Musical Hyperlapse: A Multimodal Approach to Accelerate First-Person Videos

Diognei de Matos, Washington Ramos, Luiz Romanhol, Erickson R. Nascimento
Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, Minas Gerais, Brazil
{diogneimatos, washington.ramos, luizromanhol, erickson}@dcc.ufmg.br

***Abstract*—With the advance of technology and social media usage, the recording of first-person videos has become a widespread habit. These videos are usually very long and tiring to watch, bringing the need to speed-up them. Despite recent progress of fast-forward methods, they generally do not consider inserting background music in the videos, which could make them more enjoyable. This paper presents a new methodology that creates accelerated videos and includes the background music keeping the same emotion induced by visual and acoustic modalities. Our methodology is based on the automatic recognition of emotions induced by music and video contents and an optimization algorithm that maximizes the visual quality of the output video and seeks to match the similarity of the music and the video's emotions. Quantitative results show that our method achieves the best performance in matching emotion similarity while also maintaining the visual quality of the output video when compared with other literature methods.**

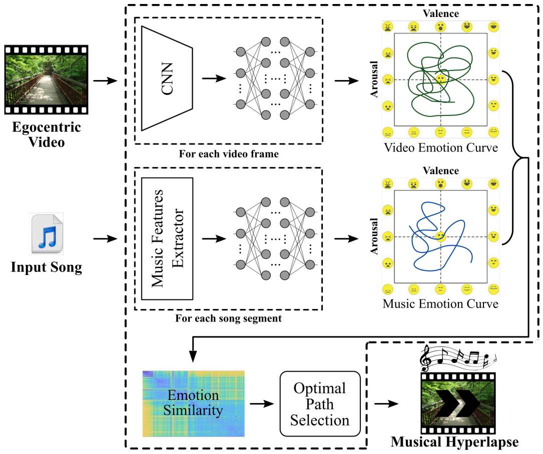

Fig. 1. **Music-driven video acceleration.** After computing the emotion similarity between the video and in the valence-arousal plane, our method accelerates the input video by removing frames according to an optimization algorithm that seeks the best matches between the video and the song.

I. Introduction

Recent years revealed an increasing volume of audio-visual data on the Internet due to the ease in people's access and usage of new digital technologies. The cost of multimedia mobile devices such as wearable cameras and smartphones constantly decreases while their storage capacity increases. As a result, many people start recording videos of their daily activities from an egocentric perspective, resulting in long and untrimmed streams. Usually, egocentric videos are tiring to watch since they contain redundant segments, and post-edition is commonly disregarded. Consequently, there has been a great interest in the computer vision community in reducing the total length of the videos to speed-up browsing and creating a pleasant watching experience.

Over the past several years, many works have been proposed to create a shorter accelerated version of egocentric videos using different strategies and under different restrictions to reduce the burden of watching the videos entirely [1]–[10]. The accelerated video is commonly called hyperlapse, where the goal is to optimize the output number of frames and the visual smoothness [5]. An extension of the classic hyperlapse is the semantic hyperlapse, which consists of hyperlapse videos with an additional restriction that strives to retain semantic information by exhibiting the most relevant parts at a lower speed-up rate [2], [3], [5]–[7], [9], [10]. Although both visual and sound streams play a major role in the video watching experience, generating videos that include audio is usually overlooked by the current hyperlapse techniques. Adding background music into an accelerated video is non-trivial. The music content must be associated with the video content considering different emotions induced by both to produce a final video that maintains the video and music contents' emotion similarity.

In this paper, we introduce a novel problem called *Musical Hyperlapse*, where the goal is to accelerate a video to the length of a song while matching the visual and audio signals to trigger, continuously, the same emotions during the output video exhibition. To tackle this problem, we propose a new multi-modal method to create hyperlapse videos based on synchronizing the feelings in the video scenes and background song. Specifically, given the predicted emotion curves for the video and audio streams, our approach seeks the best set of frames to be discarded in the video stream restricted to preserve the smoothness in the visual continuity and the matching between the emotion induced by segments of the video and audio (Figure 1).

