Interactive Musical Setting with Deep Learning and Object Recognition

Mário Cardoso[1][a] and Rui Pedro Lopes[2][b]
[1]*Research Center for Basic Education, Instituto Politécnico de Bragança, Portugal*
[2]*Research Center for Digitalization and Industrial Robotics, Instituto Politécnico de Bragança, Portugal*

Keywords: Deep Learning, Object Recognition, Musical Setting, Musical Textures, Musical Education.

Abstract: The SeMI - Interactive Musical Setting, explores the possibilities of joining machine learning, the physical and the sound world. In this context, a machine learning algorithm and model was used to identify physical objects through image processing. Each physical object is associated with a student's produced musical texture that starts playing when the object is recognized by the device. This allows defining use cases in which students have to develop diverse although interrelated sound textures and combine them with a physical world, in both a fake orchestra, that reacts to people and objects in front of it, and mood rooms, for example. The application was developed for iPad and iPhone, using Swift programming language and the iOS operating system and used in the classes of the masters on Teaching of Musical Education in the Basic School.

1 INTRODUCTION

Over the last twentieth century, the nature of music itself has changed dramatically (Hargreaves and Lamont, 2017). Several musicologists document many revolutions and transformations: (i) the fresh and innovative music of Beethoven, Wagner, Debussy, Stravinsky, Schoenberg and Messiaen; (ii) the developments of electronic sound production; (iii) the transformations and discussions of iconoclastic composers (Stockhausen, Boulez, Cage, Glass, Reich) and the impact in music performance and composition; (iv) the influence and power of the different genres and styles; and (v) the digital revolution (to produce, record and transmit) and the new sounds and effects. This last aspect had deep effects in the people's lives and in particular in the way musicians (and non-musicians) perform and compose music (Cardoso et al., 2019; Ruthmann and Mantie, 2017).

Nowadays, being a music teacher involves more skills: working knowledge of music software and hardware; arranging or improvising; arts technology (interactive art, computer programming, virtual and augmented reality). For Brown (2015), it is necessary to reinterpret the nature of musical experience. The digitisation of music means that we have to change and reinvent the paradigm of teaching and learning music. For Peppler (2017), this positive vision radically shifts the lines between performer, listener and composer. In music education, technology can be a catalyst that contribute to expand the process of teaching and learning music into a more comprehensive, creative, innovative and imaginative experience (Cardoso et al., 2019; Brophy, 2001). In this framework, musical learning can benefit from the development and adoption of innovative devices and tools that foster their autonomous work and help them overcome the difficult task of learning to play and compose music. It is important that the process puts great emphasis on the student, encouraging higher education institutions and academic staff to place students at the center of their thinking and to help them manage their expectations and be able to consciously and constructively design their learning paths throughout their higher education experience (Lopes et al., 2019; Tenorio et al., 2018). If the education is one of the most important aspects of human development, greatly influencing the path of professional development and success (Mesquita et al., 2015, 2014), it is important to increase the training at the higher-education level, that contributes to the scientific and technological qualification of youth and adults, towards the "creative, innovative and competitive development, with high productivity standards" (Correia and Mesquita, 2006, 166).

This paper describes the development of a machine learning based application, that runs in an iOS

[a] https://orcid.org/0000-0003-3645-9641
[b] https://orcid.org/0000-0002-9170-5078

Cardoso, M. and Lopes, R.
Interactive Musical Setting with Deep Learning and Object Recognition.
DOI: 10.5220/0009856406630667
In *Proceedings of the 12th International Conference on Computer Supported Education (CSEDU 2020)*, pages 663-667
ISBN: 978-989-758-417-6

device such as an iPhone or iPad, that performs image analysis and reproduces student defined musical textures or planes. This tool was used in both secondary school students, in a science summer school, and higher education students, in the masters on Teaching of Musical Education in the Basic School. It is through experimentation and careful reflection that the student will learn the principles underneath sonification, try art installations, interactive performances, and others.

