

Computer Vision Based Freeze Concealment for Digital Multimedia Broadcasting

Siamak Khatibi*, Hans-Jürgen Zepernick*, Muhammad Imran Iqbal[†], Thi My Chinh Chu*, Yao Yong[‡],
Fatemeh Tavakoli*, Momin Jamil[§], Rolf Esslinger[§], and Trung Q. Duong[¶]

*Blekinge Institute of Technology, Karlskrona, Sweden

[†]Axis Communications AB, Lund, Sweden

[‡]DozeroTech AB, Karlskrona, Sweden

[§]Harman/Becker Automotive Systems GmbH, Karlsbad, Germany

[¶]Memorial University, St. John's, Canada

Abstract—Broadcasting of compressed video to mobile devices has to cope with time-varying channels and stringent latency constraints. These impairments can cause packet loss, frame drops, and frame freeze, which may lead to severe visual distortions. In this paper, computer vision techniques together with lightweight statistical learning are proposed to construct an approximate geometric scene representation from the last correctly received frame. The underlying concepts of this freeze concealment approach resort to normalized cuts and multi-hypothesis region grouping. The perceptual effectiveness of the proposed approach is evaluated using subjective tests for digital multimedia broadcasting (DMB) videos recorded in a live DMB system in South Korea. The obtained perceptual quality has been improved for all frozen video clips after freeze concealment regarding mean opinion scores (MOS), MOS change of video clips after computer vision based freeze concealment, and boxplots of opinion scores of video clips before and after freeze concealment. Because audio and video data of DMB is encapsulated in the prominent MPEG transport stream, the proposed approach can be easily adapted to other digital television transmission or streaming systems.

Index Terms—Digital Multimedia Broadcasting, Freeze Concealment, Normalized Cuts, Virtual 3D Scenes, Subjective Tests.

I. INTRODUCTION

Broadcasting systems increasingly target mobile devices operating under time-varying channel conditions and stringent latency constraints. In such deployments, transmission impairments commonly induce packet loss, frame drops, and frame freezing, where the same video frame is repeatedly displayed for several refresh cycles. Among these impairments, frame freezing is particularly disruptive because it breaks temporal continuity and is immediately noticeable to viewers, even when spatial image quality remains high. Studies have shown that playback interruptions and temporal stalls more severely impact on subjective quality than many spatial distortions [1].

Decoder-side error concealment techniques have therefore remained an important component of practical video delivery systems. These techniques operate at the pixel, block, or frame level, centered around spatial interpolation, temporal repetition, or motion-compensated prediction [2]. Although

computationally efficient and compatible with contemporary video coding standards, these methods become increasingly ineffective as freeze durations increase. Simple frame repetition quickly becomes visually unacceptable, and block-based interpolation often lacks sufficient motion or structural information to produce perceptually plausible results.

More recently, neural rendering approaches have shown remarkable ability to synthesize novel views of complex scenes by learning continuous volumetric representations from multiple observations [3]. While such methods are powerful for view synthesis, they require dense input views, extensive offline training, and substantial computational resources. These limitations have motivated increasing interest in content-aware and structure-driven concealment strategies, which aim to exploit higher-level visual information. At the same time, the computer vision community has made significant advances in image and video inpainting, where missing regions or frames are reconstructed using learned priors. Recent surveys document the success of deep generative models in hallucinating temporally and spatially consistent content, e.g., [4]. Similarly, progress in monocular depth estimation and scene understanding has demonstrated that approximate 3D scene structure can be inferred from single images or monocular video streams using deep neural networks [5].

