Towards Automatic Content Generation for Immersive Cinema Theater Based on Artificial Intelligence

David Traparic, Mohamed-Chaker Larabi, Ladjel Bellatreche

Towards Automatic Content Generation for Immersive Cinema Theater Based on Artificial Intelligence

David Traparic[1,2,3], Mohamed-Chaker Larabi[1], Ladjel Bellatreche[2]
[1] CNRS, Univ. Poitiers, XLIM, UMR 7252, France
[2] LIAS, Univ. Poitiers, France
[3] CGR Cinémas, Perigny, France

***Abstract*—Immersive display systems like the one proposed by ICE® technology aims to enhance visual immersion by widening the field of view. However, creating immersive content while maintaining immersion integrity is a challenging task due to the sensitivity of human peripheral vision to flickering and movement. Moreover, identifying elements in videos that may disrupt immersion and determining whether they can be expanded into an immersive context is a complex and time-consuming process due to the lack of automatic methodologies. In this paper, we propose a pipeline for automatically generating content for lateral displays from movies. The pipeline consists of several steps. Firstly, the input content is divided into cinematic shots, and then further segmented into snippets. Next, domain-specific features are extracted using dedicated video deep learning models. Additionally, handcrafted features are computed to provide task-specific information. These extracted features are utilized to predict the required processing steps for generating lateral content that aligns with ground-truth annotations provided by cinema experts. The results obtained from our pipeline show promising accuracy and demonstrate the potential for this specialized application.**

***Index Terms*—immersive display system, Deep Learning, extrafoveal video, Wide Field of view**

I. Introduction

From the earliest audiovisual creation dates we can find, immersion has always been sought. From Chauvet cave paintings [1], to cinema theaters, 3D movies, VR headsets until nowadays new sensory experiences such as haptic gloves, research has continuously pushed forward experiences to be closer and closer to reality.

Even when focusing only on the visual sensory system, merely fulfilling the human FoV (Field-of-View) with adapted content of view is still a complex task to achieve today. Some may choose to record a wider camera view in order to display it on a specific format, such as IMAX [2], Barco Escape [3] or ScreenX [4]. This approach is based on pre-production and can be achieved by multi-camera support to record surrounding contents, or by a very wide-lens camera. The latter lowers the quality of the overall image, as a single camera attempt to cover the same wide FoV [5]. Either way, camera will inevitably disturb lighting and microphone placement [5]. The high cost of these very specific camera setup is a major drawback of this method.

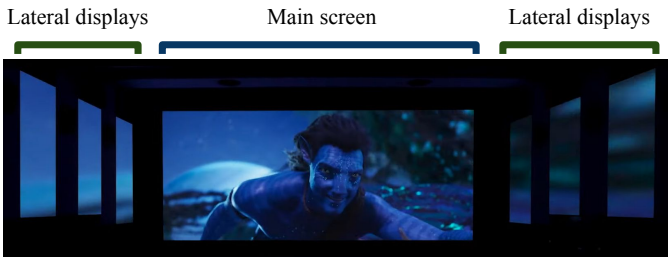

Fig. 1. Photo of an ICE immersive Theater, showing a movie with its lateral displays.

To solve pre-production problems, multiple techniques are oriented towards real-time generation of surrounding content. A few techniques work with LED lights such as the Ambilight [6], SparseLightVR [7], Ambiculus [8] and DeepDive [9] that provide ambient color lighting around the TV screen or around lenses of a head-mounted display. More advanced techniques, related to outpainting *i.e.* expanding an image content beyond its borders, focus on real-time generation. These methods such as Infinity-By-Nine [10] (inspired by the CAVE [11] system) or more recently ExtVision with deep learning techniques [12], introduce artifacts, including spatial incoherence or flickering problems due to the real-time generation constraint. While, the human peripheral vision is less sensitive to color and texture, allowing to be less demanding in terms of quality of the surrounding content (which defines the so-called Focus+Context [13] system), it is more sensitive to motion and flicker [14].

To avoid such artifacts having an impact on the user's quality of experience, post-production involving content-designers team, could be considered. However this may significantly increase the cost of content production like movies. Finding solutions to make the process automated is an important challenge of the last decade. To cope with this problem, another approach consists of dropping the real-time constraint, although visual artifacts may persist [15].