2 MACHINE LEARNING

Digital image processing has been a challenge for several decades. Nevertheless, the importance of this field has pushed research towards significant advances throughout the years. Both the technological advances and the adoption of new paradigms has been making image processing faster, more precise and more useful. There are several problems that are usually addressed through digital image processing (Egmont-Petersen et al., 2002):

1. Preprocessing/filtering - the output is the same type as the input, with the objective of improving contrast or reduce noise;
2. Data reduction/feature extraction - the resulting data usually are smaller and indicative of significant components in the image;
3. Segmentation - partition and selection of regions in the image within some criterion;
4. Object detection and recognition - determining the location, orientation and classification of specific objects in the image;
5. Image understanding - getting semantic knowledge of the image
6. Optimization - improving image characteristics.

The emergence of deep learning and convolutional neural networks (CNN) demonstrated significant advances in many of these tasks (LeCun et al., 2015). CNNs represent computational models composed of multiple processing layers, where each layer learns different levels of representations of data. An image is represented by a three dimensional array, where the first two dimensions describe the 2D distribution of pixels, and the third dimension contains information about the color of each pixel. The CNN will learn the features from the images through an intense learning process. The first layer will typically represents the presence or absence of edges at particular orientations and locations. The second layer detects arrangements of edges, and the third layer assembles these into larger combinations that correspond to parts of familiar objects. Subsequent layers would detect objects as combinations of these parts. The key aspect is that these layers of features are learned from data using a general-purpose learning procedure, instead of relying on a human engineering process.

In this paper, we are mainly interested in object detection and recognition (point 4 in the above list). Object recognition in images is a complex task, not only because of the intrinsic complexity of the process but also because of the semantic information they convey (Uijlings et al., 2013).

Humans are able to recognize, with ease, both the location of objects, their type and even their function. To replicate this process in a computer, it is necessary that the algorithm is able to mark a region in the image (segmenting the object) and perform a classification (to identify the type of the object).

There has been a huge progress over the last years in this field. One popular approach to both identifying and recognizing objects in a single operation is YOLO - You Only Look Once (Redmon et al., 2016). This approach uses a single CNN to simultaneously predict multiple bounding boxes and class probabilities for those boxes, following a regression problem. This makes the algorithm very fast and able to run in a portable device, such as a tablet or smartphone.

3 METHODOLOGY

The purpose of this work is to develop a mobile application, which we called SeMI[1], targeting iOS devices, that can detect and recognize objects in images collected with the camera. Each time an object is recognized, a sound texture is played. If multiple objects are recognized, the associated texture is mixed with the output sound.

Objects are specified by the user in a virtual setting, a representation on the screen, where the user can drag, drop, resize and move several objects. The user can also associate a sound file (WAV, MP3, AAC, and other formats) to each object.

Based on this, a software application was designed, composed of four main modules (Figure 1).

Images are captured with the camera and the image is rotated and scaled according to the orientation of the device and the requirements of the object recognition input. The YOLO module is constantly analyzing the images and outputs a list of vectors, each with the bounding box coordinates, the class of the object and the associated probability of the classifica-

[1] Acronym resulting from “Interactive Musical Setting”.

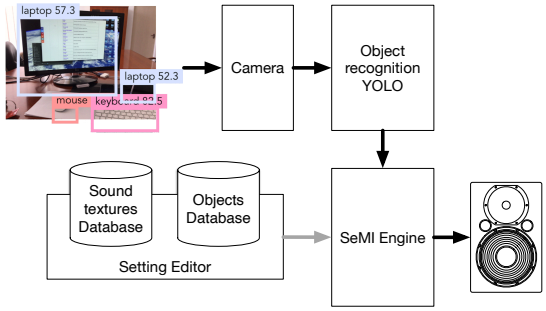

Figure 1: Main modules of the application.

tion performed. The YOLO network has 24 convolutional layers followed by 2 fully connected layers, alternating with 1x1 convolutional layers to reduce the features space. The input image is scaled to fit 224x224 pixels.

The SeMI Engine crosses the object class with the previous user assembled setting and, if the object is present in both locations (image and virtual setting), the sound texture is played.

