A CNN-based Framework for Enhancing 360° VR Experiences with Multisensorial Effects

Péter Szabó, Anderson Augusto Simiscuka, *Member, IEEE*, Stefano Masneri, Mikel Zorrilla, and Gabriel-Miro Muntean, *Senior Member, IEEE*

***Abstract*—Improving user experience during the delivery of immersive content is crucial for its success for both the content creators and audience. Creators can express themselves better with multisensory stimulation, while the audience can experience a higher level of involvement. The rapid development of mulsemedia devices provides better access for stimuli such as olfaction and haptics. Nevertheless, due to the required manual annotation process of adding mulsemedia effects, the amount of content available with sensorial effects is still limited. This work introduces an innovative mulsemedia-enhancement solution capable of automatically generating olfactory and haptic content based on 360° video content, with the use of neural networks. Two parallel neural networks are responsible for automatically adding scents to 360° videos: a scene detection network (responsible for static, global content) and an action detection network (responsible for dynamic, local content). A 360° video dataset with scent labels is also created and used for evaluating the robustness of the proposed solution. The solution achieves a 69.19% olfactory accuracy and 72.26% haptics accuracy during evaluation using two different datasets.**

***Index Terms*—Multisensory media, neural networks, image recognition, olfaction, haptics, machine learning.**

I. Introduction

ADDING multisensorial information to video content in order to increase users' Quality of Experience (QoE) is attracting attention in both industrial and academic research environments [1], [2].

Regarding smell, the human nose is very sensitive and is able to discriminate a wide range of odors and scents. Recognizing an odor can invoke different unconscious and involuntary mechanisms [3], such as defense, fight or flight, awareness, etc. Combined with visual cues, smells can be responsible for the brain to process an event in a negative or positive manner. Since scents are connected to the brain memory function, olfaction can be applied in various settings, including in marketing and product design [4] and for educational purposes [5]. Similarly, the human haptic sense can also be used in various settings, as it impacts a person's ability to adapt or respond to diverse inputs.

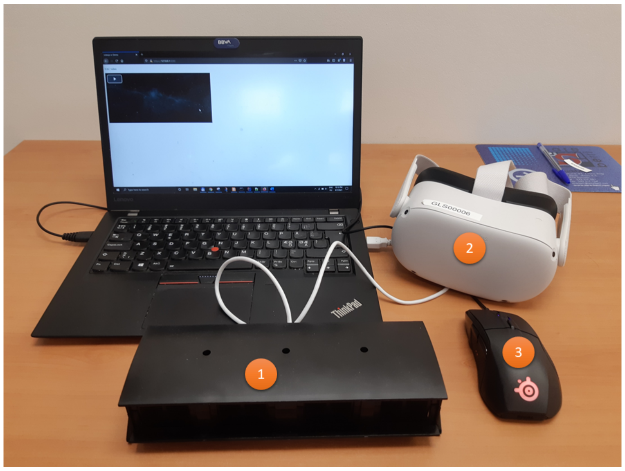


Fig. 1. The prototype of the solution. 1: olfactory machine, 2: Oculus Quest 2 VR headset, 3: haptic mouse

Combining olfactory (sense of smell) and haptic (sense of touch) cues with visual and audio cues, enables the creation of a more realistic experience in immersive environments. Using smell or tactile cues in immersive training spaces can be crucial in certain scenarios [6]. They are often the first stimuli the human body reacts to, when the senses are exposed in high pressure situations. Scent is invisible and can travel through obstacles and therefore could provide otherwise unknown information. Moreover, adding certain scents or tactile information could help neutralizing other unpleasant smells [7] and haptics, such as disturbing shaking when driving [8], thus they can be used for training of law enforcement, military or firefighters, and some manufacturers are already producing scents and odors for this purpose.

Regarding the use of human senses in audiovisual distribution, the main focus of past research was to increase the Quality of Experience (QoE) by enhancing the quality of visual and audio content, and not focusing on other senses, such as touch and smell. In recent years, the adoption of XR technologies has seen both an increase in the quality of video representation and added interactivity. However, as suggested in [9], only managing the resolution of video content results in a very limited approach, especially as the human eye fails to perceive the difference in video content displayed above a certain resolution and refresh rate.

This work addresses the problem of enhancing the viewer QoE by proposing an innovative solution that automatically

This work was supported by the European Union's Horizon 2020 Research and Innovation program under Grant Agreement no. 870610 for the TRACTION project. The support of the Science Foundation Ireland (SFI) Research Centres Program Grant Numbers SFI/12/RC/2289_P2 (Insight) and SFI/16/SP/3804 (ENABLE), co-funded by the European Regional Development Fund, is gratefully acknowledged. *(Corresponding author: Anderson Augusto Simiscuka.)*

Péter Szabó, Stefano Masneri and Mikel Zorrilla are with the Department of Digital Media, Vicomtech, 20009 Donostia-San Sebastian, Spain (e-mail: pszabo@vicomtech.org, smasneri@vicomtech.org, mzorrilla@vicomtech.org).

Anderson Augusto Simiscuka and Gabriel-Miro Muntean are with the Performance Engineering Laboratory, School of Electronic Engineering, Dublin City University, Glasnevin, Dublin 9, Ireland (e-mail: anderson.simiscuka2@mail.dcu.ie, gabriel.muntean@dcu.ie).

adds to video sequences multiple sensorial content targeting other senses, beyond sight and hearing. The proposed solution provides a mechanism for generation of olfactory and haptic stimuli to complement the audio-visual ones in order to increase users' level of enjoyment, immersion and sense of reality. The goal is to contribute actively towards the design of future solutions for distribution of immersive multi-sensorial media, also known as mulsemedia. In this context, the proposed solution is designed to work with Virtual Reality (VR) content and employs machine learning technologies.

As the current approach to adding mulsemedia effects to videos involves a manual annotation process which is time-consuming, the amount of content with sensorial effects available is very limited. Furthermore, authors in [10] point out that most of the automatic haptic annotations are generated based on handcrafted audio features, but video-based solutions are still immature, as they do not use the capabilities of modern neural network architectures.

The work of Sexton et al. [11] proposed a first attempt to automatically generate haptic and olfactory effects based on 360° content using both video and audio. Scents are generated via scene recognition performed by neural networks while haptic content relies on audio cues. The system, however, can only detect a limited range of scents, which strongly reduces the generality of the solution. Each scent is predicted solely based on the current scene, failing to capture further dimensions of information, such as actions involving different objects and people. The authors tested a small number of neural network architectures and did not consider action detection. Furthermore, the methodology used for measuring olfactory performance did not consider important metrics for image detection, such as inference time, giga floating point operations per second (GFLOP), number of network parameters, false positive rate (FPR) and false negative rate (FNR). Although authors provide a small dataset for measuring the olfaction accuracy, the dataset is limited to five scent categories.

This paper fills the aforementioned gaps by making the following contributions:

- We propose a mulsemedia effect generation solution based on multimedia content analysis. The proposed solution generates a wide range of scent categories based on local (i.e., action-based) and global (i.e., scene-based) content. Action-based categories are also used for haptics-enhancement generation. The broad support of content categories makes the solution more general and widely applicable.
- A thorough performance evaluation of the proposed solution when deployed with several neural network architectures is performed in this paper, including novel architectures and approaches. Based on the analysis of results of 11 Convolutional Neural Networks (CNNs), it is possible to recommend the best deployment approach for an end-to-end system with automatic scene and haptic content generation involving 360° videos. Additional performance enhancements are investigated: handling scenes that do not require any scents, performance of scents predicted directly instead of from scenes; and increasing olfactory accuracy by merging diverse probabilities (i.e., different associations can be detected related to the same scent).
- As currently there are no benchmark standards for measuring olfactory detection or recognition accuracy, a 360° video olfactory dataset is proposed in this paper, containing approximately 170 clips, which were used for evaluation of results. These clips feature 14 scent categories which are related to 45 scene and action classes. A label dataset was also created consisting of relevant labels for olfactory and haptic effects. The dataset includes reduced versions of the Places and Kinetics scene and action datasets, with 62 and 48 classes, respectively. A total of 54 distinct scent categories and 44 haptic events can be identified by the solution proposed in this paper.
- A real life prototype was built as shown in Fig. 1. It includes a USB-based olfaction dispenser which releases scents to users. Both audio and video features are utilized for haptic generation, incorporating handcrafted audio and deep-learning action-based video features. Users receive haptic feedback (i.e., vibrations) via a haptic mouse when an action with some sort of impact happens, and the 360° content is displayed on an Oculus VR headset.

The proposed solution for automatic generation of mulsemedia content can also be used as a tool for enhancing artistic projects (e.g., 4D cinemas, video games and opera), both for the creators and users.

This paper is organised as follows. Section II describes related works and section III introduces the architecture, components and implementation details of the solution. Section IV presents the testing setup, performance analysis and results. Section V concludes the paper.

II. Related Work

The works related to this paper were divided in three categories. First, mulsemedia systems are analyzed, with a focus on haptics and olfaction. Next, CNN architectures that perform scene and action classification are presented, as well as zero-shot solutions. Finally, existing datasets for action and scene recognition are evaluated.

A. Mulsemedia

VR comprises interactive 3D computer environments that monitor and react to users' positions and actions, providing a sense of immersion in the simulation [12]. There are several works investigating if additional stimuli could benefit the QoE in immersive applications. Some experiments conclude that simply adding more stimuli could overwhelm the senses of the users and even negatively impact their experience and performance if they are not consistent with the content [13].

However, according to a systematic review of 105 studies covering multisensory VR and the impact of haptics, olfactory, and taste cues over audiovisual VR, 84% of the reports demonstrate a positive impact of multisensorial additions, as sensorial components usually cause a larger benefit towards a realistic user experience [10]. The authors also highlight that the vast majority of the studies focus on the use of haptics and a smaller portion examines the impacts of olfaction.

