (Table II) shows the average IMI scores per participant shown in Figure 7.



IMI Item	Mean	Ecart-Type
Interest/Enjoyment	6.2	0.9
Perceived competence	5.21	1.01
Effort/Importance	6.09	0.9
Value/Usefulness	6.58	0.89

TABLE II: Overview of IMI scores (n=11) [5].

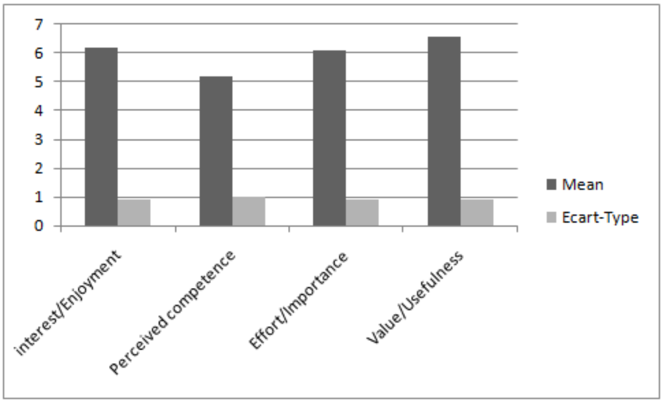


Figure 7. Overview of IMI scores [5].

Discussion

For 11 patients, the score obtained on the SUS questionnaire is 71.81, so according to [8], our system is usable and acceptable in the grade C scale, with a good adjective score. Thus, the average IMI subscores range from 6.09 to 6.58 (out of 7), with the exception of the "Perceived Competence", perhaps due to limited exposure to the system. Participants were enthusiastic and provided promising feedback, highlighting the benefits of the system for hand rehabilitation. Users emphasized the system's positive impact on motor and brain function. Motivation and commitment were observed, with some patients expressing a desire for frequent use.

Design and development of VR-based exergames for functional hand rehabilitation after stroke

A. Bouatrous[1], N. Zenati[2], A. Meziane[3], C. Hamitouche[4], (1) Computer Science Faculty, USTHB University, (2)CDTA, (3)CERIST, (4) MT Atlantique Bretagne, France.

The study showed that the system did not require therapeutic support, suggesting partial autonomy for patients undergoing hand rehabilitation after stroke. Despite the positive results, limitations were identified, particularly with regard to the perception of 3D movements in the virtual environment. Patients, often elderly and cognitively impaired, found it difficult to perceive 3D movements, resulting in both motor and brain fatigue.

Conclusion

This article presents a system designed for functional hand rehabilitation, using two VR-based exergames. The games involve harvesting fruit in a virtual grove and incorporate measurements to quantitatively assess the patient's performance. The system uses an innovative approach, relying on artificial intelligence to adapt the game to individual motor skills during rehabilitation sessions. Developed in collaboration with rehabilitation specialists, the system was tested on 11 stroke patients.

Results from subjective evaluations and standardized questionnaires indicate positive feedback on usability and acceptability.

Future improvements aim to include additional hand movements for a diverse range, introduce varied environments for controlled practice, and resolve depth perception issues through verbal cues and VR headset integration for a more realistic experience.

References

[1] Streicher, A., Smeddinck, J.D.: Personalized and Adaptive Serious Games. In: D¨orner, R., Gobel, S., Kickmeier-Rust, M., Masuch, M., and Zweig, K. (eds.) Entertainment Computing and Serious Games. pp. 332–377. Springer International Publishing, Cham (2016)

[2] Aguilar-Lazcano, C.A., Rechy-Ramirez, E.J.: Performance analysis of Leap motion controller for finger rehabilitation using serious games in two lighting environments. Measurement. 157, 107677 (2020). https://doi.org/10.1016/j.measurement.2020.107677

[3] Rechy-Ramirez, E.J., Marin-Hernandez, A., Rios-Figueroa, H.V.: A human–computer interface for wrist rehabilitation: a pilot study using commercial sensors to detect wrist movements. Vis Comput. 35, 41–55 (2019). https://doi.org/10.1007/s00371-017-1446-x

[4] Burke, J.W., McNeill, M.D.J., Charles, D.K., Morrow, P.J., Crosbie, J.H., McDonough, S.M.: Optimising engagement for stroke rehabilitation using serious games. Vis Comput. 25, 1085–1099 (2009). https://doi.org/10.1007/s00371-009-0387-4

[5] A. Bouatrous, A. Meziane, N. Zenati, and C. Hamitouche, "A new adaptive VR-based exergame for hand rehabilitation after stroke," Multimedia Systems, vol. 29, no. 6, pp. 3385–3402, Dec. 2023.

[6] A. Bouatrous, N. Zenati, A. Meziane, and C. Hamitouche, A Virtual Reality-Based Serious Game Designed for Personalized Hand Motor Rehabilitation. 2023, p. 5.

[7] Hocine, N., Goua¨ıch, A., Cerri, S.A., Mottet, D., Froger, J., Laffont, I.: Adaptation in serious games for upper-limb rehabilitation: an approach to improve training outcomes. User Model User-Adap Inter. 25, 65–98 (2015). https://doi.org/10.1007/s11257-015-9154-6

[8] Bangor, A.: Determining What Individual SUS Scores Mean: Adding an Adjective Rating Scale. 4, 10 (2009)

[9] Franck, J.A., Timmermans, A.A.A., Seelen, H.A.M.: Effects of a dynamic hand orthosis for functional use of the impaired upper limb in sub-acute stroke patients: A multiple single case experimental design study. TAD. 25, 177–187 (2013). https://doi.org/10.3233/TAD-130374

Visual Computing Magazine

Embedded Cheating Detection with NVIDIA Jetson Nano

Mohammed Kadri, Zohdi Kilani, Nassim Kaddouri, Ilyes Djebara, Alaa Guermat Computer Science Faculty, USTHB University.

1. Introduction

In the evolving landscape of education, the integration of technology presents both opportunities and challenges, with academic integrity facing new threats. As institutions embrace innovative assessment methods, the rise of academic dishonesty necessitates robust solutions. Our collaborative effort focuses on designing a Cheating Detection System (see figure 1), leveraging cameras and Convolutional Neural Networks (CNNs) to identify and deter cheating during examinations.

This paper explores the project's conception, design, and implementation stages. In the context of modern educational challenges, we highlight the significance of our Cheating Detection System. Combining NVIDIA Jetson Nano with deep learning algorithms, our approach provides a comprehensive solution to mitigate dishonest behavior during exams.