This paper tackles this problem for a very specific context, where the tolerance towards visual artifacts is very close to zero. This context is the immersive cinema and more precisely the ICE Immersive technology [16]. This theater system provides post-processed lateral contents to offer an

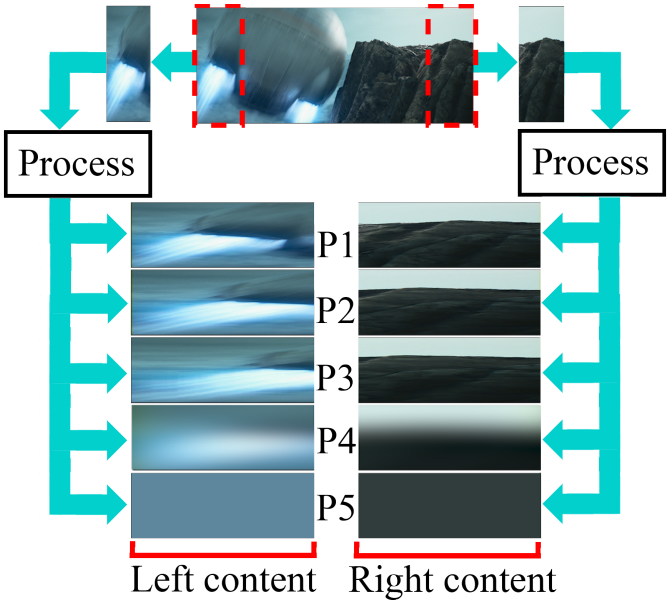

Fig. 2. Five possible processings, applied to the same frame, showing an accelerating spaceship and a rock cliff: P_1) *details*, P_2) *seam_carving*, P_3) *less_details*, P_4) *blur* and P_5) *static*.

immersive environment to the user illustrated in Fig. 1. The first milestone of ICE expert's working process is to choose what processing to use as a baseline for a given cinematic shot and this paper will be focused on this part. Experts have a panel of five processing (represented in Fig.2, described in Table I) from which they can choose. Each processing has its own benefits and disadvantages. Adequate processings are used according to shots in Fig.3.

Using the actual movie frames to extend the image in real time induces a discontinuity in the image. This phenomenon can result in an apparent object duplication, deformation, or temporal phase shift when an object moves into or out of the shot. This is why special care is needed to choose adequate lateral content. *Details* (Fig.2.P_1) is used to emphasize fast moving video sequences because the human eye is more tolerant to temporal phase shift under this condition (Example in Fig.2.P_1, left content). Landscapes can also be emphasized by detailed choices, as they tend to have few to no salient object, or to have repetitive pattern (such as Fig.2.P_1, right content and Fig.3.P_1). The human eye is more tolerant in discontinuity when it occurs on patterned image. Both can

TABLE I
DESCRIPTION OF THE DIFFERENT PROCESSINGS.

Details (P_1)	Intakes a slice of the frame and scales up its horizontal size ($\times 6.7$)
Seam_ Carving (P_2)	Applies a Seam Carving algorithm to the entire frame, *i.e.* extending image pixels of lowest density, then intakes a slice of the resulted frame and enlarge its horizontal size ($\times 6.7$)
Less_Details (P_3)	Intakes a thinner slice of the frame than "Details" and scales up its horizontal size ($\times 10$)
Blur (P_4)	After applying less_details, applies an important vertical blur to the resulted image
Static (P_5)	Applies a single color for the entire cinematic shot

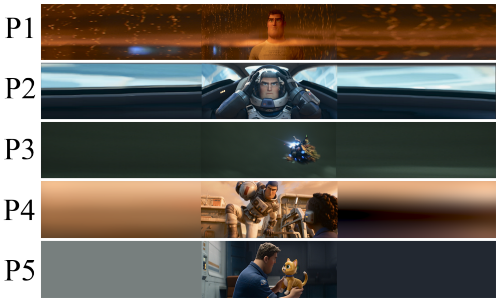

Fig. 3. Basic processing techniques when used by experts on adequate shots, same order than on Fig.2.