In contrast, the approach pursued in this paper avoids heavy deep-learning pipelines and neural rendering frameworks. Instead, it leverages classical computer vision techniques combined with lightweight statistical learning to construct an approximate geometric scene representation from the last correctly received frame. This design choice prioritizes computational efficiency, interpretability, and feasibility for deployment in broadcast receivers, while still enabling the synthesis of visually plausible motion during freeze intervals. In particular, this paper investigates a computer vision-based freeze concealment framework that integrates super-pixel segmentation, multi-hypothesis region grouping, and probabilistic geometric labeling to form a virtual 3D scene from a single frozen frame. By applying controlled virtual camera motion to this scene, a short sequence of replacement frames is rendered

to preserve perceptual continuity during video freeze until video playback resumes. The objective is not to predict the true future content, but to maintain motion cues and scene dynamics that reduce the perceptual salience of freeze events. The main contributions of this work are as follows:

- We propose a freeze concealment pipeline that constructs an approximate geometric scene representation from a single frozen frame, enabling view synthesis without relying on multi-frame input or large-scale training data.
- We introduce a multi-hypothesis region grouping strategy to improve robustness of geometric labeling and reduce sensitivity to segmentation initialization.
- We evaluate the perceptual effectiveness of the proposed approach using subjective tests for the case of Digital Multimedia Broadcasting (DMB).

II. RELATED WORK

A. Video Error Concealment and Freeze Handling

Video error concealment has been extensively studied in the context of lossy transmission. This includes codec techniques that exploit spatial and temporal redundancy for recovering lost data without the need of modifying contemporary codecs. Comprehensive surveys point to spatial interpolation, temporal frame copying, and motion-compensated concealment as dominant approaches in block-based video codecs [2], [6]. While these approaches are effective for short losses, their performance tends to rapidly degrade with increasing freeze durations. Several studies have assessed the perceptual impact of freezes and playback interruptions on the quality of experience, showing that viewers are highly sensitive to temporal disruptions even when spatial quality is preserved [1]. This has motivated concealment strategies to minimize reconstruction errors while maintaining perceptual continuity. Nevertheless, most standardized concealment solutions are limited to low-level signal processing operations due to complexity constraints in consumer-grade receivers.

B. Image Segmentation and Super-Pixel Representations

Image segmentation enables higher-level reasoning about visual scenes. Graph-based segmentation has gained increased attention exploiting normalized cuts as part of a global optimization framework balancing inter-region dissimilarity and intra-region coherence [7]. Subsequent work focused on graph simplifications and hierarchical clustering to improve efficiency and scalability to support practical deployment for larger images [8]. Super-pixel-based representations were introduced to further reduce computational complexity while preserving structural information. Super-pixels group pixels into locally homogeneous regions that serve as intermediate representations for segmentation and scene analysis, e.g. [9].

C. Learning-Based Segmentation

Learning-based approaches have been proposed to infer semantic and geometric properties of image regions. Early work demonstrated that classifiers trained on appearance features could be used to assign geometric labels to image regions,

enabling coarse scene layout recovery from a single image [10]. These methods showed that even limited supervision can significantly improve region labeling consistency. With the rise of deep learning, monocular depth estimation and scene understanding have seen substantial progress. Surveys summarize how convolutional and transformer-based architectures can infer dense depth maps and surface orientations from a single image or video [5]. These methods achieve impressive accuracy but typically require large annotated datasets and substantial computational resources which limit their applicability in real-time broadcast receivers.

D. Image and Video Inpainting, Neural Rendering

Image and video inpainting using deep generative models aim to reconstruct missing regions or frames by learning strong priors over spatial and temporal content. These approaches potentially produce visually compelling results even under severe corruption [4]. Further, neural rendering techniques such as neural radiance fields have introduced a powerful framework for view synthesis by learning continuous volumetric scene representations from multiple input views [3]. However, their use is currently more aligned with offline rendering, media content creation, and high-end systems rather than real-time freeze concealment in broadcast receivers.