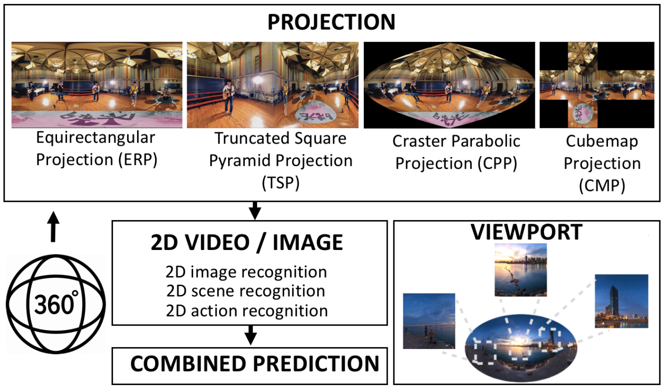


Fig. 2. Flowchart of processing 360° image/video

1) Haptics: In recent years, a number of works described the perspectives and relevance of integrating haptics with audiovisual material to enhance user immersion, creating what is known as Haptic AudioVisual (HAV) content [14], [15].

F. Danieau et al. [16] proposed models to enhance the quality of the video-viewing experience by automatically recognizing and adding haptic content. The authors propose two models: a cinematic model, which aims to make the user feel the camera's movement; and semantic model which renders a haptic effect that is related to the cinematographic effect. The user study shows that the cinematic model creates a significant improvement in QoE compared to no haptic feedback, whereas no haptic feedback was preferred over semantic models.

Several other studies focused on manual haptic content generation (i.e., the haptic effects are synchronized to the main application based on specific actions or timestamps), such as the solution proposed by Darren et al. [17] for educational purposes, and the work from Shafiq et al. [18], which adds haptic content for mobile phones to keep track of soccer matches. In the study of Mazzoni et al. [19], the authors explore the possibility of using haptic feedback to enrich the emotional aspects of a film experience by amplifying the viewer's emotion, recreating specific body reactions through haptic stimulation to convey the feeling of a certain emotion to the viewer (e.g., shivers down the spine to induce fear). A prototype is tested on a manually annotated dataset with emotions, concluding that the intensity of haptic feedback could enhance different sensations. Another solution based on manual annotation is a touchable 3D video system by Cha et al. [20], called Touchable 3D. It allows users to explore the various haptic properties of video scenes, such as stiffness, static friction, dynamic friction and roughness.

Finally, the work of Israr et al. [21] employs manual annotation for haptic content creation. The authors created a plugin for playback and authoring of 360° video content with haptic feedback events. The plugin connects the VR engine to the haptic device and renders the haptic content.

2) Olfaction: Several experiments analyze the impact of adding olfactory stimuli to audiovisual experiences. Adding scents to viewers during playback have reportedly improved user immersion [22]–[25], increased QoE [26] and helped improve user memorization [5].

Comsa et al. [9] proposed a framework based on the PlaySEM platform[1] that consists of a 360° mulsemedia capture and delivery solution on the server-side; and a 360° mulsemedia player prototype on the user-side, enabling users to experience mulsemedia content, including visual, audio, olfactory and wind feedback. Another prototype based on the PlaySEM platform was proposed in [27], extending the scent emitter with a spiral conic shape, so that the scent can flow towards the nose in a more controlled way. The solution is tested on 3 different scenes with a subjective quality assessment evaluating different encoding formats (HD, Full HD, 2.5K, 4K). The results show a statistically significant benefit for the presence of odor and wind in the QoE.

Sexton et al. [11] developed an algorithm to automatically add multisensorial information to 360° videos, combining *hapics and olfaction*. A playback system was designed to improve on the works of Comsa et al. [9] and Bi et al. [28]. The olfaction effects are generated based on automatic scene recognition on 360° videos using Equi-Angular Cubemap (EAC) projection, and the haptic generation is based on digital signal processing of audio cues. For scent generation a ResNet18 network pre-trained on Places365 dataset is used to predict the probability of 365 scene categories, associated with five different scent types (i.e., ocean, oak, candy, chocolate and diesel). The authors exclude the top and bottom video tiles during the prediction, and achieve 61.35% top-1-accuracy and 72.67% olfactory accuracy with ResNet18, using a dataset of MPEG-4 Youtube EAC videos created by the authors. Also, according to authors, users take 1.9 second to detect a change in scents. The haptic content is predicted by a digital signal processing algorithm, which uses a Root Mean Square (RMS) to locate the loudest parts of videos, which trigger haptic feedback.

B. Convolutional Neural Networks for Video Processing

Recent CNN architectures, such as AlexNet, ResNet, DenseNet, VGGNet, RegNet and EfficientNet, have been improving the process of image classification both in terms of accuracy and efficiency. They have evolved to contain more layers and make better use of computational resources.

The RegNet family of architectures aims to improve the effectiveness of neural networks by understanding and creating design spaces that contain a high density of good models that are more likely to be robust and generalize well. The techniques of manual network design and later Neural Architecture Search (NAS) have been used for improving the parametrization and generalization of CNN networks [29].

Traditional CNN models (e.g., AlexNet, ResNet, DenseNet) tend to underperform in comparison to newer models (e.g., RegNet, EfficientNet), but they offer a wider range of publicly available pre-trained architectures. In relation to the work presented in this paper, it is necessary to evaluate the performance of different older and newer CNNs (both pre-trained and not) in the process of image classification for multisensorial output.

Cohen et. al. [30] proposed the theoretical concept of Spherical CNNs for 360° content. Authors focus on minimizing the effects of video distortion, classifying 3D objects and consider

[1] https://github.com/lprm-ufes/PlaySEM

the rotations of the immersive space. However, the main pipeline for working with 360° content consists of classifying 2D faces or tiles of the main image separately, as CNNs for 2D content are widely available. As seen on Fig. 2, initially, either a view-port or a projection based pre-processing is utilized to obtain the 2D content. The 2D images are then processed to extract the predictions, and they are finally combined to obtain the 360° merged prediction [31]–[33].

Scene recognition is relevant for the generation of olfactory effects, as several scene classes correspond to olfactory-related concepts. Zhou et al. [34] propose 4 state-of-the-art scene recognition networks (i.e., Places-CNN) trained on the Places205 and Places365-Standard datasets. The four networks are based on AlexNet, GoogLeNet, VGG16 and ResNet152.

The earliest works on *action recognition* considered actions as spatiotemporal objects, and tried to capture them via handcrafted spatiotemporal filters, such as histograms of gradients and cuboids [35]. Another approach for action recognition relies on optical flow [36]. It similarly uses handcrafted features, such as histograms of flow and motion boundary.

Deep learning methods started to be used for action recognition, with the use of 3D convolution and 3D pooling in CNN layers to capture features in spatial and temporal dimensions. 2D optical flow maps have also been applied to extract motion features as an input for predictions [37].

Feichtenhofer et al. [38] proposed a video recognition model called SlowFast, which consists of two different pathways, a slow pathway that operates at a lower framerate and captures high spatial and semantic details and a fast pathway that operates at a high framerate and can capture rapid motion at a finer temporal resolution. SlowFast provides several network architectures with different input sampling and different backbones (ResNet50, ResNet101 and Nonlocal). SlowFast 16x8 on ResNet101 achieved a 78.9% top-1-accuracy and 93.5% top-5-accuracy on Kinetics400, using 213 GFLOPs for 30 views. On Kinetics600, the same architecture achieved a 81.8% top-1-accuracy and 95.1% top-5-accuracy, which outperformed the top-1-accuracy of the previous winner of 2018 ActivityNet Challenge by 2.1%.

Zero-shot learning is the study of generalizing and recognizing object categories that were unseen during training in image classification. One example of a zero-shot network is Contrastive Language-Image Pre-Training (CLIP), which jointly relies on vision and language, using natural language for learning concepts in the vision domain. The model consists of a robust image and text encoder that was trained on a large-scale dataset created by the authors, called WebImageText (WiT). The network predicts which encoded label has the highest likelihood with the encoded image by comparing the similarity of the encoded latent vectors. This way, the feature extractor does not need to be retrained for every dataset, only the output must be restricted to the possible classes. CLIP can outperform the baseline Visual-N-Grams zero-shot predictor by a large margin. Tests on the ImageNet dataset showed a 76.2% of accuracy of CLIP compared to 11.5% of Visual N-Grams; and tests on the SUN397 dataset indicated a 58.5% of accuracy compared to 23.0% of Visual N-Grams [39].

At the time of writing, CLIP is the highest performing zero-shot classifier both on scene (SUN397) and action (Kinetics700) recognition tasks. Therefore, by combining the olfactory-relevant scene and action labels, CLIP is one of the possible networks that can be used for the proposed multisensory solution without extensive training.

C. Datasets

1) Scene datasets: Places [34] is a large scale dataset with more than 10 million images, with 365 scene semantic categories. SUN397 [40] is another scene-centric dataset offering 397 scene classes, but for many of them it fails to provide enough samples per class for large scale training of deep neural network. In this work, the Places365 is used in conjunction with its publicly available pre-trained networks provided by the authors (e.g., AlexNet, GoogLeNet, VGG16, ResNet18 and ResNet50) and by the research community (e.g., SqueezeNet1.0, MobileNetV2 and DenseNet161[2]).

2) Action datasets: The UCF101 [41] dataset was one of the first dataset providing a wide range of human activities. However, it only contains a few samples per class. Charades [42], a dataset that contains videos of indoor activities, has more classes than the UCF101 and has enough samples per class for successful training of convolutional neural networks.

Something-something v2 [43] provides classes of humans performing basic actions with everyday objects, with sufficient samples for training complex models. However, the dataset contains very few classes that could be associated to scents.

Finally, the Kinetics [44] dataset contains a large collection of videos, with a broad range of categories. Kinetics400 was the first version of the dataset, and is the most widely used for training networks. Kinetics600, the second version of the dataset, is commonly used for testing and recent research focuses on extending the training using this model. Kinetics700-2020, which improved categories balancing, is the newest and most complex benchmark, however, there are very fewer pre-trained networks available.

The works presented in this section indicate that the addition of multisensory output to immersive content enhances viewer experience. CNNs can accurately perform action and scene recognition for the automatic generation of multisensory effects from multimedia inputs, as manually adding these effects is a lengthy process. The work described in [11] proposed an initial CNN-based solution for multisensory systems, but left several questions to be answered: how action detection can improve the effects dispensed, what other metrics can be analyzed to indicate the feasibility of the solution (e.g., GFLOPs, FPR, FNR), and how the solution can be generalized to work with a large number of videos, categories, scents and datasets. Several CNNs were also described in this section and a thorough evaluation needs to be performed for the identification of a suitable network to work with the proposed solution and help achieve best results in terms of performance, complexity and accuracy.

