Virtual reality for training: Continental's case study

Paulo Rodrigues
UTAD
Vila Real, Portugal
al63984@utad.eu

Guilherme Gonçalves, Luís Barbosa,
Maximino Bessa
INESC TEC & UTAD
Porto & Vila Real, Portugal
guilhermeg@utad.pt

Miguel Melo
INESC TEC
Porto, Portugal
mcmelo@inesctec.pt

***Abstract*— Virtual Reality (VR) applications designed to train professionals are becoming more common, due to more companies searching for ways to reduce costs by improving the efficiency and efficacy of their training programs. Several training applications are found in the literature, but few consider a trainer's observation features, performance scoring systems, and digital twin integration. With that in mind, a solution is being developed under the R&D project entitled "Continental - Factory of the Future". This article describes the corresponding research line of the project and presents an immersive training application prototype designed to train professionals in antenna production lines. The prototype was evaluated through performance tests with different system specifications, revealing its viability.**

Keywords*— *virtual reality, digital twin, immersive training.

I. INTRODUCTION

Currently, technology has reached a point where Virtual Reality (VR) is no longer out of reach of the general public, with head-mounted displays (HMD) price drops, better quality [1] and an increasing number of applications (apps) being developed for these devices. There are apps like games [2][3], virtual meeting apps, and business-oriented apps. This work will focus primarily on virtual training.

VR training apps are becoming more important and useful to various businesses. Factors like the reduction of VR cost of devices and development are catalysts to reduce training time and increase in trainee's safety as training has a controlled scenario. For instance, adopting VR training in production lines can reduce downtime and avoid wasting components only for training purposes [4][5].

VR can benefit from other technologies such as digital twin (DT), where a virtual representation of a real environment exists coordinated with its real-world counterpart, allowing an exchange of data and inputs between the virtual and real elements. Although DT is commonly used for data visualization, it is underexplored in VR. In some apps it could prove helpful, for the privileged perception available in an immersive VR environment, and a regular display cannot provide a good notion of depth, or more than one display may be needed for said representation, causing a confusing and costly setup, which can be simplified with VR. Of course, this does not come without some trade-offs, like inflated cost, the need for digital machine output and input capabilities, delay between real and digital components, among many others [6][7].

This paper presents the R&D project "Continental – Factory of Future" that exploits the technologies mentioned above and describes a real case study in a production line to maximize production, minimize waste and create more resilient training. The next section describes pertinent related work, section 3 has a brief presentation and evaluation of the project. The main conclusions and future research are presented to promote the use of VR technologies for sustainability and market competitiveness.

II. STATE OF ART

VR training apps already exist in the market, and some of these apps will be discussed in the next few paragraphs. Siemens-Gamessa [8], a German-Spanish wind power company that provides services both on land and at sea, decided to improve the training efficiency of its collaborators in specialized, high-cost facilities, by using a VR training solution. The development company Kanda designed an application that allows standard training procedures, and environment habituation. This way, it allows several users to be connected simultaneously, without timing (to avoid pressure), allowing errors to be committed whilst providing a visual aid to correct them.

Virtual Assembly Line Training[a], designed by Opel, is a customized car assembly line simulator created to reduce up to 50% of the critical errors in training in an assembly line, reducing assembly costs and reducing worker training time by about 40%. This software allows the customization of the assembly line for each scenario (custom cars, workstations, and other details) by analysing the DTs of these lines. Each operator has login data, so each one has custom content available, tailored for their post. It is also compatible with mouse and keyboard or an Xbox Kinect.

SimSpay[b] from VRSim is a 3D VR training App designed for training in paint/coverage of objects. The main purpose of the App is to learn how to properly handle a paint gun, and the correct work posture. It also allows customization of the experience, like the object to cover, paint details like thickness or colour, among others. This App also allows performance evaluation and has visual cues and pre-built tutorials (for common paint operations). It uses a modified Oculus Rift headset with a custom paint gun and allows one user at a time.

Grundfos [2], a Danish company, also decided to improve its water pump production via immersive training, to reduce operator training costs, production line halting costs (for training in the real environment), to provide better security to these operators, and teach them correct equipment handling. This app, by the Danish company Unity Studios, has several training scenarios: installation, measuring, assembly, verification, packing, transport preparation and transportation of Grundfos' water pumps.

He et al. [2] proposed an application related to the construction equipment industry, with a favourable evaluation by the users who participated in study about this app. Due to the necessity to increase production and product quality, this VR solution is very beneficial to a sector that is

978-8-3503-4573-57/22/$31.00 ©2022 IEEE

slowly switching over to more precise, digital equipment, reducing training time and component waste.

A comparison of functionalities and purposes between training apps can be found in Table I.

III. Continental - Factory Of The Future

The project "Continental Factory of the Future" was launched to contribute to technological advancements in the industry. This project, led by four companies (Continental AA, Up Motion, Neoception and Follow Inspiration) and four academic entities (University of Trás-Os-Montes e Alto Douro, Faculty of Engineering of the University of Porto, University of Minho, and INESC TEC). The consortium has different research lines, such as production floor efficiency improvement, product waste reduction, better worker safety and formation, automation, cybersecurity, or human-machine symbiosis. This paper presents an immersive VR training system that overcomes conventional training constraints by production line down-times due to training activities and component waste for training purposes. At the same time, the envisaged solution aims to contribute to faster time to competency and more resilient training. In addition, DT integration is envisaged.

IV. Immersive VR Training App - Requirement Analysis

The case study and requirement analysis was defined in collaboration with Continental staff responsible for the corporate training. The case study of a production line at Continental consists of assembling a car antenna, where the trainee must learn the following seven step process:

TABLE I. Summary of the analysed VR applications.

Name	Description	Advantages	Disadvantages
Siemens-Gamesa	App that allows operator training in Wind power working without need to visit the real workplace.	•Highly faithful simulation; •Better worker performance[c]; •Wide range of interactions; •Visual cues help identify the steps the user should follow; •Allows user error; •Reduction in training time; •Less procedure errors; •Less work accidents; •Several possible training scenarios; •Possibility for multiple users at a time.	•No timer; •No anomaly[d] introduction possibility; •No scoring system.
Virtual Assembly Line Training	Vehicle assembly line simulator, highly reliable and customizable to each operator.	•Highly faithful simulation; •Better worker performance; •Wide range of interactions; •Visual cues help identify the steps the user should follow; •Reduction in training time; •Less assembly mistakes; •Less work accidents; •Several possible training scenarios, including customizable scenarios;	•No timer; •Does not allow user error; •No anomaly introduction possibility; •No scoring system.
SimSpray	VR based training for surface coverage / paint, where students can learn the equipment configuration, proper posture and correct spray gun movements. The simulation can be customized to the intended purpose	•Accurate haptic feedback; •Better worker performance; •Real time performance feedback; •Visual cues help identify the steps the user should follow; •Reduction in training time; •Less work mistakes; •Customizable scenarios; •Several pre-built scenarios; •Has a timer; •Has a score system; •Custom spray gun for better immersion.	•Does not allow user error; •No anomaly introduction possibility; •The custom equipment necessary for its use is costly.
Grundfos Pump Training	Water pump assembly line simulator, highly reliable, built specifically for the intended purpose and so, does not require further customization.	•Better worker performance; •Highly faithful simulation; •Reduction in training time; •Less assembly mistakes; •Several pre-built scenarios; •Customizable scenarios; •Less work accidents.	•No timer; •Does not allow user error; •No anomaly introduction possibility; •No scoring system.
Construction Equipment Assembly Training	Construction equipment assembly line simulator, highly faithful, highly customizable.	•Better worker performance; •Highly faithful simulation; •Comfortable simulation environment; •Wide range of interactions; •Visual cues help identify the steps the user should follow; •Several pre-built scenarios; •Reduction in training time; •Less assembly mistakes; •Less work accidents.	•No timer; •Does not allow user error; •No anomaly introduction possibility; •Affected immersion by representation of controllers instead of virtual hands; •No scoring system.

[a](Serious Games Interactive. 2020. Virtual Assembly Line Training. Available at: <https://www.seriousgames.net/en/portfolio/opel-virtual-assembly-line-training/>)
[b](VRSim. 2020. SimSpray. Available at: https://www.simspray.net/)
[c](as compared to those who did not receive VR training)
[d](as in faulty components, purposely introduced to test the user's attention)

1) Joining the baseplate, prefixation, foam, two cables and a tag and associating these components with a printed circuit board (PCB);
2) Welding the cables and the PCB;
3) Bolting together the PCB and the baseplate and wrapping the cables in duct tape;
4) Adding a radiator to a vertical PCB and welding it to the first PCB;
5) Adding the base ring and cup and bolting this cup to the base;
6) Adding a fixation and a bolt to the base of the antenna;
7) Testing of the antenna and packing.

The virtual environment replicates the real production line with the same manufacturing steps to create an immersive VR app capable of effectively training professionals. To improve the training process, visual cues and other assistance are included. A Trainer's screen is displayed on an outside monitor so he can assist and rate a trainee's performance. This system includes manual anomaly introduction, simulation data and different points-of-view (POV). A scoring system is also important for proficiency level evaluation. The functional requirements go as follows:

- Must have a realistic visual representation of the production line machines and steps;
- Must have a trainer's POV, where an outsider may observe, guide, or grade a worker's performance;
- Must have a scoring system;
- Must record the worker's score;
- Must have a timer;
- Must record the worker's completion time;
- Must allow teleportation movement.

The non-functional requirements are:

- Must have interactions easy to understand;
- Must have interactions easy to use;
- Must have accurate interactions to reality;
- Must be developed in Unity (VR).

The app's requirements do not include the required HMD system or the app's system requirements as these have not been set in stone, although the Oculus Quest was used, and the standard system requirements are based around the development system.

V. Prototype Proposal

A proof of concept was developed based on the Oculus Quest HMD, connected to a computer (Fig. 1 (a)).

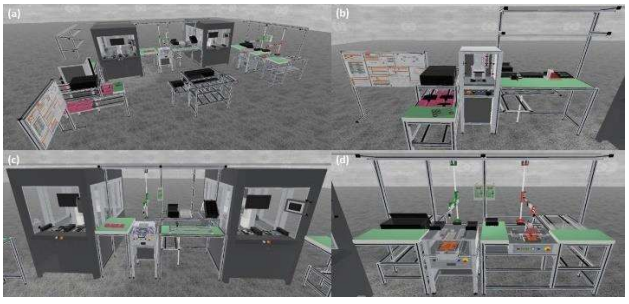


Fig. 1. (a) A complete view of the area developed so far, (b) the first post of the production line, (c) the Welding Posts (left and right machines), along the first screw and tape post (table in the middle of the two welders), (d) the last two bolters in the factory line.

The prototype developed allows the training of the first work post (Fig. 1 (b)). The machinery is mostly in place, except for the last post, the antenna tester.

The environment was built having as basis CAD models of the production line for straightforward integration of industry standards. Fig. 1 (b) shows the current state of the completed portion of the project, Fig. 1 (c) has the central portion of the factory line, where work has started. In Fig. 1 (d) we can see the last portion of the factory line, where only modelling work has started, except for the bolter arms, wich already function as intended.

The production steps developed so far are as accurate as possible to reality, but some may be simplified for easier understanding of the production steps, without sacrificing quality.

The trainers' point of view (Fig. 2) was developed to be shown on separate second screen, outside VR.

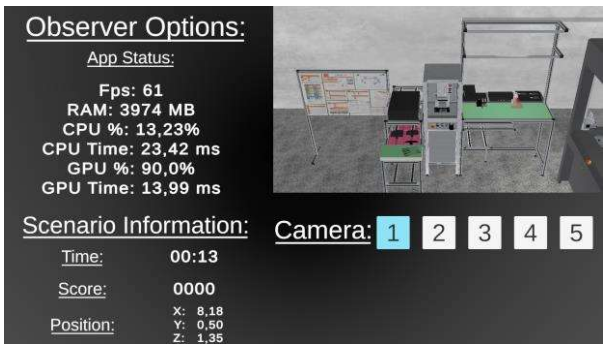


Fig. 2. Trainer's screen, with different Points of view, app performance evaluation and other relevant information.

The prototype has a provisional version of the scoring, calculated by completion time. Worker performance evaluation can prove especially useful, and it will be done with a mix of completion time and points based on the completed production steps, detected malfunctions and defects. The timer and a save system (to a file) are already built. The interactions attempt to be as simple as possible. The teleportation system can move a user within the boundaries of the scenario, by pointing and clicking at the destination. A free movement option, using the controller's analogic component is also in place. The proposed prototype was developed using the Unity Engine and its editor tools.

VI. Performance Evaluation

The CAD's complexity and the need for better performance in immersive VR apps for better user experience makes it important to evaluate the prototype's performance [9]. There is no standard for the performance of a VR app, although some studies refer to an ideal of 90 frames-per-second (FPS) for the best experience [9]. In contrast, others refer than an acceptable app runs at a minimum of 50 fps [10]. The tested variables were maximum/minimum/average FPS, average CPU/GPU load and delay and the RAM usage. Noting that testing with other users is impossible due to the app's state and the project's data sensitivity.

The tests used two different computers, one with an i7-5820k Intel CPU, 16GB RAM, and an RTX 2070 GPU (A)

and another with an i7-8700k Intel CPU, 32GB RAM, and an RTX 2080ti GPU (B). The tests were conducted by observing each computer for approximately five minutes (a period that allows testing all possible heavy loads), looking at different scene locations, camera movement, and different observer camera selections, as seen in Table II.

TABLE II. PERFORMANCE RESULTS OF THE VR APPLICATION IN THREE DIFFERENT QUALITY SETTINGS.

Variables	Very Low		Medium		Ultra	
Computer Used	A	B	A	B	A	B
Maximum FPS	78	80	78	75	55	56
Minimum FPS	69	67	40	42	34	35
Average FPS	77	78	70	72	50	50
Average CPU Usage (%)	12	14	30	22	38	25
Average CPU Time (ms)	14	14	14	14	27	32
Average GPU Usage (%)	95	99	69	77	55	55
Average GPU Time (ms)	2.5	2.0	3.5	3.0	5.0	4.0
Average RAM Usage (MB)	488	486	496	499	503	505

A. Discussion

The HMD used in the experiment had a frequency of 72hz, meaning anything above 72 fps will not be perceived by the user. The results were acceptable and similar, although two different computers were used, the app will still need improvements like better texture attribution, simpler of 3D models, or code improvement for better performance.

The testing settings used were the standards in Unity version 2021.2.19f1. Using the app in Ultra settings is not recommended, as it causes low framerates. However, in the other settings it feels good and responsive. The FPS, GPU and CPU loads observed were acceptable except in Ultra settings. The RAM usage was low. The minimum registered fps were sporadic spikes and not continuous/noticeable.

VII. CONCLUSION

This article presents an immersive training app, its research line and is designed to train professionals in antenna production lines. The prototype includes trainer's observation, and performance scoring which contributes to the state of the art, aiming to lessen ambiguity in the performance scoring and a possible future DT integration.

Some of the app's benefits include the accuracy of the virtual representation compared to reality, the options given to the observer to control the experiment and the scoring system, which will have significant use in determining how ready a worker is to try the real factory line.

Some of this project's challenges are creating a not necessarily linear assembly line, allowing for user error. The materials of the objects are not perfect and will need further work for a more natural feel. The app needs a performance improvement, for a better and smoother experience.

As for future work, the missing components will be implemented. A non-assisted variant of the app is envisaged, where the workers are challenged with a point system, little to no visual cues and part failure. The score is intended to be an indicator of preparedness. The DT is also a future research topic, for better prediction systems and easier system diagnosis.

ACKNOWLEDGMENT

This work was supported by the R&D Project "Continental Factory of Future, (CONTINENTAL FoF) / POCI-01-0247-FEDER-047512", financed by the European Regional Development Fund (ERDF), through the Program "Programa Operacional Competitividade e Internacionalização (POCI) / PORTUGAL 2020", under the management of AICEP Portugal Global – Trade & Investment Agency.

REFERENCES

[1] L. Kugler, "The state of virtual reality hardware", in Communications of the ACM 64.2, February 2021, pp. 15-16, doi: 10.1145/3441290.

[2] L. He, R. Wang, X. Shi, Q. Liang, K. Fang and J. Li, "VR Educational Game Design and Research Based on Multi-Modal Interaction from the Perspective of Embodied Cognition", 2020 4th Annual International Conference on Data Science and Business Analytics (ICDSBA), 2020, pp. 329-331, doi: 10.1109/ICDSBA51020.2020.00091.

[3] J.V. Christensen, M. Mathiesen, J.H. Poulsen, E.E. Ustrup, M. Kraus, "Player experience in a VR and non-VR multiplayer game", ACM International Conference Proceeding Series, 2018, pp. 373-384, doi: 10.1117/12.317451.

[4] S.C. Mallam, S. Nazir, "Effectiveness of VR Head Mounted Displays in Professional Training: A Systematic Review", Technology, Knowledge and Learning, 2021, pp. 999–1041, doi: 10.1007/s10758-020-09489-9.