4 SeMI

The SeMI application starts by presenting, to the user, an empty virtual setting (Figure 2).

Figure 2: Initial screen, with an empty setting.

The user is now able to touch anywhere on the screen to call the object database. This object database is composed of 80 different objects, that the YOLO module can recognize (Figure 3).

After placing, resizing and moving all the required objects, the user can select the sound files to associate with each object (Figure 4).

In the end, the virtual setting, with all the required objects and the associated sound files, is complete (Figure 5). In this example, the user decided to include a plant (in a vase), a bottle, a bicycle, a chair and a person in the virtual setting. When the SeMI Engine is running and any of these objects is detected in the physical scene, the predefined sound texture is played (Figure 6).

In the top left corner, the user can see the images captured by the camera and the object detection and recognition results from the YOLO module. The bounding box is drawn with a label indicating the

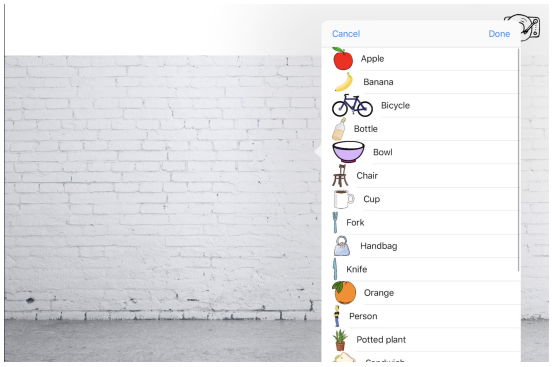

Figure 3: Placing objects on the virtual setting.

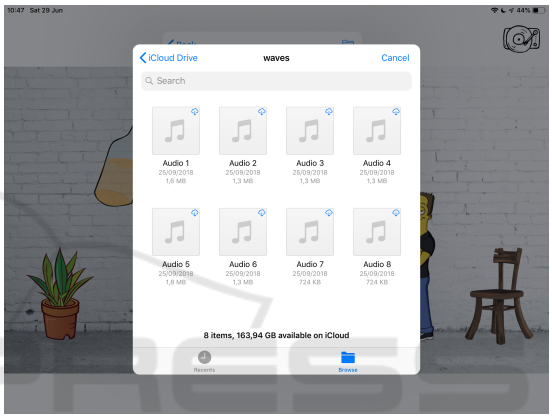

Figure 4: Selecting a sound file for an object.

Figure 5: Example of a finished virtual setting.

Figure 6: Running SeMI Engine.

class and a number that reveals the confidence level on the recognition.

The application can be used directly by the user, exploring the physical space when moving or different use cases can be considered.

5 USE CASES

In the context of music learning, the objective of SeMI is to provide an environment where sound textures are mixed in different planes according to the objects perceived in an image.

Since an object can be detected at unpredictable times, even when the device is already producing sound, the textures should be designed considering different parameters of sound and music features. The combination and the overlap of melodic, rhythmic and harmonic material represent a fundamental aspect in the process of composition and choice of the music motives that can be associated with objects. It is important to ensure complementarity and a certain process of randomization and self-generation between the music motives.

This paves the way for some use cases, starting with the design and creation of compatible sound textures. Mainly, the device can be made static or mobile. In the former situation, the device is placed in a specific location, with camera facing an area where people can enter and place physical objects. The ideal scenario is to build a stage with fake or non-functional musical instruments as a phantom orchestra. The device would be hiding inside one of the instruments and broadcasting the audio output to a set of speakers, through AirPlay, for example.

The fake orchestra reacts to people and objects standing or moving in front of it, thus building a dynamic and involving experience. Without holding an instrument, people in the audience change the sound of the orchestra by moving in or changing objects in front of it.

Another use case is the development of mood rooms (or classrooms). By transporting the smartphone or tablet, people wander through the spaces and, depending on the decoration and on the objects available in each room, the sound will change.

More possibilities can be foreseen, always keeping in mind the three main variables: static or mobile device, combination of physical objects and design and playing of sound textures.