[2] https://github.com/HuaizhengZhang/scene-recognition-pytorch1.x/blob/master/model_zoo.md

III. Solution Design

This section describes the proposed solution, which is capable of generating a wide range of scent categories based on local (action based) and global (scene based) content, as well as introducing haptics components based on action detection.

A. Prototyping environment

The prototype contains a laptop with an Intel i7-7500 CPU with 8GB RAM and running Windows 10 and a desktop PC with an AMD Ryzen 2070 GPU running Ubuntu.

The end-to-end solution consists of a VR headset, an olfactory dispenser, and a haptic device. The Oculus Quest 2[3] VR headset is used for visualizing the 360° videos. The headset needs to be connected to the same network of the server that controls the haptic and olfaction devices.

The haptic feedback was provided by the Rival 700 haptic USB mouse[4] by SteelSeries, controlled with the HTTP API of the SteelSeries 3 Engine.

The olfactory effects are generated by the SBi4v2 USB scent dispenser manufactured by Inhalio[5]. The device contains 4 slots for scent cartridges, each with a fan for diffusing scents towards users. An API is used to control the fans with the desired scents, the duration of the effects, their intensity and the recurrence of the scents. The device API is controlled by a Java server with the Olfactory API, created for the prototype.

During the experiments, a Python 3.7 environment was developed, with PyTorch 1.8, Keras, and Tensorflow 2.1. The end-to-end system uses Python for the web server side and JavaScript and Java (32-bit v8.241) for the mulsemedia player. The OlfactionAPI that operates the scent dispenser uses Java, and the web server is based on JavaScript and requires HTTPS security (implemented via OpenSSL[6]). The player employs the WebVR[7] JavaScript library. Youtube-dl[8] is used for downloading the dataset, while the videos pre-processing was done via FFmpeg[9].

The next subsections describe the methodology employed in this work. The paper presented in [45] introduced a methodology for evaluating and designing mulsemedia applications, and its guidelines were followed during the evaluation of the capacity of the proposed solution. The development of the approach started with selecting datasets with classes related to scents and actions, as well as choosing a subset of existing CNN architectures to train on the selected datasets (III-B). Once datasets were prepared (III-C) and the network architectures were chosen, the hardware and software stack of the solution was defined (III-D) and the training process began. Following that, it was specified how the solution should employ CNNs to generate scents and haptic feedback based on the scenes and actions detected (III-E). Finally, evaluation metrics were defined in order to assess the performance and accuracy of the approach (III-F).

B. Selection of CNNs and Datasets

As reviewed in the related works section, 360° videos can initially be projected into two-dimensional images, where traditional CNN methods can be used for prediction. In the proposed solution, the EAC projection is employed, since it minimizes video distortions, leading to faster fine-tuning on pre-trained networks. The choice of EAC projection can have a negative impact on the accuracy of action detection for both haptics and scent, in case the relevant action happens at the boundary between two tiles. The double cube-map projection proposed in [46] avoids this problem by projecting the image to two cubes, where one of the cube centers is tilted 45° in the horizontal and vertical plane, but it would actually double the amount of samples to process. In order to prevent an increase in complexity, EAC projection is adopted, as this offers the best compromise between accuracy, simplicity and data processing requirements. Action and scene detection are complementary to each other and applicable to two-dimensional frames, with scene information providing a general perspective of the environment in a passive manner, and actions being localized and based on dynamic changes.

The Places dataset contains relevant categories for scent detection. Therefore, the baseline networks provided by the authors of Places (i.e., ResNet, AlexNet and DenseNet) are tested for the proposed solution as well as other pre-trained networks, such as SqueezeNet and MobileNet.

Additionally, three recent network architectures from the RegNet family were trained "from scratch" on the high-resolution version of the Places dataset for scene recognition, in order to explore the networks' speed and accuracy trade-off, and compare the performance of these novel architectures with their classical counterparts. The three selected RegNet models consist of a small model, RegNetY-800, a medium one, RegNetY-1.6GF, and a larger model, RegNetY-3.2GF, and they achieved a top-1 accuracy of 35.22%, 38.82% and 35.56%, respectively. The RegNet models were adjusted with a process of freezing the body and fine-tuning only the head of the network. Due to these models achieving similar accuracies in this test, RegNetY-800 will be selected for further tests with the proposed solution, due to its lower complexity.

In Section II-C2, Kinetics600 was identified as a suitable action-centric dataset for the proposed solution. The pre-trained baseline models from the SlowFast family provide efficient high-accuracy models.

The construction of scent categories followed four steps:

1) Choosing the useful datasets that contain many scent-related classes based on the research of the market of scent providers.
2) Select the labels from the datasets that correspond to a scent and discard the ones that have no correspondence (e.g., army base, boardwalk) or are too general (e.g., cafeteria, dinette).
3) Select the scents that can be detected by computer vision tools. Exclude scents that express a general feeling or

[3] https://www.oculus.com/quest-2/

[4] https://steelseries.com/gaming-mice/rival-710

[5] https://inhalio.com/

[6] https://www.openssl.org/

[7] https://webvr.info/

[8] https://youtube-dl.org/

[9] https://github.com/FFmpeg/FFmpeg

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3157556, IEEE Transactions on Multimedia

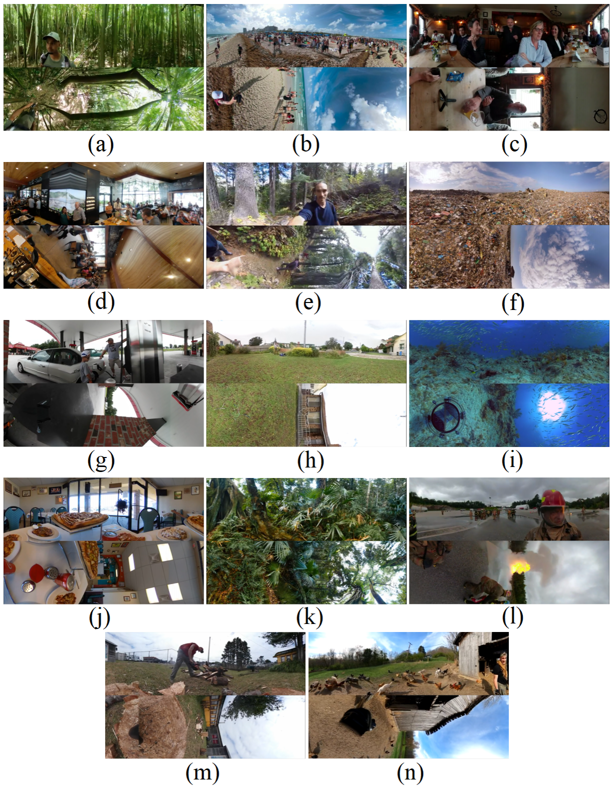


Fig. 3. Thumbnails from each class of the video dataset.

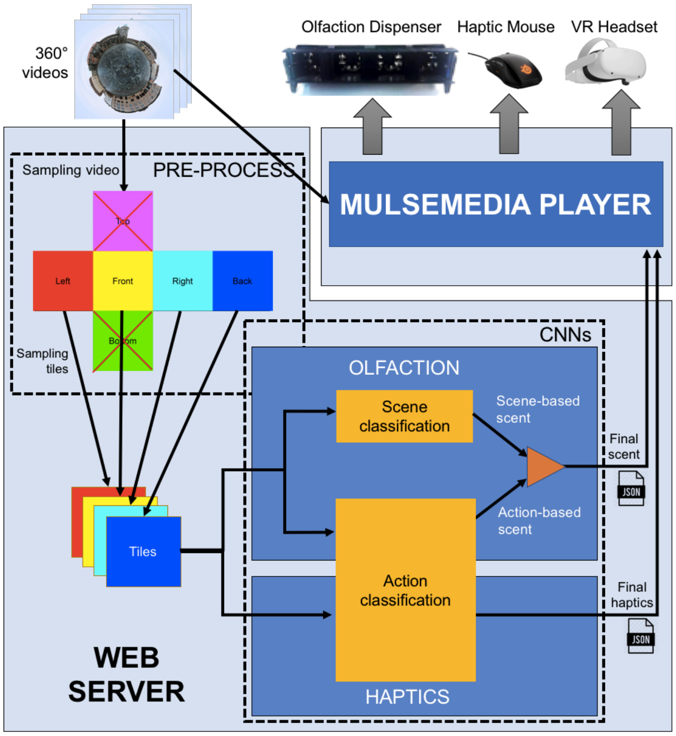


Fig. 4. Scene and action CNN-based architecture for mulsemedia.

mood, which are hard to define, or represent classes that are hard to differentiate from each other (e.g., Boston cream pie scent, buttercream pecan scent, etc.). The mixture of smells were also excluded (e.g., blackberry and lavender smell, banana leaf and açaí, etc.).

4) Assign scents (i.e., the ones available in cartridges for the olfaction device) to labels that can be detected by CNNs (one scent category can be assigned to multiple class labels).

The subjective process of selecting scents led to Places365 being reduced to a dataset[10] of 62 classes and Kinetics being reduced to 48 classes, with a total of 54 distinct scent categories. After that, the candidate networks' performances are evaluated on the reduced dataset (discussed in Section IV-A), which only contains olfactory relevant classes selected previously. The metrics used for evaluation are discussed in Section III-F, and based on the results it is possible to recommend which CNN is suitable for the multisensorial solution.

In addition, the zero-shot prediction accuracy of CLIP was tested with crafted prompts for the labels to increase the accuracy on the reduced dataset.

In Section II, different works related to haptics were presented, in which the effects are provided based on user emotions, optical flow and sound. None of these works consider action detection-based haptics generation, which is possible in the proposed solution due to the adoption of action detection with an approach such as SlowFast.

Haptic events are be classified into two categories in an additional dataset[11]: constant stimuli and single stimuli. In constant stimuli, effects are triggered throughout the entire action, while in single stimuli, the haptic feedback happens at a key moment. In both scenarios the whole action is processed by the network, but for single stimuli, audio cues are used for the recognition of the key moment. A total of 44 haptic events can be detected.

C. Configurations of the Dataset and Labels

The testing dataset containing 360° video segments[12] and a setup guide[13] are publicly available. The dataset was created to evaluate the performance of end-to-end solutions with 360° content, as such a dataset was not previously available. It contains 170 ten-second clips of 360° videos from YouTube, with manually labeled scent categories for 14 olfactory classes, as shown in Fig. 3.