[5] F. Tao, H. Zhang, A. Liu, A. Y. C. Nee, "Digital Twin in Industry: State-of-the-Art", in IEEE Transactions on Industrial Informatics, vol. 15, no. 4, pp. 2405-2415, April 2019, doi: 10.1109/TII.2018.2873186.

[6] V. Kuts, T. Otto, T. Tähemaa, Y. Bondarenko, "Digital Twin based synchronized control and simulation of the industrial robotic cell using Virtual Reality", Journal of Machine Engineering, vol. 19, pp. 128-144, February 2019, doi: 10.5604/01.3001.0013.0464.

[7] U. Radhakrishnan, F. Chinello, K. Koumaditis, "Immersive Virtual Reality Training: Three Cases from the Danish Industry", Proceedings - 2021 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops, 2021, pp. 581-582, doi: 10.1109/VRW52623.2021.00008.

[8] R. Barkokevas, C. Ritter, V. Sirbu, X. Li, M. Al-Hussein, "Application of virtual reality in task training in the construction manufacturing industry", 36th International Symposium on Automation and Robotics in Construction, 2019, pp. 796-803, doi: 10.22260/isarc2019/0107.

[9] D. Kanter, "Graphics Processing Requirements for Enabling Immersive VR", AMD: Santa Clara, CA, USA, 2015.

[10] C. Zhang, "Investigation on Motion Sickness in Virtual Reality Environment from the Perspective of User Experience", 2020 IEEE 3rd International Conference on Information Systems and Computer Aided Education (ICISCAE), 2020, pp. 393-396, doi: 10.1109/ICISCAE51034.2020.9236907.

Supporting Human Operators in an Industrial Shop Floor through Pervasive Augmented Reality

Rafael Maio
IEETA, DETI
University of Aveiro
Aveiro Portugal
rafael.maio@ua.pt

André Santos
IEETA, DETI
University of Aveiro
Aveiro, Portugal
andrembs@live.ua.pt

Bernardo Marques
DigiMedia, IEETA
University of Aveiro
Aveiro, Portugal
bernardo.marques@ua.pt

Duarte Almeida
Bosch
Thermotechnology
Aveiro Portugal
duarte.almeida@pt.bosch.com

Pedro Ramalho
Bosch
Thermotechnology
Aveiro, Portugal
pedro.ramalho@pt.bosch.com

Joel Baptista
Bosch
Thermotechnology
Aveiro, Portugal
joel.baptista@pt.bosch.com

Paulo Dias
IEETA, DETI
University of Aveiro
Aveiro, Portugal
paulo.dias@ua.pt

Beatriz Sousa Santos
IEETA, DETI
University of Aveiro
Aveiro, Portugal
bss@ua.pt

***Abstract*—Augmented Reality (AR) has been applied in Industry 4.0 contexts for training, assistance, maintenance, assembly or quality control. This work describes the use of Pervasive AR to support human operators in a shop floor through pervasive experiences, while performing logistics operations. A Human-Centered Design methodology with partners from the industry was used to identify operators' difficulties, challenges, and define requirements, leading to the creation of a Pervasive AR prototype to support the operators' task or/and the initial training in such scenarios. An initial user study with 12 participants was conducted in a simulated environment, comparing three conditions: C1 - Head-Mounted Display (HMD), C2 - Handheld Device (HHD) and C3 - Paper manuals. Later on, a second user study with 26 participants took place in a real shop floor, to collect preliminary feedback from individuals with different expertise. Results from both studies suggest advantages in using AR, in particular for the training of operators not familiar with the task. Condition C1 was preferred and considered more useful to support the operator's task by the majority of participants.**

***Index Terms*—Industry 4.0, Human Operators, Shop Floor, Pervasive Augmented Reality, Human-Centered Design.**

I. Introduction

Industry 4.0 has been proposed as a new stage in industrial maturity, integrating smart sensors, embedded systems, cyber-physical systems and Internet-of-Things (IoT) into the manufacturing processes [1]. Augmented Reality (AR) is one of its pillars, given its ability to provide solutions for supporting operators during their daily tasks. Prior studies support the added value that AR can have in industrial scenarios, integrating digital information in the human-operators workspace [2], helping them in assembly tasks [3], context-aware assistance [4], data visualization and interaction (acting as a Human-Machine Interface (HMI)) [5], indoor localization [6], maintenance applications [3], quality control [7] or material management [5]. Literature identifies several benefits of using AR, like increased work safety, effective learning and training, as well as error and task-time reduction [8].

The use case considered in this work consists in picking materials to assemble kits. In a *mixed-model assembly*, the contents of the kits differ, meaning that operators need to identify the kit materials during their preparation [9]. This information can be conveyed through paper manuals [9], pick-by-voice systems [10], pick-by-light strategies [10], Head-Up Displays (HUD) [11] or by AR [9]. Paper manuals can be mentally and physically demanding to some operators, after hours of labor, resulting in mistakes and less efficiency [9]. Hence, the learning and training phases are not straightforward and often novice operators require a long preparation time.

This paper showcases a Human-Centered Design (HCD) methodology used to create a Pervasive AR prototype for training operators in an industrial shop floor. Plus, two user studies (laboratory/real-life production line) are described and their results discussed.

II. Methods and Materials

This section describes the industrial scenario considered, including operator tasks and the industrial context they work on. Then, the conceptualization and development of a Pervasive AR prototype is presented.

A. Shop Floor Scenario

The shop floor is centrally located in the factory, surrounded by corridors where vehicles, such as counterbalance forklift trucks, circulate and by several other industrial lines, containing for example presses and other heavy and loud machines, making the environment extremely noisy. The production line is composed of seven large shelves, arranged in an "*L*" shape. Each shelf is organized from a bottom-right to a top-left labelling, with the format *"shelf_number-row_number-column_number"* (e.g., E02-01-06) (Figure 1).

978-8-3503-4573-57/22/$31.00

The components relative to the label remain static during large periods (6 months or larger).

Fig. 1. Shelves of the industrial shop floor with different boxes of components, typically denominated as 'supermarket'.

The operators' tasks consist in assembling kits of predefined components (documentation, spare parts, etc.) to be merged with the final product before being packed for distribution worldwide. Overall, there are several different kits, which are produced according to the needs of the final product. The list of materials composing each kit is printed in a paper sheet. Each component of the list has the following information:

- *Component:* The reference of the component;
- *Denomination:* Component common name;
- *Quantity:* The number of pieces to include in the kit;
- *Picking location:* The position of the component (using the labelling system explained above (e.g., E02-01-06).

The kit assembly requires the selection of components by hand, and putting them into a cart with 16 sections. Operators report that frequently, to save up time, the list is not followed strictly sequentially; if they spot that two or more components are close to each other, they collect those components successively. At the end, the packing process is finalized and the kit is sent to the next production line. It is noticeable that operators perform the task using their mental capabilities and task expertise, drawing a shorter path from reading the labels in the list while memorizing which materials were already picked and those who were not. This behaviour leads to errors, such as skipping components in the list, resulting in incomplete kits. Also, the learning and training process is not straightforward and takes a significant amount of time to reach the task performance of experienced operators.

B. Methodology and Requirements

To understand the operators' needs and challenges, leading to the elicitation of requirements, we considered a HCD methodology with partners from the industry in an ongoing research project. Hence, various on-site and remote meetings occurred, as well as brainstorming sessions and visits to the shop floor. To address the scenario considered, it is necessary to select a pre-defined kit, read its content, and inform the location of each component. This imply continuous movement along the production line. The goal is to allow operators to walk less throughout the entire shift, by doing so in a more efficient manner, with higher productivity and reduced error (by validating correct picking). Another relevant topic, is to validate the picking, reducing the possibility of grabbing a component from the wrong box or forgetting to pick it. In this context it is also important to evaluate hardware alternatives, being able to have a solution that can run in different devices, to better comprehend which can be used on the long term.

C. Pervasive AR Tool

To fulfill the previous requirements, a Pervasive AR tool was developed (figure 2). Pervasive AR extends the AR concept for experiences that are continuous in space, being aware of and responsive to the user's context and pose [12]–[14].

Two distinct methods were considered: Head-Mounted Display (HMD) and Handheld Display (HHD). For both methods, the tool is divided into two modules: *configuration* and *visualization* (figure 2). The *configuration* of the AR content over the real-world is mandatory, but only needs to occur once (unless the context is changed later on). This module maps the points of interest of the real-world and allows to add virtual content, in the form of green cubes over the real-word boxes, arranged on the shelves. These cubes can be translated, rotated, scaled and copied, providing the required geometric transformations to perform the entire AR configuration efficiently and adaptable to various layouts and box sizes. Then, from a list containing the existing components, each cube is associated to the desired component. After having the virtual information correctly placed and labelled, it is stored for later use. Regarding the *visualization* module, it uses the stored information to automatically present the AR content in the correct pose over the real environment. The tool compares the stored real-world mapping with the the camera viewpoint. When a match occurs, the associated AR content appears at the configured pose. Every cube representing the component that needs to be picked is identified as such and as the process is performed. When a component is collected, the corresponding cube changes its color from green to gray. This way, displaying visual feedback, while identifying the remaining components that must be picked to complete the kit.

Although the two methods are quite identical and were built using the Unity game engine, different technologies were considered. For the HMD, we considered the Mixed Reality Toolkit (MRTK) with its local world anchors feature, allowing local persistence. Concerning the HHD, it uses the ARCore cloud anchors feature, requiring internet connection for storing the real-world information in an external API. Beyond the technological difference, the way that the material picking validation is performed is also distinct. While the HMD, detects if a human hand entered the virtual cube (while grabbing components in the real boxes), the HHD requires the user to press the virtual cube on the device screen, before or after collecting the component.

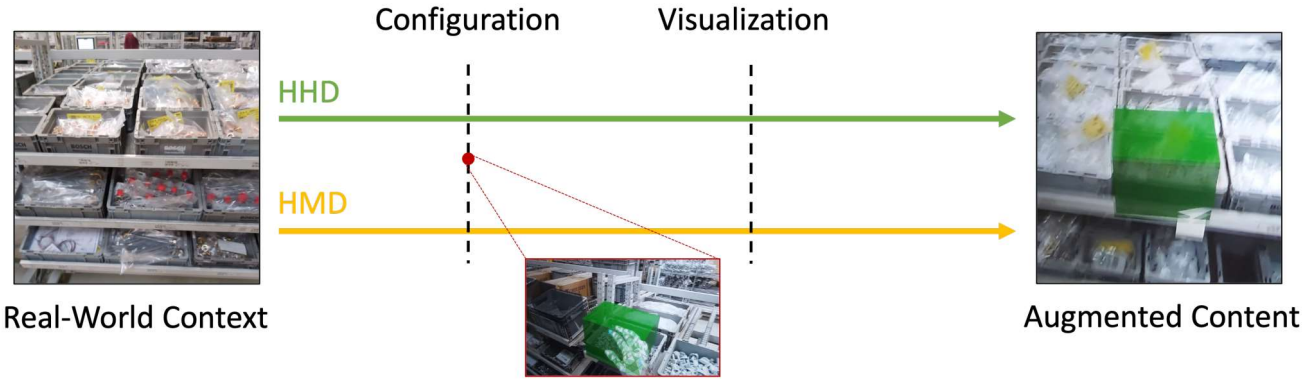


Fig. 2. Prototype architecture. Supports two display technologies: Head-Mounted Display (HMD) and Handheld Display (HHD). The AR environment configuration is performed over the real-world scenario and stored. It is possible to visualize the augmented content several times using the stored information.

III. User study

Next, we describe the two initial users studies conducted to evaluate and compare different methods. First a user study performed at a laboratory, which simulates the real scenario, and a second one at the industrial environment.

A. Initial User Study

A laboratory study with 12 participants (4 female, 8 male) was conducted to compare three conditions during a realistic learning procedure (figure 3): C1 - HMD (Microsoft HoloLens 2); C2 - HHD (Asus Zenfone AR); C3 - Paper manual. This study occurred in a laboratory with 28 square meters, with two shelves facing each other on opposite sides of the room. To simulate the industrial task, each shelf contained twelve paper boxes (using a labelling scheme similar to the one used by our industrial partner), containing different Lego pieces. Three distinct picking lists were used, requiring physical movement between shelves. A card box resting at the center of the room was used to place the picked components. The picking had to be performed sequentially from the lists order.

After participants have given their informed consent, the experience setup and the task was explained. Then, they were presented with each condition. A within-group experimental design was used and the order of conditions was alternated to minimize learning effects. At the end, participants answered a post-task questionnaire to compare the conditions and gave their opinion and suggestions.

B. First Impressions on the Shop Floor

The prototype was also tested in an industry shop floor during two visits, with an average 6-hour duration each. The goal was to evaluate the technology in a real scenario, as well as to gather first impressions about the use of AR (figure 4).

During the first visit, six participants (3 female, 3 male), including target-users, assembled a kit using the proposed prototype through the same hardware as before. Before performing the task, a brief introduction was given. At the end, an informal interview occurred to collect subjective data. All feedback collected was used to improve the prototype

Fig. 3. Participant performing the picking task using AR to visualize the location of the next material at the laboratory environment.

Fig. 4. Participant performing the picking task using AR to visualize the location of the next material at the industry shop floor.

before the second visit, during which 20 participants (7 female, 13 male) from multiple departments (e.g., operators, line managers, logistics, ergonomics, maintenance, process engineering, engineering manager, production manager) tested the prototype. The procedure was identical to the previous

visit, collecting subjective data through post-task interviews.

IV. Preliminary Results and Discussion

Overall, the AR prototype was deemed robust and accurate to be applied in a real industrial setup. It was rated as being easy to use and having great potential to support the operator's task, specially for inexperienced operators. Likewise, the AR conditions were considered as having lower cognitive effort when compared to the paper alternative. All participants, except one, preferred the use of the HMD condition.

Participants preference was mainly due to being easier to maintain a good visual-motor coordination, while having a hands-free setting. Also, they better perceived the real-world surroundings while visualizing the virtual information in relation to the real world. Moreover, they stated that the virtual content was more stable using the HMD, which is advantageous versus the HHD conditions that may cause some disorientation. Besides, it was unanimous that a brief adaption period is required, although with training, performance can easily increase. Another relevant insight was uncovered during the task execution of an operator, where it was found that the HMD condition poorly detects hands covered with black and shiny gloves. Notwithstanding, the following improvements were suggested:

- The AR content should be sorted by the shortest path to complete the kit;
- The gray cubes, which do not need to be picked, should be hidden.
- In the HMD condition, the green color is easily confused with the environment. Other colors should be tested;
- The HHD condition could also validate the picking process through a hand detection feature (as in the HMD).
- The tool should be prepared for color blind people;
- Directional arrows could appear to guide operators on the next step, avoiding wasting time looking around;
- The percentage of components gathered should be displayed to provide the status of the task.
- The virtual information should not instantly disappear after validating the picking of a specific material;
- The HMD position on operators head requires an adjustment for ergonomic purposes before being used.

V. Concluding Remarks and Future Work

This work proposed a Pervasive AR prototype using two distinct methods (HHD & HMD), aimed at training/supporting operators in a shop floor scenario. A Human-Centered Design (HCD) methodology was used in collaboration with partners from the industry sector. This was essential to identify needs, challenges and requirements from domain experts and target-users. Results from a user study with 12 participants in a simulated environment, along with a second study with 26 participants at the industrial environment demonstrated the high potential of adopting AR to support operators during their daily tasks, in particular for the training of new operators. All in all, the larger majority of participants preferred the HMD condition, being considered more useful and efficient to support the operators in a hands-free setting.

This study is being expanded by integrating the feedback received during the tests and by planning a new study, having participants conducting longer and more complex tasks.

VI. Acknowledgments

We thank everyone involved for their time and expertise, in particular Professor Carlos Ferreira from IEETA, DEGEIT, University of Aveiro This research was developed in the scope of the Augmented Humanity project [POCI-01-0247-FEDER-046103 and LISBOA-01-0247-FEDER-046103], financed by ERDF through POCI. It was also supported by IEETA, in the context of project [UIDB/00127/2020].

References

[1] S. S. H. Al-Maeeni, C. Kuhnhen, B. Engel, and M. Schiller, "Smart retrofitting of machine tools in the context of industry 4.0," *Procedia CIRP*, vol. 88, pp. 369–374, 2020.

[2] E. Marino, L. Barbieri, B. Colacino, A. K. Fleri, and F. Bruno, "An augmented reality inspection tool to support workers in industry 4.0 environments," *Computers in Industry*, vol. 127, p. 103412, 2021.

[3] N. Gavish, T. Gutiérrez, S. Webel, J. Rodríguez, M. Peveri, U. Bockholt, and F. Tecchia, "Evaluating virtual reality and augmented reality training for industrial maintenance and assembly tasks," *Interactive Learning Environments*, vol. 23, no. 6, pp. 778–798, 2015.

[4] S. Aromaa, A. Väätänen, I. Aaltonen, V. Goriachev, K. Helin, and J. Karjalainen, "Awareness of the real-world environment when using augmented reality head-mounted display," *Applied Ergonomics*, vol. 88, p. 103145, 2020.