E. Positioning of the Presented Work

In contrast to deep generative inpainting and neural rendering approaches, the work presented in this paper focuses on a lightweight, receiver-feasible computer vision framework for freeze concealment. By combining super-pixel segmentation, multi-hypothesis region grouping, and probabilistic geometric labeling, the proposed method constructs an approximate 3D scene representation from a single frozen frame, avoiding the need for multi-frame input or extensive training. This positions the approach between classical low-level concealment techniques and modern data-intensive learning-based methods, offering a practical trade-off between perceptual quality, robustness, and computational complexity.

III. COMPUTER VISION APPROACH FOR FREEZE CONCEALMENT

A promising computer vision approach to commence concealment of freezes in video clips is to use the last best received frame. If such a frame has been determined, it is possible to create a virtual environment according to the information of that last frame. Then, a virtual camera movement can be utilized allowing to render new frames which replace the original video stream for the period of freezing. While other reconstruction techniques rely on multiple images, manual labeling, or specific scenes, the computer vision approach suggested in this paper aims at estimating the 3D geometrical properties of a scene from a frozen frame. The 3D creation process using the frozen frame has two major parts, i.e., generation of a super-pixel image and generating of a categorized (labeled) image. One way to estimate a 3D geometry of a frozen frame is to look at the problem as a statistical learning process instead

of deducing a sufficiently comprehensive set of geometrical parameters from the frozen frame. Given either indoor or outdoor scenes, it is possible to find common geometrical information associated with such images to furnish information about the frozen frame implicitly via recognition.

A. Generation of a Super-Pixel Image

This preprocessing groups a frozen frame's pixels into super-pixels, an approach motivated as follows [7], [10], [11]:

- Pixels are a consequence of a discrete image representation but not natural entities.
- Optimization on pixel level is intractable due to the high number of pixels even at moderate resolutions.

Super-pixels maintain essential spatial structures required for segmentation at the targeted scale while being local and coherent. As such, the normalized cuts algorithm proposed in [11] is used here to produce a so-called super-pixel map.

B. Normalized Cuts

In this section, normalized cuts and image segmentation proposed in [7] are reviewed to support the understanding of computer vision freeze concealment. Normalized cuts segmentation characterizes points of a feature space as a weighted undirected graph $G = (V, E)$. The graph G comprises of a set of vertices V with its vertices i and j being connected by bidirectional edges E_{ij} of the set E . Each edge E_{ij} is associated with a weight or similarity function w_{ij} measuring, e.g., distance, cost, or other relevant parameters. The objective of image segmentation is to form disjoint subsets $V_1, V_2, \dots, V_m$ of vertices such that similarity between the vertices of V_i with respect to a selected measure is high compared to their similarity with the vertices residing in the other sets $V_{j \neq i}$.

For the sake of exposition, let us consider the optimal partitioning of graph G into two disjoint subsets A of nodes u and B of nodes v through minimizing their cut value

$$\text{cut}(A, B) = \sum_{u \in A, v \in B} w(u, v) \quad (1)$$

where $\text{cut}(\cdot, \cdot)$ represents the dissimilarity. As minimizing the dissimilarity between two subsets using (1) tends to partition out small sets of points, a disassociated measure, referred to as normalized cut, is defined to alleviate this problem [7]:

$$\text{Ncut}(A, B) = \frac{\text{cut}(A, B)}{\text{assoc}(A, V)} + \frac{\text{cut}(A, B)}{\text{assoc}(B, V)} \quad (2)$$

where

$$\text{assoc}(X, V) = \sum_{u \in X, t \in V} w(u, t), \quad X \in \{A, B\} \quad (3)$$