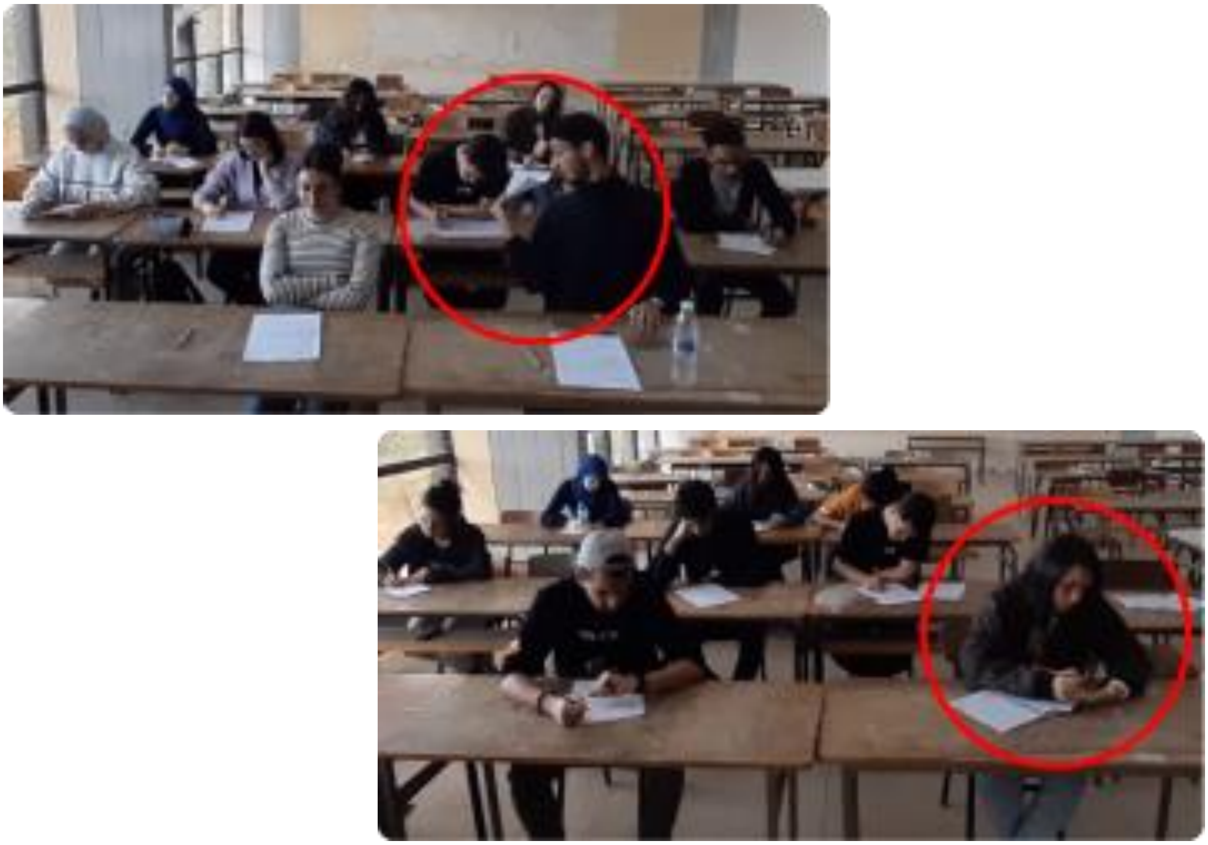


Figure 1. Cheating by passing paper, sheet on hand

This paper not only addresses immediate concerns of cheating detection but contributes to the broader discourse on maintaining academic integrity in the technological age. As technology continues to reshape education, our commitment to upholding assessment sanctity remains steadfast, exemplified by our Cheating Detection System—a stride towards fortifying academic integrity.

2. Method

2.1 Dataset preparation

In the realm of single-board computers crucial to the development of our project, both the Jetson Nano and Raspberry Pi Camera play integral roles. Developed by NVIDIA, the Jetson Nano is a high-performance single-board computer tailored for Artificial Intelligence (AI) and deep learning applications. Its robust computational power makes it well-suited for tasks demanding advanced processing capabilities.
The Raspberry Pi Camera, designed explicitly for use with Raspberry Pi single-board computers, boasts essential features for our project, including an 8-megapixel camera capable of capturing photographs at 3280 x 2464 pixels and video capture

To capture a diverse range of scenarios involving both cheating and non-cheating actions, a multitude of images were meticulously captured during various simulated scenarios.
The next phase involved processing these images using YOLO (You Only Look Once) (see figure 2). Each frame, depicting instances of cheating or non-cheating actions, underwent YOLO to identify and establish bounding boxes around the students. Subsequently, these bounding boxes were used to crop the images, and the cropped results were locally saved.

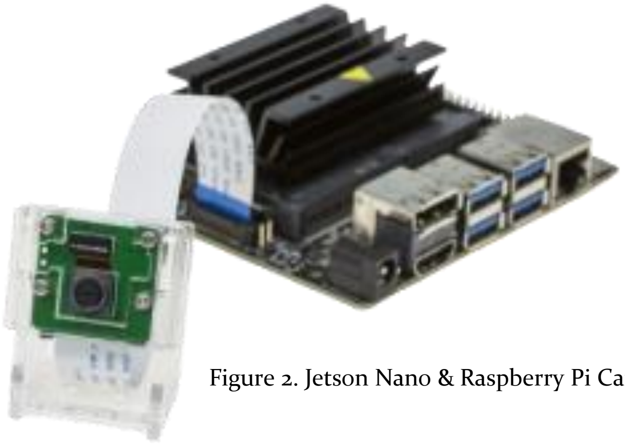


Figure 2. Jetson Nano & Raspberry Pi Camera

Embedded Cheating Detection with NVIDIA Jetson Nano

Mohammed Kadri, Zohdi Kilani, Nassim Kaddouri, Ilyes Djebara, Alaa Guermat Computer Science Faculty, USTHB University.

Following the image cropping process, the dataset underwent meticulous labeling. The cropped students' images, identified using YOLO, were manually separated into distinct folders dedicated to cheating and non-cheating instances (see figure 3).

Augmentation

While the initial dataset provided a solid foundation, we recognized the need to augment the data for increased variability. To achieve this, a data augmentation process was implemented to introduce another batch of variation into our dataset.

The data augmentation process was a strategic play on diverse transformations. Techniques such as applying a median filter, blurring, and rotating the images were systematically employed. These augmentation methods not only expanded the dataset but also introduced variations in lighting, orientation, and other factors, ensuring a more robust and comprehensive training set for our Cheating Detection System. This section explores the significance of data augmentation in fortifying the dataset and, consequently, the overall efficacy of our innovative cheating detection solution.

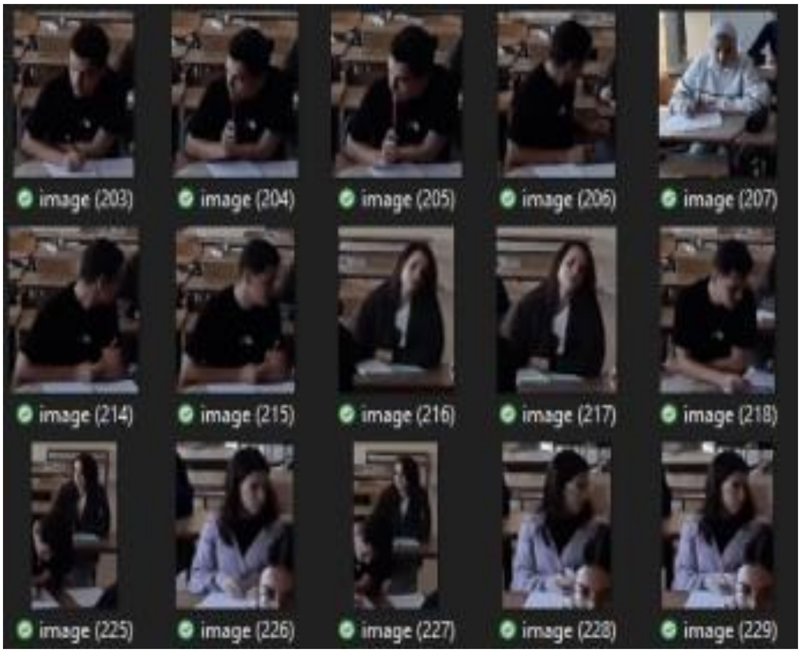


Figure 3. Cheating images - results of cropping

Embedded Cheating Detection with NVIDIA Jetson Nano

Mohammed Kadri, Zohdi Kilani, Nassim Kaddouri, Ilyes Djebara, Alaa Guermat Computer Science Faculty, USTHB University.