[5] P. Fraga-Lamas, T. M. FernáNdez-CaramáS, Ó. Blanco-Novoa, and M. A. Vilar-Montesinos, "A review on industrial augmented reality systems for the industry 4.0 shipyard," *IEEE Access*, vol. 6, pp. 13358–13375, 2018.

[6] S. Saeedi, B. Bodin, H. Wagstaff, A. Nisbet, L. Nardi, J. Mawer, N. Melot, O. Palomar, E. Vespa, T. Spink, C. Gorgovan, A. Webb, J. Clarkson, E. Tomusk, T. Debrunner, K. Kaszyk, P. Gonzalez-De-Aledo, A. Rodchenko, G. Riley, C. Kotselidis, B. Franke, M. F. O'Boyle, A. J. Davison, P. H. J. Kelly, M. Luján, and S. Furber, "Navigating the landscape for real-time localization and mapping for robotics and virtual and augmented reality," *Proceedings of the IEEE*, vol. 106, no. 11, pp. 2020–2039, 2018.

[7] D. Segovia, M. Mendoza, E. Mendoza, and E. González, "Augmented reality as a tool for production and quality monitoring," *Procedia Computer Science*, vol. 75, pp. 291–300, 12 2015.

[8] E. Bottani and G. Vignali, "Augmented reality technology in the manufacturing industry: A review of the last decade," *IISE Transactions*, vol. 51, no. 3, pp. 284–310, 2019.

[9] R. Hanson, W. Falkenström, and M. Miettinen, "Augmented reality as a means of conveying picking information in kit preparation for mixed-model assembly," *Computers & Industrial Engineering*, vol. 113, pp. 570–575, 2017.

[10] D. Battini, M. Calzavara, A. Persona, and F. Sgarbossa, "A comparative analysis of different paperless picking systems," *Industrial Management & Data Systems*, vol. 115, pp. 483–503, 04 2015.

[11] A. Guo, X. Wu, Z. Shen, T. Starner, H. Baumann, and S. Gilliland, "Order picking with head-up displays," *Computer*, vol. 48, no. 6, pp. 16–24, 2015.

[12] J. Grubert and S. Zollmann, "Towards Pervasive Augmented Reality: Context- Awareness in Augmented Reality," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 6, pp. 1706–1724, 2017.

[13] B. Marques, R. Carvalho, J. Alves, P. Dias, and B. S. Santos, "Pervasive Augmented Reality for Indoor Uninterrupted Experiences: a User Study," in *UbiComp/ISWC'19*, p. 141–144, 2019.

[14] M. Neves, B. Marques, T. Madeira, P. Dias, and B. S. Santos, "Using 3D Reconstruction to create Pervasive Augmented Reality Experiences: A comparison," in *IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pp. 712–713, 2022.

Session 3: Computer Vision and Image Processing

chair: António Ramires

Trios: A Framework for Interactive 3D Photo Stylization on Mobile Devices

Ulrike Bath
Hasso Plattner Institute
University of Potsdam, Germany
ulrike.bath@student.hpi.uni-potsdam.de

Sumit Shekhar
Hasso Plattner Institute
University of Potsdam, Germany
sumit.shekhar@hpi.uni-potsdam.de

Hendrik Tjabben
Adobe Systems Engineering GmbH
Hamburg, Germany
tjabben@adobe.com

Amir Semmo
Digital Masterpieces GmbH
Potsdam, Germany
amir.semmo@digitalmasterpieces.com

Jürgen Döllner
Hasso Plattner Institute
University of Potsdam, Germany
juergen.doellner@hpi.uni-potsdam.de

Matthias Trapp
Hasso Plattner Institute
University of Potsdam, Germany
matthias.trapp@hpi.uni-potsdam.de

***Abstract*—For decades, Image-based Artistic Rendering (IB-AR) has been successfully employed to simulate the appeal of traditional artistic styles for enhanced visual communication. Recently, 3D photography has emerged as a new medium that provides an immersive dimension compared to 2D photos. The possibility to change the viewpoint, depicting parallax effects is mesmerizing. We present *Trios*, an interactive mobile app that combines the vividness of IB-AR with immersive 3D photos. *Trios* implements an end-to-end pipeline for the acquisition of data, generation of a 3D photo in the form of a Layered Depth Image (LDI), and its artistic rendering. The app allows the user to either capture the input data or load existing data from the device. As part of the generation step, users can set the number of layers used for representation of 3D photos. Finally, with different artistic filters and their parameterization users can stylize either an individual semantic layer or all layers simultaneously. The complete pipeline runs at interactive frame rates and the final output is obtained as a compact video, which can easily be shared. Thus, it serves as a unique interactive tool for digital artists interested in creating immersive artistic content.**

***Index Terms*—3D photos, image stylization, mobile devices**

I. Introduction

A. Motivation

Traditional 2D photo captures a scene as a frozen moment in time. Recently 3D photos have emerged as a new medium to make such moments more immersive [1]. We refer to 3D photo as a representation which introduces scene parallax – difference in the apparent position of scene-objects due to change in viewpoint (and not the traditional stereo-pair images for perceived 3D effect). The ability to explore parallax effects, especially on the flat-screen of a mobile device, is compelling [2]. On the other hand, with the advancement in mobile graphics advanced stylistic rendering techniques have been successfully deployed on mobile devices [3], [4]. In this work, we aim to extend the visual-richness of image-stylization approaches for 3D photos which are deployable on a mobile-device. As compared to traditional stylization techniques, 3D photos offer new possibilities with respect to stylization and visualization. Moreover, a mobile-based approach will have implications in terms of ease of use.

Fig. 1: *Trios* is a mobile app that enables users to render and stylize 3D photos. The user interface provides interactive control over a variety of artistic filters, both classical and neural, as well as rendering aspects of 3D photos.

B. Problem Statement & Challenges

To this end, we develop a mobile-based framework for generation and stylization of 3D photos given RGB or RGB-D input data (Fig. 1). We identify and address the following challenges:

a) Interactivity: Most of the existing approaches for creating 3D photos do not involve users in the generation step. The end user only views the final output while the generation pipeline remains a black-box. We offer a simple user-interface to interact with the generation and the stylization aspects.

b) Consistency: For traditional 2D photos any editing should be spatially consistent across semantic similar regions for visually aesthetic output. In case of 3D photos, the above requirement becomes paramount as any spatial inconsistencies also reflect temporally while viewing/exporting the 3D photo. We address the above via (semantic-)segment-wise stylization of the image and smooth virtual-camera movement while viewing/exporting the 3D photo.

978-8-3503-4573-57/22/$31.00

c) Throughput: To achieve consistent output while maintaining interactivity, we require fast throughput on a mobile device. We achieve the above by making use of GPU-aligned data structures and Apple's proprietary graphical processing (Metal) Application Programming Interfaces (APIs).

C. Approach & Contributions

To address these challenges, our proposed framework adopts the following approach. As the first step, if no depth data is given, we estimate the depth for the given input image using a mobile version of a state-of-the-art depth-estimation technique [5]. The given/estimated depth data is used to decompose the RGB image into layers at different depth levels and is represented as a LDI [6], stored in a 2D texture array for Graphics Processing Unit (GPU)-based processing. Unlike Kopf *et al.* [2] we do not convert the LDI representation into a 3D-mesh, as LDI-space allows for efficient image-based stylization. For each layer, potential dis-occlusions which might get visible due to viewpoint change are identified and inpainted. The inpainted layers can be visualized as a 3D photo within the app via viewpoint variations induced due to device movement or via traversing a particular trajectory by the virtual camera. Subsequently, the generated 3D photo can be exported as a video file. For spatially consistent stylization, the input image is divided into (semantic-)segments and the user performs stylization on a per-segment basis. Similar to input-image, the stylized image is decomposed into depth-based layers which can then be exported as a 3D photo.

Our contributions are summarized as follows, we propose (1) a novel framework for 3D photo generation and stylization on mobile devices, and (2) a respective user-interface for interactive editing of stylization parameters and camera animations.

II. Background & Related Work

A. 3D Photo Generation

Hedman *et al.* [1] introduce the concept of 3D photography by using a set of input photos to construct a 3D panorama. In a follow-up work, Hedman and Kopf [7] propose a novel optimization that speeds up the panorama generation step by two orders of magnitude. The focus of both of these work is to capture large panoramas to be viewed in a Virtual Reality (VR) setup. The approaches are inconvenient regarding usability and requires capturing multiple photos. Mildenhall *et al.* [8] propose an algorithm that guides users to capture a grid of sampled views, wherein each sample is expanded to a local light-field which are then fused for virtual scene exploration. Similar to previous techniques, the above requires capturing of multiple images for light-field fusion. Shih *et al.* [9] propose the first method for converting a single RGB-D image into a 3D photo employing a multi-layer representation for novel view synthesis. The authors make use of LDI as a multi-layer representation for efficient editing [6], [10]. We also employ this representation for novel view synthesis and exploration. Jampani *et al.* [11] utilize depth-aware inpainting for improved segmentation and layering, thereby preserving fine image details in the foreground. Kopf *et al.* [2] for the first time proposed a method wherein the entire 3D photo generation pipeline is carried out on a mobile device. Our framework is also implemented on a mobile-device, however, unlike previous approach we provide user control in the generation process. Further, we allow interactive stylization of 3D photos for enhanced visual aesthetics.

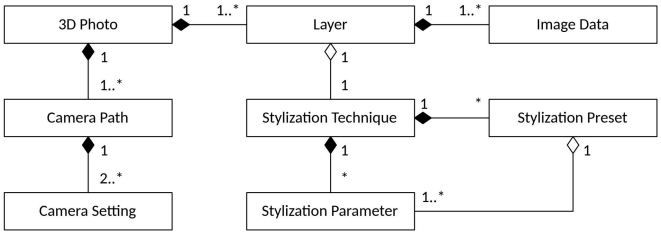


Fig. 2: Data model for 3D photo stylization. A 3D photo basically resembles an Layered Depth Image (LDI) comprising multiple color and depth layers, each referencing a stylization technique that is parameterized by a number of parameters, which are grouped to presets. For rendering and export, a number of virtual camera animations can be used.

B. Image Stylization on Mobile Devices

Due to advances in mobile graphics hardware, on-device image stylization is not only becoming feasible but also increasingly popular via casual creativity apps. Durschmid *et al.* [12] present a generic GPU-based app that allows to design on-device stylization components by reusing building blocks. Pasewaldt *et al.* [4] showcase a broad range of IB-AR techniques running on a mobile device. The above mobile-based methods consider only RGB data as input. Recently, Shekhar *et al.* [13] demonstrate depth-based stylization methods running interactively on a mobile device. As part of our approach we employ depth data only for 3D photo generation and as future work would also incorporate it for stylization. Note that there are already consumer mobile applications (e.g., Loopsie, PopPic, Parallax, DazzCam) that allow on-device 3D photo generation and limited editing. However, unlike these, we provide a broad range of advanced IB-AR techniques for editing the 3D photo. Further, we provide more control to the user in terms of generation and editing.

III. Data Model for 3D Photo Stylization

Prior to focusing on 3D photo synthesis and stylization (Sec. IV), this section briefly describes the basic data model used by our approach. Fig. 2 shows an overview of the respective data structures.

3D Photo: A 3D photo is represented as an LDI [6], i.e., a number of layers, as well as respective camera animation data required for viewing and export.

Layer: Fundamentally, a layer associates required (color and depth) and optional (normal, mask) image data with reference to a depth-level.

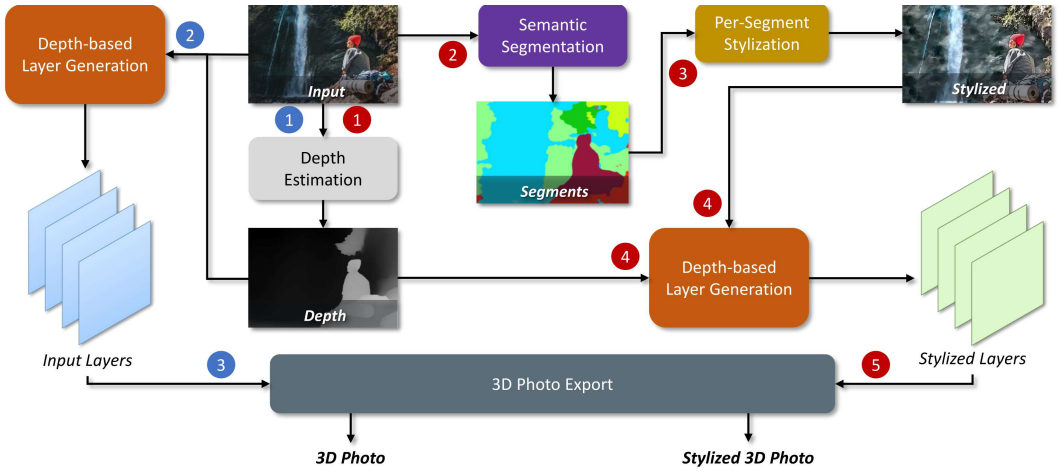


Fig. 3: Schematic overview of the processing pipeline implemented in our app *Trios*, where steps 1-3 depict the creation of a *3D Photo* and another set of steps 1-5 showcase the creation of a *Stylized 3D Photo*.

Image Data: In addition to color and depth images, image data instances can comprise masks and intermediate representations required for processing.

Stylization Technique: An instance of a stylization technique defines the order of algorithms applied to a semantic segment yielding an abstraction or stylized image.

Stylization Parameter & Preset: Stylization techniques are parameterized by a number of parameters whose values can be controlled directly by a user or defined via presets.

Camera Path & Setting: To represent animations of the virtual camera, for 3D photo synthesis, a camera path instance stores an ordered list of camera settings. A camera setting comprises a 2D camera position, 2D look-to vector, and a zoom parameter.

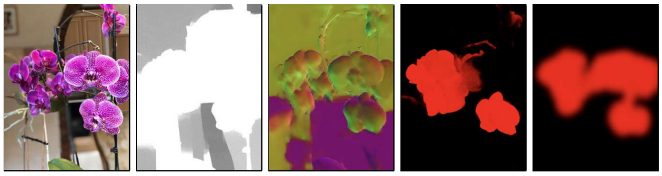
IV. Method for 3D Photo Stylization

Fig. 3 shows an overview of the presented framework. Its modular design comprises the following conceptual processing stages, which are described in the remainder of this section.

Input Data Acquisition & Preprocessing: This stage acquires the data required to synthesize and stylize a 3D photo. It can consume RGB and RGB-D data and optionally compute additional raster data used within the framework and perform any preprocessing if required (Sec. IV-A).

LDI Generation & Inpainting: For 3D photo synthesis, our approach is based on the concept of LDIs. For it, this stage separates individual colors layers and perform inpainting necessary to generate a plausible parallax effect (Sec. IV-B).

Per-Segment Stylization: The depth-based layer generation do not respect image semantics. To address this issue and produce a spatially consistent output, we divide the image into semantic-segments. This core stage enables the above and further application of stylization techniques on a per-segment basis (Sec. IV-C).



(a) Input (b) Depth (c) Normal (d) Matte (e) Salience

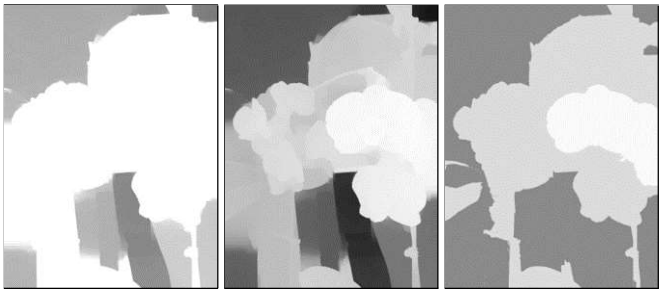
Fig. 4: Examples of additional raster data that can be automatically derived based on a color input image and used as input for image stylization and 3D photo rendering.

Rendering & Export: This stage exports the synthesized 3D photo animation as a video based on the virtual camera path specified by the user (Sec. IV-D).

A. Acquisition and Preprocessing of Input Data

This stage prepares the input data to be used in the subsequent stages of our framework.

a) Input Data Acquisition and Generation: Fig. 4 shows examples of data our framework can consume for 3D photo stylization. Besides depth, it computes normal vectors for surface orientation [14], as well as matte and saliency data; the last two acquired using the Apple Vision framework. To support a potentially wide range of mobile devices, this stage can handle two types of input data: (1) RGB-only images and (2) RGB-D images, depending on the respective mobile hardware available to a user. In case of RGB-only input image, the pre-processing stage uses MiDaS [5] to compute relational depth information based on color values. In case the depth data is provided by the device depth-sensors, it is usually of lower spatial resolution than the respective color image. For it, the depth map is upsampled to the color image resolution using a standard upsampling filter, e.g., joint-bilateral upsampling [15].



(a) Acquired depth D (b) Norm. depth D_N (c) Quantized depth D_Q

Fig. 5: Visualization of exemplary depth data during different stages of pre-processing. First, the acquired depth values (a) are normalized to span the complete range of possible depth values (b). Subsequently, normalized values are then quantized uniformly into a number of bins specified by the user (c).