Music plays an essential role in society, especially in our digital age. Since many music files are scattered across storage media, a need has arisen to classify them by different emotions. There are many works in the field of music emotion recognition [11]–[15], which consist of estimating the induced emotion by a specific piece of music. A classic representation of emotion is given by Thayer's model [11], where songs are classified with different labels around two axes: the valence and the arousal. By using this model, it is possible to represent emotions such as angry, delighted, calm, bored, etc., numerically. Since images also affect our affective states, our method applies the same model to classify the emotions induced by images and uses this audio-visual classification to synchronize an input video with background music when accelerating the video. Experiments in a variety of first-person videos and music showed that our method achieves the best performance in matching emotion similarity while conveying the original message in the untrimmed video and maintaining the visual quality of the hyperlapse when compared with other methods present in the literature.

Our contributions can be summarized as follows: *i)* a novel optimization algorithm to create hyperlapse videos whose function is to reduce the video by matching its emotion curve to the music emotion curve; and *ii)* a new dataset comprising several first-person videos and songs with different genres, sizes, and rhythms.

II. Related Work

A. Hyperlapse

Over the past decade, hyperlapse methods have been proposed to reduce the length of long egocentric videos. The evolution of these works is focused on improving the quality of the output video by keeping it as smooth and pleasant as possible, with the desired short length, and without losing essential information.

Kopf *et al.* [16] present a classical work in creating hyperlapse from first-person videos. The video is accelerated by using techniques based on image rendering, such as projecting, stitching, and blending after the optimal trajectory of the camera poses is computed. As a drawback, their method has a high computational cost and requires camera motion and parallax to compute the 3D model of the scene. Joshi *et al.* [1] presented a real-time hyperlapse creation algorithm that uses feature tracking to recover the camera motion and compute the optimal path with an algorithm inspired by dynamic programming and Dynamic Time Warping (DTW). Our work shares similarities with the work of Joshi *et al.* since our optimal path selection also draws inspiration from dynamic programming. However, unlike Joshi *et al.*, which consider only the visual modality during the optimization process, we handle two: the input visual stream and the output audio stream.

B. Semantic Hyperlapse

Recent approaches in hyperlapse include visual semantics as part of the optimization process. These methods, referred to as semantic hyperlapse, aim to accelerate the input video optimizing the camera stability, target speed-up rate, and semantics jointly.

Ramos *et al.* [2] introduced a new adaptive frame sampling process that considers the semantic information during the optimization. Their approach assigns a semantic score for each video frame and split the video into temporal segments according to their relevance. The authors applied different playback rates such that more relevant segments are exhibited at a lower rate. Their optimization balances the semantics and traditional hyperlapse objectives using energy cost minimization in a graph representing the frames' transition. Ramos *et al.*'s work was later extended by Silva *et al.* [17], where a homography-based stabilization was included in the process. Lai *et al.* [3] presented a system capable of converting a 360° video into a normal field-of-view hyperlapse. After determining the per-frame viewing directions to the regions of interest, their approach produces a saliency-aware frame selection that considers denser sampling at attractive regions and attending the target speed-up rate. Silva *et al.* [6], [9] modeled the adaptive frame sampling as a weighted minimum sparse reconstruction problem. Similar to the work of Ramos *et al.*, the Silva *et al.* split the video temporally using frame-wise levels of relevance. Then, each segment is represented as a dictionary from which the output video frames are sparsely selected, aiming to reduce abrupt camera motions.

Unlike previous works, which are mainly focused on visual information, Furlan *et al.* [7] proposed to use the input sound information. Their approach uses psychoacoustic metrics extracted from the video soundtrack to set the frames' importance. The original video's soundtrack is segmented, and for each segment, the Psychoacoustic Annoyance (PA) [18] is computed. The PA values guide the semantic hyperlapse creation since they are used as semantic scores. Although using the source audio in the optimization process, Furlan *et al.* ignored the audio in the output video, making their problem fundamentally different from ours. Our main goal is to create a hyperlapse with background music where both visual and acoustic signals induce similar emotions during the exhibition.

C. Emotion Recognition

Significant progress has been made by researchers in the field of music emotion recognition. Some of these works aim to classify an entire song to a specific emotion, such as happy, sad, angry, *etc.* [11]–[13]. Others focus on the prediction of arousal and valence emotional values from segment-wise continuous features extracted from the song [14], [15], [19].

According to Lu *et al.* [19], the features of music such as rhythm, melody, harmony, pitch, and timbre play an essential role in human physiological and psychological functions, altering their mood. With these features, the music mood can be divided into different types of moods. Some of these features, specifically the intensity, timbre, pitch, and rhythm, are acoustic features.