2.2 Model Architecture

To find an optimal model architecture we began with the utilization of the VGG Net architecture. However, confronted with the practical constraints of embedded systems, particularly a 1 GB model deemed impractical for our project, we embarked on a strategic shift. The chosen solution was a more lightweight approach—a Convolutional Neural Network (CNN) with two layers comprising 32 and 64 convolutions, aligning with the project's embedded system requirements.

In response to the limited dataset, we introduced a Max Pooling layer and a Dropout layer to enhance the model's generalization capabilities. The final refinement involved incorporating a dense layer with 256 neurons, culminating in a binary decision neuron. This adjustment in the model architecture represents a delicate balance between computational efficiency and performance robustness, ensuring optimal functionality within the constraints of our Cheating Detection System. This section delves into the rationale behind these adaptations, elucidating how the refined model strikes a harmonious balance to meet the specific demands of our innovative cheating detection solution.

Augmenting Data Resources and Streamlining Annotation

Additional videos and images were systematically captured to bolster our resources. This expansion aimed to ensure a diverse and comprehensive set of data for training and refining our Cheating Detection System.

To streamline the annotation process for object detection, we employed Roboflow—an efficient tool in our workflow. Focusing on two distinct classes, namely "cheating" and "not_cheating," we undertook the manual annotation task meticulously. This section provides insights into the rationale behind enriching our dataset and the strategic use of Roboflow for effective annotation, contributing to the robustness and accuracy of our innovative cheating detection solution.

Seamless Data Integration and Enhanced YOLO Architecture

To seamlessly integrate our annotated data into the Darknet framework, we leveraged Roboflow, a versatile tool that supports the download of data in various formats. Crucially, we adopted the YOLO Darknet format. With this approach, every image corresponds to a text file sharing the same name, housing crucial information such as class labels and the corresponding object box coordinates in the format: <class> <x> <y> <w> <h>. This labeling process sets the stage for optimal utilization of the Darknet framework in our Cheating Detection System.

Embedded Cheating Detection with NVIDIA Jetson Nano

Mohammed Kadri, Zohdi Kilani, Nassim Kaddouri, Ilyes Djebara, Alaa Guermat Computer Science Faculty, USTHB University.

Optimized YOLO Architecture

Our commitment to refining the YOLO architecture led to a strategic optimization with a tailored approach. The enhanced YOLO architecture comprises 38 layers, featuring two detection layers with distinct scales. A key innovation is the incorporation of a CSP connection, further enhancing the model's capacity for detecting and classifying instances. This illuminates the rationale behind these optimizations, shedding light on how the streamlined YOLO architecture enhances the precision and efficiency of our innovative cheating detection solution

Empowering YOLO Detection

Harnessing the power of our dataset and the Darknet framework, we embarked on the training phase for our Yolov4-tiny model (see figure 4). The model underwent 6000 iterations, aligning with the recommended minimum for effective training. A discerning analysis of the training process reveals a consistent decline in the average loss, a pivotal indicator of the model's adept learning. This phenomenon is vividly illustrated on the training chart, showcasing the model's progressive improvement.

Remarkably, the mean Average Precision (mAP) achieved an impressive 98% accuracy, further affirming the robustness and efficacy of our Yolov4-tiny model (see figure 5). This section unfolds the comprehensive journey of training and validation, emphasizing key performance metrics and the remarkable learning curve observed during the training process for our innovative cheating detection solution (see figure 6).

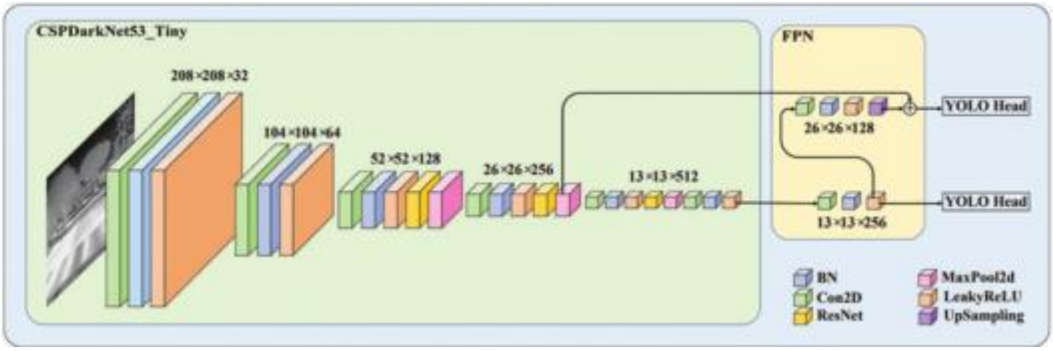


Figure 4. YOLO v4 tiny architecture

Embedded Cheating Detection with NVIDIA Jetson Nano

Mohammed Kadri, Zohdi Kilani, Nassim Kaddouri, Ilyes Djebara, Alaa Guermat Computer Science Faculty, USTHB University.

3. Results

We tested our model using full frames instead of cropped ones on our custom data. The results turned out pretty good, encouraging us to broaden our testing scope by including online images.

The model performed well in diverse testing conditions, showing adaptability and accuracy. Expanding our evaluation to include online images allowed us to further validate its effectiveness in real-world scenarios. This section provides insights into our testing process, highlighting the model's performance across different datasets and affirming its practical applicability in various situations.

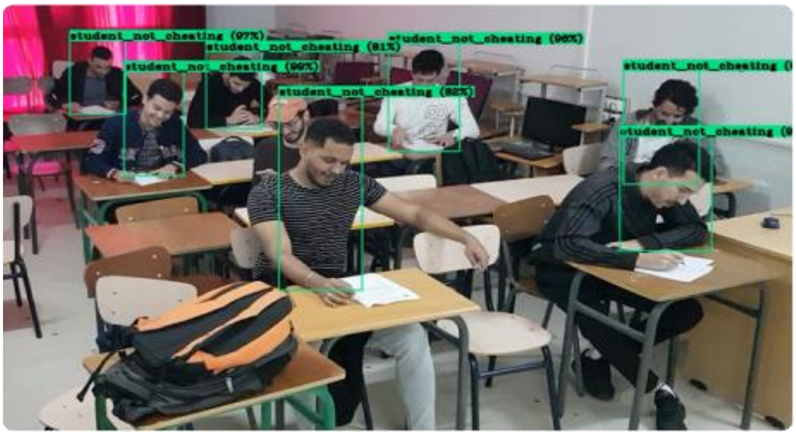

student_not_cheating (97%)
student_not_cheating (81%)
student_not_cheating (96%)
student_not_cheating (99%)
student_not_cheating (82%)
student_not_cheating (
student_not_cheating (9

Figure 5. Results - Few students not cheating

Conclusion

In conclusion, our cheating detection project employed two methods—CNN and YOLOv4 Tiny using Darknet—revealing that the latter surpassed the former in terms of effectiveness. The utilization of YOLOv4 Tiny showcased superior results, highlighting its efficacy as a robust approach for cheating detection compared to the CNN method. This outcome underscores the significance of leveraging advanced object detection techniques, particularly within the framework of YOLOv4 Tiny, to enhance the accuracy and reliability of cheating detection systems.

To enhance the camera's performance, we opted for C language over Python, prioritizing speed and smooth operation. The decision to use C language contributed to a faster and smoother camera operation, ensuring real-time efficiency in our cheating detection system.

Embedded Cheating Detection with NVIDIA Jetson Nano

Mohammed Kadri, Zohdi Kilani, Nassim Kaddouri, Ilyes Djebara, Alaa Guermat Computer Science Faculty, USTHB University.