The video dataset is used in conjunction with a configuration file[14], which is a CSV containing the scent or haptic labels, links to videos, and start and end times of the segments. The setup guide also contains the path of the Python script which uses the CSV file for locating and downloading the videos in mp4 format and converting them into EAC projections. Once segments are downloaded, the scripts[15] for the scene and action recognition can be executed.

[10] https://github.com/Fjuzi/traction_base/blob/main/scents.pdf

[11] https://github.com/Fjuzi/traction_base/blob/main/haptics.pdf

[12] https://anon.to/FkB86G

[13] https://github.com/Fjuzi/traction_base/blob/main/data/DATA.md

[14] https://github.com/Fjuzi/traction_base/blob/main/data/DATASET.csv

[15] https://github.com/Fjuzi/traction_base

TABLE I
COMPLEXITY OF DIFFERENT ARCHITECTURES

	Values from literature				Our measurement of inference	
Architectures	params [M]	GFLOPS [B]	Inference [s] [29]	ImageNet top-1 error	CPU [s]	GPU [s]
RegNetY-3.2GF	19.4	3.2	0.070	21.0	0.1500	0.0133
RegNetY-1.6GF	11.2	1.6	0.039	22.0	0.0928	0.0130
RegNetY-800MF	6.3	0.8	0.022	23.7	0.0576	0.0079
RegNetY-600MF	6.1	0.6	0.019	24.5	0.0488	0.0077
RegNetY-400MF	4.3	0.4	0.019	25.9	0.0438	0.0077
RegNetY-200MF	3.2	*0.2*	**0.011**	29.6	**0.0277**	0.0064
EfficientNet_b3	12.0	1.8	0.114	22.5 (**18.9**)	0.1513	0.0125
EfficientNet_b2	9.2	1.0	0.068	23.4 (20.2)	0.5205	0.0107
EfficientNet_b1	7.8	0.7	0.052	24.1 (21.2)	0.4030	0.0097
EfficientNet_b0	5.3	0.4	0.034	24.9 (23.7)	0.2683	0.0068
MobileNet v2	4.2	0.6	-	29.4 (27.1)	0.1826	0.0049
ResNet50	22.6	4.1	0.053	35.0 (22.2)	0.1938	0.0125
ResNet18 [47]	11.0	2.0	-	28.2	0.0741	**0.0032**
DenseNet161 [47]	28.7	8.0	-	23.8	0.4063	0.0210
SqueezeNet 1.0 [47]	*1.3*	0.8	-	-	0.0549	0.0033
ViT-B-32 [47]	**88.0**	**13.0**	-	26.6	0.1431	0.0067

The haptic and scent labels related to actions and scenes are also configured via CSV files[16]. These files are uploaded to the server with the desired effects to be executed when a certain scene or action is detected. The file contains the names of events (e.g., cutting apple, rainforest), event ID, effect name (e.g., apple, wood), effect ID, and event type (i.e., action or scene). By changing the effect names and IDs, it is possible to customize the effects according to user preference. Future work may consider adding personalization to automatically generate such content based on user preferences.

D. Architecture of the Proposed System

The architecture of the proposed solution, presented in Fig. 4 consists of two parts: The *web server* and the *mulsemedia player*. The web server is responsible for the haptic and olfactory annotation, and the mulsemedia player distributes generated content. The web server processes 360° videos in advance, and once it generates the annotations, they are sent to the mulsemedia player.

The web server contains 360° videos in the EAC format and pre-processes them. The pre-processing step consists of sampling the video frames and separating the different tiles based on the cubemap projection. From each frame, the front, back, right and left tiles are used, and the top and bottom are discarded, as they usually do not contain relevant information for the predictions. The bottom tile usually shows the ground while the top tile usually contains the sky or a ceiling. The CNNs, then, perform scene and action recognition, generating annotations for the given frame, stored in JSON files.

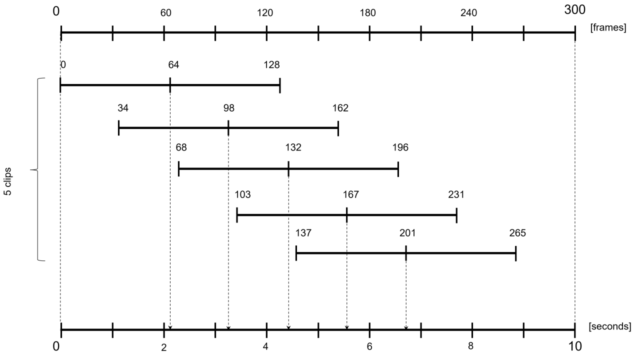


Fig. 5. The sampling of clips for action and scene prediction.

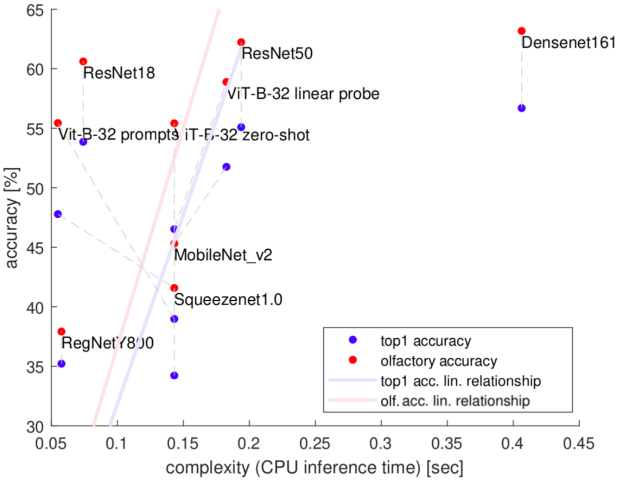


Fig. 6. Relation of top-1 accuracy (blue) and olfactory accuracy (red) with model complexity (measured in CPU inference time) on the reduced dataset.

The *olfactory content generator* of the web server consists of two parallel networks, one recognizing the scene (i.e., general or background content) and the other the actions (i.e., local or foreground content), and based on that, predicts a scene-related scent and an action-related scent, if detected. When detected, action-related scents have a higher priority than scene-related scents, as they tend to grasp users' visual attention due to their dynamic, fast-changing nature over the static, scene-related background.

The *haptics generator* of the web server detects actions associated to haptic stimuli. Based on the nature of the action, it can either be a constant vibration, or just a short vibration

[16] https://github.com/Fjuzi/traction_base/blob/main/configs/labels.csv

TABLE II
THE TOP-1, TOP-5 AND OLFACTION ACCURACIES [%] OF SEVERAL ARCHITECTURES ON THE REDUCED PLACES DATASET

Architectures	Top-1 Accuracy	Top-1 Error	Top-5 Accuracy	Top-5 Error	Olfactory acc.	FPR	FNR
RegNetY-800MF	29.13	70.87	58.16	41.84	37.92	11.2	55.57
ResNet18	48.63	51.37	80.90	19.10	60.61	*7.43*	32.95
AlexNet	43.79	56.21	74.90	25.10	55.52	8.53	38.26
ResNet50	*50.29*	*49.71*	*81.93*	*18.06*	*62.22*	**7.11**	*31.80*
CLIP (ViT-B-32) zero shot	30.50	59.50	57.66	42.34	41.58	9.51	54.05
CLIP (ViT-B-32) prompts	36.92	63.08	67.89	32.11	45.31	8.81	48.89
CLIP (ViT-B-32) linear probe	42.19	57.81	-	-	55.41	9.20	37.40
SqueezeNet1.0	46.66	53.34	77.14	22.85	55.43	9.06	37.80
MobileNet_v2	50.03	49.97	81.32	18.68	58.89	8.27	34.73
DenseNet161	**54.60**	**45.41**	**85.35**	**14.64**	**63.18**	**7.11**	**30.73**

that is enhanced by audio signal.

The mulsemedia player, built with JavaScript, plays the 360° videos on the VR headset synchronously with the olfactory and haptic content, activating the scents dispenser and haptic mouse via USB. The player forwards the JSON files received from the web server with olfactory and haptic information through WebSockets to the Olfactory API, that operates the scent dispenser, and to the API of the haptic mouse. The player is WebVR-based and users access it on the browser of the VR headset. Once playback starts, the server synchronously sends the haptic and olfactory information to the USB devices.

E. Scene and Action Detection Process

Action recognition in videos requires temporal windows, while scene recognition can be performed in each frame. The length of the temporal segments is based on the findings of Sexton et al. [11], as well as the constraints imposed by the SlowFast architecture. In [11] an experiment suggested that users took, on average, 2 seconds to notice a change in scents generated by the olfaction dispenser. Therefore, predictions must be performed at least once every 2 seconds. The original implementation of the SlowFast architecture requires segments to be spaced by 34 frames (i.e., approx. 1.13s for 30 fps videos such as the ones in the dataset). Therefore, 1.13s is the interval employed in this solution, as it also fulfills the requirement of detections being less than 2 seconds apart.

Furthermore, the SlowFast network employed in action detection requires as input a batch of 64 frames, with a sampling rate of 2. Therefore, each prediction on SlowFast covers a time range of 128 frames (i.e., 4.27s for the videos in the dataset). Finally, the final second of each video clip is discarded to avoid the impact of fading effects on the detection. With these constraints, and using video clips of 300 frames (i.e., 10 seconds), such as the ones in the testing dataset, this leads to splitting the input video into 5 separate and overlapping views, as shown in Fig. 5.

The scene prediction process does not require processing a batch of frames over a temporal window, as it only requires one frame. The middle frame of each of the 5 views is the one selected to be processed for two reasons: first, this allows running the same number of predictions for both scene and action recognition, and secondly, choosing the middle frame of each view guarantees that selected frames are evenly distributed across the clip, while also avoiding frames at the beginning or end of the video sequence, which may contain scene transitions.