There are some emotion models used in music classification, such as Russell's model [20] and Thayer's model [21]. Russell's valence-arousal emotion plane is a widely used model in works about music emotion recognition. His emotion plane has as the x-axis the valence, which represents how pleasant is the feeling, and as the y-axis the arousal, which represents how exciting is the feeling. A similar and more intuitive emotion model is the EmojiGrid [22], which also considers valence and arousal as axes and with extreme locations represented with emojis. Figure 2 depicts the EmojiGrid diagram.

Yang *et al.* [11] formulated musical emotion recognition as a regression problem to predict the arousal and the valence

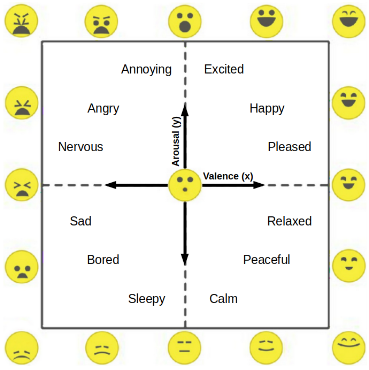

Fig. 2. **EmojiGrid emotion representation**. The x-axis represents how pleasant is the emotion (valence), and the y-axis represents how exciting is the emotion (arousal). The center of the plane represents a neutral state.

values of the music samples. Each music sample results on a point in the valence-arousal plane, and then the users can obtain the music sample by specifying a desired point in the plane. Panda *et al.* [12] introduced another approach to generate audio features to improve the classification performance. They reviewed the existing audio features obtained by state-of-the-art and their relationships with the musical concepts. The authors rely on clues like melodic lines, notes, intervals, and scores to access higher-level musical concepts such as harmony, melody, articulation, or texture. They also assigned importance to the determination of musical notes, frequency, and intensity contours mechanisms to capture the music information. Chowdhury *et al.* [13] aimed to create a model to provide a musically meaningful and intuitive explanation for its predictions. They proposed a VGG-style deep neural network to obtain emotional features from a music piece through human interpretable mid-level perceptual features, using the audio spectrogram as input.

Several researchers also seek to predict valence and arousal emotional values from segment-wise continuous features extracted from the song. Thammasan *et al.* [14] proposed a continuous music emotion recognition approach based on brainwave signals. Their experiment included self-reporting and continuous emotion annotation in the valence-arousal space. However, their approach only classifies valence and arousal into two classes: high/low valence and high/low arousal. Dong *et al.* [15], proposed a method to classify the songs continuously in time using segments of 0.5 seconds. They proposed a weighted hybrid binary representation (WHBR) method to convert the regression prediction process into a weighted combination of multiple binary classification problems, reducing the computational complexity.

There has also been a significant progress in image emotion recognition context. In the Affective Sciences, emotional scenes and facial expressions are some of the most essential stimuli [23]. Datasets relating images to emotions such as GAPED [24] have been created for research purposes both for attention and emotion. In image emotion recognition, the problem consists of retrieving the emotional content from an image. In general, the categorization of such images is made upon human annotation or automatically by using learned representations that rely on high and low-level features.

Joshi *et al.* [25] and Zhao *et al.* [26] explored the use of psychology and art-theory knowledge to determine which emotions may be evoked by a picture. However, as shown by Jia *et al.* [27], the use of high-level features like social network data when analyzing images is much more effective than raw low-level features such as primary colors in the image. Descriptive data also play a role in several solutions to recognizing the emotion induced by the image. For instance, the work of Borth *et al.* [28] uses pairs of adjectives and nouns to classify each picture. Mittal *et al.* [29], for their turn, take a wider range of objects in the scene later to sort the most important ones regarding the induced emotion.

Despite the progress of both emotion induced by music and images, it is worth noting that none of the works investigate the interplay between acoustic and visual signals regarding the induced feeling. Conversely, in this paper, we propose to apply both visual and acoustic data to accelerate a video by aligning segments with the emotion induced by the frames and music.

III. Methodology

We model the problem of accelerating a video according to the emotion induced by visual and acoustic information as a time-series matching problem. Formally, given a long first-person video $V = [v_1, v_2, \ldots, v_F]$ with F frames and a target song $M = [m_1, m_2, \ldots, m_S]$ with S segments, we aim at creating a shorter video $\hat{V} = [\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_S]$ by maximizing the matching between valence-arousal emotion curves $X \in \mathbb{R}^{F\times 2}$ and $Y \in \mathbb{R}^{S\times 2}$ of the video and audio, respectively. Figure 3 shows an overview of our method, divided into two main steps: *i)* Emotion Curves Creation and *ii)* Optimal Path Selection.