In [7], it was shown that an optimal partitioning can be obtained by solving the following minimization problem:

$$\min_{\mathbf{x}} [\text{Ncut}(\mathbf{x})] = \min_{\mathbf{y}} \left[\frac{\mathbf{y}^T (\mathbf{D} - \mathbf{W} \mathbf{y})}{\mathbf{y}^T \mathbf{D} \mathbf{y}} \right] \quad (4)$$

where the diagonal matrix $\mathbf{D} = \text{diag}(d_i)$ contains the total connections from node i to all other nodes and the symmetrical

weight matrix $\mathbf{W} = [w_{ij}]$ indicates the likelihood of two nodes belonging to the same object. Furthermore, the relationship between the indicator vectors $\mathbf{y}$ and $\mathbf{x}$ is given by

$$\mathbf{y} = (\mathbf{1} + \mathbf{x}) - b(\mathbf{1} - \mathbf{x}), \quad b = \frac{\sum_{x_i > 0} d_i}{\sum_{x_i < 0} d_i} \quad (5)$$

where the elements x_i of the indicator vector $\mathbf{x} = [x_i]_{1 \times N}$ are $x_i = 1$ if node i is in A and $x_i = -1$ otherwise.

C. Modified Weights

An essential part of the normalized cut concept is the association or weighting function which quantifies the similarity or dissimilarity of the data. Since the weighting function is application dependent, modification to the specific target data needs to be considered. The original weighting function represents the weights for each edge of the graph as [7]

$$W = \exp \left(-\frac{d_k}{\sigma} \right)^2 \quad (6)$$

where d_k denotes the distance from node i to node j and σ is a scaling factor taken here in the range 5 – 10% of the maximum distance of the original work. After analyzing the histograms of distances among same points, it becomes clear that W needs to give more weight to the closer points than to points that are farther away. This requires a modified weighting function with adaptive σ for each node.

To obtain better clustering of the point data, an updated weighting function W_p is used which can emphasize closer points more than the ones that are farther away. This weighting improves the effect of minimization in the Ncut technique. The new weighting function with median μ , taken as the center of gravity, of each node is given as

$$W_p = \exp \left[-\frac{(d_k - \mu)^2}{\sigma_p^2} \right] \quad (7)$$

where $\sigma_p = k\mu$ with $k = 0.25$ is used in our study. This choice allows the scaling factor to reshape the weighting function adaptively to the specific input data.

D. Generation of a Categorized (Labeled) Image

To retrieve the coarse properties of the 3D geometric structure of an image, each super-pixel is assumed to correspond to a single category or label [12]. Super-pixels carry spatial information about image features such as color and texture which allows to compute their associated first order statistics. To estimate the 3D orientations of large-scale image regions, more complex geometric features need to be considered and evaluated over such larger regions. This task may be solved by using a super-pixel image to assess all possible segmentations to eventually find the best segmentation. Based on simple cues from sampled super-pixels, it is possible to generate potential multiple segmented regions according to some hypotheses in having certain geometrical surfaces. As the initial hypothesis for making such regions cannot be the final, each hypothesis' increased spatial support is used to better evaluate its quality.

Apart from inducing errors in different image regions, different hypotheses vary in the number of segments. It becomes therefore challenging to identify which parts of the various hypotheses could likely be considered as correct and accurately determine the label for those regions. By implementing a learning approach regarding segmentation, it is possible to produce multiple segmentations of an image into geometrically homogeneous regions. For this purpose, the likelihood of two super-pixels belonging to the same region is estimated with similarity being quantified through features such as color, texture, and spatial proximity.

Once the features of a region have been computed, it is possible to estimate the homogeneity likelihood (probability that all super-pixels have the same label) and label likelihood (confidence in each geometric label). The label confidence C of super-pixel i assigned label v (e.g., ground, vertical, non-planar), given image data $\mathbf{x}$, can be calculated as [12]

$$C(y_i = v|\mathbf{x}) = \sum_{j=1}^{n_h} P(y_j = v|\mathbf{x}, \mathbf{h}_{ij}) P(\mathbf{h}_{ij}|\mathbf{x}) \quad (8)$$

where n_h is the number of segmentation hypotheses, y_j denotes the label of the j th region, and $\mathbf{h}_{ij}$ defines the region containing the i th super-pixel given the j th segmentation hypothesis. Furthermore, $P(y_j = v|\mathbf{x}, \mathbf{h}_{ij})$ represents the label likelihood and $P(\mathbf{h}_{ij}|\mathbf{x})$ denotes the homogeneity likelihood.