(a) Input (b) Background (c) Midground (d) Foreground

Fig. 6: Example visualization of separating an input image (a) into individual LDI layers (b)-(d), based on quantized depth data (Fig. 5c). White indicates pixels associated to other layers.

b) Preprocessing of Depth Data: Following to that, further depth data processing is performed as depicted in Fig. 5. After acquisition (Fig. 5a), the depth map (D) is normalized (Fig. 5b) based on a pre-computed depth histogram. The histogram yields the minimum ($d_{\min}$) and maximum ($d_{\max}$) depth values. Normalization is then performed:

$$D_N(x,y) = \frac{D(x,y) - d_{\min}}{d_{\max} - d_{\min}}$$

Subsequently, the normalized values are thresholded (Fig. 5c) based on uniform quantization:

$$D_Q(x,y) := \begin{cases} \lfloor D_N(x,y) \cdot b \rfloor & D_N(x,y) < 1 \\ b-1 & \text{otherwise} \end{cases}$$

The number of bins $b \in \mathbb{Z}^{0+}$ represents layers of unique depth value, and can be set by the user as a parameter. In general, the number of depth layers account for the resulting rendering quality and processing performance, i.e., lower enables faster rendering, higher increases visual quality – depending on the distribution of depth values. In our experiments, we found that $b = 3$, i.e., the separation between background, midground, and foreground is sufficient for most scenes (Fig. 6).

(a) Input color (b) Coverage (c) Inpainting

Fig. 7: Visualization of the layer processing stages performed by the framework prior to the stylization and rendering of 3D photos. For the input color layer (a), regions are inpainted that would dis-occlude during camera animation (c). All other pixel data will remain unchanged, visualized by the coverage (b) (white: requires inpainting, gray: no inpainting required).

B. LDI Computation and Inpainting

Based on the pre-processed depth data, this stage first performs LDI generation by separating the RGB-D data into individual layers of unique depth complexity. Following to that, for each layer the RGB regions that possibly become subject to disocclusions during 3D photo rendering are inpainted. These steps yield a number of RGB-D layers that can be stylized individually in an art-directed way (Sec. IV-C).

a) Layer Separation: Fig. 6 shows an example of the separated LDI layer, based on the pre-processing results depicted in Fig. 5c. The layer segmentation is performed based on a per-pixel level using depth data as follows. For each depth bin $i \in 0, \ldots, b-1$, the input image I is copied into a respective layer L_i. The transparency value of pixels is set to zero if the depth value at that location does not belong to the respective bin.

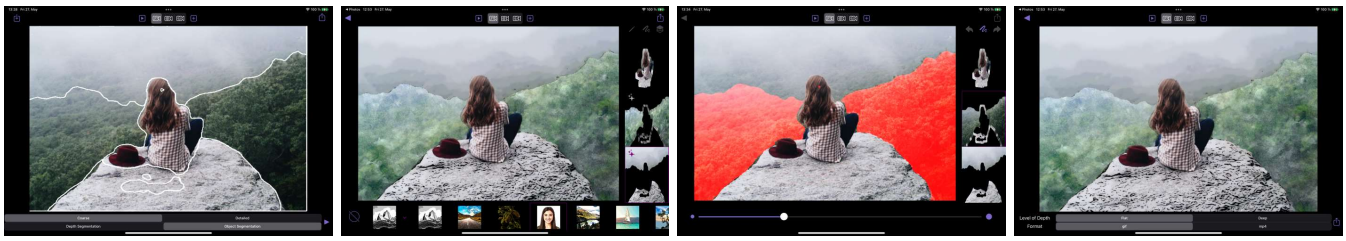
b) Inpainting: Subsequent to the layer separation, inpainting is performed for each layer L_i with $i = 0, \ldots, b-2$, i.e., the foreground layer remains unaffected. Inpainting is required to hallucinate color information that was occluded during acquisition and becomes visible during 3D photo rendering. For each layer L_i, a coverage value c at each pixel position (x, y) is determined based on the depth values in D_Q:

$$c(x,y) := \begin{cases} 1 & D_Q(x,y) > i \\ 0 & \text{otherwise} \end{cases}$$

With the coverage data c (c.f. Fig. 7b), we make use of Bilateral Filter for inpainting similar to Shekhar *et al.* [13]. The filter is applied on a per-layer basis for the image regions that are qualified with $c(x,y) = 1$ and within a specified distance to the visible pixels. For bilateral filter parameters, we use $r = 4$, $\sigma_s = 5$ for spatial as well as $\sigma_r = 12$ for the range. The above is an efficient approach which gives visually plausible output. However, as part of future work it can be improved using learning-based techniques.

(a) Original image (b) Depth map (c) Semantic segmentation map (d) Per-Layer Stylization. (e) Per-Segment Stylization

Fig. 8: For an input image (a) depth-map (b) is used to break image into multiple layers. For a spatially consistent stylization we perform semantic segmentation (c). The stylization can be performed using a per depth-based layer approach (d) or a per semantic-segment approach (e). Note how semantic segmentation reduces spatial inconsistencies due to stylization.



(a) Segmentation Screen (b) Stylization Screen (c) Touch-Up Screen (d) Export Screen

Fig. 9: Overview of the main user interface screens provided by our prototypical application for a stylization session. Starting with the segmentation screen (a), a user can load or capture an image and define the basic properties of the 3D photo, e.g., the segmentation method and desired number of segment-layers. The stylization screen enables the user to select and adjust stylization effects (b). If a user is not satisfied with layer borders, it can be adjusted by directly drawing on the image (c). After finishing the editing process, the export screen provides result preview and allow the user to export different file formats (d).

C. Per-Segment Stylization

The layers generated based on depth do not respect the image semantics, (Fig. 8). To obtain a consistent output, semantically similar regions should be stylized in a similar fashion. Thus, we divide the input image into segments based on a semantic segmentation model. We integrate different artistic rendering effects for stylizing these segments [16]. Specific to *Trios*, these comprise variants of Cartoon [17], Watercolor [18], Oil paint [19], and Hatching [20]. In case, integrated stylization techniques require depth data, it is facilitated via our framework. Sec. VI-A describes the possibilities of the per-layer stylization approach.

D. Rendering for Preview and Export

This stage synthesizes a 3D photo animation for preview and exports based on the (stylized) LDI layer and camera settings. For rendering a single frame of 3D photo animation, we implement the approach given by Shade *et al.* [6]. To achieve interactive performance, we make use of GPU-aligned implementation based on custom data structures for image storage and representation. Our framework allows setting the number of LDI layers generated during export to achieve varying intensities of parallax.

V. User Interface

We prototypically implement the proposed framework based on iOS and iPadOS. The code is based on Swift, UIKit, CoreImage, CoreML, and Metal APIs. However, the implementation methodology is not device-specific and can also be extended for other high-end mobile devices. Our prototypical app reflects our method's structure by offering a dedicated screen for each step. The modular structure allows for easy transitions between these (Fig. 10). Further, the user experience is designed to accommodate editing on multiple levels-of-control [21].

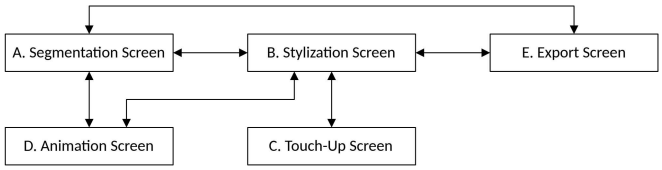


Fig. 10: Diagram showing the control flow between the individual screens of our prototypical application.

Fig. 9 shows an overview of the main views, which functionalities are briefly outlined in the remainder of this section. As a main feature for visual feedback during editing, every screen allows for selecting and playing a 3D photo animation. The animation is described via a virtual camera path wherein a user can select from a number of pre-defined paths.

A. Segmentation Screen

Fig. 9a shows an example of the segmentation screen, the first screen provided after on-boarding. It allows the user to

load a RGB(-D) image or acquire one using built-in camera functionality. The screen allows for previewing the synthesized 3D photo using pre-defined or custom animation modes (Sec. V-D). Further, the user can choose between a "coarse" ($b = 3$) or "detailed" ($b = 5$) LDI representation. In addition to depth-based layers (IV-C) a semantic-segmentation-map is also created here. To provide visual feedback to the user, the boundaries of the respective layers are depicted using white lines. This separation is only used for the stylization phase.

At this point, a user can already choose to export the 3D photo animation (Sec. V-E) which represents the standard functionality of existing 3D photo apps.

B. Stylization Screen

For design reasons, we assume that the target audience is familiar with raster-image editing apps and therefore decide to re-use Graphical User Interface (GUI) concepts from common image-editing applications [22]. It offers different levels-of-control, ranging from choosing stylization presets (high-level) to adjusting individual parameters (low-level) [21]. An icon indicates, if a layer has a stylization technique applied.

For it, the stylization screen (Fig. 9b) presents the individual layers on the right screen side using an ordered list, descending from background (top) to foreground (bottom). Upon layer selection, the user can choose from various artistic effect presets displayed below the preview image. Thus, a user can rapidly switch between stylization variants and control the overall results in an art-directed manner. Additionally, the stylization screen offers access to further low-level layer-management operations, such as parameter control (Fig. 11a), touch-up for the layer mask (Fig. 11b, Sec. V-C), as well as merging selected layers. Specific parameter controls can be used to make adjustments to the selected preset. The touch-up button opens a screen described in the next paragraph.

(a) (b) (c)

Fig. 11: Layer controls.

C. Touch-Up Screen

Fig. 9c shows the screen that offers layer-based touch-up functionality. If the user is not satisfied with the generated segments, the borders can be corrected by simply drawing on the image. The respective pixels are added to the mask of the selected segment. The current segment is highlighted with a red overlay. The drawn path is displayed directly and added to the path after the touch is finished. The radius of the brush can be adjusted with the slider at the bottom.

D. Camera Controls & Edit Screen

In order to enable easy exploration of preliminary editing results, our GUI offers control over the virtual camera in every screen, located above the 3D photo preview. It enables a user to play a camera animation and select from three different predefined animation modes as follows. The pre-selected camera animation mode (Fig. 12b) interpolates between a fixed number of virtual camera positions and orientations.

(a) (b) (c) (d) (e)

Fig. 12: Camera controls.

Further, for fast exploration of the result and detection of possible artifacts due to discontinuities or disoccluded areas, the GUI enables the traversal along a spiral path (Fig. 12c) – usual for 3D photo exploration. Finally, the position and orientation of the virtual camera can be directly controlled by the device gyroscope or accelerometer data (Fig. 12c). To allow for additional control over the virtual camera, a user can specify custom camera animations by switching to the camera edit screen (Fig. 12e). Here, the user can specify their own camera animation path by replacing the predefined one. For it, camera settings are generated by storing positions tapped on the image view. Subsequently, the camera settings can be manipulated with respect to the look-to vector, zoom level, traversal timing, and interpolation functions.

E. 3D Photo Export Screen

Finally, the user can export the 3D photo stylization results. For this, *Trios* offers a dedicated screen (Fig. 9d) for settings regarding the "level-of-depth" as well as the output file "format". The number of layers reflects the perceived intensity of the resulting parallax effect. We choose $b = 3$ for a "flat" and $b = 7$ for the "deep" option. This way, the user does not need to edit all layer that contribute to the parallax effect during final 3D photo rendering and can focus on the major composition elements (e.g., background, foreground, etc.). With respect to the export-file format, our prototype currently supports videos and as future work we would also like to include animated images.

VI. Results & Discussion

This section evaluates our approach regarding runtime performance (Sec. VI-B) and discusses limitations (Sec. VI-D) by means of different application examples (Sec. VI-A).

A. Application Examples

Fig. 13 shows exemplary results generated using our framework and a prototypical mobile application. In average, the users required $1\,\mathrm{min}$ to $3\,\mathrm{min}$ for stylization. The per-segment stylization approach offers a high degree of flexibility. For example, all segments can be stylized with the same stylization technique but using a single or multiple different presets. Usually, users tends to apply more aggressive stylization on background segments and maintain high detail in the foreground. Further, users are allowed to use different stylization techniques per-segment or do not even apply any stylization to a particular segment.

B. Performance Evaluation

a) System & Setup: We test the performance of *Trios* using the following setup. Tests on mobile were executed using an iPad Pro 3^{rd} generation equipped with an Apple A12X Bionic and $4\,\mathrm{GB}$ Random Access Memory (RAM). With respect to the test data, we perform runtime analysis using

(a) *Cartoon stylization* with more abstraction towards the background.

(b) *Watercolor stylization* using different accent colors per layer.

(c) *Pencil-hatching stylization* applied on each layer except the main focus.

Fig. 13: Images stylized by *Trios* using different presets and stylization parameter configurations on a per-layer basis.

TABLE I: Runtime performance of the individual stages in our framework different input image resolutions.

Input Image Resolution	Pre-processing	#Layers	Segmentation	Stylization per layer	Rendering	Overall
HD (1280 × 720 pixels)	0.15 s	3	0.14 s	0.4 s (0.15 s × 3)	0.11 s	0.85 s
		5	0.31 s	0.75 s (0.15 s × 5)	0.15 s	1.21 s
FHD (1920 × 1080 pixels)	0.17 s	3	0.15 s	0.57 s (0.19 s × 3)	0.16 s	1.05 s
		5	0.32 s	0.95 s (0.19 s × 5)	0.24 s	1.68 s
QHD (2560 × 1440 pixels)	0.19 s	3	0.15 s	0.63 s (0.21 s × 3)	0.17 s	1.14 s
		5	0.33 s	1.05 s (0.21 s × 5)	0.27 s	1.84 s

images of three different resolutions: High Definition (HD) (1280×720 pixels), Full High Definition (FHD) (1920×1080 pixels), and Quad High Definition (QHD) (2560×1440 pixels).

b) Run-time Performance Results: Tab. I shows the runtime performance results for each pipeline stage with increasing image resolutions. We record the processing time for the steps of segmentation, stylization of all layers, and the final rendering. One can observe, that the runtime performance of each step scales with the image resolution and the number of layers to stylize. The type of effect, selected for stylization, has negligible impact on the overall performance. During the export, the rendered 3D photo is displayed on the screen and is simultaneously written to the memory. Thus, saving the 3D photo takes approximately as long as the final result visualization. Note, that for depth estimation and/or object segmentation we use trained neural-network models. These models are only loaded once and have an initial loading overhead of approx. 5 s.

c) Memory Consumption: The prototypical app itself has a storage size of 1.8 GB on the iPad. The memory consumption of our prototype scales linearly with the resolution of the used image. For an image of spatial resolution of 1920×1080 pixels, the memory usage is approx. 35 MB without stylization and approx. 135 MB with stylization applied. The final 3D rendering step increases the memory usages to 275 MB. The exported 3D file itself, e.g., M4V, of 5 s has a size of 7 MB. Thus, the application has a reasonable memory footprint.

C. Usability Evaluation

The prototype was presented at an international conference on computer graphics using the same setup described in Sec. VI-B. We gave a brief introduction to *Trios* to approximately 40 people, who then choose example images or took photos and edited these accordingly. A user spend on average 2 min to 5 min to create a stylized 3D photo. Most users were familiar with the general concepts of 3D photos and stylization, thus immediately understood the concept of layer-wise combination of both. The working modes were well understood, however some functionalities had to be pointed out repeatedly, e.g., multi-selection and layer merging.

The device motion was mostly used to view the parallax effect since it had the most immersive effect for the users. When handed the device, users often directly tried to move the photos using device rotation. However, the responding transformation of the 3D photo was often found to be slightly contra-intuitive or not always reliable.

Of particular interest to most users was switching between different stylization presets rather than using the fine-tuning option. Generally, mostly two different stylizations were chosen – background and foreground. Regarding the type of stylization, either strongly varying styles were chosen for more contrast or similar stylization techniques with different level-of-abstraction to increase depth sensation. Overall the user's feedback was positive. Especially people from non-technical background were excited to have on-device 3D stylization. However, as per the feedback, the experience can be further improved by more reliable device motion and better layer separation.

D. Discussion and Future Work

Our goal is to develop a framework for interactive stylization of 3D photos on mobile devices. To this end, we deploy depth-estimation and semantic-segmentation neural-network models on-device. We observe that the optimized mobile-based models perform significantly worse than their desktop counterparts,

while still giving plausible results for our purpose. For high-quality results, better depth-quality and inpainting techniques are required. However, increased layer numbers and a sophisticated inpainting algorithm impacts interactivity.

Although, our app shows the feasibility of our framework and achieve sufficient interactive characteristics, we plan to address several aspects as future work. The presented modular framework provides the basis for straightforward integration of alternative or additional image processing operations. For example, we plan to use live-photos or videos to improve the resulting inpainting quality, e.g., using the work of [9]. With a combination of depth- and semantic-estimation, the resulting depth map can be further improved to avoid objects being split to different layers [23]. Further, depth-map upsampling can be improved using guided filtering [24] and can form the basis to implement stylized atmospheric effects [13].

VII. Conclusions

In this work, we present a framework for implementing 3D photo stylization techniques on mobile devices. Our approach is based on layered depth-images and proposes a modular concept for data acquisition, pre-processing, stylization, and rendering of 3D photos. We demonstrate and evaluate the feasibility by providing an initial implementation based on Apple consumer devices. Our integrated approaches enable users to rapidly create stylized variants of 3D photos, which can easily be shared using common interchange file formats such as animated images or videos.