It's worth noting that the first method, utilizing CNN, was primarily applied to images of one person, whereas the second method, employing YOLOv4 Tiny, delivered superior results when applied to full frames instantaneously. This distinction highlights the versatility and instantaneity offered by the YOLOv4 Tiny approach, making it a preferred choice for comprehensive cheating detection.

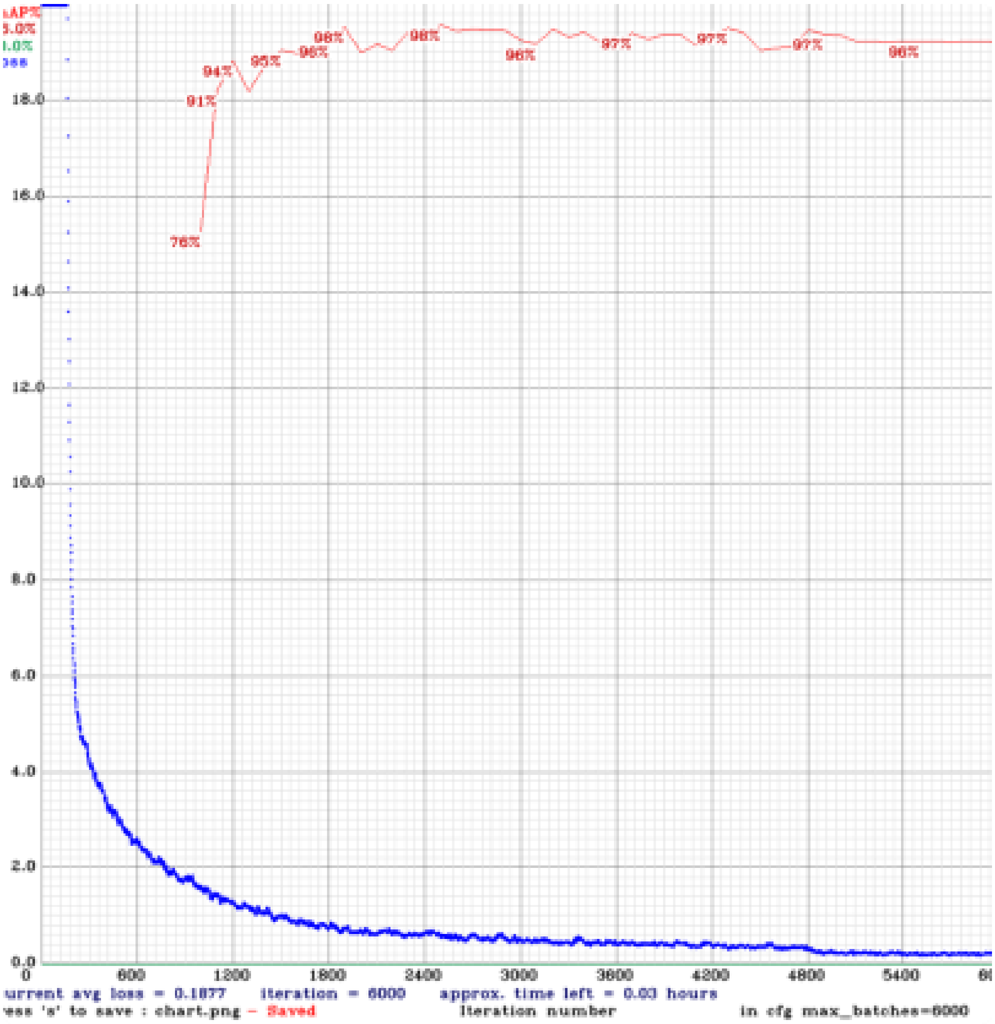


Figure 6. Loss and accuracy of the model

Embedded Cheating Detection with NVIDIA Jetson Nano

Mohammed Kadri, Zohdi Kilani, Nassim Kaddouri, Ilyes Djebara, Alaa Guermat Computer Science Faculty, USTHB University.

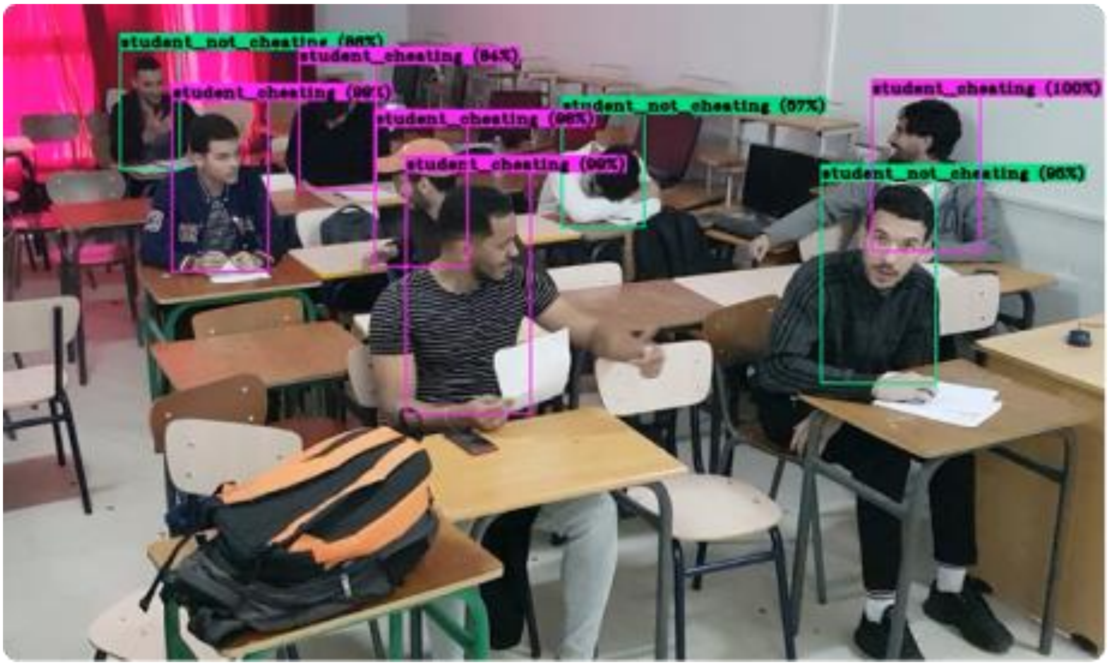


Figure 7. Results - Few students cheating

References

[1] Jacob Solawetz, Samrat Sahoo.
Train YOLOv4-tiny on Custom Data - Lightning Fast Object Detection.
https://blog.roboflow.com/train-yolov4-tiny-on-custom-data-lighting-fast-detection/
[2] NVIDIA Jetson Nano
https://www.nvidia.com/fr-fr/autonomous-machines/embedded-systems/jetson-nano/
[3] https://www.raspberrypi.com/documentation/accessories/camera.html

Autoencoders for Anomaly detection in exam surveillance

BENATALLAH Rayan Ibrahim, SAYOUD Lynda , MALLEK Lina, BENAZZOU Fatima. Visual Computing Master's students,Computer Science Faculty, USTHB University

1. Introduction

Nowadays, the issue of cheating in exams is becoming a growing concern in the field of education, casting doubt on the fairness of exams and challenging the reliability of academic results. Faced with this disconcerting reality, it is imperative to explore innovative solutions that enhance academic standards and contribute to a more precise evaluation of students.

It is from this perspective that our approach takes shape, built upon the integration of a specific deep learning technique, we shall distinctly define cheating from normal cases within an exam room through the use of anomaly detection with autoencoders.