E. Creation of a Virtual 3D Scene

A 3D scene based on plane surfaces is created by grouping the pixel information of the frozen frame according to the categorized/labeled image. The planes can represent the ground, vertical, horizontal, and frontal surfaces. The original frozen image data is mapped to the plane surfaces according to their geometrical position estimation (see example in Fig. 1).

F. Rendering new frames

By having the information of plane surfaces and creation of the virtual scene, it is possible to choose the position of a virtual camera and render a new frame. To offer a continuation of visual stimulation after a frozen frame, we need to estimate the camera position and its movement as smooth as possible for the viewer. For this purpose, we have used the camera movement pattern shown in Fig. 2. The initial position of the camera and the range of left and right as well as the depth can be estimated automatically by the information from the existing planes in the virtual scene. In this way, it is also possible to move straight forward by using the range of horizontal axis as zero. In addition, the speed of movement can be defined by depth range and number of required frames.

IV. STRUCTURE OF THE COMPUTER VISION APPROACH USED FOR FREEZE CONCEALMENT

This section provides an overview of the connection between the different functions used here for freeze concealment in digital television transmission and streaming systems.

Fig. 3(a) shows a block diagram of the overall procedure of freeze concealment using computer vision methodologies. The

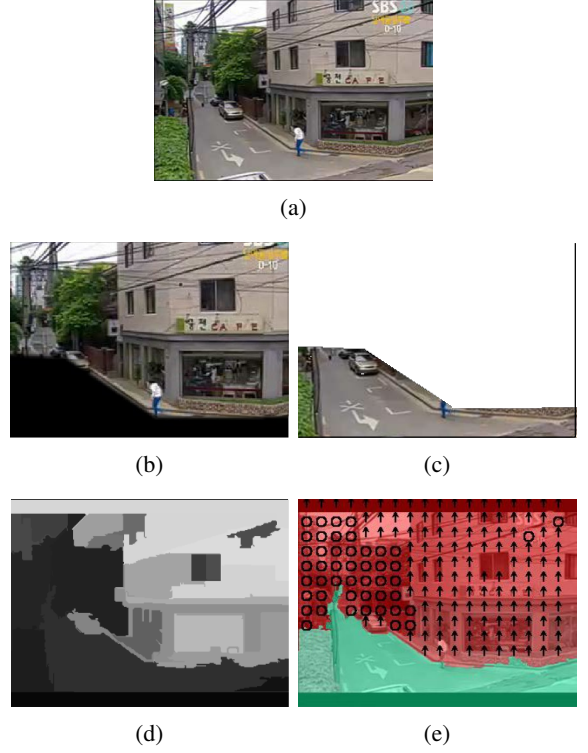


Fig. 1: Generating a virtual 3D scene: (a) Original frame, (b) Vertical planes, (c) Ground plane, (d) Super-pixel image, (e) Labeled frame (Green: ground; Red: vertical regions; Arrows: planar orientations; X: non-planar solid; O: porous).

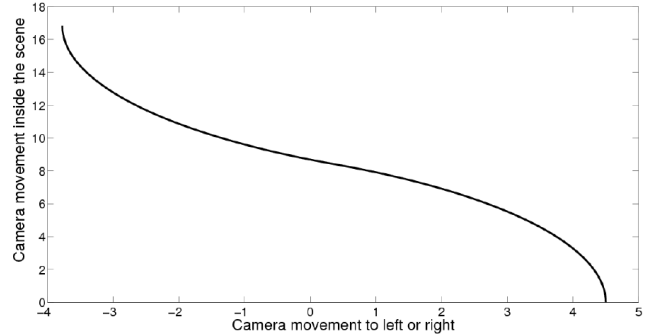


Fig. 2: Estimation of horizontal camera movement (left or right) and camera movement inside the scene (depth) as a basis for generating a 3D scene [13].