A. Emotion Curves Creation

In the first step, our method creates two emotion curves, one for the video stream and another for the audio stream. The values in these curves reflect the induced emotion at a certain timestep. Image and audio classifiers are used to estimate each value, as illustrated in Figure 3-left. Next, we detail the classifiers and the estimation of these curves.

1) Video Emotion Curve: To create the video emotion curve, frames of the video stream V are used to feed an image emotion classifier as $X' = \phi(V)$. The classifier ϕ outputs the valence and arousal values for each frame composing a discrete two-dimensional emotion curve $X' = [x'_1, x'_2, \ldots, x'_F]^T \in \{0,1\}^{F\times 2}$ (video emotion curve in Figure 3-left). We decomposed the curve into separated values of valence $X'_v = [x'_{v1}, x'_{v2}, \ldots, x'_{vF}]^T \in \{0,1\}^F$ and arousal $X'_a = [x'_{a1}, x'_{a2}, \ldots, x'_{aF}]^T \in \{0,1\}^F$. Thus, the video frame v_i has the coordinates x'_{vi} and x'_{ai} that represent it in a quadrant in the valence-arousal plane. The frame is classified as inducing a positive valence if $x'_{vi} = 1$ and negative valence otherwise, and classified as inducing a high arousal if $x'_{ai} = 1$ and low otherwise.

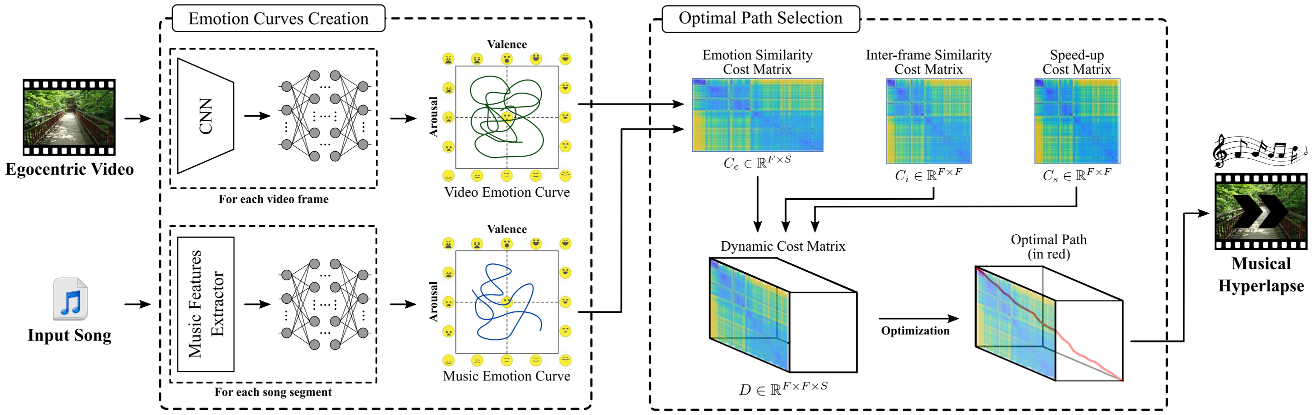

Fig. 3. **Methodology Overview.** After extracting features from each video frame and each song segment and classify them to obtain their induced emotion in the first step, we use the classification to create continuous emotion curves in the valence-arousal plane. In the second step, we calculate inter-frame and cross-modal cost matrices to create a dynamic cost matrix used to compute an optimal path that aligns the emotion induced by a song with the emotion induced by the frames while preserving the visual and temporal continuity.

We use a pretrained 2D-CNN (ResNet-50 [30]) as a backbone network topped with a fully-connected network to approximate the function ϕ. To train the network, we use the MVSO dataset [31]. This dataset comprises about 7 million images and their respective concepts defined in the form of adjective-noun pairs such as *colorful-clouds*, *tiny-dog*, *old-books*, *crying-baby*, and others. Each of these adjective-noun pairs is associated with a distribution over the 24 emotion categories (*e.g.*, joy, anger, sadness) from Plutchik's Wheel of Emotions [32]. We converted these categories to the valence-arousal plane to create the final valence-arousal labels for the images in the MVSO dataset. For each image, we took the predominant emotion out of the 24 and use its location in the plane as label. Finally, we randomly split the final set into training, validation, and test sets in the proportion 70:15:15 and perform the training using the cross-entropy loss. During the training, the feature extraction layers were kept frozen.