provide solutions when a detailed choice would emphasize the action, but a salient element is too close to the border of the screen and thus, would be duplicated in lateral screens, breaking the immersion. In Fig.3.P_2, *seam carving* gets rid of the elbow of the pilot, whereas Fig.3.P_3, *less_details* avoids the ship, both located too close to a border. *Blur* option is employed in situations where the shot exhibits a medium-speed motion or features a prominent object situated close to the frame's edge, disrupting the viewer's immersion (refer to Fig.3.P_4). This includes scenarios such as a character leaping out of a car or a person's head appearing duplicated on the lateral screen.

option is used when the shot is medium-speed or with a too salient object near the border, breaking immersion (Fig.3.P_4), with a character jumping out of car, or a head that duplicates itself in the lateral screen). At any time, the side panels should not draw the spectator's attention more than the initial sequence does. This is why *static* choice is useful for low-speed video sequences, such as the discussion between characters on Fig.3.P_5.

Whereas this is oversimplified for the sake of this paper, the general ideas given here remain. Decisions on each shot are made on a case-by-case basis, with trial and error expertise. In this paper, As the actual ICE working process is so tedious, the goal is to push further the automation of it. To realize it, deep learning models with similar problematics than ours are explored. Building upon this comprehensive exploration, we have successfully incorporated existing models tailored to extract intricate temporal features into our video datasets. A benchmark of these different models applied to our problem as feature extractor with their performance is provided, taking special care of their respective original setups (training video framerate, resampling, etc.). Different classifier configurations are also tried to better retrieve the obtained feature wealth. Ablation studies are conducted on this classifier and on data preparation to make sure the model proposed is coherent as a whole. Results and discussion show that a deep learning model for our very specific context is feasible and promising.

II. METHODOLOGY

Choosing the best processing to each of the cinematic shots of a movie to offer an expansion covering the lateral displays is tedious. The aim of the following paper is to explore solutions

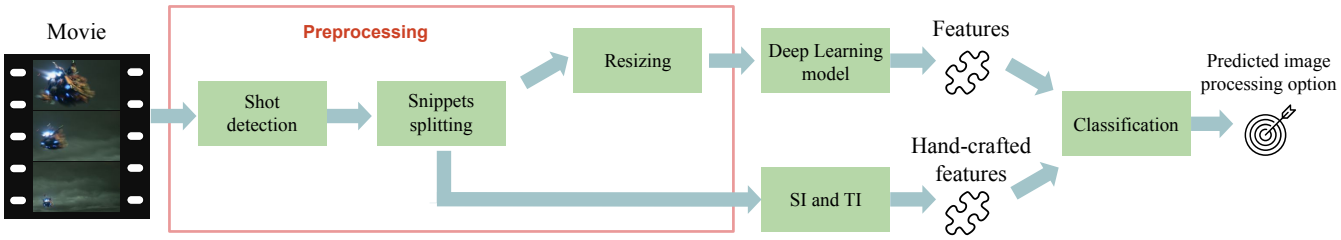

Fig. 4. Pipeline used to predict lateral contents of a given snippet, for movie expansion purpose.

to perform this task in an automatic manner. In this section, we describe the adopted solution based on artificial intelligence. The pipeline of this solution is given on Fig. 4.

The input data, *i.e.* a movie, is split into cinematic shots, with a ground truth prediction associated to each, provided by ICE experts. The aim is to predict, for any given shot, which is the most adequate processing in order to offer the user an immersive and comfortable experience. To perform this task, the aim is to extract representative features that could help to characterize the shot, video-specific features are extracted by adopted deep learning models from the literature. As our problematics include not only subjective classification targets but also immersion and extrafoveal vision, handcrafted features are computed to provide some specific data adapted to our model aiming to learn how to make the best choice for peripheral immersive content.