inputs to the respective algorithm are the frames f_{k-1} and f_k which may undergo an enhancement process and subsequent impairment detection. The outcome of the impairment detection may drive the texture concealment algorithm [14] and also serve as input to the freeze concealment path. If frame f'_{k-1} , identified by the impairment detection, does not indicate a freeze, the sequence of frames $f''_{k-1}, f''_k, \dots, f''_{k-1+N}$, will be streamed out until a freeze is detected. In the case of a freeze, the algorithm will commence with the creation of a 3D environment and then generate the sequence of N frames

to replace the frozen frames by releasing the sequence of frames $f''_{k-1}, f''_k, \dots, f''_{k-1+N}$. This new sequence shows then movements within the scene provided by the last frame before freezing. In this way, the illusion of dynamic content over time is produced which joins the video clip streaming to the point at which the freeze is over.

Fig. 3(b) shows the functions involved in the creation of the 3D environment based on the content and structural information of the last frame f'_{k-1} before freezing. It starts with the super-pixel processing as outlined in Section III-A. Then, geometries are computed as well as related labels and the corresponding confidences. After these functions have been done, the virtual modeling of the 3D environment is performed.

Fig. 3(c) shows the three processing steps needed for the generation of the N new frames. Given a cluster of N consecutive frames, the parameters of the actual 3D environment are estimated. In this work, we have used $N = 350$ to 400 frames but could use other values as well if it turns out that this would be more beneficial for an implementation on a target platform. On the basis of having an actual 3D environment generated, it is possible to estimate the camera movement for synthesizing the continuation of temporal dynamics until the freeze is over. The algorithm concludes with the rendering and release of the N new frames $f''_{k-1}, f''_k, \dots, f''_{k-1+N}$.

V. SUBJECTIVE TESTS

A. Test Material, Test Method, and Statistical Analysis

The subjective quality of video clips with freezes and after freeze concealment was assessed. The 8 video clips of duration 19 to 44s were recorded in a live DMB system containing freezes at different times and with different durations. The absolute category rating was chosen for the quality assessment using a 5-point Likert scale: (5) Excellent, (4) Good, (3) Fair, (2) Poor, (1) Bad. The arithmetic means over the opinion scores given to each test clip were calculated resulting in mean opinion scores (MOS). Boxplots are provided that reveal the distribution, median, and outliers of the opinion scores.

B. Test Panel

Demographic information of the 18 non-expert participants is shown in Table I. They were given the task to rate the quality of the video clips with freezes and after freeze concealment.

TABLE I: Demographic Information of the Test Panel

Size	18 non-expert participants
Age	23 to 50 years with an average of 26.7 years
Gender	4 females and 14 males
Occupation	Bachelor, Masters, Ph.D. students, Postdoc, Teacher
Nationality	5 China, 5 Vietnam, 2 Pakistan, 1 Sri Lanka, 4 Rwanda, 1 Italy

C. Test Results

Fig. 4 provides a quality comparison between the video clips with freezes (input) ordered according to decreasing MOS and the clips where the freezes have been concealed (output). As can be clearly seen from these results, the quality of all