In the inference, the discrete video emotion curve is converted to a continuous emotion curve as $X = f(X') \in \mathbb{R}^{F\times 2}$, where $f : \{0, 1\} \to \mathbb{R}$ is a smoothing function that applies a quadratic interpolation to the sequential values.

2) Music Emotion Curve: Similarly, to create the music emotion curve for an audio stream M, we use the music emotion classifier $Y' = \psi(M)$ that provides the discrete curve $Y' = [y'_1, y'_2, \ldots, y'_S]^T \in \{c_1, c_2, \ldots, c_N\}^{S\times 2}$, where N is the number of discrete categories, in which a song segment can be classified in the valence-arousal plane (music emotion curve in Figure 3-left). We decompose Y' as valence and arousal one-dimensional curves $Y'_v = [y'_{v1}, y'_{v2}, \ldots, y'_{vS}]^T \in \{c_1, c_2, \ldots, c_N\}^S$ and $Y'_a = [y'_{a1}, y'_{a2}, \ldots, y'_{aS}]^T \in \{c_1, c_2, \ldots, c_N\}^S$. Thus, given a song segment $m_k, k \in \{1, \ldots, S\}$, (y'_{vk}, y'_{ak}) is represented as one of the $N \times N$ points of a grid in the valence-arousal plane, where higher y'_{vk} values indicate a more positive valence and higher y'_{ak} values indicate a higher arousal.

Our music emotion classifier ψ is composed of a feature extractor topped with two fully-connected networks, one for each dimension (valence and arousal). To extract the features, we create a window of size $\alpha = 6$ seconds and slide it over the audio stream with a stride of $\delta = 0.5$ seconds to extract the mel-spectrogram. Then, following Panda *et al.* [33], we extract from each spectrogram a d-dimensional feature vector $\hat{\mathbf{m}}_k \in \mathbb{R}^d$ dedicated to the song. Finally, we feed each $\hat{\mathbf{m}}_k$ to the classifiers to obtain the discrete curves Y'_v and Y'_a.

We use the DEAM dataset [34] to train the music emotion classifier. The DEAM dataset comprises about 1,802 songs of various styles, such as rock, classic, country, and others, with durations between 45 and 400 seconds. For each song, some raters (10 in most cases) annotated its valence and arousal values in a range of $[-1, +1]$ at each step of 0.5 seconds, starting from the 15^{th} second of the song. There are approximately 126,000 annotated song segments in the entire dataset. To define the song segment label, we averaged the raters' annotated valence and arousal values after filtering all values distant by 0.5 standard deviations from the mean. Then, to create the pairs of segments and labels used in our training procedure, we discretize the valence and arousal annotations provided in the DEAM dataset into N classes. Similar to our image emotion recognition classifier, we train the music emotion recognition classifiers using training, validation, and test splits in the same proportion.

Note that, by using a stride of $\delta = 0.5$ seconds, during inference, we only obtain 2 samples per second, while the video stream operates at a higher rate, usually 30 frames per second. To match the video's sampling rate, we apply a linear interpolation in the valence and arousal values before applying a smoothing function that creates the final continuous curve $Y = g(Y') \in \mathbb{R}^{S\times 2}$. By smoothing the curves, we avoid creating abrupt transitions in time of the induced emotions.

B. Optimal Path Selection

After creating the emotion profile of the video and audio streams, we aim to find the optimal path that matches the emotion induced by the video frames and the song.

As stated, when shrinking the video size, besides of aligning the emotions in both modalities, we also need to produce a visually continuous video, *i.e.*, a video that presents a smooth motion during the exhibition. To attend to both objectives, we draw inspiration from the optimization process proposed in the work of Joshi *et al.* [1], which creates a hyperlapse video with smooth transitions between frames using dynamic programming and DTW based algorithm. However, unlike Joshi *et al.*, which optimize the output video path regarding only inter-frame transitions and on visual modality, in our work, we must consider not only the inter-frame transitions but also the audio-visual relation regarding the induced emotion. Therefore, our algorithm creates inter-frame and cross-modal cost matrices to perform the optimization.