Acknowledgments

This work was partially funded by the German Federal Ministry of Education and Research (BMBF) through grants 01IS18092 (“mdViPro”) and 01IS19006 (“KI-LAB-ITSE”) and the Research School on “Service-Oriented Systems Engineering” of the Hasso Plattner Institute.

References

[1] P. Hedman, S. Alsisan, R. Szeliski, and J. Kopf, “Casual 3d photography,” *ACM Trans. Graph.*, vol. 36, no. 6, nov 2017. [Online]. Available: https://doi.org/10.1145/3130800.3130828

[2] J. Kopf, K. Matzen, S. Alsisan, O. Quigley, F. Ge, Y. Chong, J. Patterson, J.-M. Frahm, S. Wu, M. Yu, P. Zhang, Z. He, P. Vajda, A. Saraf, and M. Cohen, “One shot 3d photography,” *ACM Trans. Graph.*, vol. 39, no. 4, jul 2020. [Online]. Available: https://doi.org/10.1145/3386569.3392420

[3] J. E. Kyprianidis, J. Collomosse, T. Wang, and T. Isenberg, “State of the ”art”: A taxonomy of artistic stylization techniques for images and video,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 5, pp. 866–885, 2013.

[4] S. Pasewaldt, A. Semmo, J. Döllner, and F. Schlegel, “Becasso: Artistic image processing and editing on mobile devices,” in *SIGGRAPH ASIA 2016 Mobile Graphics and Interactive Applications*, ser. SA ’16, 2016.

[5] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, “Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer,” *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020.

[6] J. Shade, S. Gortler, L.-w. He, and R. Szeliski, “Layered depth images,” in *Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques*, ser. SIGGRAPH ’98. New York, NY, USA: Association for Computing Machinery, 1998, p. 231–242. [Online]. Available: https://doi.org/10.1145/280814.280882

[7] P. Hedman and J. Kopf, “Instant 3d photography,” *ACM Trans. Graph.*, vol. 37, no. 4, jul 2018. [Online]. Available: https://doi.org/10.1145/3197517.3201384

[8] B. Mildenhall, P. P. Srinivasan, R. Ortiz-Cayon, N. K. Kalantari, R. Ramamoorthi, R. Ng, and A. Kar, “Local light field fusion: Practical view synthesis with prescriptive sampling guidelines,” *ACM Trans. Graph.*, vol. 38, no. 4, jul 2019. [Online]. Available: https://doi.org/10.1145/3306346.3322980

[9] M.-L. Shih, S.-Y. Su, J. Kopf, and J.-B. Huang, “3d photography using context-aware layered depth inpainting,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8028–8038.

[10] M. Trapp and J. Döllner, “Efficient Representation of Layered Depth Images for Real-time Volumetric Tests,” in *Theory and Practice of Computer Graphics*, I. S. Lim and W. Tang, Eds. The Eurographics Association, 2008.

[11] V. Jampani, H. Chang, K. Sargent, A. Kar, R. Tucker, M. Krainin, D. Kaeser, W. T. Freeman, D. Salesin, B. Curless, and C. Liu, “Slide: Single image 3d photography with soft layering and depth-aware inpainting,” in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 12 498–12 507.

[12] T. Dürschmid, M. Söchting, A. Semmo, M. Trapp, and J. Döllner, “Prosumerfx: Mobile design of image stylization components,” in *SIGGRAPH Asia 2017 Mobile Graphics & Interactive Applications*, ser. SA ’17. New York, NY, USA: Association for Computing Machinery, 2017. [Online]. Available: https://doi.org/10.1145/3132787.3139208

[13] S. Shekhar, M. Reimann, M. Mayer, A. Semmo, S. Pasewaldt, J. Döllner, and M. Trapp, “Interactive Photo Editing on Smartphones via Intrinsic Decomposition,” *Computer Graphics Forum*, 2021.

[14] A. R. Zamir, A. Sax, N. Cheerla, R. Suri, Z. Cao, J. Malik, and L. J. Guibas, “Robust learning through cross-task consistency,” in *CVPR*. Computer Vision Foundation / IEEE, 2020, pp. 11 194–11 203.

[15] J. Kopf, M. F. Cohen, D. Lischinski, and M. Uyttendaele, “Joint bilateral upsampling,” *ACM Trans. Graph.*, vol. 26, no. 3, p. 96–es, jul 2007. [Online]. Available: https://doi.org/10.1145/1276377.1276497

[16] A. Semmo, M. Trapp, J. Döllner, and M. Klingbeil, “Pictory: Combining neural style transfer and image filtering,” in *ACM SIGGRAPH 2017 Appy Hour*, ser. SIGGRAPH ’17. New York, NY, USA: Association for Computing Machinery, 2017. [Online]. Available: https://doi.org/10.1145/3098900.3098906

[17] H. Winnemöller, S. C. Olsen, and B. Gooch, “Real-time video abstraction,” *ACM Trans. Graph.*, vol. 25, no. 3, p. 1221–1226, jul 2006. [Online]. Available: https://doi.org/10.1145/1141911.1142018

[18] A. Bousseau, M. Kaplan, J. Thollot, and F. X. Sillion, “Interactive watercolor rendering with temporal coherence and abstraction,” in *Proceedings of the 4th International Symposium on Non-Photorealistic Animation and Rendering*, ser. NPAR ’06. New York, NY, USA: Association for Computing Machinery, 2006, p. 141–149. [Online]. Available: https://doi.org/10.1145/1124728.1124751

[19] A. Semmo, J. Döllner, and F. Schlegel, “Becasso: Image stylization by interactive oil paint filtering on mobile devices,” in *Proceedings SIGGRAPH Appy Hour*. New York: ACM, 7 2016, pp. 6:1–6:1.

[20] A. Semmo and S. Pasewaldt, “Graphite: Interactive photo-to-drawing stylization on mobile devices,” in *ACM SIGGRAPH 2020 Appy Hour*, ser. SIGGRAPH ’20. New York, NY, USA: Association for Computing Machinery, 2020. [Online]. Available: https://doi.org/10.1145/3388529.3407306

[21] T. Isenberg, “Interactive NPAR: What Type of Tools Should We Create?” in *Proc. NPAR*, ser. Expressive ’16. Goslar, DEU: Eurographics Association, 2016, p. 89–96.

[22] M. Klingbeil, S. Pasewaldt, A. Semmo, and J. Döllner, “Challenges in user experience design of image filtering apps,” in *SIGGRAPH Asia 2017 Mobile Graphics & Interactive Applications, Bangkok, Thailand, November 27 - 30, 2017*, M. Billinghurst and W. Rungjiratananon, Eds. ACM, 2017, pp. 22:1–22:6. [Online]. Available: https://doi.org/10.1145/3132787.3132803

[23] S. Niklaus, L. Mai, J. Yang, and F. Liu, “3d ken burns effect from a single image,” *ACM Trans. Graph.*, vol. 38, no. 6, nov 2019. [Online]. Available: https://doi.org/10.1145/3355089.3356528

[24] K.-L. Hua, K.-H. Lo, and Y.-C. F. Frank Wang, “Extended guided filtering for depth map upsampling,” *IEEE MultiMedia*, vol. 23, no. 2, pp. 72–83, 2016.

Building Portuguese Sign Language datasets for computational learning purposes

Carlos Mayea and Dibet García González
Center for Computer Graphics
University of Minho
Guimarães, Portugal
carlos.mayea@ccg.pt, dibet.gonzalez@ccg.pt

Miguel Guevara
Setúbal School of Technology
Polytechnic Institute of Setubal
Setúbal, Portugal
Centro ALGORITMI, University of Minho
Guimarães, Portugal
miguel.lopez@estsetubal.ips.pt

Emanuel Peres
Eng. Dept, School of Science and Technology
University of Trás-os-Montes e Alto Douro
Vila Real, Portugal
eperes@utad.pt

Luís Magalhães
Centro ALGORITMI, University of Minho
Guimarães, Portugal
lmagalhaes@dsi.uminho.pt

Telmo Adão
Center for Computer Graphics
Centro ALGORITMI, University of Minho
Guimarães, Portugal
telmo.adao@ccg.pt

***Abstract*— Communication consists of a set of linguistic signs and rules that establish a common base for understanding between two or more interlocutors. However, many are the conditions that can affect this process (cultural, idiomatic, environmental, etc.), being one of the most challenging associated to the interaction between deaf and hearing persons. In such regard, this work presents an approach proposal to establish a dataset for the Portuguese Sign Language (PSL), envisaging the application of computational learning methodologies capable of recognizing communication gestures, not only to complement human PSL interpreters as major players to enable the bidirectional communication between deaf and hearing persons, but also to increase the autonomy and confidence of those who have congenital or acquired auditory perception deficiency. Data augmentation techniques were applied to increase the number of available samples, as well as variability. A set of preliminary tests was carried out relying in long short-term memory (LSTM), from which 86% of accuracy was reached.**

***Keywords*— Portuguese Sign Language (PSL), Sign Language Recognition (SLR), Sign Dataset, Deep Learning, Long Short-Term Memory, Artificial Intelligence (AI).**

I. INTRODUCTION

Inclusion aims to ensure that all persons – regardless of background, gender, culture, age, condition, etc. - are able to fully participate in all aspects of life, from social to professional contexts. However, there are many challenges that still need to be addressed, being one of them the development of effective strategies/approaches/tools to shorten the communication barriers between deaf and hearing communities.

The Portuguese deaf community performs communication through the Portuguese Sign Language (PSL), with origins in Casa Pia, Lisbon, 1823 [1]. It combines gestures involving hand configurations and movements, facial expressions and even body posture and its interpretation occurs through observation. Considering the extremely visual characteristics of PSL, deep learning (DL) stands out as a pertinent, modern, popular, and effective umbrella of algorithms to approach gesture classification. The sequences of signs involved in such communication are mappable through recurrent neural networks, and more specifically, LSTM architectures, due to their ability for dealing with action-oriented flow notion imagery. Having that said, it is necessary to meet the specificities of PSL - e.g., the order of the words in grammar construction, context-dependent gesture nuances, etc. -, to establish representative datasets of the problem under analysis, on which the performance of the LSTM network depends. To this extent, this paper proposes a preliminary approach for the translation of sign language into text, focusing on the collection and organization of datasets, as a support to design DL models.

Regarding this paper organization, section II presents a short literature review, followed by section III, wherein a preliminary set of considerations regarding PSL and some of its rules is presented, as a result from meetings that were carried out with members of the official Portuguese Association of the Deaf (Associação Portuguesa de Surdos, APS). Next, in section IV, a proposal of a general DL-based architecture to convert image streaming of PSL messaging into text will be presented. Dataset gathering and organization strategy will be detailed in section V, right before the preliminary experiments and tests done with two LSTM architectures (section VI). Conclusions and future work are provided in the end this paper, i.e., section VII.

II. PREVIOUS WORKS

Sign language translation has been widely addressed in the literature review, based on LSTMs. For example, in [2] a set of 26 signs were collected through a sensor glove and used to train an LSTM. In [3], an hierarchical LSTM is proposed to predict gestures based on images focusing the region of interest (ROI) associated to hands. Based not only in the ROIs of hands but also in the relative position of them together with the face, a spatial-temporal LSTM (ST-LSTM) information fusion technology was proposed in [4]. 3-D convolution in combination with long-short-term-memory (LSTM) networks was a strategy followed in [5], which learns gesture patterns in two main stages related to short- and long-term spatiotemporal features.

Besides the proposal of artificial learning architectures, several datasets have been created around the world to train and evaluate other Sign Language Recognition (SLR) systems, which can be classified into the following groups: a) single word datasets, where each video corresponds to a word (gesture); and b) multiple words, where each video represents a sentence communicated through a combination of gestures (words). While INCLUDE [6], WLASL [7], LSA64 [8], RWTH-PHOENIX-Weather2014T [9] How2Sign [10] and LIBRAS-UFOP [11] are among the datasets usable for SLR that can be found in literature, Table 1 sums up their

978-8-3503-4573-57/22/$31.00

characterization in terms of targeting language, involved number of contributors and technologies/features, as well as supported number of words/sentences. These datasets consider conditions variability (e.g., users and illumination) and have associated text files with metadata providing information such as gesture classes, bounding boxes, segmentation masks for hands, skeleton pose coordinates, among others. Inspired by the previously addressed contributions for sign language and considering the scarcity of structured gestures data concerning the Portuguese context, a methodology for PSL dataset construction was defined, applied, and preliminarily tested, as it will be shown in the remainder of this paper.

Table 1 – Summary of available sign language dataset

Datasets	Vocabulary			#C	T&E
	#W	#S	Language		
INCLUDE	263	-	Indian	7	RGB videos
WLASL	2000	-	English US	100	RGB videos
LSA64	64	-	Argentinian	10	RGB videos, Hands Gloves, dark clothes
RWTH-PHOENIX-Weather2014T	-	+1980	German	7	RGB videos, Dark clothes, artificial grey background
How2Sign	-	16k (~17 words per sentence)	English US	11	RGB videos, 3D key-points, green background
LIBRAS-UFOP	56	-	Brazilian	5	RGB videos, pose estimation, depth

Legend: #W – number of words; #S – number of sentences; #p – number of contributors; T&E – technologies and environment.

III. PSL CONSIDERATIONS

To develop a PSL dataset, a couple of major actions was taken: 1) the observation and study of the multilingual online dictionary "SpreadTheSign" [12] source, which includes the Portuguese signs set; and 2) presential meetings with APS. While the former was used for a preliminary analysis on how to perform certain Portuguese signs to produce a set of words and sentences, the latter was resorted for corroboration purposes and, also, to gather additional and more accurate information regarding PSL. In the specific case of the meetings carried out with APS, sessions were conducted based in a document that was created having in mind simple and direct quotidian dialogs, considering the following contexts: bakery, butcher, fruit shop. The following general questions were posed to APS: "How to formulate affirmations sentences?"; "How to formulate questions?"; "How to use verb tenses (simple present, simple past, future)?"; "What is the most common syntactic structure?"; "How to distinguish gender?"; and "Do the gestures vary according to context?" According to APS, the PSL has a grammatical structure formed, in a general way, by three elements: Subject (S), Verb (V), and Object (O). These grammatical elements are organized to form sentences and the most used form is O-S-V. This communication aspect is also corroborated in other literary sources [13]. Regarding verb tenses conjugated with day, week, month year, signs following a logic that promotes intuitions are defined. For example, to express "next/last year" in PSL, indicator fingers mimic the translational movement of earth around sun. The movements are executed in different (opposite) directions, depending on the desired time indication. "Next year" is expressed by flowing the indicator in a certain rotation sense, while "last year" is precisely the other way around. The gender distinction is not often made, being masculine the dominant one. However, there is a sign that can be done in the beginning of the sentence to refer feminine gender, which consists in sliding the stretched indicator from the corner of an eyebrow, until the bottom of the cheek, respecting a vertical trajectory. In what concerns to the context, it can indeed involve variable signs for different situations, as it is the case of the verb "open", which has specific nuances for "open bottle", "open door", "open box", "open book", etc. Such variations were also defined to be intuitively relatable with the respective objects in focus.

Next section will address an architecture proposal for PSL-to-text translation.

IV. PSL-TO-TEXT OVERALL ARCHITECTURE PROPOSAL

After earning some awareness regarding PSL requirements, an overall architecture was proposed (Figure 1), envisaging the participation of deaf and hearing persons, as well as sign language interpreter. It is composed by the following main components and modules: a portable smart device with camera, processing, rendering and connectivity capabilities, communication layers that interact with an environment responsible for converting PSL into text, and/arms/face landmarks pre-processor, an LSTM for estimating words out of the PSL, a tokenizer and outlier filter for sanitizing LSTM's output, an OVS to SVO converter to prepare the loose terms for being processed by a Natural Language Processing (NLP) module that finally produces a meaningful sentence. Incorrect outputs can be reported by the deaf persons and later analysed by a PSL interpreter, who is responsible for improving the dictionary of gestures/signs of the system.

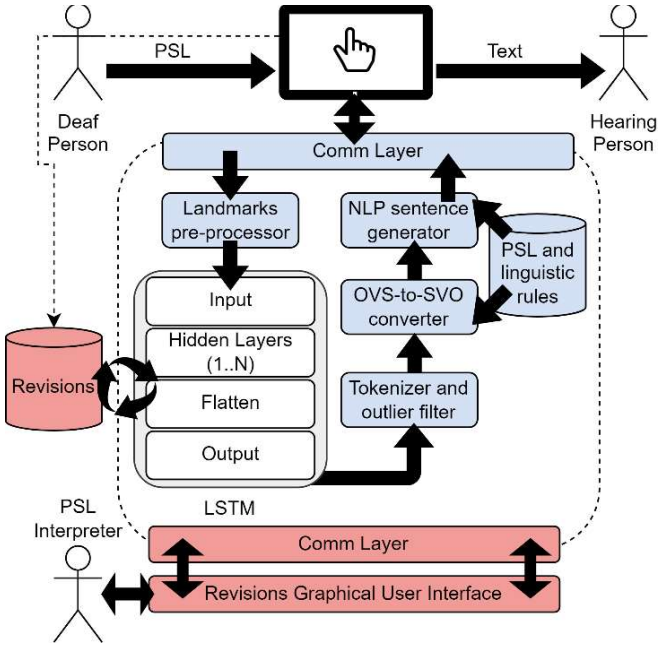


Figure 1 - General architecture proposal for a PSL-to-text system. A deaf person inputs PSL to a mobile device that communicates with an intelligent environment composed of an LSTM and sentence building components, which are responsible for translating the gesture-based message into text, delivered afterwards to the hearing user. Also, a PSL interpreter contributes for fine-tuning the model continuously, through a human-in-the-loop strategy.