To delve deeper into the specific mechanisms of autoencoders, the upcoming sections will explore their application in the realm of cheating detection in exams.

2. Method

The conventional approach to mitigating cheating in exams (see figure 1) often involves enumerating known cheating instances, a daunting task given the myriad ways in which students may engage in dishonest behavior. In contrast, our methodology adopts a novel perspective by employing autoencoders for anomaly detection without the need to explicitly define all cheating instances.

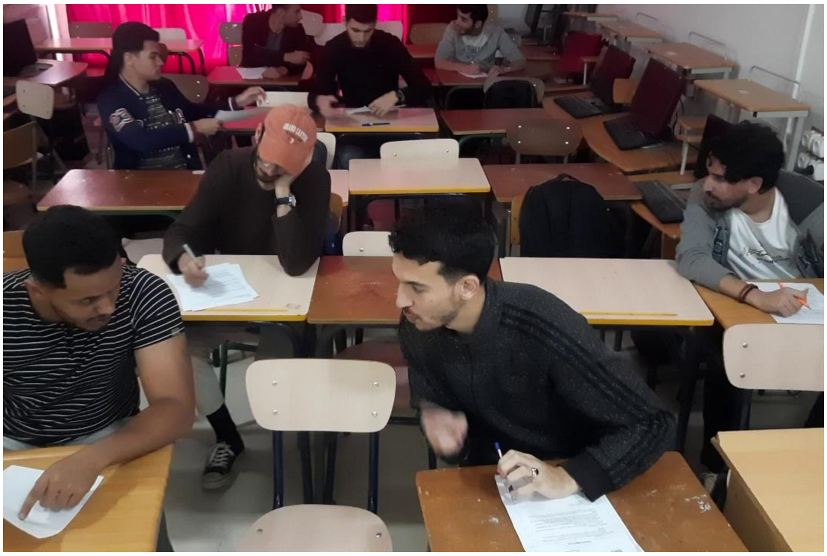


Figure 1. Sample image from our dataset

Autoencoders for Anomaly detection in exam surveillance

,BENATALLAH Rayan Ibrahim, SAYOUD Lynda , MALLEK Lina, BENAZZOU Fatima. Visual Computing Master's students,Computer Science Faculty, USTHB University

Data Preparation

We commenced our study by assembling a dataset comprising exclusively normal, non-anomalous exam instances by capturing a 2-minute video of students in a class where cheating didn't take place, extracting its frames to end up with over 3500 pictures. This deliberate choice was grounded in the recognition that obtaining a comprehensive set of cheating instances is impractical. The autoencoder was trained exclusively on this pristine dataset to learn the inherent patterns of legitimate exam behavior.

Model architecture and training

Our autoencoder architecture is tailored for anomaly detection in the context of exam surveillance (see figure 2). The model, implemented using the Keras framework, follows a symmetric encoder-decoder structure. The encoder progressively reduces the spatial dimensions of the input, capturing key features, while the decoder up-scales the encoded representation to faithfully reconstruct the original input. The choice of the mean squared error (MSE) as the loss function and the Adam optimizer underpins the training process, aligning with our objective of minimizing reconstruction errors for anomaly detection.

Dual-Threshold Anomaly Detection

However, after delving into further research, we discovered that depending only on the reconstruction error may not be adequate for effective anomaly detection. This insight was particularly emphasized in the "Robust Anomaly Detection in Images using Adversarial Autoencoders" [6], where our exploration deepened, leading us to uncover another factor – Kernel Density Estimation (KDE) from DigitalSreeni on YouTube [7]. This addition aims to further enhance anomaly detection efficiency, complementing the reconstruction error.

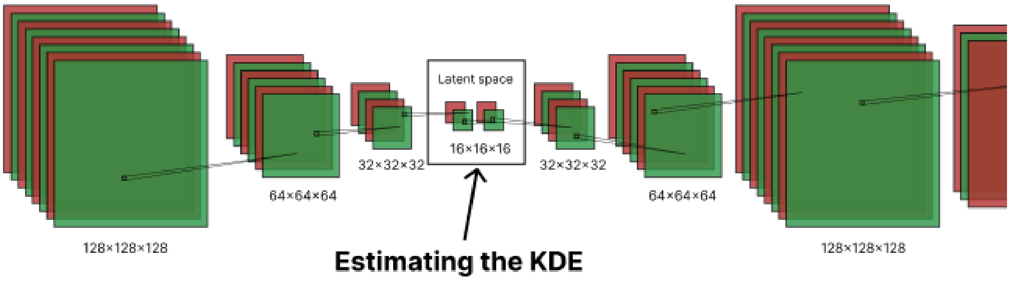


Figure 2. Model architecture

Autoencoders for Anomaly detection in exam surveillance

BENATALLAH Rayan Ibrahim, SAYOUD Lynda , MALLEK Lina, BENAZZOU Fatima. Visual Computing Master's students,Computer Science Faculty, USTHB University

Observation

After splitting our trained autoencoder, using our trained encoder helped us estimate the KDE of the latent representation of both normal and anomalous data.

Simultaneously, we perform a reconstruction error analysis on both normal and anomalous data using the trained autoencoder. This step ensures a comprehensive assessment, considering both the spatial characteristics in the latent space and the fidelity of reconstructed instances.

Thresholds setting

Guided by the observations from the KDE analysis and reconstruction error metrics, we establish dual thresholds. The first, a soft threshold, identifies instances with a moderate deviation from the expected pattern, prompting a closer examination. The second, a hard threshold, serves as a more stringent criterion, unequivocally flagging instances with a significant deviation as anomalies.

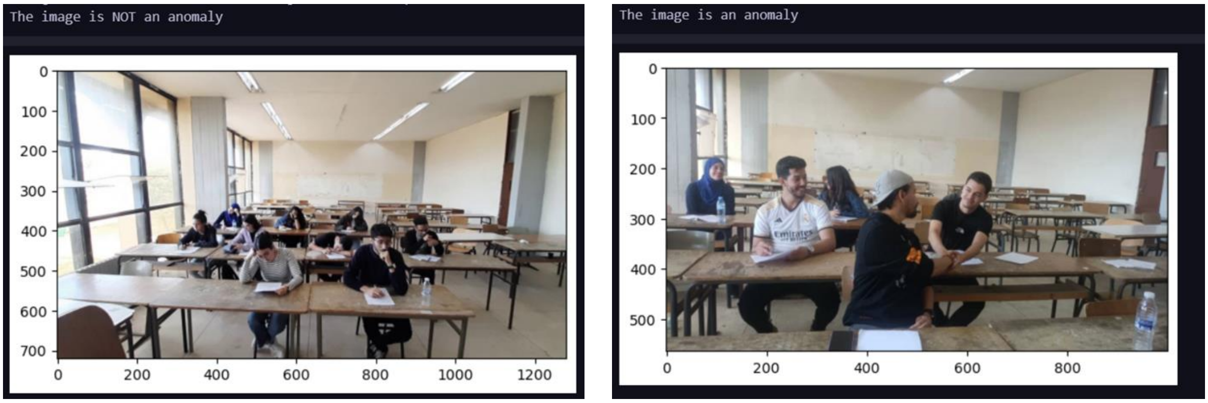
3. Results

Testing the model on images

In this section, we delve into the practical implementation of our proposed approach and showcase the results obtained through the figure 3.