A. Data preprocessing

To preprocess accordingly a given movie, the movie is divided into shots by Adobe Scene Detection tool, which are then sub-divided into fixed length neighbouring frames called snippets (16 frames long). If the last snippet of a shot has a length of less than 16 frames, this snippet is ignored. Whereas backbones have been trained on square cropped images, decisions of our task rely more on the left and right borders of the frames as this is where a discontinuity will occur between the frame and its extension. Thus, it is thought that center-cropping from the video data will seriously impair the model training. Frame size of video data will be resized as 112×267 (cinema ratio preserved) without cropping.

B. Deep learning based features

In the context of our work, deep learning models that handle correctly both spatial and temporal information are required. This allows to solve some situations such as the case of having salient objects appearing only on the first frames of the shot. Furthermore, as shots can last for hundreds of frames, long-term temporal information becomes important. Some models, by the nature of the tasks they were designed for, require longer term temporal information. These tasks can be grouped under the umbrella of video understanding [17], which includes action localization, action classification and so on. C3D [18], I3D [19], TSP [20] and SlowFast [21] represent examples of such models. Our application shows similarities with the aforementioned tasks. This led us to select three largely available and recognized models, namely, C3D (Convolutional 3D), I3D (Inflated 3D ConvNet), and TSP (Temporally-Sensitive Pretraining) for spatial-temporal deep features extraction.

TABLE II
ORIGINAL INPUT FRAME SIZE OF EACH MODEL.

Model	Frame resized to:	During Train:	During Test:	#Params
C3D	smallest side to 128, while keeping ratio	random crop 112×112	center crop 112×112	78.0M
I3D	smallest side to 256, while keeping ratio	random crop 224×224	center crop 224×224	12.3M
TSP	smallest side to 128, while keeping ratio	random crop 112×112	center crop 112×112	31.3M

The latter models, often used as backbones for more complex architectures, show some differences in terms of training procedures. For instance, C3D has been trained with 16 RGB frame snippets, at 25 fps. I3D has been trained using two streams composed of 64 RGB frame snippets and the corresponding optical flow at 25 fps. When videos were shorter than 64 frames, they were looped to reach this requirement. Finally, TSP has been trained using 16 RGB frame snippets temporally subsampled from 30 fps to 15. These training procedures showcase different temporal length of the input snippet *i.e.* 0.64s for C3D, 2.56s for I3D and 1.875s for TSP. Regarding the frame size, all three models perform a similar procedure by resizing the input video and randomly crop or center crop during training and test stages (see Table II). Within the framework of our problem, 16 frame snippet is represented by 4096 features, using C3D, 1024 features using I3D and 512 features using TSP.

C. Handcrafted features

Spatial and temporal characteristics of a given cinematic shot are paramount attributes during the classification stage to determine which method should be used with regards to the ground truth. To account for these characteristics and feed the classifier with additional handcrafted features, we selected Spatial Perceptual Information (SI) and Temporal Perceptual Information (TI) respectively related to details level and movement intensity at the frame scale.

SI and TI are defined in ITU-T Recommendation P.910 (”Subjective video quality assessment methods for multimedia applications”). SI is expressed as:

$$SI_n = \sigma^*(\text{Sobel}(F_n)) \tag{1}$$

based on the content edges, where SI_n is the spatial information of a single frame at time n, $\sigma^*(\cdot)$ the standard deviation

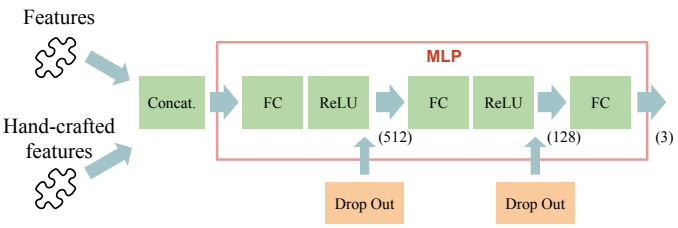

Fig. 5. Structure of the classifier placed after feature extraction by the different temporal backbones.

along the pixels of an image, $\mathrm{Sobel}(\cdot)$ being the Sobel filter applied to a matrix and F_n the video frame luminance plane at time n as defined in the recommendation.