V. Dataset construction

Given the existence of this grammatical structure and rules associated to PSL, an approach for building a dataset formed by independent videos of gestures/signs was specified, in which each video represents a word or short expression in PSL. Each word in the dataset will be classified grammatically as a verb, adjective, adverb, pronoun, etc. This way, the mapping of terms sorted by O-S-V structure will allow the application of NLP for setting up of entire sentence.

Encompassing quotidian handiness, the set of signs specified for being collected are listed as follows:

- Verbs: “open”, “give”, “help”, “want”, “talk”, “arrive”, “pay”, “have”;
- Adverbs: “more”, “yes”, “no”, “tomorrow”, “when”;
- Objects: “money”, “cake”, “week”, “pineapple”, “chicken”, “bread”, “wing”, “card”;
- Subjects: “I”, “you”;
- Adjectives: “expansive”, “cheap”;
- Pronouns: “something”;
- Short expressions: “Good afternoon”, “Good morning”, “How much is it?”, “Thank you!”, “How?”, “Help me.”, “See you later.”, “Hello!”.

It is necessary to clarify that in PSL there are short sentences/expressions that are represented by a simple gesture such as "something (*alguma coisa*)" or "see you later (*até logo*)". Other gestures were selected based in similarity criteria regarding the configuration of arms/hands trajectory, being differentiated simply by the facial expression, as it is the case of words like "money (*dinheiro*)" and expressions such as "how much is it (*quanto custa*)?". As previously mentioned in section III, words were chosen considering quotidian contexts such as bakery, butcher, fruit shop.

A. Video collection

The videos are recorded with a resolution of 1028x770 at 20fps in an office environment, reflecting the conditions of everyday life. Background was not modified. Sensor gloves were not used. Also, there was no restrictions regarding colour-specific clothing.

For the recording of the videos, there was the collaboration of 8 different participants, only 1 of them was an expert in sign language. Each person repeated each defined gesture 5 times for a total of 1360 videos recorded (34 words * 8 signers * 5 repetitions/signer). To address the video collection, a Python application that allows to verify the presence of the skeleton in each frame was developed. The coordinates of this skeleton were extracted resorting to MediaPipe algorithms [14] and then used to train LSTM-based neural networks that allow the recognition of sign language gestures.

B. Augmentation techniques

The amount of data is a recurrent issue that must be considered when training supervised machine/deep learning models. As such, data augmentation techniques were applied upon the gathered dataset, i.e., collected videos, to increase the number of samples and variability.

To implement these techniques, the characteristics of each sign language should be considered, within the PSL context. When it comes to gesturing, there is a dominant hand and a non-dominant hand, which provide movement information that may or may not combined with facial expression or body posture. It is quite common to find individuals whose the dominant hand is right one, but it is also possible to find who use the opposite configuration, with the left hand as the dominant. Therefore, besides techniques that change general light parameters of the image, horizontal mirroring operations were also applied, considering dominant hand factor.

More specifically, a set of transformations were set up to induce colour, brightness, contrast, saturation, and blur variations, but without affecting the image geometry. In addition, we also applied Y-axis flip. These augmentation techniques were applied, not only to increase dataset and variability, but also to balance the different classes, resulting in the same number of samples per word.

VI. Preliminiary implementation and experiments

For the initial implementation of an SLR decoder prototype, a few approaches presented in the literature were considered, varying both in the feature extraction techniques and in the algorithms used to infer gesture-to-sentences in a video sequence. A continuous Indian Sign Language Recognition system is proposed in [15]. The authors use a Leap Motion Sensor that provides a feature vector with the coordinates of both hands. These features feed a Conv-LSTM network that analyses the inputs as a sequence of time where the time component means passing the features corresponding to a sequence of frames that allows capturing the motion by taking advantage of the memory capacity of Recurrent Neural Networks. Other examples are presented in [16], wherein a comparison of methods for human activity recognition is carried out and applied to American Sign Language Recognition. As a result, the authors obtained higher accuracy in systems involving neural networks mixture of convolutional layers with LSTM layers.

Supported by previous research, the decision was to create models based on LSTM architectures since they allow, not only to perform simple classifications, but also to maintain sequence awareness, which match the dynamics of gesturing for deaf people.

Two LSTM-based architectures were adopted for performing preliminary tests with the collected datasets:

- a network composed only of LSTM layers with a Dense layer in the output (Figure 2), having as input a feature matrix in which each column represents a coordinate vector of the hands, face and pose obtained with MediaPipe and the length of the matrix represents the frame analysed at time T;
- a network integrating convolutional layers followed by LSTM layers with a Dense layer in the output (Figure 3), which expects as input a sample of 40 frames, selected per video and pre-handled with image processing algorithms for background removal.

The models were built using the Tensorflow/Keras API, which allows the creation of sequential models. The training was performed on a computer with 16 GB RAM and NVIDIA GeForce RTX 2070 graphics card. For each training session, 200 epochs were defined with initial learning rate of 0.0001 using ADAM optimizer for both model architectures. The dataset was split into 70% for training, 15% for validation and 15% for final testing. No augmentation technique was applied to the test set of videos.

The results show that models relying in the face, pose and hand coordinates seem to be more sensitive to changes in lighting conditions (synthetically induced in the image), while those with convolutional layers at the input, followed by LSTM layers managed to learn better. Table 2 shows higher accuracy for models combining Convolutional and LSTM layers in the augmented dataset group.

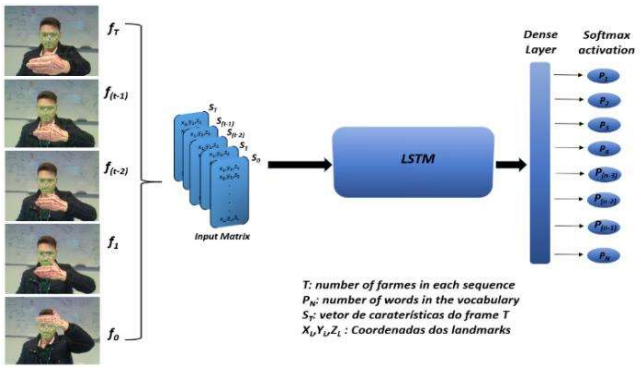


Figure 2 - Architecture with simple LSTM layers.

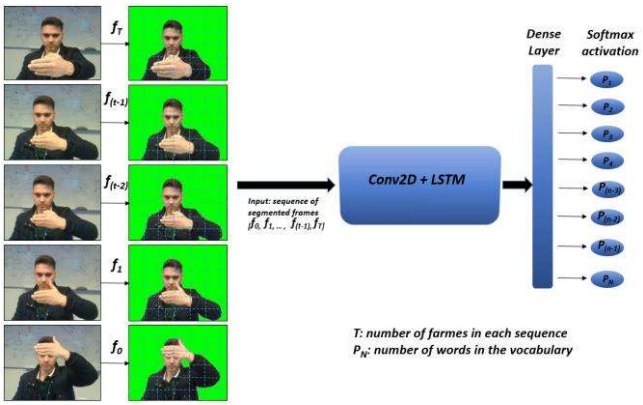


Figure 3 - Architecture with ConvLSTM2D Layers.

Table 2 - Training results.

Model architecture	Sign word recognition accuracy	
	No augmented data	Augmented data
LSTM	66%	79%
Conv + LSTM	74%	86%

CONCLUSIONS

In this paper, a simple strategy for dataset creation with the objective of training or testing Portuguese SLR systems was presented. The strategy is based on previous experience investigated in the literature. Only simple videos representing a gesture corresponding to a word or phrase in PSL were tested in the experiments; longer videos with more complex phrases were not evaluated.

The results show that the construction of models combining convolutional layers with LSTM allows to obtain results with a promising accuracy, especially with the application of some video augmentation techniques.

Future work will encompass the following challenges: 1) the extension of the current PSL dictionary, with more participants performing a wider range of signs; 2) a benchmark with more LSTM/CNN approaches, seeking the most suitable model in terms of accuracy; 3) the completion of the pipeline shown in Figure 1, namely, the integration of natural language processing techniques for shaping sentences out of the extracted sign-based tokens and the implementation of active learning approaches to provide PSL system with capabilities for improving the recognition of known signs, as well as to support the learning of new ones.

ACKNOWLEDGMENTS

This work was financed by the project "IVLinG - Interprete Virtual de Língua Gestual" (N° POCI-01-0247-FEDER-068605), financed by Portugal 2020, under the Competitiveness and Internationalization Operational Program (POCI), and by the European Regional Development Fund (ERDF).

REFERENCES

[1] P. Escudeiro et al., "Virtual Sign - A Real Time Bidirectional Translator of Portuguese Sign Language," in Procedia Computer Science, 2015, vol. 67, pp. 252–262. doi: 10.1016/j.procs.2015.09.269.

[2] E. Abraham, A. Nayak, and A. Iqbal, "Real-Time Translation of Indian Sign Language using LSTM," 2019. doi: 10.1109/GCAT47503.2019.8978343.

[3] D. Guo, W. Zhou, H. Li, and M. Wang, "Hierarchical LSTM for sign language translation," 2018. doi: 10.1609/aaai.v32i1.12235.

[4] Q. Xiao, X. Chang, X. Zhang, and X. Liu, "Multi-Information Spatial-Temporal LSTM Fusion Continuous Sign Language Neural Machine Translation," IEEE Access, vol. 8, 2020, doi: 10.1109/ACCESS.2020.3039539.

[5] G. Zhu, L. Zhang, P. Shen, and J. Song, "Multimodal Gesture Recognition Using 3-D Convolution and Convolutional LSTM," IEEE Access, vol. 5, 2017, doi: 10.1109/ACCESS.2017.2684186.

[6] A. Sridhar, R. G. Ganesan, P. Kumar, and M. Khapra, "INCLUDE: A Large Scale Dataset for Indian Sign Language Recognition," in MM 2020 - Proceedings of the 28th ACM International Conference on Multimedia, Oct. 2020, pp. 1366–1375. doi: 10.1145/3394171.3413528.

[7] D. Li, C. R. Opazo, X. Yu, and H. Li, "Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison," 2020. doi: 10.1109/WACV45572.2020.9093512.

[8] F. Ronchetti, F. Quiroga, and L. Lanzarini, "LSA64 : An Argentinian Sign Language Dataset," Congreso Argentino de Ciencias de la Computacion (CACIC), 2016.

[9] N. C. Camgoz, S. Hadfield, O. Koller, H. Ney, and R. Bowden, "Neural Sign Language Translation," 2018. doi: 10.1109/CVPR.2018.00812.

[10] A. Duarte et al., "How2Sign: A Large-scale Multimodal Dataset for Continuous American Sign Language," 2021. doi: 10.1109/CVPR46437.2021.00276.

[11] L. R. Cerna, E. E. Cardenas, D. G. Miranda, D. Menotti, and G. Camara-Chavez, "A multimodal LIBRAS-UFOP Brazilian sign language dataset of minimal pairs using a microsoft Kinect sensor," Expert Syst Appl, vol. 167, 2021, doi: 10.1016/j.eswa.2020.114179.

[12] "Dicionário de língua gestual | SpreadTheSign." https://www.spreadthesign.com/pt.pt/search/?cls=1 (accessed Jul. 26, 2022).

[13] Fernanda, "Maria Fernanda da Silva Bettencourt A ordem de palavras na Língua Gestual Portuguesa," 2015.

[14] "Home - mediapipe." https://google.github.io/mediapipe/ (accessed Jul. 26, 2022).

[15] A. Mittal, P. Kumar, P. P. Roy, R. Balasubramanian, and B. B. Chaudhuri, "A Modified LSTM Model for Continuous Sign Language Recognition Using Leap Motion," IEEE Sens J, vol. 19, no. 16, 2019, doi: 10.1109/JSEN.2019.2909837.

[16] V. Hernandez, T. Suzuki, and G. Venture, "Convolutional and recurrent neural network for human activity recognition: Application on American sign language," PLoS One, vol. 15, no. 2, 2020, doi: 10.1371/journal.pone.0228869.

Dance and Machine Learning: A study in human-pose detection to generate new visual approaches

Maria Rita Nogueira
Institute of Systems and Robotics
University of Coimbra
Coimbra, Portugal
0000-0001-8783-3619

Paulo Menezes
Dept. of Electrical and Computer Engineering
University of Coimbra
Coimbra, Portugal
0000-0002-4903-3554

José Maçãs de Carvalho
Centre for Interdisciplinary Studies
University of Coimbra
Coimbra, Portugal
0000-0002-2569-303X

***Abstract*—Human pose detection is a research field that has provided different works on the analysis of body movement. Contemporary dance is a performing art that integrates beautiful human poses resulting from simple and complex movements. In this study, we present a new digital art concept that results from the intersection between dance movement and machine learning technique, in particular, human pose detection. This intersection provides the creation of visual results which enrich the dance performance, in real-time. For this purpose, a framework has been developed to provide new different visual forms through the human pose detection from dancers. The main objective of this intersection is to enrich the perception of dance movement, through abstract or literal visual models that combine geometric and curvilinear forms. The present work culminates with a real case study presented in a contemporary dance performance on the stage. The performance was created with professional young dancers that explored their movement, through the framework developed and enriched the experience of the public.**

***Index Terms*—dance, digital art, machine learning, human-pose detection, human computer interaction, creativity.**

I. Introduction

Dance is an art that connects and communicates with its audience through body language and body movement [1]. Computational art has contributed to new results of interpretation, analysis, and understanding of art concepts, namely in dance. Since the mid-1960s, the interest in the area of dance movement, and computing technology has united professionals from different research fields [2]. The integration of computer machines to generate dance sequences was firstly researched in 1964, by Jeanne Beaman and Paul Le Vasseur [3]–[5]. After this work, different developments emerged in movement notation [6], [7] and later other works presented innovations for stage performances, like "SWAP" by Rudolfo Quintas [8] or "Co:Lateral" by Né Barros and João Martinho Moura [9]. Through the evolution of computer systems, such as Artificial Intelligence (AI) we can now research developments that apply techniques to detect body movement and estimate the human pose [11]–[14]. In the last two decades, human pose estimation has become a topic of much interest, especially after the development of Convolutional neural networks (CNN's) and the integration of human pose datasets for benchmark evaluation [13], [15]. CNN's have been implemented to the challenge of human pose estimation and applied to directly regress the 2D cartesian coordinates of body joints in a holistic manner [16], [17]. Through heat maps, the system has an intermediate representation for all joints to refine the joint positions in 2D pose estimation [18], and to use depth images to regress joint positions for 3D pose estimation [19]–[21].

Contemporary dance is a dance technique non-bodily formalistic, non-symbolic, or being representational in some sense [22]. In terms of the focus of its technique, contemporary dance tends to combine strong but controlled legwork with upper-body contraction, horizontal movements, contract-release, floor work, and improvisation, among other attributes which define contemporary dance [23]–[25]. This dance technique integrated with other mechanisms of body detection, such as motion capture, would imply the dancer's movement once there is more difficulty in movement execution due to the technological apparatus used. In this sense, the present article merged contemporary dance and human pose detection, only using one camera for a new visual approach that enriches the perception of the dance movement. The human pose detection is already used for artistic creation and choreography support [11], [26], [27]. Conversely, to develop new approaches between dance and machine learning mechanisms, it is crucial to improve new visual paradigms, such as visually exploring the graphical component resulting from the connection between dance and human pose detection. The visual representation that corresponds to the movement of the dancer or its skeleton is still very rigid. In real life, when the dancer performing tries to express softness, movement fluidity, a relationship with all space and with the personal space around his body(*kinesphere*) [28]. The *kinesphere* is a synonym for reach-space and how around the body in which the movements take place [29]. Space treats the spatial extent of the dancer's *kinesphere* and what form is being revealed by

This research was funded by Fundação para a Ciência e Tecnologia (FCT) through grant number 2020/09137/BD and partially supported by FCT, under the project FCT/MCTES UIDB/00048/2020.

978-8-3503-4573-57/22/$31.00

the spatial pathways of the movement [29]. In related works, the visual graphics should also express these characteristics of dance movement. In this sense, the presented work explores new visual approaches to represent dance movement, giving a new layer to the performance itself. Thus, the audience perceives in more detail the connection between human body parts during the dance performance. And at the same time, the dancers can explore and see their movement through a new visual perspective which provides progress of their movement.