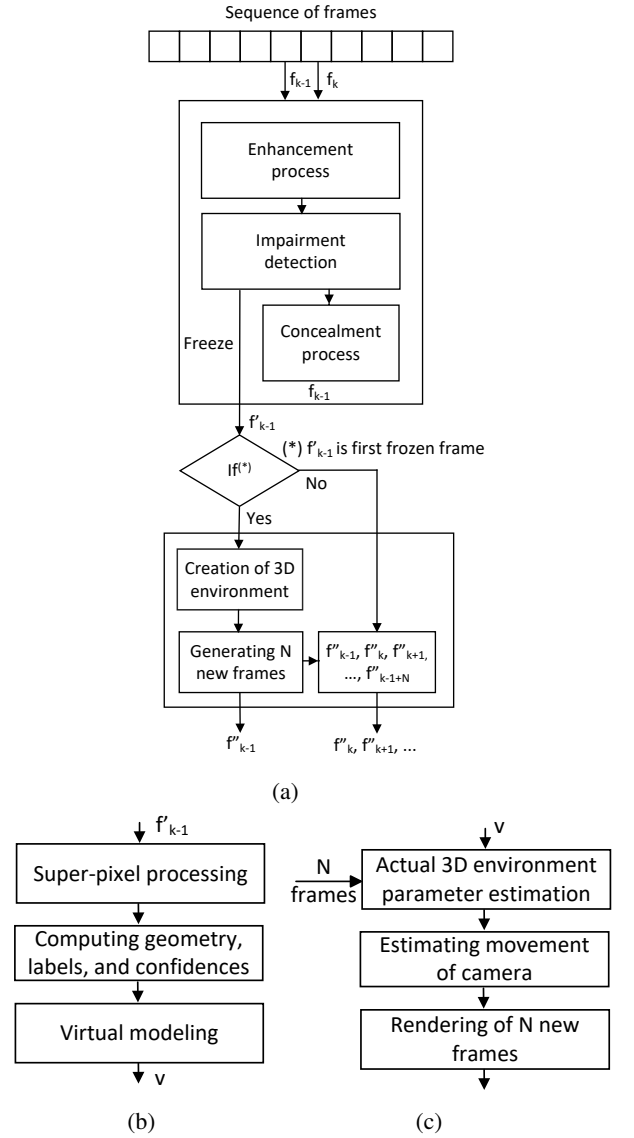


Fig. 3: Block diagram of the computer vision based freeze concealment approach: (a) Overall procedure, (b) Generating a 3D environment, (c) Generating a number of N new frames.

clips has been increased after computer vision processing (see also Fig. 5). The results support the conjecture that continuing with dynamics in the sequence of frames is preferable by the test subjects compared to leaving the clip frozen without concealment. Further quality improvements may be obtained in future work by considering alternative frame synthesis approaches in terms of percentage of viewing area covered.

Fig. 6 shows boxplots of the opinion scores for the video clips with freezes and after freeze concealment. It can be observed that the opinion score distribution for the majority of the freeze-concealed video clips shifts toward higher quality compared to the video clips with freezes. Significant improvements are obtained for the video clips C4, C5, and C8.

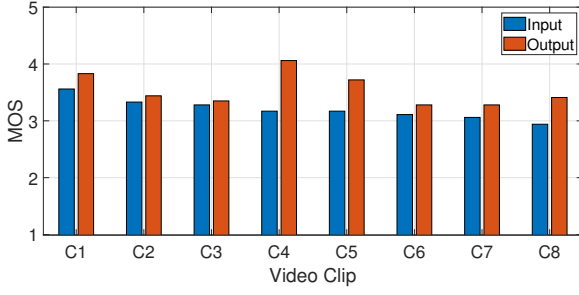


Fig. 4: MOS before and after freeze concealment.

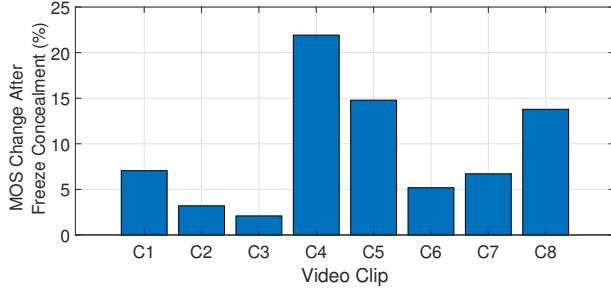


Fig. 5: MOS change after freeze concealment in percent.