Regarding TI, it is calculated on successive frames, as the difference between pixel values of the present frame and the successive one.

$$M_n(i,j) = F_n(i,j) - F_{n-1}(i,j) \quad (2)$$

$$TI_n = \sigma^*(M_n) \quad (3)$$

with $F_n(i,j)$ being the pixel value of the video frame F at time n, i^{th} row and j^{th} column, and M_n being the motion difference (pixel-wise) between two successive frames.

Our handcrafted features have one SI value and one TI value for each frame in snippet *i.e.* handcrafted features are 32 scalars long as input for the classifier.

D. Classification

The baseline for our classifier which is a multilayer perceptron (MLP) is represented by Fig.5 where FC stands for fully connected layers. Handcrafted features represented by *SITI* inputs of the classifier are an option attempted to account for the variability of spatial and temporal content to improve model performance: the same holds for the *drop out* layers. Drop out layers are common to avoid overfitting, and forces the model to attach importance to every input feature. This is especially useful to learn from our handcrafted features. Other attempts to improve the model have been made, like using an additional convolutional layer before the MLP or additional FC layers. For the latter case, the additional FC layer is followed by ReLU activation function or drop out, to keep coherence with the classifier as a whole. Another option lies in the fact of feeding the model with handcrafted features as a separate feature stream. Drop out layers on this second stream are optional to see the variation of performance with and without them. Predictions of this classifier are for snippets, because shots have variable length that would not be compatible with our classifier. To get a prediction for an entire shot, a majority vote between snippets predictions is computed.

III. Experiments

A. Dataset

Gaining access to cinema movies content is extremely challenging due to the strict restrictions imposed by the producers. As of the time of writing this paper, access to such content had not been granted. In order to validate our architecture, we assembled a dataset consisting of movie trailers, which are readily accessible from CGR, without any restrictions. These trailers provided us with the necessary ground truth information regarding the processes applied to the frames to obtain the left and right content for the lateral displays.

TABLE III
Distribution of processing chosen by experts.

Process	Name	in trailers	in movies	Fused in trailers
P_1	Details	13.48%	22.19%	17.66%
P_2	Seam	0.69%	0%	
P_3	Less_details	3.49%	1.22%	
P_4	Blur	55.70%	36.27%	55.70%
P_5	Static	26.64%	40.32%	26.64%

Our data set is composed of 30 trailers with a resolution of 858×2048, with labels provided by ICE experts. These trailers have a duration between 40 and 176 seconds. C3D could not work with our frame size of 112×267, therefore, the original center-cropping was specifically kept for this model.

It is acknowledged that the temporal activity of trailers are quite different from the actual movies. Additional tuning would be probably needed once the content is available.

We measured the occurrence of each possible processing to apply to a given shot in the constructed dataset as well as from the full movies, shown in Fig. III. It can be noticed that P_2 and P_3 are rarely or never used whatever the explored dataset. With the aim to avoid difficulties at the training stage, we opted for the fusion of P_2 and P_3 with P_1 because of the similarity of the applied processing.

As data is quite limited for this application, since only 30 trailers are available, data augmentation appears as an important solution to mitigate the problem. With the help of experts, we validated a list of image transformations that can be applied to cinematics shots without any impact on the ground truth, as for instance: Hue shifting in the HSV color space, saturation shifting (HSV), value shifting (HSV), contrast increasing, rotation from $-5°$ to $+5°$ and vertical flipping. Precise limit values have been empirically found with the help of the experts.

B. Implementation details

Our experiments were run on an NVIDIA Quadro RTX 4000 with 8Go of VRAM. After conducting empirical trials, we determined that the optimal learning rate for the classifier is 1×10^{-3}, except in the case of C3D. For the latter, the learning rate was adjusted to 1×10^{-4} to ensure proper learning. We employed the Adam optimizer and a scheduler was utilized to decrease the learning rate at each epoch using the formula $lr = lr \times 0.95$. The chosen loss function for classification is the cross-entropy loss, which is widely recognized as a standard loss function for this task.