II. Generating forms through Dance and ML

The present section explains the working process that led the research to the final result, and for this reason, the current section also involves the applied methodology, although the term is not used explicitly throughout the text.

A. Proposed System

The proposed system is based on the detection of the human body movement and its visual representation in the physical space. What is the body connection during the contemporary dance performance? Through the integration of ML mechanisms, namely human-pose detection, it is possible to visually answer the previous question. The proposed model integrates the OpenPose system [13]. OpenPose has represented the first real-time multi-person system to jointly detect human body keypoints (in total 135 keypoints) [13]. The OpenPose model presents an accuracy very close to that expected in dance movements. We obtain detailed movement data from each body part on the physical space and spatial coordinates which connect the movement drawn on the space. In this sense, the development algorithm integrates this model (i.e. pose-detection) to detect the human body skeleton during the performance. As shown in Figure 1, each element identified in the skeleton corresponds to each keypoint detected in real-time.

Fig. 1. Posture detection using PoseNet in real-time and keypoints number identification.

By detecting the different keypoints from the skeleton of the human body, we initially explored the possibility of these points connecting with each other, but through different visual forms. For this purpose, we studied conceptually, the representations to explore visually, and after that, we implemented the algorithmic component that allows us to present the visual graphics, associated with real-time detection.

B. Translating dance movement into visual forms

By studying the representation of the human skeleton, we started by sketching the visual representations, as shown in Figure 2, which better translated the bodily connection that exists when a body dances in space. According to the study already done in works involving dance and real-time body detection, we chose the connection between hands and feet as the most simplified visual representation. From this representation geometric and bézier visual forms were explored, as presented in figure 2.

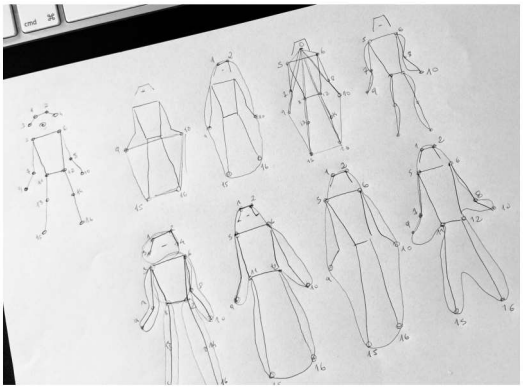


Fig. 2. Sketch of the visual forms using the keypoint references.

From these minimalist forms, more complex ones were developed, as shown in Figure 3, such as the total representation of the body through an abstract, or literal representation of the human body. We divided the representations into three different groups:

- **Relationship between extremities of the body**. The movement of the dancer in its spatial form has different representations, from contained movements to more expansive movements, such as geometric shapes in space. By visually representing the connection of the dancer's extremities, in real-time, while dancing, the observer can see in more detail this body connection that is required in dance, as show in Figure 3 (left picture).
- **Complex connection between different points**. A connection between different points generates a complex form which visually represent the body relationship. This visual model has the particularity of generating unexpected results, both for the performer and the spectator,as show in Figure 3 (right picture).
- **Representation of the whole body (literal representation)**.In this form, the visual representation is similar to the body skeleton. Although it has very interesting applicability when the dancer's body is not present behind the visual form, as shown in Figure 4.

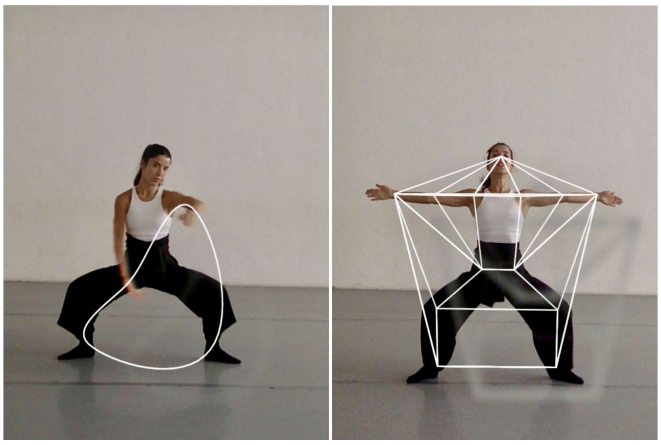


Fig. 3. Usability tests through the framework developed. Left picture presents the connection between extremities (i.e. hands and feet). Right picture presents a complex connection between the different keypoints.

C. Applicability

The framework developed has numerous exhibit possibilities, as it substantially enriches the dance movement and gives a new visual perspective on the movement. On the other hand, the interactive character that this work provides allows the creation of an interactive installation open to the whole public, in an exhibition format. According to previous works [30], [31], the audience greatly appreciates the interactive experiences and personalized actions in which the audience sees their reflection or personalization of themselves. On the other hand, this framework could result in an application itself that is accessible to anyone, to experience anywhere. Involving the public as performers and being spectators of themselves is one of the main objectives. Conversely, there is another possibility of being presented on a stage, with live dance, and the tool contributing to a new dimension of a performance show, leading to the involvement of the dancers and the audience. Finally, the pedagogical contribution that this work may have in teaching dance will provide, from an early age, for young dancers to look at their movement in a more mature, organic way and obtain a greater body awareness. This last contribution is presented in the next section as a case study, already presented to the public.

III. Real Case Study

Creating a contemporary dance performance with the integration of sensing mechanisms of the human body is a challenge. For this dance performance, the author (also choreographer) created a piece entitled *"Perspectives"*. As the name suggests, *"Perspectives"* contemplated a narrative of movement that sought to translate the different ways of interpreting and understanding the same situation.

A. Concept and Development

The concept of the piece focused on the way we look at the world, and how often we are only interested in our perspective, not thinking about the view of others. In essence, the visual form that each dancer chose as their representation also translated this thought, that we should look at each other from various perspectives. For the performance moment, in which the dancers danced live with the technology, we worked with seven young professional dancers over a month. The dancers chose within the set of possibilities the visual form that would best represent their movement. For this moment, the choreographic composition had as its first part the execution of contained movements and only later the total expansion of the body. Each dancer developed their choreographic phrase, through the chosen visual form, and after this choreographic composition was created, the structure was tested, in real-time. This process was tested and rehearsed until the stage performance.

B. Final Result, on the stage

In the stage performance, each dancer danced individually, while the other dancers were in a static position. The work between the different elements (i.e. real-time projection, dancers, light, and stage environment) was complex. It was necessary to ensure, together with the light technicians, that the light incident on the dancer does not interfere with the design of the projection. To simplify this process, when the dancers were performing, they were positioned on the stage in the opposite side of the visual projection, as shown in Figure 4. In this sense, the audience looking at the stage observed a stage, which in fact was translated as a picture composed of different elements that completed each other, throughout the performance. Of the group of seven participants, three dancers chose their visual representation through curved shapes and four dancers chose straight shapes. Two straight shapes and one curvilinear shape had their abstract representation through a square, geometric shape, or ellipse, most similar to Figure 3. The remaining chosen representations had humanoid shapes. That is, the visual shape representation was close to the shape of the human skeleton, as presented in the Figure 4.

Fig. 4. Photograph of the performance on stage, where the dancer's movement is detected through the framework. The visual form presents the whole body (literal representation) with curvilinear lines.

C. Dancers evaluation

At the end of the stage experience, through the developed system, the dancers were asked to answer a short questionnaire to evaluate their experience. Each dancer signed informed consent before the anonymized survey. The age range of the young dancers is between 17 and 24 years old, with an average of 18.7 years (SD = 2.56). At the end of the survey, the dancers were asked to rate their experience from 0 to 10 (i.e. Weak = 0, Excellent = 10). Accordingly, to the results, the total average was 9.7 (SD = 0.48).The same classification was applied to understand if the dancers considered this tool to enrich performance, and the total average was 9.85 (SD = 0.37). As an open-ended question, the dancer was asked to describe whether this tool facilitated the perception of his movement, and the following answers were obtained: "When I observed my movement through technology, I realized that I have to improve the range of my movement", "After seeing myself in another perspective, I could understand that my movement has to be clearer and more precise, as well as thinking more about the extremities of my body", "The movement of my arms can be more expressive and wider". The dancers' own assessment of technology was very positive, as the integration of this work with dance provides a greater body awareness to each one.

IV. Future Work and Conclusion

The present work results from the intersection of different areas, but each area interconnected with the others produces a very enriching final result for art and technology. The result presented in the real case study contributed positively to the present investigation, and future deployments will be developed based on this experience. One of the points to consider is the detection and visual composition of choreographies with large groups. On the other hand, making the experience of this interaction available to the general public can provide gratifying results of well-being and increased body awareness.

Acknowledgment

We thank our colleagues from the Institute of Systems and Robotics and the College of Arts, both institutions from the University of Coimbra, Portugal. We thank the young dancers from Dance N'Arts School of Coimbra. We thank gratefully support of "Fundação para a Ciência e Tecnologia" (FCT – Portugal), through the Research Grant (2020/09137/BD). Special thanks to the Teatro Académico de Gil Vicente (TAGV) for their support on the performance space and infrastructures.

References

[1] Bartenieff, I., Lewis, D. (2013). Body movement: Coping with the environment. Routledge.

[2] Herbison-Evans, D. (1991). The Dance and the Computer: A Potential for Graphic Synergy. University of Sydney, Basser Department of Computer Science.

[3] Schiphorst, T. (1993). A case study of merce cunningham's use of the lifeforms computer choreographic system in the making of trackers (Doctoral dissertation, Arts and Social Sciences: Special Arrangements).

[4] Warner, M. J., Stitt, N. S. F. (1994). Leaping into the 21st century: computer technologies for dance. Canadian Theatre Review, (81), 36.

[5] Plone, A. (2019). The Influence of Artificial Intelligence in Dance Choreography.

[6] Calvert, T. W., Chapman, J. (1978, January). Notation of movement with computer assistance. In Proceedings of the 1978 annual conference-Volume 2 (pp. 731-736).

[7] Camurri, A., Morasso, P., Tagliasco, V., Zaccaria, R. (1986). Dance and movement notation. In Advances in Psychology (Vol. 33, pp. 84-124). North-Holland.

[8] Quintas, R., Dionísio, T. (2006). SWAP project presentation. Performance Research, 11(4), 152-154.

[9] Moura, J. M., Barros, N., Ferreira-Lopes, P. (2019, May). From real to virtual embodied performance-a case study between dance and technology. In Proceedings of the 25th International Symposium on Electronic Art (ISEA 2019) (pp. 370-377). Gwangju: ISEA International.

[10] Pettee, M., Shimmin, C., Duhaime, D., Vidrin, I. (2019). Beyond imitation: Generative and variational choreography via machine learning. arXiv preprint arXiv:1907.05297.

[11] Crnkovic-Friis, L., Crnkovic-Friis, L. (2016). Generative choreography using deep learning. arXiv preprint arXiv:1605.06921.

[12] Chan, C., Ginosar, S., Zhou, T., Efros, A. A. (2019). Everybody dance now. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 5933-5942).

[13] Cao, Z., Simon, T., Wei, S. E., Sheikh, Y. (2017). Realtime multi-person 2d pose estimation using part affinity fields. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7291-7299).

[14] Du Sautoy, M. (2020). The creativity code: art and innovation in the age of AI. Harvard University Press.

[15] Zhang, W., Liu, Z., Zhou, L., Leung, H., Chan, A. B. (2017). Martial arts, dancing and sports dataset: A challenging stereo and multi-view dataset for 3d human pose estimation. Image and Vision Computing, 61, 22-39.

[16] Toshev, A., Szegedy, C. (2014). Deeppose: Human pose estimation via deep neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1653-1660).

[17] Moon, G., Chang, J. Y., Suh, Y., Lee, K. M. (2017). Holistic planimetric prediction to local volumetric prediction for 3d human pose estimation. arXiv preprint arXiv:1706.04758.

[18] Cao, Z., Simon, T., Wei, S. E., Sheikh, Y. (2017). Realtime multi-person 2d pose estimation using part affinity fields. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7291-7299).

[19] Huang, J., Altamar, D. (2016). Pose estimation on depth images with convolutional neural network.

[20] Mehta, D., Sridhar, S., Sotnychenko, O., Rhodin, H., Shafiei, M., Seidel, H. P., ... Theobalt, C. (2017). Vnect: Real-time 3d human pose estimation with a single rgb camera. Acm transactions on graphics (tog), 36(4), 1-14.

[21] Kim, Y., Kim, D. (2018). Real-time dance evaluation by markerless human pose estimation. Multimedia Tools and Applications, 77(23), 31199-31220.

[22] Stevens, C., McKechnie, S. (2005). Thinking in action: thought made visible in contemporary dance. Cognitive Processing, 6(4), 243-252.

[23] Louppe, L. (1997). Poétique de la danse contemporaine.

[24] Craine, D., Mackrell, J. (2010). The Oxford dictionary of dance. Oxford University Press.

[25] Duffy, M., Atkinson, P. (2014). Unnatural movements: Modernism's shaping of intimate relations in Stravinsky's Le sacre du printemps. Affirmations: of the modern, 1(2).

[26] Jordan, J. (2021). AI as a Tool in the Arts. URL: https://amt-lab.org/blog/2020/1/ai-as-atool-in-the-arts [07 April 2022].

[27] Jacob, M., Magerko, B. (2015, June). Interaction-based Authoring for Scalable Co-creative Agents. In ICCC (pp. 236-243).

[28] Brooks, L. M. (1993). Harmony in space: A perspective on the work of Rudolf Laban. Journal of aesthetic education, 27(2), 29-41.

[29] Rett, J., Dias, J., Ahuactzin, J. M. (2010). Bayesian reasoning for laban movement analysis used in human-machine interaction. International Journal of Reasoning-based Intelligent Systems, 2(1), 13-35.

[30] Nogueira, M.R, Menezes, P., Patrão, B. (2019). Painting with Movement. In The 17th International Conference on Virtual-Reality Continuum and its Applications in Industry (VRCAI '19). Association for Computing Machinery, New York, NY, USA, Article 53, 1–2, https://doi.org/10.1145/3359997.3365750

[31] Nogueira, M.R, Menezes, P., Patrão, B. (2021). "Understanding Art through Augmented Reality: Exploring Mobile Tools for Everyone's Use," 2021 9th International Conference on Information and Education Technology (ICIET), 2021, pp. 410-414, doi: 10.1109/ICIET51873.2021.9419620.

Session 4: Computer Graphics & Games

chair: António Coelho

Digital Fishes

David Pérez
CIIC, ESTG, Polytechnic of Leiria
Leiria, Portugal
dacert@gmail.com

Nuno Rodrigues
CIIC, ESTG, Polytechnic of Leiria
Leiria, Portugal
nunorod@ipleiria.pt

Rita Ascenso
CIIC, ESTG, Polytechnic of Leiria
Leiria, Portugal
rita.ascenso@ipleiria.pt

***Abstract*—This paper describes the development of critical elements during the creation of the Digital Fishes Interactive Inirtual Aquarium. Here the solution to simulate the movement of marine flora and fauna added to the virtual aquarium is described. Furthermore, an implementation of the fish behavior, to move in groups using the Boids algorithm, is created. In addition to autonomous behaviors such as chasing prey and fleeing from predators, interactive behaviors include following the finger on the screen and going towards the food the user may throw in the pond.**

***Keywords*—** Digital Fishes, Boids Behavior, User Interaction, Ubiquitous Interfaces, Computer Graphics

I. Introduction

Digital Fishes is an interactive virtual aquarium with a physical representation of a cubic fish tank, which can be configured remotely. Also, the visual information, and animated elements, such as fish and vegetation, are synchronized between the lateral faces of the cubic tank. The solution aims to create fishes to appear as natural as possible, swimming in groups or alone and reacting to user interactions through contact with the device's screen. The reactions can vary between fleeing or approaching the contact's source and reacting to the user's feeding action. In addition, it can be used for educational purposes and thus contribute to a better understanding of marine life in Portugal.

Inspired by the development of the Digital Fishes Interactive Virtual Aquarium, this paper aims to describe critical elements such as fish behavior and the creation of some visual elements like fishes, seaweed, water surface, and caustic using shaders.

The rest of this work is organized into fourth sections. First, section II shows the background. Then, section III describes the implementation. Finally, section IV presents the conclusions of this work.

II. Background

According to J. R. Li [6], Virtual reality (VR) is a synthetic, three-dimensional, interactive environment typically created by a computer. It provides a unique avenue to enhance the visualization of complex three-dimensional objects and environments with real-time, more interactive, and spatial ability. Furthermore, as Robertson [11] says, a Desktop VR system uses animated interactive 3D graphics to build virtual worlds with desktop displays and without head tracking.

So we can say that an Interactive Virtual Aquarium is a 3D representation of a natural aquarium with which we can interact. In our case, the goal is to build a virtual aquarium with the immersive characteristics of a Desktop VR system with which we can interact through a device's screen.

Lee and colleagues [4] presented a method to build a virtual aquarium environment system that can be controllable in real-time by expressing the movement of algae using virtual fish and control points based on the spring-mass model. In addition, a fluid expression generator is presented based on fluid mechanics and control fluid flow. In our case, the solution will be more straightforward, without fluid dynamics, and the movement of the algae and fish is performed using vertex displacement along the direction axis of the fish with a sinusoidal function to simulate the movement of locomotion. Finally, Lypovyi and Montusiewicz [7] use Unity as a platform to develop simulation experiments with different configurations of force coefficients that control a model based on Boid behavior.