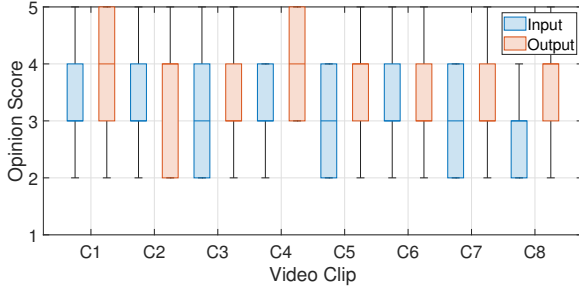


Fig. 6: Boxplots before and after freeze concealment.

VI. CONCLUSIONS

A computer vision based freeze concealment approach has been proposed allowing to construct an approximate geometric scene representation from the last correctly received frame. The approach along with lightweight statistical learning has been centered around normalized cuts and multi-hypothesis region grouping. The results of subjective tests have illustrated the perceptual effectiveness of the proposed approach for the visual test stimuli, including video clips recorded in a live DMB system. In particular, the statistical analysis has shown quality improvements for all frozen videos after freeze concealment. Since DMB data are encapsulated in the prominent MPEG transport stream, which is a standard for compressing digital audio and video, the proposed approach is well suited for adaptation to other digital television transmission or streaming systems.

ACKNOWLEDGMENT

We thank the volunteers for taking part in the subjective tests. R. Esslinger was with Harman/Becker Automotive Sys-

tems GmbH at the time the work was completed.

REFERENCES

- [1] J. Allard, A. Roskuski, and M. Claypool, "Measuring and modeling the impact of buffering and interrupts on streaming video quality of experience," in *Proc. Int. Conf. on Advances in Mobile Computing and Multimedia*, Chiang Mai, Thailand, 2020, pp. 153–160.
- [2] Y. Wang and Q.-F. Zhu, "Error control and concealment for video communication: A review," *Proceedings of the IEEE*, vol. 86, no. 5, pp. 974–997, 1998.
- [3] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NeRF: Representing scenes as neural radiance fields for view synthesis," in *Proc. European Conf. on Computer Vision*, Glasgow, UK, 2020, pp. 405–421.
- [4] W. Quan, J. Chen, Y. Liu, D.-M. Yan, and P. Wonka, "Deep learning-based image and video inpainting: A survey," *Int. J. of Computer Vision*, vol. 132, pp. 2367–2400, 2024.
- [5] U. Rajapaksha, F. Sohel, H. Laga, D. Diepeveen, and M. Bennamoun, "Deep learning-based depth estimation methods from monocular images and videos: A comprehensive survey," *ACM Computing Surveys*, vol. 56, no. 12, pp. 1–51, 2024.
- [6] S. Wenger, "Video redundancy coding in H.264/AVC," *IEEE Signal Process. Mag.*, vol. 20, no. 1, pp. 61–72, 2003.
- [7] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, 2000.
- [8] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *Int. J. of Computer Vision*, vol. 59, no. 2, pp. 167–181, 2004.
- [9] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC Superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, 2012.
- [10] D. Hoiem, A. A. Efros, and M. Hebert, "Recovering surface layout from an image," *Int. J. of Computer Vision*, vol. 75, no. 1, pp. 151–172, 2007.
- [11] J. R. Siddiqui, M. Havaei, S. Khatibi, and C. A. Lindley, "A novel plane extraction approach using supervised learning," *Machine Vision and Applications*, vol. 24, no. 6, pp. 1229–1237, 2013.
- [12] X. Ren and J. Malik, "Learning a classification model for segmentation," in *Proc. IEEE Int. Conf. on Comp. Vision*, Nice, France, 2003, pp. 10–17.
- [13] S. Khatibi, H.-J. Zepernick, and M. I. Iqbal, *Device and method for processing a stream of video data*, US Patent 9948959-B2, Apr. 2018.
- [14] T. M. C. Chu et al., "QoE of syntax, freeze, and texture repair techniques for digital multimedia broadcasting," in *Proc. IEEE Metacom*, Seoul, Republic of Korea, 2025, pp. 292–297.