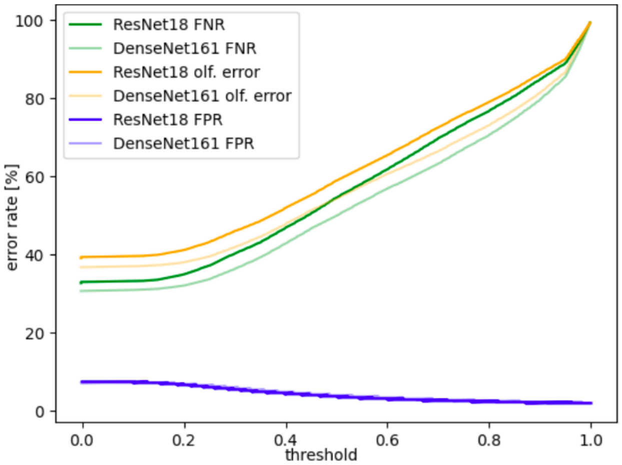


Fig. 7. FPR, FNR and olfactory accuracy for DenseNet161 and ResNet18

Afterwards, both the action and scene classification networks classify each of the 5 segments. Each segment contains 4 horizontal tiles to be processed, resulting in 20 predictions for each clip, both from the action and scene classifiers. The probabilities generated by the scene and action classifiers are summarized for each tile, and the tile with the highest probability above the selected threshold serves as the final prediction for the scene and action. If an action is detected, its prediction always overwrites the prediction of the scene. The final scent and its timestamp are encoded into a JSON file, which is forwarded to the mulsemedia player.

F. Evaluation Metrics

A number of metrics are used for the evaluation of the proposed approach. *Top-1 and top-5 accuracies* indicate if the prediction matches the true label, or if the 5 highest probability predictions contain the true label, respectively. The *olfactory accuracy* determines the number of correctly identified scent categories, which can be higher than the top-1 accuracy, as similar scene labels can be related to the same scent.

A confusion matrix was created in a one-vs-all manner and the following metrics are used for result analysis: the *False Positive Ratio (FPR)* is the percentage of scents diffused wrongly, when there should be a different one or no scent. The

False Negative Ratio (FNR) is the ratio of scents not being diffused, when there should be a scent.

IV. Testing and Results

In order to test the performance and feasibility of the CNN-based generation of mulsemedia effects, a number of experiments were performed. Multiple CNN architectures and datasets were tested in terms of accuracy and complexity for scene and action recognition, including a zero-shot approach.

A. Evaluation of Scene-Based Scent Recognition

A number of CNN architectures were tested for scene detection. Table I contains performance measurements from several CNNs on the ImageNet validation dataset. Some architectures were trained from scratch (e.g., RegNetY), while others had available pre-trained weights (e.g., ResNet18, AlexNet, ResNet50, SqueezeNet1.0, MobileNetv2, DenseNet161).

Regarding the values provided in Table I, the inference times are taken from the work presented in [29], using an Nvidia V100 GPU. Additionally, we calculated ImageNet top1-errors for RegNet, EfficientNet, MobileNet and ResNet50 architectures, while the remaining values, including the ones in parenthesis, are from [47], where the state-of-the-art performances on ImageNet are collected.

Additional accuracy tests were performed on a reduced version of the Places365 dataset (which contains olfactory relevant classes, as described in Section III-B). As indicated in Table II, RegNet underperformed in comparison to other models. This is likely due to the training being performed by us from scratch on a more modest GPU than the ones used for pre-trained models. However, this is helpful to demonstrate the performance of a newly trained network on widely available hardware.

Table II also indicates that CLIP zero-shot (employing the ViT-B-32 model proposed in [48]) achieves a similar accuracy to RegNetY-800MF, which shows the potential of this network. When using zero-shot and prompts, CLIP's performance (i.e., 30.5% and 36.92% top-1 accuracy, respectively) exceeds RegNetY-800MF. Prompts are better-designed class names to describe scenes, which improved accuracy by approx. 6%. Certain modifications were applied to these prompts, such as replacing underscores with spaces; using articles at the beginning of classes, use of full stop, and label rephrasing (e.g., the class label "florist_shop/indoor" is replaced by "an indoor photo of a florist shop."). Furthermore, using a linear probe on CLIP features notably increases the top-1 accuracy to 42.19%, almost matching the performance of a model such as AlexNet, trained using millions of images. The linear probe uses a Logistic Regression function that requires a hyperparameter sweep of the regularization strength. The highest accuracy was achieved with a Regularization Parameter Value of 0.0492.

The best performing models, however, are still the traditional ones, which were trained with a large amount of data. More complex models, such as Densenet161, perform better. Fig. 6 illustrates in blue the relation of top-1 accuracy and complexity. DenseNet161 is an outlier, achieving the highest accuracy but at the cost of a significantly higher complexity. The light blue and light red lines represent the regression of the data points (excluding DenseNet161) and it shows a linear relationship between complexity and accuracy.

Table II indicates that olfactory accuracy is higher than the top-1 accuracies. This can also be seen on Fig. 6, where the relationships between the model inference times and olfactory accuracies are marked in red, and indicate a similar correlation to the top-1 accuracy.

ResNet18 demonstrates competitive performance compared to DenseNet161, while having significantly better inference time. Employing this model over DenseNet requires a sacrifice of approx. 2.5% olfactory accuracy but it allows approx. 6 times faster inference time on a GPU. RegNet provides lower performance with a similar inference time, and might not be the best choice for the olfaction generation approach.

1) False Postive and False Negative Rates: The changes in FPR and FNR were also investigated with the use of different cut-off thresholds for classification. In the proposed solution, FPR and FNR are important metrics since the FPR represents the percentage of incorrectly generated scents, and FNR the percentage of times that no scent was generated when there should have been a scent.

Generally, it is desired that the FPR is as low as possible as it can disturb the user and diminish the general experience, whereas false-negative detections, while not enhancing the user experience, do not lower it either.

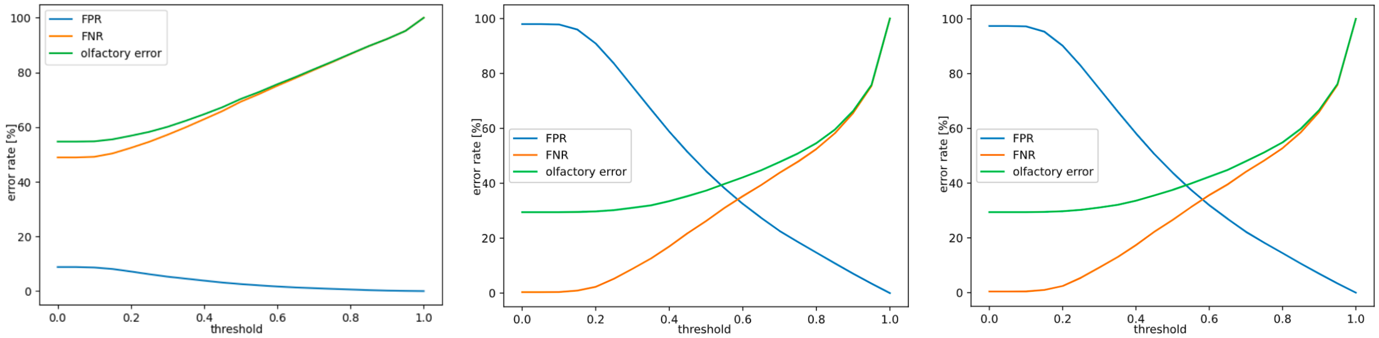
Fig. 7 illustrates the FPR, FNR and olfactory accuracy of DenseNet161 and ResNet18 architectures. The figure indicates that the FPR is less than 10% in both cases with zero threshold (e.g., 7.11% for DenseNet161 and 7.43% for ResNet18), which demonstrates the robustness of these networks. The confidence of false-positive classes is approx. 0.2, so this is the threshold when the FPR starts to decrease and the FNR starts to increase.

In terms of FNR and olfaction error, ResNet18 performs slightly worse than DenseNet161. However, considering the significantly lower inference time, ResNet18 is the most suitable network for the solution, among the ones tested.

B. Additional Experiments For Scene Detection

In order to thoroughly test and achieve a feasible solution, further experiments were performed. These experiments are related to scenes that do not require any scents in CLIP, scents being predicted directly instead of scenes and increasing olfactory accuracy by merging classes probabilities.

1) Scenes with no relevant scent: CLIP zero-shot was tested with different 'no scent' definitions on the reduced Places dataset, as seen in the plots from Fig. 8. In Fig. 8 (a), all classes from the model are predicted, with the ones not corresponding to any scent being classified as "no scent". In (b), only scene-related classes are predicted and a threshold is defined. If no class is detected with a probability above the given threshold, the scene is classified with "no scent". In (c), only scent-related classes are predicted and an extra class is labeled as "no scent". Based on the FPR and FNR values, scenario (a) is the best performing one, with an 8.8% FPR, 48.9% FNR and 45% olfactory error at threshold 0, while the other scenarios



(a) Irrelevant scene categories are classified as ‘no scent’ (all scene classes predicted)

(b) Low probability classes are classified as ‘no scent’ (only olfactory relevant classes predicted)

(c) Additional ‘no scent’ class (only olfactory relevant classes predicted + ‘no scent’)

Fig. 8. Scene prediction with different ‘no scent’ definitions for CLIP.

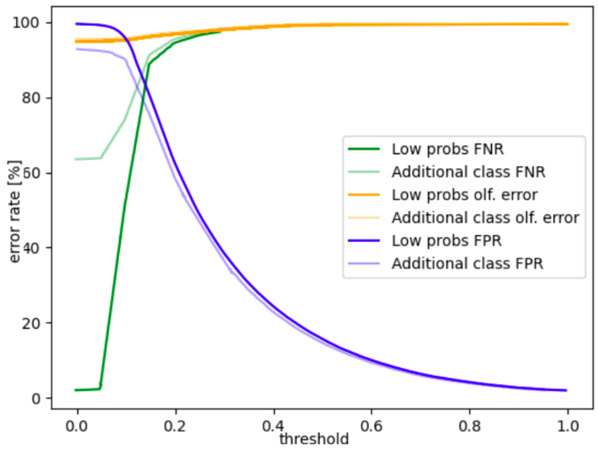


Fig. 9. Direct prediction of scent with two different no scent class definitions: low probabilities are classified as ‘no scent’ and additional ‘no scent’ class.