Testing the model on video

Our program processes input videos by extracting individual frames and analyzing each one to check if it's an anomaly or not. Frames depicting anomalies are specifically identified and saved in a designated 'anomalies' folder (see Figure 4).



```
The image is NOT an anomaly
```

```
The image is an anomaly
```

Figure 3. Non anomaly (left) and anomaly detection (right)

Autoencoders for Anomaly detection in exam surveillance

BENATALLAH Rayan Ibrahim, SAYOUD Lynda , MALLEK Lina, BENAZZOU Fatima. Visual Computing Master's students,Computer Science Faculty, USTHB University

Conclusion

In conclusion, while our approach with autoencoders for exam surveillance offers unparalleled advantages in precision and adaptability, ongoing vigilance and refinement is essential to address the ever-changing landscape of educational environments and student behaviors. This system represents a crucial step toward maintaining the integrity of examinations in an era of technological advancements and evolving academic challenges.

As part of our future work, we aspire to enhance the capabilities of our current model. Our upcoming focus involves the introduction of segmented anomaly detection to refine the results further. This segmentation will allow us to pinpoint and visualize specific individuals engaged in anomalous activities during exams, providing a more granular and insightful analysis. This expansion aims to improve the precision and interpretability of our cheat detection system, contributing to its effectiveness.

References

[1] Subash Palvel, Exploring Autoencoders in Deep Learning, Sep 12, 2023,[en ligne], Available on: https://subashpalvel.medium.com/exploring-autoencoders-in-deep-learning-2dd41a689104

[2]Pouya Hallaj, Anomaly Detection with Autoencoders, Sep 26, 2023, [online], available on: https://medium.com/@pouyahallaj/anomaly-detection-with-autoencoders-956893b60aef

[3] Subham Sarkar, Anomaly Detection in Images — AUTOENCODERS, Analytics Vidhya, Jun 13, 2021, [online], available on: https://medium.com/analytics-vidhya/anomaly-detection-in-images-autoencoders-b780abf88f51

[4] ANURAG SINGH CHOUDHARY, Unveiling Denoising Autoencoders, Jul 06, 2023, [online], available on: https://www.analyticsvidhya.com/blog/2023/07/unveiling-denoising-autoencoders/

[5] Anay Dongre, Overview of Autoencoders, Jan 1, 2023, available on: https://dongreanay.medium.com/overview-of-autoencoders-52c777418937

[6] Laura Beggel , Michael Pfeiffer , Bernd Bischl ,Robust Anomaly Detection in Images using Adversarial Autoencoders, Jan 18, 2019, page 10, available on: https://arxiv.org/pdf/1901.06355.pdf

[7] DigitalSreeni, 260 - Identifying anomaly images using convolutional autoencoders, march 9,2022, available on: https://www.youtube.com/watch?v=q_tpFGHiRgg&t=268s

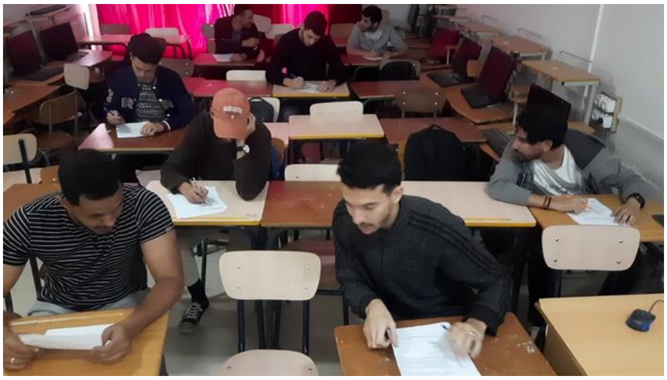


Figure 4. Cheating Detected Frame in 'Anomalies' Folder

RGB-D Segment Captioning

K. Delloul, Prof. S. Larabii., Computer Science Faculty, USTHB University

1. Introduction

In recent years, image captioning and image segmentation have emerged as crucial tasks in computer vision, with applications ranging from autonomous driving to content analysis. While several solutions have emerged to textually describe the content of a given image, few AI models are capable of generating fully detailed descriptions of each panoramic segment within an image.

In our conducted research, we propose an approach based on a deep learning model that generates a descriptive phrase for each segment present in the image. The uniqueness of our research lies in the fact that the generated captions are enriched with region positions relative to the user (left, right, front) and position relationships between the regions. This solution is applied to our TS-RGBD dataset, consisting of images collected in the auditorium using the Kinect.

Therefore, the goal is to create a deep learning model for translating images into text, where the output takes the form of textual descriptions of all segments in the input image. The textual descriptions of the segments are enhanced by their positioning relative to the user (left, right, in front) and relative to other segments. The solution is intended to assist visually impaired individuals.

2. Method

Before embarking on any scientific work, a state-of-the-art study is essential. We first collected a set of publications related to topics relevant to our subject, such as Image Captioning, Image Segmentation and Depth Estimation. On the other hand, a state-of-the-art study also includes studying datasets that could be used to complete our research.

Analyzing different published papers and proposed solutions from laboratories and firms from all around the globe allowed us to build a survey and to present the findings of our analysis in [1]. We classified different image captioning models according to their architectures, from single sentence to paragraph captioning. We presented different existing datasets and their lack of diversity as well as their shortcomings regarding depth information and detailed annotations. We also highlighted the limitations of existing solutions regarding our research objectives and thus the necessity to start building our own model.

RGB-D Segment Captioning

K. Delloul, Prof. S. Larabii., Computer Science Faculty, USTHB University

One of the main pillars to our research is a the DenseCap model for dense captioning proposed by a Ph.D. student of Prof. Fei Fei Li, a pionneer in Computer Vision. It takes an RGB image as input and returns a number of regions of interest with their descriptive sentences. We generated dense descriptions for the Visual Genome Dataset on which DenseCap was trained, then we applied another model for depth estimation of said images. Both depth and RoI boxes were used to generated egocentric descriptions of images, which constitute the topic of our second paper [2].

For each image from the VG dataset, we used the AdaBins model to generate a depth map as seen in the figure 1. Note that depth is the distance between the pixel and the camera. AdaBins was trained to estimate this depth from monocular images.

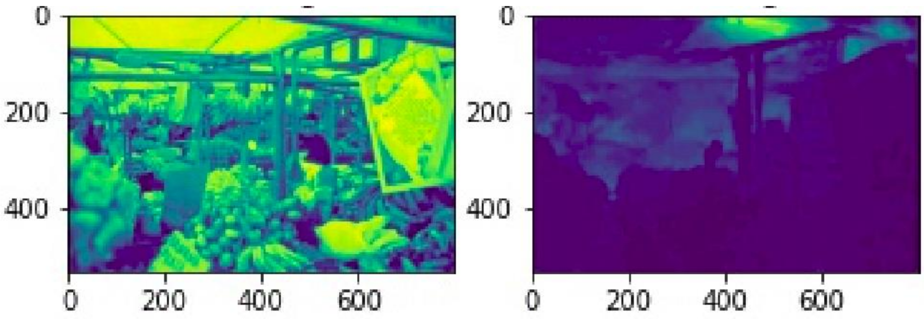


Figure 1. Image from VG Dataset and the computed depth map.

Then using DenseCap model we extracted bounding boxes of regions of interest with generated captions (see figure 2). The same process was applied on frames we extracted from theatre shows that were available on YouTube for free (see figure 3).