Training is performed for 50 epochs involving 9 fixed random splits of the dataset including training and validation sets. To ensure sufficient validation coverage, a minimum of 20% of the trailers is reserved for validation in each split. The dataset is carefully partitioned to avoid any overlap between the trailers used in the training set and those used in the

TABLE IV
ACCURACY OF DIFFERENT MODELS AS TEMPORAL FEATURE EXTRACTORS, IN DIFFERENT CONFIGURATIONS OF CLASSIFIER

Configurations	TSP	C3D	I3D
Default	**68.02%**	54.99%	65.44%
Data Augmentation	**68.85%**	52.69%	68.18%
Convolution	**66.92%**	54.24%	66.91%
Drop Out	**70.32%**	46.97%	66.42%
SITI	**69.95%**	56.97%	68.94%
Two-Stream	**69.53%**	57.42%	68.48%
red2nd stream dropout	**69.66%**	55.14%	67.58%

validation set. This precautionary measure ensures that the training data and validation data are not correlated, thereby preserving the integrity of the evaluation process.

Despite the fusion of three processings, there is a need to address the issue of label imbalance during training. To achieve a balanced representation of labels, a sampling strategy is employed on each epoch. This strategy involves randomly dropping snippets with over-represented labels from both training and validation sets. This process continues until a state of perfect equilibrium is reached, where label occurrences are evenly distributed across the dataset.

C. Results and discussion

Table IV presents the individual performance of models with one configuration at a time, while table V showcases the performance when multiple configurations are combined. In both tables, green cells indicate better performance compared to the default classifier, while red cells indicate worse performance. The best feature extractor for a given classifier configuration is highlighted in bold.

A notable observation from these tables is that C3D, serving as a temporal model for our specific task, consistently underperforms compared to the other two models, regardless of the classifier configuration. On the other hand, TSP emerges as the most effective model for our task, followed by I3D. This discrepancy could be attributed to the fact that I3D is trained on 64-frame snippets, which may not align well with the shot lengths present in our dataset (*see* Fig. 7).

When testing the different configurations separately (Table IV), we found that data augmentation, SITI, and dropout

TABLE V
ACCURACY OF DIFFERENT MODELS AS TEMPORAL FEATURE EXTRACTORS, IN DIFFERENT COMBINED CONFIGURATIONS

Data aug.	Drop Out	SITI	Two Stream	2nd stream dropout	TSP	C3D	I3D
✓	✓				**69.56%**	47.78%	67.93%
	✓	✓			**71.72%**	53.03%	69.12%
	✓	✓	✓		**71.21%**	50.11%	67.42%
	✓	✓	✓	✓	**69.89%**	48.99%	67.16%
✓		✓			**71.43%**	56.40%	68.69%
✓		✓	✓		68.21%	57.88%	**68.85%**
✓		✓	✓	✓	**68.28%**	54.90%	68.14%
✓	✓	✓			**70.88%**	54.48%	68.87%
✓	✓	✓	✓		69.17%	51.52%	**70.54%**
✓	✓	✓	✓	✓	69.28%	47.69%	**69.70%**

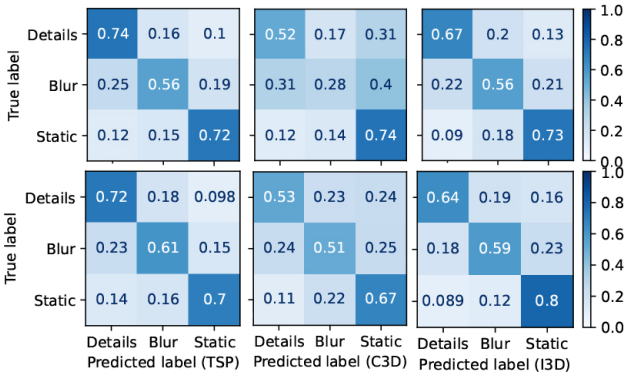

Fig. 6. Confusion matrices between predicted labels and ground truth: default MLP configuration (top), best MLP configuration (bottom).

techniques significantly improved the performance of the model when using TSP and I3D as feature extractors. Dropout showed the best improvement for TSP features, SITI input as a separate input improved C3D features the most, and concatenating SITI input with I3D features yielded the best performance increase. These findings highlight the effectiveness of these techniques in enhancing the model's performance for different feature extraction architectures.