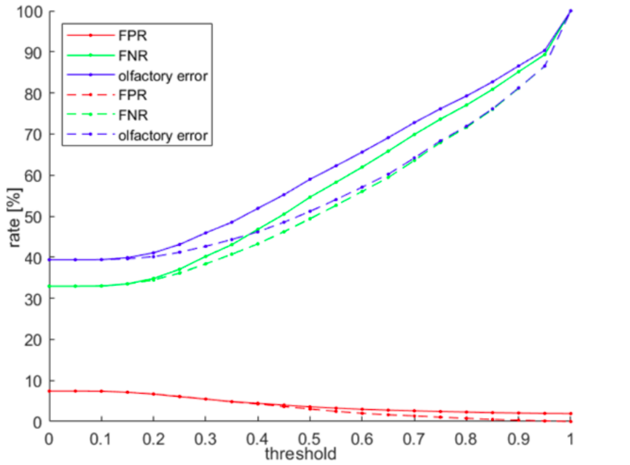


Fig. 10. The impact of merging prediction probabilities of classes that correspond to the same scent category.

yield a high FPR, which can be improved by increasing the threshold, affecting negatively the FNR and olfactory error. Furthermore, keeping all the classes is beneficial, as it requires fewer modifications on the original networks.

2) Predicting scents directly: An additional test with CLIP zero-shot was also preformed, with scents being detected directly instead of scenes. As indicated by Fig. 9 an additional ‘no scent’ category has a negligible impact on the performance, similarly using low probabilities to indicate that there is ‘no scent’ in a certain scene. The FPR rate and olfactory error are high in both approaches regardless of the definition of the ‘no scent’ category. Increasing the threshold successfully reduces the FPR, but fails to decrease the olfactory error to an acceptable level. The low performance of predicting scents directly is likely due to the abstract nature of the labels. The authors of CLIP also acknowledge that zero-shot CLIP has a limitation of being biased by the distribution of the training data, affecting the prediction of new unknown categories, such as predicting scent labels which were not present in the original training data.

3) Increase Olfactory Accuracy by Merging Class Probabilities: In order to reduce the classification of scenes as ‘no smell’, it is possible to observe all probabilities of the different detected labels, as multiple predicted labels might correspond to the same smell. In that case, the probability of the scene labels that correspond to the same scent must be summed, and if the summed confidence is higher than the given threshold, the scene is considered to contain that scent. Fig. 10 shows the impact of merging prediction probabilities that correspond to the same scent categories on ResNet18. The straight lines stand for the original scenario’s error rates, and the dashed lines show the error rate of the merged classes probabilities. The figure indicates that the merging successfully decreases the olfactory error rate by approx. 10%, especially in the probability threshold between 0.3 and 0.9. This proves that individual class predictions were incorrectly classified as ‘no scent’ due to their low probability.

C. Evaluation of Action-Based Scent Recognition

Table III presents the performance of action detection for SlowFast64x2 and CLIP zero-shot the reduced Kinetics dataset. The reduced dataset based on Kinetics600 contains 48 classes related to olfaction, as described in Section III-B. The table indicates that SlowFast performs very well on the reduced dataset, with high olfactory accuracy and low FPR. The top-1 and top-5 accuracies are 75.56% and 93.12%, respectively. In this experiment the ‘no scent’ class was defined accordingly to our findings in Section IV-B1. CLIP zero-shot with prompts maintains a promising performance for a zero-shot predictor, with a 75.26% top-5 accuracy.

TABLE III
PERFORMANCE OF ACTION RECOGNITION NETWORKS WITH THE REDUCED KINETICS DATASET FOR SCENTS

	Top1 acc.	Top1 err.	Top5 acc.	Top5 err.	Olf. acc.	FPR	FNR
SlowFast64x2	75.56	24.44	93.12	6.88	76.45	1.55	18.86
CLIP (prompts)	49.31	50.69	75.26	24.74	53.38	3.97	38.81

D. Evaluation of Scent Prediction Based on Scene and Action

A number of tests evaluated the performance of the end-to-end system on the proposed dataset. Fig. 11 illustrates the performance of a smaller and faster network (i.e., ResNet18) and a slower but more accurate network (i.e., DenseNet161) for scene detection with SlowFast64x2 as their parallel counterpart for action detection. As DenseNet161 was the highest-performing network in the earlier experiments, it indicates how accurate the solution can be when not constrained by computational power, while ResNet18 offers the best balance of accuracy and complexity trade-off among the tested networks. When the prediction probability threshold is set at 0, DenseNet161 achieved a 74.27% olfactory accuracy, with 6 false positive detections out of 172 samples. ResNet18 achieved a 69.19% olfactory accuracy with 10 false positive detections. These mismatches are inherently coming from the underlying dataset and the different distribution of the training and test dataset. In addition, not all false positive detections are equally poor. ResNet18 confused forest with rainforest three times, pizza with smoke twice, coffee with burger and beer with burger. These categories were also present in the same videos and thus they could be a fitting scent.

The impact of different combinations of action and scene thresholds was also evaluated. The results can be seen in Fig. 12. As the probability threshold increases, the FPR slightly decreases, but the FNR and thus the olfactory error strongly increases. With a low threshold, the FPR is sufficiently low, therefore both the action and scene probability thresholds are recommended to be set to 0.

The class-wise olfactory accuracy of ResNet18 was also evaluated. Fig. 13 presents the performance of the network on some of the classes, and indicates a strong performance on most of the classes. In some specific classes (e.g., smoke and coffee), the accuracy is significantly lower.

1) Using a Zero-shot solution instead of a parallel approach: Figs. 14 and 15 show the performance of CLIP zero-shot with prompts classifier on the evaluation dataset. Fig. 14 illustrates the olfactory error and the FNR (left axis) and the FPR (right axis) as a function of the probability threshold.

The threshold was set to 0.6 (matching FPR and olfactory error) for comparisons with the parallel system (i.e., SlowFast and ResNet18). With this threshold the olfactory accuracy of CLIP is 63.17% with 33% FNR. Even though the average olfactory accuracy is approx. 6 percentage points lower than the parallel system, as seen in the yellow and purple horizontal dashed lines in Fig. 15, this performance is remarkable for a completely zero-shot predictor.

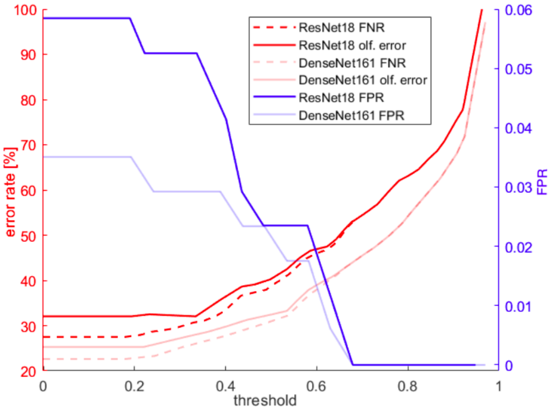


Fig. 11. The FNR and olfactory error (left axis) and FPR (right axis) as a function of probability threshold of ResNet18 and DenseNet161 on the evaluation dataset.

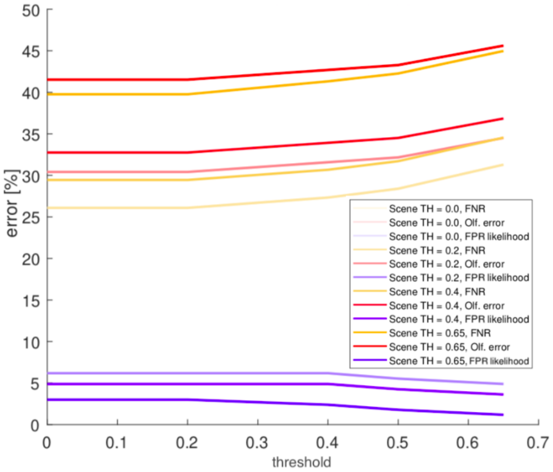


Fig. 12. Probability thresholds of the parallel system.

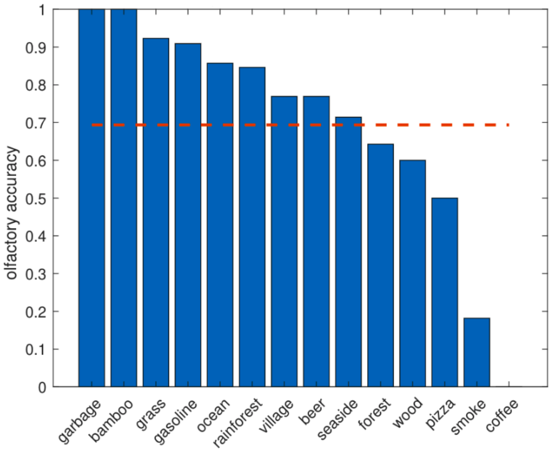


Fig. 13. Class-wise olfactory accuracy of the parallel scent prediction system with ResNet18 and SlowFast64x2 on our 360° video scent dataset. The red dashed line shows the average olfactory accuracy.

Fig. 15 presents class-wise performances of CLIP and the parallel system. CLIP performs well in certain classes where the parallel system has low performance (i.e., coffee) and poorly in others (i.e., village and wood). There is also a class (i.e., pizza) in which both networks do not perform well. Examples of unsuccessful classification are shown in Fig. 16, where from left to right three frames are presented from the classes coffee, smoke and pizza, which are classified as burger, grass and smoke, respectively. The most likely explanation for

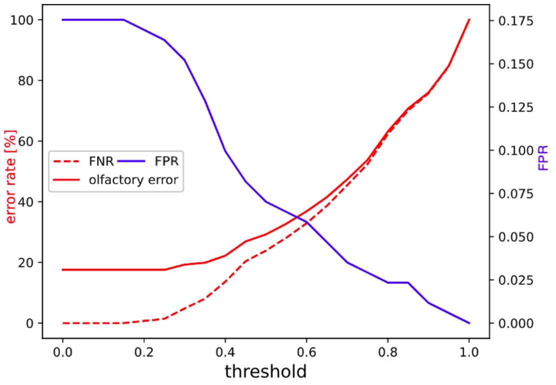


Fig. 14. FNR and olfactory error (left axis) and FPR over all samples (right axis) as the function of probability-threshold of CLIP on the proposed dataset

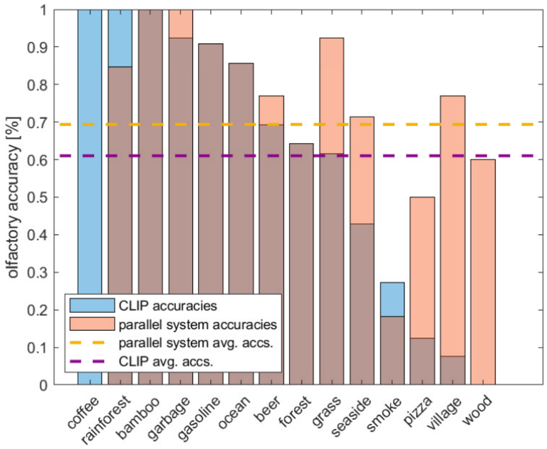


Fig. 15. Class-wise performance of CLIP zero-shot classifier, with probability threshold 0.6, when its FPR matches the FPR of the parallel system.