6 CONCLUSIONS

In this paper we have explored the development of an application for smartphones and tablets that can react to physical objects recorded by the video-camera with playing and mixing sound textures. Students are required to design and record the music textures and assemble a virtual setting in the device. Whenever the device recognizes the objects in the field of vision, the audio flows according to the textures that the student decided to use.

This application can be used to foster music learning and create scenarios that correlate the physical world with the sound world, such as a fake orchestra that reacts to the physical environment and mood rooms that reproduce the sound based on the objects they contain.

The application was developed in Swift for iOS and resorts to YOLO (You Only Look Once) for object detection and recognition.

ACKNOWLEDGEMENTS

This work has been supported by FCT – Fundação para a Ciência e Tecnologia within the Project Scope: UIDB/05757/2020.

REFERENCES

Brophy, T. S. (2001). Developing Improvisation in General Music Classes. *Music Educators Journal*, 88(1):34–53.

Brown, A. R. (2015). Engaging in a sound musicianship. In McPherson, G. E., editor, *The Child as Musician*, pages 208–220. Oxford University Press.

Cardoso, M., Morgado, E., and Silva, E. (2019). Musical creation experiences in primary education. In García-Valcárcel, A., Gonçalves, V., Meirinhos, M., Patrício, M. R., Rodero, L., and Sousa, J. S. d. P. C., editors,

ieTIC2019: livro de atas, pages 228–237, Bragança, Portugal. Instituto Politécnico de Bragança.

Correia, A. M. R. and Mesquita, A. (2006). *Novos públicos no ensino superior desafios da sociedade do conhecimento*. Edições Sílabo, Lisboa. OCLC: 494458369.

Egmont-Petersen, M., de Ridder, D., and Handels, H. (2002). Image processing with neural networks—a review. *Pattern Recognition*, 35(10):2279–2301.

Hargreaves, D. and Lamont, A. (2017). *The psychology of musical development*. OCLC: 1055841873.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

Lopes, R. P., Mesquita, C., de Góis, L. A., and dos Santos Júnior, G. (2019). Students' Learning Autonomy: A Systematic Literature Review. pages 5958–5964, Palma, Spain.

Mesquita, C., Lopes, R. P., García, J. l., and de la Cruz del Río Rama, M. (2015). First Contact with the Word of Work: The Competence Built in the Teaching Practices. In Peris-Ortiz, M. and Merigó Lindahl, J. M., editors, *Sustainable Learning in Higher Education SE - 6*, Innovation, Technology, and Knowledge Management, pages 75–87. Springer International Publishing.

Mesquita, C., Lopes, R. P., García, J. l., and Rama, M. d. l. C. d. R. (2014). Pedagogical Innovation in Higher Education: Teachers' Perceptions. In Peris-Ortiz, M., Garrigós-Simón, F. J., and Gil Pechuán, I., editors, *Innovation and Teaching Technologies: New Directions in Research, Practice and Policy*, pages 51–60. Springer International Publishing, Cham.

Peppler, K. (2017). Interest-Driven Music Education: Youth, Technology, and Music Making Today. In Mantie, R. and Ruthmann, A., editors, *The Oxford Handbook of Technology and Music Education*, pages 190–202. Oxford University Press.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You Only Look Once: Unified, Real-Time Object Detection. pages 779–788. IEEE.

Ruthmann, A. and Mantie, R., editors (2017). *The Oxford handbook of technology and music education*. Oxford handbooks. Oxford University Press, New York, NY.

Tenorio, M., Reinaldo, F., Esperandim, R., Lopes, R., Gois, L., and Dos Santos Junior, G. (2018). Céos: A collaborative web-based application for improving teaching-learning strategies. In *Advances in Intelligent Systems and Computing*, volume 725, pages 107–114.

Uijlings, J. R. R., van de Sande, K. E. A., Gevers, T., and Smeulders, A. W. M. (2013). Selective Search for Object Recognition. *International Journal of Computer Vision*, 104(2):154–171.