Confusion matrices for the default classifier is given on Fig.6, top matrices. It is clear that C3D does not output relevant enough features (according to our task) for the classifier to distinguish *blur* labels from others. Even though *Blur* labels are the most difficult label to predict correctly for any feature extractor with 56% accuracy for TSP and I3D.The main difference between the accuracies of I3D and TSP lies in their ability to correctly predict *details* labels, with TSP achieving 74% accuracy compared to 67% for I3D.

The best configurations of the models, as listed in Table V, involve combination of configurations. For TSP, the best combination is SITI+Drop Out achieving an accuracy of 71.72%. For I3D, the best combination is SITI+Data Augmentation+Drop Out+2stream with an accuracy of 70.54%. The confusion matrices for these best classifier configurations can be seen in Fig. 6, bottom matrices. It is worth noting that, in these configurations, the prediction of the *blur* label improves for all feature extractors.

D. Ablation study

Deeper multi-layers perceptron: The addition of more fully connected layers, along with corresponding ReLU and dropout layers if necessary, did not lead to any improvement in the performance of the models tested in this study. This suggests that the existing three fully connected layers in our MLP are sufficient to capture all possible feature associations, at least within the number of epochs the model was trained for.

Traditional regression classifiers: Instead of using a MLP as a classifier, traditional regression models were tested using TSP features concatenated with our handcrafted features. The results showed an accuracy 60.40% with Gradient Boosting, 65.70% with Random Forest and 57.90% with SVR. These performances were found to be worse compared to training the MLP as a classifier.

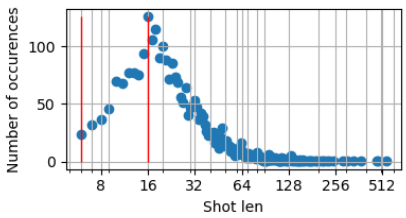

Fig. 7. Shot length in the trailer dataset built. First red line is the shortest shot (6 frames). Second red line is the most occuring shot length (16 frames).

Padding: Instead of dropping the remaining frames after snippets splitting, zero-padding could be used until completion of 16-frame snippets. This padding resulted in worse performance, no matter the configuration of the classifier or the used model. From TSP features, The decrease of accuracy was of 0.93% for default MLP, 1.87% for Data Augmentation MLP, 2.59% for Drop Out MLP, and 2.95% for SITI MLP.

Resampling: Our original data is based on movies standard frame-rate, *i.e.*, 24 frames per second.

Considering that TSP sub-samples its 30 fps videos and is trained on a 15 fps basis, a possible approach would be to re-sample our own movies from 24 fps to 15. However, as depicted in Fig. 7, the distribution of shot lengths in our dataset reveals that 16-frame shots are quite prevalent.

This observation can be attributed to the fact that the available data consists of movie trailers rather than full-length movies. Hence, when re-sampling the data to 15 fps, the resulting snippets need to be padded to reach a length of 16 frames. Without this padding, the dataset would have insufficient data. To assess the impact of this added padding, an additional test involving TSP without re-sampling but with padding is included in the ablation study mentioned earlier.

TSP 15 fps with padding performs most of the time better than TSP 24 fps with padding on a given configuration: -0.04% accuracy on default MLP but $+1.77\%$ on Data Augmentation MLP, $+0.20\%$ on Drop Out MLP, $+2.93\%$ on SITI MLP. This observation is probably due to the fact that TSP is trained on 15 fps videos in the original paper. TSP 15 fps with no padding would possibly be an option to explore, except that too many shots are too short in our trailer dataset i.e. 24 frames long or fewer, giving a single snippet after resampling (Fig. 7). If possible, re-sampling accordingly to the trained frame rate is a viable option to improve performance.

IV. Conclusion

This work introduces a pipeline that automates the generation of lateral content for immersive display systems using movies. By leveraging temporal classification models as backbones, the method effectively determines the optimal processing techniques for different cinematic shots. Through extensive experiments, the potential of temporal deep learning models for generating immersive contexts is demonstrated, showcasing promising accuracy and highlighting the advantages over existing approaches like generative deep learning models or patch matching. The exploration of different classifier configurations yields some improvements compared to the baseline, indicating the potential for further enhancement. However, due to restrictions imposed by producers, the study was constrained to a limited dataset of 30 trailers. Future directions include expanding the dataset to include full-length movies, allowing for increased quantity and diversity in the training data. Additionally, the architecture can be improved by incorporating perceptual models and diverse features, which would contribute to further advancements in performance. Moreover, A subjective experiment involving both naive and expert observers is planned to study validity of the predicted results.