RGB-D Segment Captioning

K. Delloul, Prof. S. Larabi,, Computer Science Faculty, USTHB University

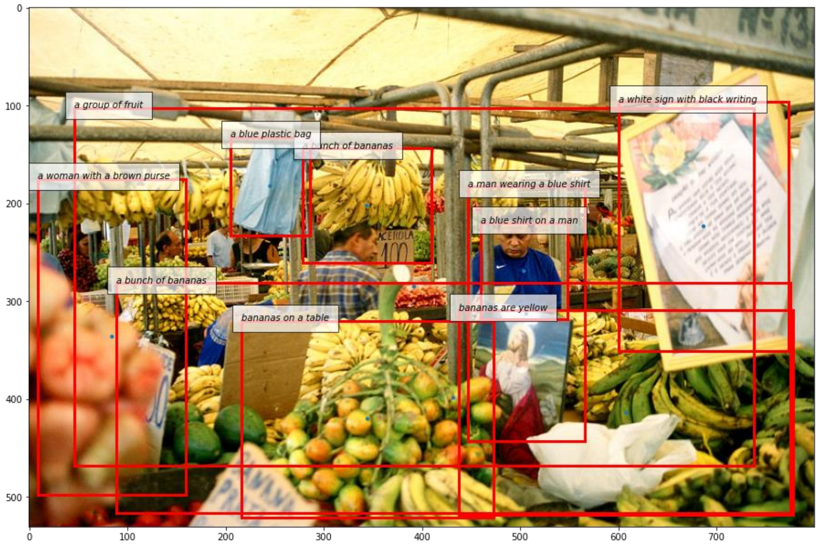


Figure 2. Extracted bounding boxes.

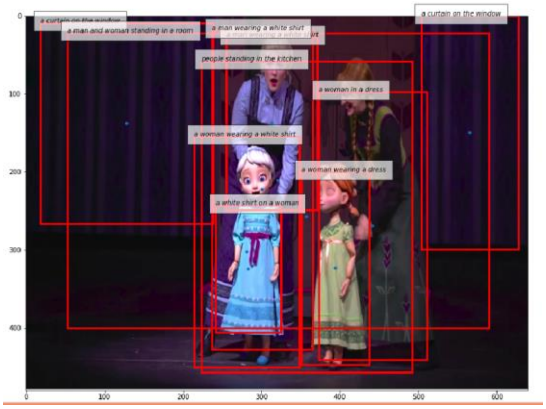
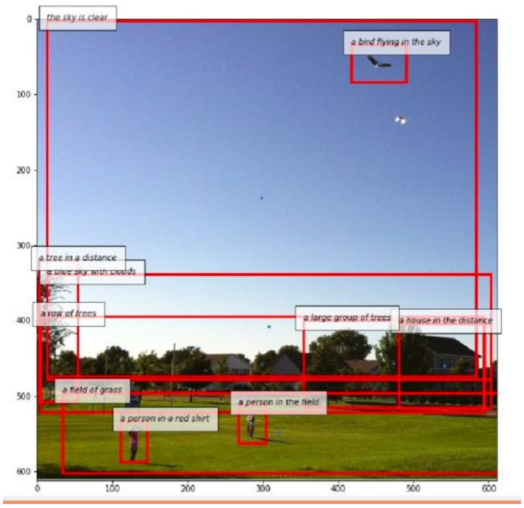

0
100
200
300
400
0
100
200
300
400
500
600
a curtain on the window
a man and woman standing in a room
a man wearing a white shirt
a man wearing a white shirt
a curtain on the window
people standing in the kitchen
a woman in a dress
a woman wearing a white shirt
a woman wearing a dress
a white shirt on a woman

Figure 3. Image from theatre shows .

RGB-D Segment Captioning

K. Delloul, Prof. S. Larabii., Computer Science Faculty, USTHB University

Then we passed results through our proposed algorithm [4] to get egocentric descriptions.



In front of you there is a row of trees, a field of grass, the sky is clear, a person in the field, a blue sky with clouds

On your right a bird flying in the sky, a house in the distance, a large group of trees

And on your left a person in a red shirt, a tree in a distance

Figure 4. Labelling with egocentric description.

Evaluation of the achieved results on 20 images from VG and theatre scenes is presented in the table below:

Dataset	Captions N°	Correct Directions	Incorrect Directions	Accuracy
VG	196	175	21	89%
Theatre	200	174	26	87%

RGB-D Segment Captioning

K. Delloul, Prof. S. Larabii., Computer Science Faculty, USTHB University

The proposed method for image captioning with a focus on blind guidance and scene understanding has identified several limitations. Firstly, the method's applicability is limited to specific places due to the lack of diversity in the VG dataset, particularly affecting visually impaired individuals in varied environments. The depth estimation achieved good metrics, but the generated map lacks real depth values, hindering the output of actual distances and angles between objects and users. The AdaBins model also has restrictions on image sizes and formats, suggesting the use of RGB-D image sensors for better performance. Additionally, reliance on the DenseCap model introduces redundancies and inaccuracies in captioning, especially that it was not trained on theatre images. To address these issues, we emphasize the necessity of collecting and annotating RGB-D images of theatre scenes, recognizing it as a crucial step for further research.

So in the next step we proceeded to collect our own dataset using RGB-D images and theatre-like scenarios. The image capturing took place in the university amphitheater using two Microsoft Kinects v1, with volunteer students.

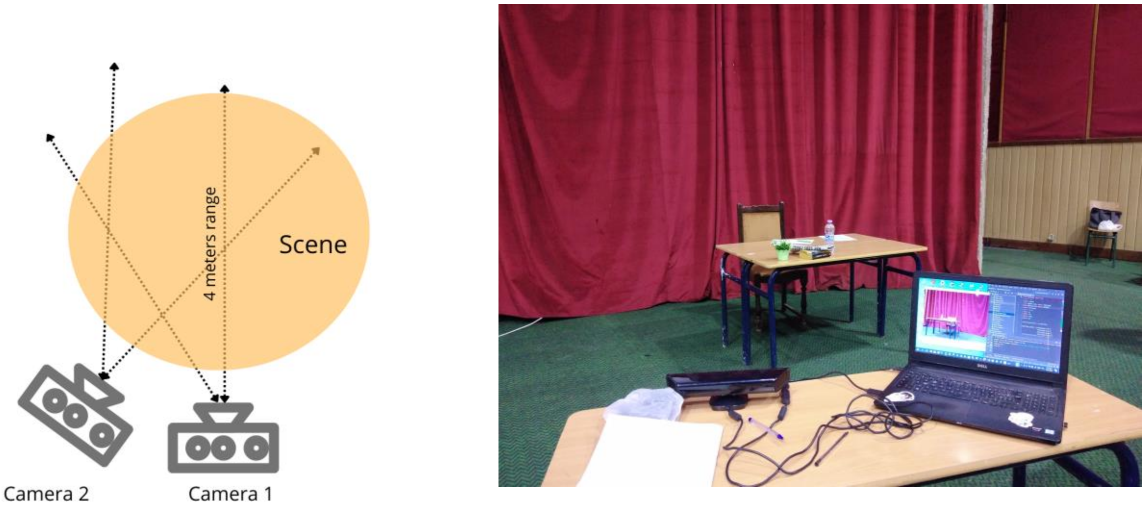


Figure 5. Collect of RGB-D images.

RGB-D Segment Captioning

K. Delloul, Prof. S. Larabii., Computer Science Faculty, USTHB University

Images from the dataset were annotated in such a way that for each panoptic segment of the image, there is a descriptive phrase. Annotations were made using the open source program LabelMe.

Finally, DenseCap model was modified to be capable of providing multiple sentences per image, each corresponding to a segment, instead of extracting the regions of interest itself, the model is fed with panoptic segments that were generated by a segmentation model OneFormer.