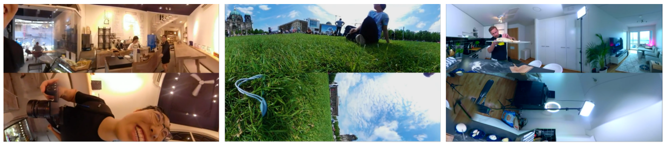


Fig. 16. Three frames incorrectly classified by the scent prediction system. From left to right, the frames are instances of the classes “coffee”, “smoke” (i.e., a person smoking) and “pizza”.

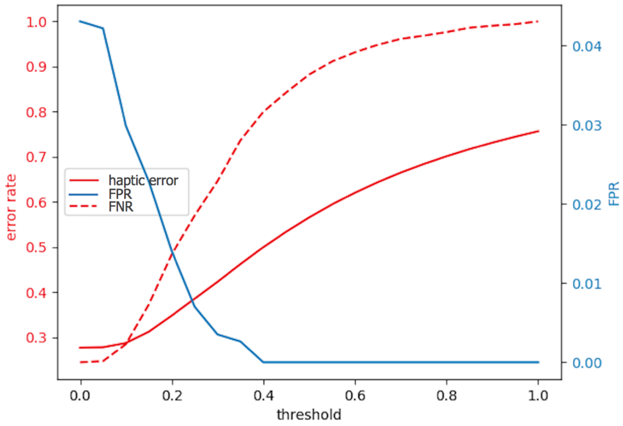


Fig. 17. FNR and haptic error (left axis) and FPR over all samples (right axis) as a function of probability-threshold of SlowFast64x2 on the reduced Kinetics dataset only with haptic related labels.

the unsuccessful classification related to the class smoke, is that the objects associated with this specific class occupy a very small part of the image (e.g., the second image in Fig. 16 is an extreme example, as the cigarette and the smoke only cover a few pixels of the frame). Regarding the other two classes, the errors are probably caused by the similarities with other classes: many frames of the class pizza are showing pizza restaurants, which also contain smoke, while frames labeled as coffee are very similar to those containing bars or restaurants.

If we allow a higher FPR for CLIP, by setting the probability threshold to 0.4, it achieves 77.78% olfactory accuracy, however, with double the number of false positives compared to the scenario with a probability threshold of 0.6.

E. Evaluation of Action Detection for Haptics

SlowFast64x2 was also evaluated on Kinetics classes that are associated with haptics, described at the end of Section III-B. The network achieved a 72.26% top-1 accuracy, 89.52% top-5 accuracy, an FPR of 2.6% and an FNR of 2.1%, which is a similar performance to the scent-related classes. The impact of different thresholds on the haptic-related labels of the reduced Kinetics dataset can be seen on Fig. 17. It can be seen that SlowFast64x2 has slightly higher performance when comparing haptic-related labels to scent-related labels (as shown on Fig. 11). As a result, the recommended threshold is also 0. Differences in the results can be explained by the nature of the classes related to haptics, as some actions (e.g., slapping a face) are easier to be detected than scent-related actions.

This section presented extensive testing of the proposed solution. The complexity of CNN architectures was evaluated comparatively on the ImageNet validation dataset, indicating that ResNet18 provides a low inference time, while RegNetY-3.2GF and EfficientNets b2 and b3 achieved the lowest top-1 errors. The accuracy associated with a number of networks was also investigated, with ResNets 18 and 50 achieving low FPR and DenseNet161 achieving the best results in terms of top-1 and top-5 accuracy and error, but also a much higher inference time in comparison to ResNets. Experiments with CLIP, a zero-shot CNN, demonstrated it is also a feasible network to be used by the proposed solution, with up to 42.19% top-1 accuracy. Additional experiments demonstrated that multiple predicted labels might correspond to the same smell, and these predictions can be merged, decreasing olfactory error rate by approximately 10%.

Using the proposed video dataset for testing the solution, the best results regarding the generation of scents based on action recognition was recorded by the SlowFast64x2 network, which achieved a top-1 accuracy of 75.56%. Scene-based olfaction effect generation was tested with DenseNet161 and ResNet18 which achieved 74.27% and 69.19% olfactory accuracy, respectively. ResNet18 however offers a better complexity trade-off, as DenseNet161 requires more computational time. SlowFast64x2 was also tested for the generation of haptic effects based on action detection. It achieved a 72.26% top-1 haptic accuracy.

V. Conclusion and Future Work

This work describes an innovative solution for enhancing 360° videos with automatically generated mulsemedia content in order to increase viewer experience. The paper starts with a thorough analysis of state-of-the-art deep learning solutions and discusses how they could be applied in this solution. We selected 11 different neural network architectures and created a test dataset consisting of 170 360° video clips with various scent categories to evaluate the networks' performance. Scene and action recognition datasets (Places and Kinetics, respectively) were adapted to support olfaction and haptics-related labels. The proposed solution supports 54 scent categories, a significant improvement to the existing baseline solution that only recognizes 5 scent categories, triggered by 62 scene classes and 48 action classes. Scents are generated based on both scene and action detection, while haptics are triggered by actions and are synchronized via audio cues.

Testing shows how the proposed solution achieves a 69.19% olfactory accuracy on the Places dataset and a 72.26% top-1 haptic accuracy on the Kinetics dataset. The olfactory and haptic accuracies were achieved while having a low number of false positive detections (10 out of 171 samples). The accuracy values indicate that the solution is applicable to realistic use cases and can replace the lengthy process of manually annotating content with multisensorial effects.

Future work considers a real-time implementation of the solution with powerful GPU servers and involving saliency-based pre-processing techniques to identify and process only relevant tiles, enhancing processing times. Haptic content generation can also be enhanced with additional sensors and include user emotions.

References

[1] T. Bi, A. Yaqoob, L. Zou, and G.-M. Muntean, "A Study of Learning Experience during Olfaction-enhanced Adaptive Rich Media Delivery," in *Proc. IEEE Int. Symp. on Broadband Multimedia Syst. and Broadcast. (BMSB)*, 2020, pp. 1–5.

[2] L. Jalal, M. Anedda, V. Popescu, and M. Murroni, "QoE Assessment for IoT-Based Multi Sensorial Media Broadcasting," *IEEE Trans. Broadcast.*, vol. 64, no. 2, pp. 552–560, 2018.

[3] E. Calvi, U. Quassolo, M. Massaia, A. Scandurra, B. D'Aniello, and P. D'Amelio, "The Scent of Emotions: A Systematic Review of Human Intra- and Interspecific Chemical Communication of Emotions," *Brain and Behavior*, vol. 10, no. 5, pp. 1–19, 2020.

[4] M. Bordegoni and M. Carulli, "Evaluating Industrial Products in an Innovative Visual-Olfactory Environment," *J. Comput. Inf. Sci. Eng.*, vol. 16, no. 3, 2016.

[5] A. Tomono, K. Kanda, and S. Otake, "Effect of smell presentation on individuals with regard to eye catching and memory," *Electronics and Communications in Japan*, vol. 94, pp. 9–19, Mar. 2011.

[6] H. Engelbrecht, R. W. Lindeman, and S. Hoermann, "A SWOT Analysis of the Field of Virtual Reality for Firefighter Training," *Front. Robot. AI*, vol. 6, pp. 1–14, 2019.

[7] S. Heba, M. Sczesny-Kaiser, K. Sucker, J. Bünger, T. Brüning, M. Tegenthoff, and T. Schmidt-Wilcke, "Pain Perception, Brain Connectivity, and Neurochemistry in Healthy, Capsaicin-Sensitive Subjects," *Neural Plasticity*, pp. 1–11, 2020.

[8] Z. Yan, K. Yang, Z. Wang, B. Yang, T. Kaizuka, and K. Nakano, "Intention-Based Lane Changing and Lane Keeping Haptic Guidance Steering System," *IEEE Trans. Intell. Veh.*, pp. 1–12, 2020.

[9] I. Comşa, E. Saleme, A. Covaci, G. Assres, R. Trestian, C. Santos, and G. Ghinea, "Do I Smell Coffee? The Tale of a 360° Mulsemedia Experience," *IEEE MultiMedia*, vol. 27, no. 1, pp. 27–36, 2020.

[10] M. Melo, G. Gonçalves, P. Monteiro, H. Coelho, J. Vasconcelos-Raposo, and M. Bessa, "Do Multisensory Stimuli Benefit the Virtual Reality Experience? A Systematic Review," *IEEE Trans. Vis. Comput. Graphics*, pp. 1–20, 2020.

[11] J. P. Sexton, A. A. Simiscuka, K. McGuinness, and G. Muntean, "Automatic CNN-Based Enhancement of 360° Video Experience With Multisensorial Effects," *IEEE Access*, vol. 9, pp. 133 156–133 169, 2021.

[12] A. Craig, W. Sherman, and J. Will, *Developing Virtual Reality Applications: Foundations of Effective Design*. Morgan Kaufmann, 2009.

[13] P. G. de Barros and R. W. Lindeman, "Performance Effects of Multi-Sensory Displays in Virtual Teleoperation Environments," in *Proc. Symp. on Spatial User Interaction (SUI)*, 2013, pp. 41–48.

[14] F. Danieau, A. Lecuyer, P. Guillotel, J. Fleureau, N. Mollet, and M. Christie, "Enhancing Audiovisual Experience with Haptic Feedback: A Survey on HAV," *IEEE Trans. Haptics*, vol. 6, no. 2, pp. 193–205, 2013.

[15] D. Villamarín and J. M. Menéndez, "Haptic Glove TV Device for People with Visual Impairment," *Sensors*, vol. 21, no. 7, 2021.