References

[1] D. E. Novy, "Computational immersive displays," Ph.D. dissertation, Massachusetts Institute of Technology, 2013.
[2] The IMAX Difference — IMAX. [Online]. Available: https://www.imax.com/content/imax-difference
[3] Barco Escape Review. [Online]. Available: https://www.slashfilm.com/533916/barco-escape-review-maze-runner/
[4] J. Lee, S. Lee, Y. Kim, and J. Noh, "ScreenX: Public Immersive Theatres with Uniform Movie Viewing Experiences," *IEEE Trans Vis Comput Graph*, vol. 23, no. 2, pp. 1124–1138, Feb. 2017.
[5] S. Lee, J. Lee, B. Kim, K. Kim, and J. Noh, "Video Extrapolation Using Neighboring Frames," *ACM Trans Graph*, vol. 38, no. 3, pp. 1–13, 2019.
[6] P. Seuntiens, I. Vogels, A. van Keersop, and AE. Eindhoven, "Visual Experience of 3D-TV with pixelated Ambilight," 2007.
[7] R. Xiao and H. Benko, "Augmenting the Field-of-View of Head-Mounted Displays with Sparse Peripheral Displays," in *Proceedings of ACM CHI 2016*, May 2016.
[8] P. Lubos, . Bruder, O. Ariza, and F. Steinicke, "Ambiculus: LED-based low-resolution peripheral display extension for immersive head-mounted displays," in *Int. Conf. Virtual Reality* , New York, NY, 2016, pp. 1–4.
[9] N. Kimura, M. Kono, and J. Rekimoto, "Deep dive: Deep-neural-network-based video extension for immersive head-mounted display experiences," in *8th ACM Int. Symp. on Pervasive Displays*, Palermo Italy, 2019, pp. 1–7.
[10] Infinity-by-Nine. MIT Media Lab. [Online]. Available: https://www.media.mit.edu/projects/infinity-by-nine/overview/
[11] C. Cruz-Neira, D. J. Sandin, T. A. DeFanti, R. V. Kenyon, and J. C. Hart, "The CAVE: Audio visual experience automatic virtual environment," *Communications of the ACM*, vol. 35, no. 6, pp. 64–73, Jun. 1992.
[12] N. Kimura and J. Rekimoto, "ExtVision: Augmentation of Visual Experiences with Generation of Context Images for a Peripheral Vision Using Deep Neural Network," in *ACMCHI Conf. on Human Factors in Computing Systems*, New York, NY, 2018, pp. 1–10.
[13] P. Baudisch, N. Good, and P. Stewart, "Focus plus context screens: Combining Display Technology with Visualization Techniques."
[14] F. L. Kooi and M. Mosch, "Peripheral Motion Displays: Tapping the Potential of the Visual Periphery," *50th ANNUAL MEETING*, 2006.
[15] L. Dehan, W. Van Ranst, P. Vandewalle, and T. Goedeme, "Complete and temporally consistent video outpainting," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. New Orleans, LA, USA: IEEE, Jun. 2022, pp. 686–694.
[16] Ice Theaters by CGR Cinémas. ICE Theaters. [Online]. Available: https://www.icetheaters.com/ice-theaters
[17] H. Xia and Y. Zhan, "A Survey on Temporal Action Localization," *IEEE Access*, vol. 8, pp. 70 477–70 487, 2020.
[18] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning Spatiotemporal Features with 3D Convolutional Networks," Oct. 2015.
[19] J. Carreira and A. Zisserman, "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset," Feb. 2018.
[20] H. Alwassel, S. Giancola, and B. Ghanem, "TSP: Temporally-Sensitive Pretraining of Video Encoders for Localization Tasks," Aug. 2021.
[21] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "SlowFast Networks for Video Recognition," Oct. 2019.