The abbreviation "Boid" is an abbreviated version of the term "bird-oid object" ("bird-like object") introduced by Craig W. Reynold to describe the Boids algorithm [9]. This algorithm allows simple rules to define the movement in groups of fish without implementing other artificial intelligence techniques.

III. Development

First, as the central technology for developing the solution, we use Unity[1]. It allows us to compile our app for web platforms with WebGL 2.0[2] graphics API. Unity is a cross-platform game engine developed by Unity Technologies. The engine has been gradually extended to support a variety of desktop, mobile, console, and virtual reality platforms. The engine can create three-dimensional (3D) and two-dimensional (2D) games, interactive simulations, and other experiences.

Unity was chosen for the proposed solution to the requirements for developing this particular Interactive Virtual Aquarium. In summary, Unity was chosen for the proposed solution to the requirements for developing this particular Interactive Virtual Aquarium. Despite this, the development described below can be carried out in other development technologies for 3D applications.

The rest of the section describes the development of fish behavior and the main shaders of Digital Fishes.

A. Fish behavior

We use the Boid algorithm created by Craig W. Reynold to implement the school of fish behavior. In his paper [9], Reynolds describes three levels of control used to create "autonomous characters" and simulate their movements within a flock of birds or a school of fish. The control levels can be split into; Action selection, Steering, and Locomotion.

1) Steering: Steering is the middle level, where we determine the path. The basic model of a herd of swarming objects consists of three simple behaviors that govern their movement. In his article [10], Reynolds describes how individual Boid maneuvers are carried out based on the location and speed of other objects.

- Separation (Fig. 1) - it is responsible for avoiding crowds of local members of the herd, thanks to which there is no collision with nearby objects.
- Alignment (Fig. 1) - requires movement to take place about the average position of the local herd members.
- Coherence (Fig. 1) - requires objects to move toward the average position of local herd members, leading to an attempt to stay close to neighboring herd members.

[1]https://unity.com
[2]https://www.khronos.org/webgl

978-8-3503-4573-57/22/$31.00

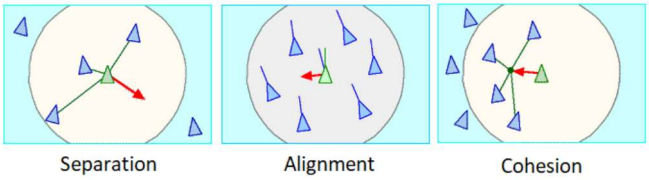


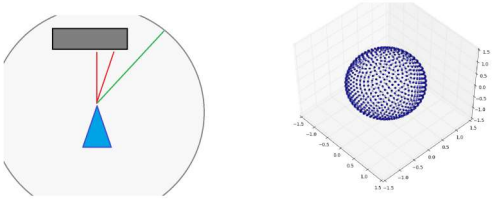
Fig. 1. Steering [10]

The key to making the school of fish work is combining **Cohesion**, **Separation**, and **Alignment**. The more steering behaviors define and use together, the more complex and "realistic" the overall effect.

Our implementation of the steering behavior is based on the pseudo-code proposed by [8] with the necessary modifications to improve performance in Unity with hundreds of Boids.

2) Obstacle avoidance: Now, add the obstacle avoidance behavior. The idea is to project rays with increasing angles until finding a free path or the farthest free path (Fig. 2a). Since we are moving in 3 dimensions, for it to work, we must project the rays to points on a sphere.

To generate the points, we use the golden spiral method on a sphere described by Chris Drost [1], which uses the golden ratio to distribute points evenly on the sphere (Fig. 2b).



(a) Avoiding collisions (b) Points on a sphere

Fig. 2. Obstacle avoidance

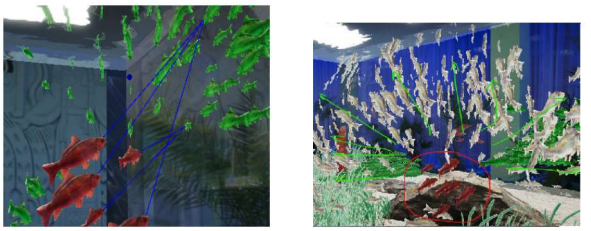
Now adding the behavior of the Boids to our fishes, we already have the movement in schools. It only remains to add other behaviors to get closer to reality and better immersion in our virtual aquarium.

3) Food chain behavior: To add food chain behavior, we separate the fishes by **species** and assign them an ascending **level**, with the lowest level being the weakest. To implement the food chain behavior, we introduce the changes in the Steering Behavior to calculate the target of the predators and the direction of the prey's flight.

Now we have different species, so we apply the cohesion and alignment vectors calculation only to the same species and the separation vector for all of them. Thus, we avoid weighing them up.

A simple solution for the target is to assign it the position of the last fish one level lower in the perception radius. Later we can chase the fish, calculating the direction from the fish to the target with the difference between the target and the position of the fish (Fig. 3a).

In the case of prey, we calculated a dispersion force decreasing the difference between the predator's and fish's positions. The predator would be all the fishes in the radius of perception of the prey with the highest level. It will result in a dispersion force vector opposite to the predators. In this case, we only calculate the cohesion and alignment when the fish does not have to flee to achieve the effect of dispersion of a school (Fig. 3b).



(a) Predators' behavior result (b) Prey dispersion behavior result

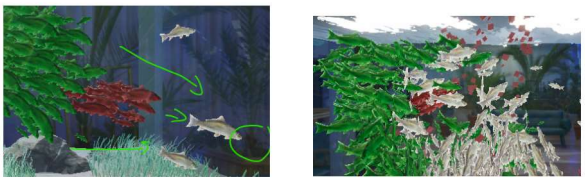
Fig. 3. Food chain behavior

4) Action selection: Action selection is the highest level and involves strategies such as target selection and planning.

To interact with our aquarium was implemented goals-tracking strategies, which are:

- The fish follow the movement of the pointer or finger (Fig. 4a).
- The fish move toward the food when fed (Fig. 4b).

For this implementation, we obtain the vectors in the direction of the objective, then add that vector to the rest of the behavior to guarantee separation, cohesion, and alignment.



(a) Pointer target result (b) Food target result

Fig. 4. Action selection

In the case of following the food, as we use a particle system for the render, we calculate the midpoint of the particles and use it as a target. Thus, the midpoint is recalculated while the particles disappear until none are left. Then, to add the synchronization effect, we synchronize the value of the random seed of the particle system across the clients.

5) Locomotion: The locomotion is the lowest level and only involves the underlying animation mechanics. In this case, the fish's swimming is carried out using the vertex displacement techniques in the shaders, which can be seen in the next section.

6) Performance optimization: Note that the straightforward implementation of the Boids algorithm has an asymptotic O(n2) complexity. Each Boid needs to consider each other if only to determine if it is not a nearby flock mate [10]. As Reynolds in his article suggests reducing this cost to almost O(n) is possible by using a proper spatial data structure that allows us to keep the Boids ordered by their location. There is no need to iterate through all the Boids to find the neighbors.

A data structure that can be used for this purpose is a **MultiHashMap**, with which we can apply the spatial hashing technique. It consists of using as an index a hash obtained from the object's location in the space subdivided into cells, see about in [5]. The objects are stored in such a way that when we query for the hash of the location, we get the list of objects in the cell in which it is located.

Unity has a collection type called **NativeParallelMultiHashMap** which we use to store the positions of our Boids and thus quickly get the list of neighbors. However, the key here is to generate the hash. For this, we use the function proposed in this article [12] by

Matthias Teschner and colleagues.

Fig. 5. Spatial hash result with active cells highlighted

To further optimize the solution, we can implement the calculation of the Boids using multi-threading. Here we have investigated two methods, one of which has been finally adopted.

The first method is to use a Compute shader [3]. Compute shaders are programs that run on the graphics card outside the regular rendering pipeline. They can be used for massively parallel GPGPU[4] algorithms or to accelerate parts of game rendering.

This method proves to be faster than using multi-hash calculation. However, in this case, as the shader uses the GPU it is inefficient in the Digital Fishes solution, which needs to do the calculation on the server in a build optimized for a dedicated server that does not need asset as audio, textures, and shaders. In addition, this would limit us to using more resources in our deployment.

The second method uses Unity Data-Oriented Technology Stack (DOTS)[5], making it possible to take full advantage of today's multi-core processors. This work uses Unity C# Job System and Unity Burst Compiler packages from the DOTS stack.

The C# Job System takes advantage of the multiple cores. In addition, C# Job System exposes Unity's internal C++ Job System, allowing C# scripts to run as jobs alongside Unity's internal components. It also protects against multi-threading pitfalls, such as race conditions.

Burst Compiler is a new LLVM-based[6] backend compiler technology that takes C# jobs and produces highly optimized machine code. In addition, the Burst Compiler optimizes the output for the compiling platform.

The characteristics of these latest technologies are ideally suited to the demands of our solutions:

- A solution to the calculation problem using multi-threads, in addition to integration with the spatial hashing method, exploits the data structures it offers us, increasing the performance even more.
- Build optimization for a dedicated server using Burst.

Let us note that in the methods tested during the development, the weight of the computation falls on the detection of collisions. If we see the last case (Fig. 6, yellow line) in which we use the Unity Physics[7] package, which allows us to perform the calculation in multi-threads, the time is significantly reduced.

B. Shaders

As mentioned above, the shaders are built using the Unity Shader Graph tool.

[3]https://docs.unity3d.com/Manual/class-ComputeShader.html
[4]https://pt.wikipedia.org/wiki/GPGPU
[5]https://unity.com/dots
[6]https://llvm.org/
[7]https://unity.com/dots/packages#unity-physics-preview

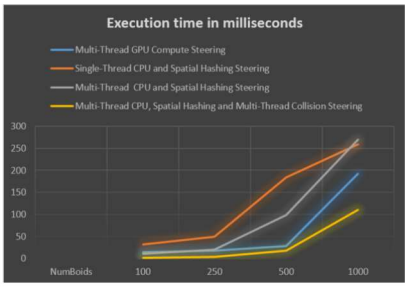


Fig. 6. Execution time in milliseconds of the movement of the Boids stressed with collisions

1) Fish animation: As mentioned before, the locomotion layer proposed by Reynolds [9] was implemented only using shaders.

The proposed solution consists of simulating the simple swimming movement of a fish by moving the vertices of the model using a sine function through its axes depending on how we have the model oriented (Fig. 7).

The solution was inspired by the article [2] by Joseph Kalathil, in which he mentions something that we were able to verify regarding the performance of this type of solution concerning using an animation if we consider that the GPU calculates the shaders.
In our case, the orientation of our models is thus the Z axis (Fig. 7), for which we will displace the vertex in X along the Z axis of the model using a sine function to simulate the undulating movement:

$$x = sin(z)$$

In order to control the animation in terms of amplitude, frequency, and speed. We calculate z as follows:

$$z = z * frequency_Z + time * speed_Z$$

$$x = x + sin(z) * amplitude_Z$$

Note that we introduced the time variable, this is necessary to guarantee continuous movement and an offset to later be able to generate a different movement for each fish.

We need to limit the amount of movement through the model. A gradient is used from the limit of the head to the tail to map the movement (Fig. 7). Them, to calculate the movement displacement through a gradient, we apply a linear interpolation[8] (lerp) between the vertices by a Sigmoid funtion[9] (smoothstep) of the limit of the head.

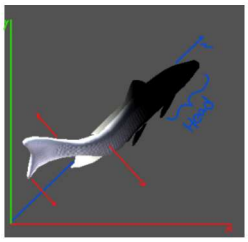


Fig. 7. Fish vertex displacement idea

[8]https://en.wikipedia.org/wiki/Linear_interpolation
[9]https://en.wikipedia.org/wiki/Sigmoid_function

2) Water surface: To create a water surface, we are based on the idea used by Yuri Kryachko in [3], which is divided into components to achieve a simple water surface shader we need: displacement vertically for low-frequency wave motion with high amplitude; details of high-frequency waves; transparency with a distortion which gives it a touch of depth.

1) *Displacement vertically for low-frequency wave motion with high amplitude*: For the movement of low-frequency waves, we also use the vertex displacement technique, but in this case, we do it based on a height map generated by a Gradient Noise (Fig. 8a).
2) *Details of high-frequency waves*: For distortion and ripples, we also use a noise gradient transformed into a normal map (Fig. 8c) but a gradient with another scale to achieve high-frequency wave details. We can combine several layers by shifting in different directions to achieve a slightly more chaotic effect.
3) *The transparency*: we use a Shader Graph Scene Color Node to obtain the transparency.

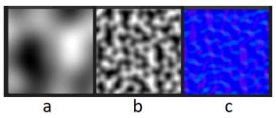


Fig. 8. Gradient Noise maps

The movement is achieved by displacing the maps (height, normal) as a function of time in such a way that when a point of the map passes through a vertex, it is displaced by the color scale input (black and white) (Fig. 8a).

3) Seaweed: In the implementation of the shader for the seaweed, the same idea of vertex displacement based on a gradient noise map is used (Fig. 8a). In this case, the displacement is horizontal, based on the position of the vertex in the world space, in this way an organic animation effect of all the seaweed is achieved (Fig. 9b).

(a) (b)

Fig. 9. Water surface and Seaweed results

The same idea to limit the movement of the fish from the head is used to anchor the movement to the root of the plant.

4) Caustics: For the caustics, Unity Decal Projector was used to project a texture onto all intersecting models. Note that this projector only correctly projects the texture on one axis. On the other hand, it stretches it (Fig. 10a). To correct this, we decompose the absolute position in the world space and add each component. We can also add distortion to the texture and displacement, combining several layers in the same way as the one used in the shader for the water surface (Fig. 10b).

IV. Conclusion

This paper has described the solution used to simulate the movement of marine flora and fauna added to the Digital Fishes Interactive Virtual Aquarium and the fish behavior implemented to move in groups using the Boids algorithm. Also, describe the development of autonomic behaviors, such as chasing prey and running away from predators. Finally, the development of interactive behaviors includes following the finger on the screen and going towards food the user can throw into the pond.

Fig. 10. Caustics projection

Also, this paper describes the analysis of solutions implemented in Unity to improve performance. However, for future improvements, it is proposed to fully adopt the DOTS architecture proposed by Unity, which we currently partially do. Another future improvement to the performance problem is to use the Spatial Hashing method to detect collisions storing the triangles of the static objects mesh and calculating the interception of these triangles with a ray casting from the head of a fish, similar to the method proposed by Matthias Teschner and colleagues [12].

Acknowledgement

This work is supported by national funds through the Portuguese Foundation for Science and Technology (FCT), I.P., under the project UIDB/04524/2020.

References

[1] Chris Drost. The golden spiral method. *Stack Overflow.(2017, May 24). Retrieved from: https://stackoverflow.com/questions/9600801/evenly-distributing-n-points-on-a-sphere*, 2017.
[2] Joseph Kalathil. How to animate a fish swimming with shaders. *Bitshift Programmer.(2018, 12 Jan). Retrieved from: https://www.bitshiftprogrammer.com/2018/01/how-to-animate-fish-swimming-with.html*, 2018.
[3] Yuri Kryachko. Using vertex texture displacement for realistic water rendering. *GPU gems*, 2:283–294, 2005.
[4] Hyun-Cheol Lee, Eun-Seok Kim, Nak-Keun Joo, and Gi-Taek Hur. Development of real time virtual aquarium system. *International Journal of Computer Science and Network Security*, 6(7):58–63, 2006.
[5] Sylvain Lefebvre and Hugues Hoppe. Perfect spatial hashing. *ACM Transactions on Graphics (TOG)*, 25(3):579–588, 2006.
[6] J.R. Li, L.P. Khoo, and S.B. Tor. Desktop virtual reality for maintenance training: an object oriented prototype system (v-realism). *Computers in Industry*, 52(2):109 – 125, 2003.
[7] Taras Lypovyi and Jerzy Montusiewicz. Simulation of boid type behaviours in unity environment. *Journal of Computer Sciences Institute*, 3:23–27, 2017.
[8] Conrad Parker. Boids pseudocode. *Website: http://www.kfish.org/boids/pseudocode.html*, 2007.
[9] Craig W Reynolds. Flocks, herds and schools: A distributed behavioral model. In *Proceedings of the 14th annual conference on Computer graphics and interactive techniques*, pages 25–34, 1987.
[10] Craig W Reynolds. Boids background and update. *http://www.red3d.com/cwr/boids/*, 2001.
[11] George Robertson, Mary Czerwinski, and Maarten Van Dantzich. Immersion in desktop virtual reality. In *Proceedings of the 10th annual ACM symposium on User interface software and technology*, pages 11–19, 1997.
[12] Matthias Teschner, Matthias Heidelberger, Brun Müller, Danat Pomerantes, and Markus H Gross. Optimized spatial hashing for collision detection of deformable objects. Technical report, Technical report, Computer Graphics Laboratory, ETH Zurich, Switzerland, 2003.