[16] F. Danieau, J. Fleureau, P. Guillotel, N. Mollet, M. Christie, and A. Lécuyer, "Toward Haptic Cinematography: Enhancing Movie Experiences with Camera-Based Haptic Effects," *IEEE MultiMedia*, vol. 21, no. 2, pp. 11–21, 2014.

[17] D. Guinness, A. Muehlbradt, D. Szafir, and S. K. Kane, "The Haptic Video Player: Using Mobile Robots to Create Tangible Video Annotations," in *Proc. ACM Int. Conf. on Interactive Surfaces and Spaces (ISS)*, 2018, pp. 203–211.

[18] S. Rehman, J. Sun, L. Liu, and H. Li, "Turn Your Mobile Into the Ball: Rendering Live Football Game Using Vibration," *IEEE Trans. Multimedia*, vol. 10, no. 6, pp. 1022–1033, 2008.

[19] A. Mazzoni and N. Bryan-Kinns, "Mood Glove: A Haptic Wearable Prototype System to Enhance Mood Music in Film," *Entertainment Comput.*, vol. 17, pp. 9–17, 2016.

[20] J. Cha, M. Eid, and A. E. Saddik, "Touchable 3D Video System," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 5, no. 4, Nov. 2009.

[21] A. Israr, Z. Schwemler, J. Mars, and B. Krainer, "VR360HD: A VR360° Player with Enhanced Haptic Feedback," in *Proc. ACM Conf. on Virtual Reality Softw. and Technol. (VRST)*, 2016, pp. 183–186.

[22] O. Baus and S. Bouchard, "Exposure to an Unpleasant Odour Increases the Sense of Presence in Virtual Reality," *Virtual Reality*, vol. 21, pp. 58–74, Jun. 2017.

[23] H. Q. Dinh, N. Walker, L. F. Hodges, Chang Song, and A. Kobayashi, "Evaluating the Importance of Multi-Sensory Input on Memory and the Sense of Presence in Virtual Environments," in *Proc. IEEE Virtual Reality*, 1999, pp. 222–228.

[24] B. G. Munyan, III, S. M. Neer, D. C. Beidel, and F. Jentsch, "Olfactory Stimuli Increase Presence in Virtual Environments," *PLOS ONE*, vol. 11, no. 6, pp. 1–19, Jun. 2016.

[25] N. Ranasinghe, P. Jain, N. Thi Ngoc Tram, K. C. R. Koh, D. Tolley, S. Karwita, L. Lien-Ya, Y. Liangkun, K. Shamaiah, C. Eason Wai Tung, C. C. Yen, and E. Y.-L. Do, "Season Traveller: Multisensory Narration for Enhancing the Virtual Reality Experience," in *Proc. Conf. on Human Factors in Comput. Syst. (CHI)*, 2018, pp. 1–13.

[26] D. Egan, C. Keighrey, J. Barrett, Y. Qiao, S. Brennan, C. Timmerer, and N. Murray, "Subjective Evaluation of an Olfaction Enhanced Immersive Virtual Reality Environment," in *Proc. Int. Workshop on Multimedia Alternate Realities (AltMM)*, 2017, pp. 15–18.

[27] E. Saleme, A. Covaci, G. Mesfin, C. Santos, and G. Ghinea, "Mulsemedia DIY: A Survey of Devices and a Tutorial for Building Your Own Mulsemedia Environment," *ACM Comput. Surv.*, vol. 52, no. 3, Jun. 2019.

[28] T. Bi, A. Pichon, L. Zou, S. Chen, G. Ghinea, and G.-M. Muntean, "A DASH-Based Mulsemedia Adaptive Delivery Solution," in *Proc. Int. Workshop on Immersive Mixed and Virtual Environ. Syst. (MMVE)*, 2018, pp. 1–6.

[29] I. Radosavovic, R. Kosaraju, R. Girshick, K. He, and P. Dollar, "Designing Network Design Spaces," in *Proc. IEEE/CVF Conf. on Comput. Vis. and Pattern Recog. (CVPR)*, Jun. 2020, pp. 10 425–10 433.

[30] T. S. Cohen, M. Geiger, J. Köhler, and M. Welling, "Spherical CNNs," in *Proc. Int. Conf. on Learning Representations (ICLR)*, 2018.

[31] M. Xu, L. Jiang, C. Li, Z. Wang, and X. Tao, "Viewport-based CNN: A Multi-task Approach for Assessing 360° Video Quality," *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1–18, 2020.

[32] K. Wang and S. Lai, "Object Detection in Curved Space for 360-Degree Camera," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, 2019, pp. 3642–3646.

[33] W. Yang, Y. Qian, J.-K. Kämäräinen, F. Cricri, and L. Fan, "Object Detection in Equirectangular Panorama," in *Proc. Int. Conf. on Pattern Recog. (ICPR)*, 2018, pp. 2190–2195.

[34] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 Million Image Database for Scene Recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1452–1464, 2018.

[35] K. Ahmed, I. M. El-Henawy, and H. A. Mahmoud, "Action Recognition Technique Based on Fast HOG3D of Integral Foreground Snippets and Random Forest," in *Proc. Int. Conf. Intell. Syst. and Comput. Vis. (ISCV)*, 2017, pp. 1–7.

[36] L. Sevilla-Lara, Y. Liao, F. Güney, V. Jampani, A. Geiger, and M. J. Black, "On the Integration of Optical Flow and Action Recognition," in *Proc. Comput. Vis. and Pattern Recog. (CVPR)*, 2019, pp. 281–297.

[37] M. Asadi-Aghbolaghi, A. Clapés, M. Bellantonio, H. J. Escalante, V. Ponce-López, X. Baró, I. Guyon, S. Kasaei, and S. Escalera, "A Survey on Deep Learning Based Approaches for Action and Gesture Recognition in Image Sequences," in *Proc. IEEE Int. Conf. on Autom. Face Gesture Recog. (FG)*, 2017, pp. 476–483.

[38] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "SlowFast Networks for Video Recognition," in *Proc. IEEE/CVF Int. Conf. on Comput. Vis. (ICCV)*, Oct. 2019.

[39] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning Transferable Visual Models From Natural Language Supervision," in *Proc. Int. Conf. on Mach. Learning (ICML)*, 2021.

[40] J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba, "Sun Database: Large-Scale Scene Recognition from Abbey to Zoo," in *Proc. Comput. Vis. and Pattern Recog. (CVPR)*, 2010, pp. 3485–3492.

[41] K. Soomro, A. Roshan Zamir, and M. Shah, "UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild," Univ. Central Florida, Orlando, FL, USA, Tech. Rep. CRCV-TR-12-01, 2012. [Online]. Available: http://crcv.ucf.edu/data/UCF101.php

[42] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta, "Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding," in *Proc. Eur. Conf. on Comput. Vis. (ECCV)*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer Int. Publishing, 2016, pp. 510–526.

[43] R. Goyal, S. Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. Mueller-Freitag, F. Hoppe, C. Thurau, I. Bax, and R. Memisevic, "The "Something Something" Video Database for Learning and Evaluating Visual Common Sense," in *Proc. IEEE Int. Conf. on Comput. Vis. (ICCV)*, 2017, pp. 5843–5851.

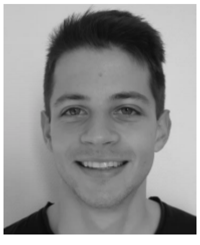
[44] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, "The Kinetics Human Action Video Dataset," *arXiv*, 2017.

[45] N. Murray, O. A. Ademoye, G. Ghinea, and G.-M. Muntean, "A Tutorial for Olfaction-Based Multisensorial Media Application Design and Evaluation," *ACM Comput. Surv.*, vol. 50, no. 5, pp. 1–30, Sep. 2017.

[46] T. Maugey, O. Le Meur, and Z. Liu, "Saliency-Based Navigation in Omnidirectional Image," in *Proc. IEEE Int. Workshop on Multimedia Signal Process. (MMSP)*, 2017, pp. 1–6.

[47] "ImageNet Benchmark (Image Classification)," 2021. [Online]. Available: https://paperswithcode.com/sota/image-classification-on-imagenet

[48] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *arXiv*, 2020.



Péter Szabó received the B.Sc. degree in Molecular Bionics Engineering from Pázmány Péter Catholic University (PPCU) and the M.Sc. degree in Computer Science with Image Processing and Computer Vision specialization in a triple degree program from UAM, PPCU and UBx in 2019 and 2021 respectively. He is working in the Department of Digital Media at Vicomtech. His research focuses on mulsemedia and virtual reality.

Anderson Augusto Simiscuka (S'17-M'20) received the B.Sc. degree in Information Systems in 2014 from Mackenzie Presbyterian University, São Paulo, Brazil, and the Ph.D. degree from the School of Electronic Engineering, Dublin City University (DCU), Ireland, in 2020. He is a Postdoctoral Researcher with the Performance Engineering Laboratory and the Insight SFI Centre for Data Analytics, in DCU. He is involved in the EU-funded project TRACTION.

Stefano Masneri received the B.Sc. degree in information technology and the M.Sc. degree in telecommunications from the Università degli studi di Brescia, Italy, in 2005 and 2008, respectively. He is with the Department of Digital Media, Vicomtech. He focuses his research in signal processing, computer vision, and interactive technologies.

Mikel Zorrilla is the Head of the Digital Media Department at Vicomtech, Spain. He studied Telecommunication Engineering at the University of Mondragon (Spain), and obtained his PhD degree in September 2016 from University of the Basque Country (Spain) entitled "Interoperable Technologies for Multi-Device Media Services". He is the overall coordinator of the EU Horizon 2020 project TRACTION (www.traction-project.eu).

Gabriel-Miro Muntean (M'04–SM'17) is a Professor with the School of Electronic Engineering, Dublin City University (DCU), Ireland, and co-Director of the DCU Performance Engineering Laboratory. He has published over 400 papers in top-level international journals and conferences, authored 4 books and 23 book chapters, and edited 8 additional books. He is an Associate Editor of the IEEE Transactions on Broadcasting, the Multimedia Communications Area Editor of the IEEE Communications Surveys and Tutorials, and a Reviewer for important international journals, conferences, and funding agencies. He is the DCU Coordinator for the EU project TRACTION www.traction-project.eu.