The architecture of our proposed solution can be illustrated as follows [4]:

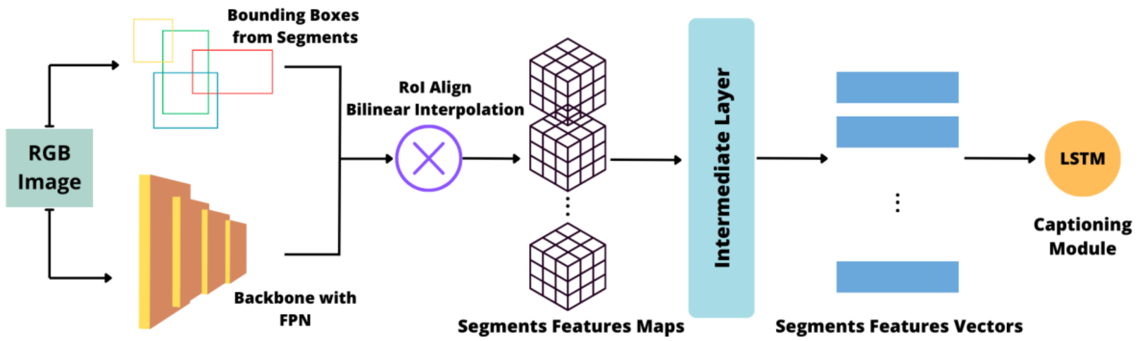


Figure 6. The used model.

The sentences are enriched by directions extracted from the point cloud, as well as the positioning relationships between different segments. The point cloud is generated using depth values that were collected using the Kinect.

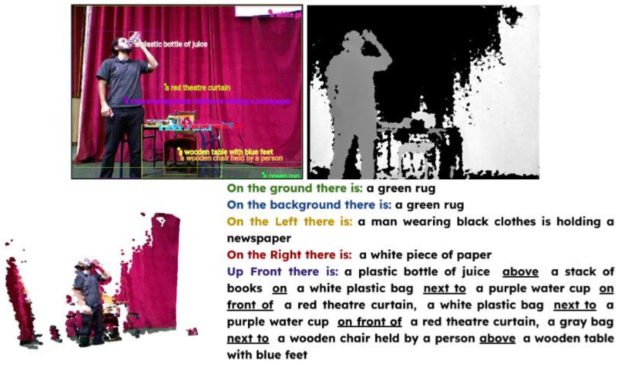
The solution is tested on our new TS-RGBD dataset of theater scenes. The solution is fast compared to dense textual description models and effective in terms of directions and relationships between segments, thanks to the depth information.

RGB-D Segment Captioning

K. Delloul, Prof. S. Larabii., Computer Science Faculty, USTHB University

Conclusion

We provided a comprehensive review of recent advancements in AI technologies for visual scene understanding, covering image captioning, image segmentation, and scene understanding. Despite notable progress, challenges such as handling occlusions, non salient regions, and generalizing to unseen scenarios persist. Our study addresses these challenges by developing a framework that enhances scene understanding through textual descriptions of image segments. The solution, applied to our novel TS-RGBD dataset of theatre scenes, outperforms other image captioning models in terms of captions per image and execution time. The approach successfully processes positional relationships using depth information, and future work includes refining ground truth captions, expanding the dataset, and incorporating more sophisticated sensors for wider applications, particularly in actual theatre plays, with plans for evaluation by blind and visually impaired users.



On the ground there is: a green rug
On the background there is: a green rug
On the Left there is: a man wearing black clothes is holding a newspaper
On the Right there is: a white piece of paper
Up Front there is: a plastic bottle of juice above a stack of books on a white plastic bag next to a purple water cup on front of a red theatre curtain, a white plastic bag next to a purple water cup on front of a red theatre curtain, a gray bag next to a wooden chair held by a person above a wooden table with blue feet

Figure 7. Result of labelling image from TS-RGBD dataset

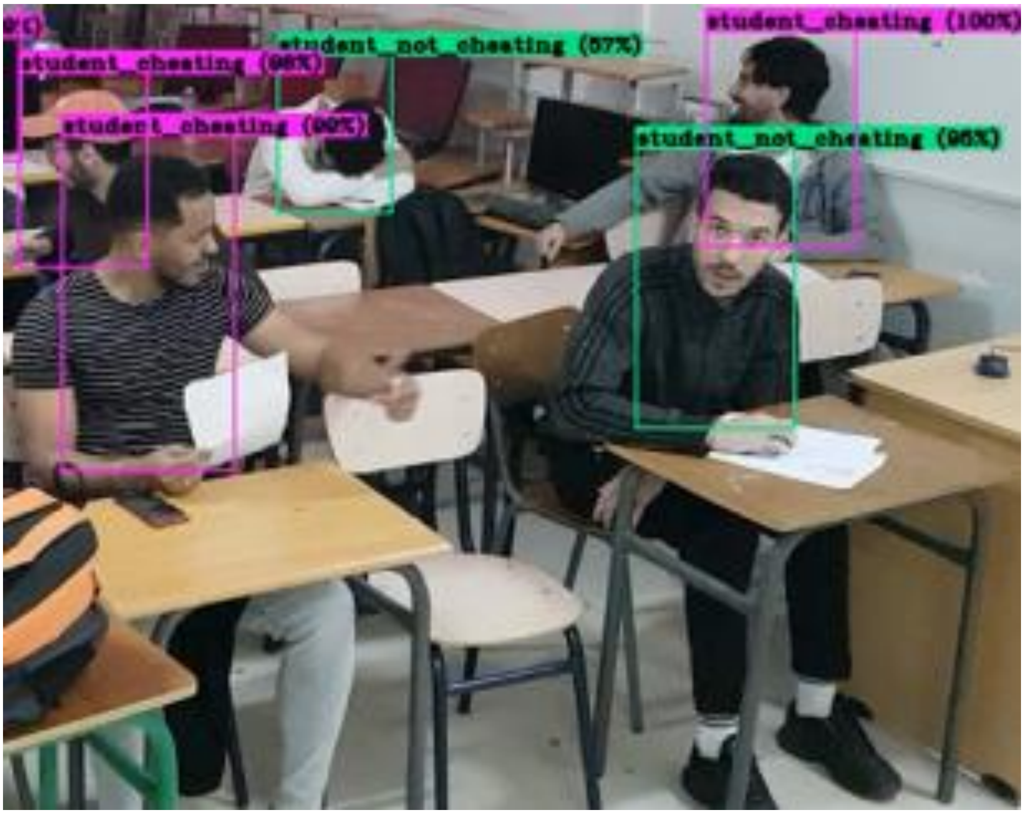
References

[1] Delloul Khadidja, Slimane Larabi.
Image Captioning State-of-the-Art: Is It Enough for the Guidance of Visually Impaired in an Environment? .
In: Senouci, M.R., Boulahia, S.Y., Benatia, M.A. (eds) Advances in Computing Systems and Applications. 17-18 May, CSA 2022. Lecture Notes in Networks and Systems, vol 513. Springer, Cham.
[2] Delloul Khadidja, Slimane Larabi.
Egocentric Scene Description for the Blind and Visually Impaired .
5th International Symposium on Informatics and its Applications (ISIA), M'Sila University, November 29-30, 2022
[3] Leyla Benhamida, Khadidja Delloul, Slimane Larabi.
TS-RGBD Dataset: a Novel Dataset for Theatre Scenes Description for People with Visual Impairments.
arXiv:2308.01035 [cs.CV], 2 Aug 2023
[4] Khadidja Delloul, Slimane Larabi.
Towards Real Time Egocentric Segment Captioning for The Blind and Visually Impaired in RGB-D Theatre Images.
arXiv:2308.13892 [cs.CV], 26 Aug 2023

Visual Computing Magazine

Vol 1, Issue 4, 2024

Content



Visual Computing Magazine