To keep the video with a continuous visual motion, we create an Inter-frame Similarity Cost Matrix, $C_i \in \mathbb{R}^{F \times F}$, with each element computed as

$$C_i(i,j) = 1 - \mathrm{SSIM}(v_i, v_j), \tag{1}$$

where $i, j \in \{1, 2, \dots, F\}$ are the frames indices in the input video and $\mathrm{SSIM}(\cdot,\cdot)$ is the structural similarity index measure [35]. Higher SSIM values indicate that the input frames are more similar to each other. The algorithm also uses a cost matrix to avoid skips that are too distant from the target speed-up rate. Specifically, let $Sp^{\star} = F/S$ be the target speed-up rate. Each element in the Speed-up Cost Matrix, $C_s \in \mathbb{R}^{F \times F}$, is given by

$$C_s(i,j) = \min(((j-i) - \lfloor Sp^{\star} \rfloor)^2, c_{min}), \tag{2}$$

where c_{min} is a threshold empirically set to 200 as in Joshi *et al.* [1].

Finally, we create a cross-modal matrix to determine the cost of skipping relevant frames regarding the video and audio stream emotion similarity. The Emotion Similarity Cost Matrix, $C_e \in \mathbb{R}^{F \times S}$, is computed as

$$C_e(i,k) = \frac{\sqrt{(x_{vi} - y_{vk})^2 + (x_{ai} - y_{ak})^2}}{d_0}, \tag{3}$$

where $k \in \{1, 2, \dots, S\}$ is the song segment index, x_{vi} and x_{ai} are coordinates that represent the video frame in the valence-arousal plane, and y_{vk} and y_{ak} are coordinates representing the song segment. d_0 is the distance between the points $(+1, +1)$ and $(-1, -1)$ in the valence-arousal plane, which is used as a normalization factor.

The cost matrices C_i, C_s, and C_e are normalized to $[0, 1]$ and further used to create a 3D Dynamic Cost Matrix $D \in \mathbb{R}^{F \times F \times S}$. Each entry $D(i,j,k)$ represents the minimal cost of the path that ends at the frame v_j and song segment k. We also create a traceback matrix $T \in \mathbb{R}^{F \times F \times S}$ that stores in $T(i,j,k)$ the index of the frame that precedes v_j in the path, given the song segment k. Next, we populate D and T by setting the first song segment slice as $D(i,j,0) = C_s(i,j)$ and the following slices recursively as

$$\begin{aligned} D(i,j,k) = {} & \lambda_i C_i(i,j) + \lambda_s C_s(i,j) + \lambda_e C_e(j,k) \\ & + \min_{h=1}^{w}(D(i-h, i, k-1)), \end{aligned} \tag{4}$$

where λ_e, λ_s and λ_i are the weights associated with each cost term and w is the maximum skip between adjacent frames in the path. When populating D, we concurrently populate the traceback matrix by computing $T(i,j,k) = \arg\min_{1 \le h \le w} D(i-h, i, k-1)$.

TABLE I

Audio-visual Dataset. LIST OF VIDEOS AND SONGS USED FOR COMPARISON WITH BASELINES.

Video Name	Duration
Berkeley1 (Self-acquisition)	17:41
Berkeley2 (Self-acquisition)	13:40
Bike3 [16]	13:10
CityWalk1 (YouTube)	10:00
MontOldCity1 (YouTube)	10:01
NatureWalk1 (YouTube)	9:50
StockHolm1 (YouTube)	24:59
Walking4 [2]	8:43
Song Name	**Duration**
In The End (Linkin Park)	3:38
Little Talks (Of Monsters And Men)	4:23
My Immortal (Evanescence)	4:32
Onward to Freedom (Trailerhead)	2:58
Last To Know (Three Days Grace)	3:28

With matrices T and D filled, we traceback the optimal path, starting from position $k = S$, and selecting, at each step, the index stored in $T(i,j,k-1)$ while $k >= 0$. The reversed order of the frames selected during this step is the final set that composes the hyperlapse video. Note that exact S frames are selected. Therefore, the video length is reduced to the song length. We add the input audio stream to the composed hyperlapse video to generate the *musical hyperlapse* video.

IV. Experiments

A. Dataset, Evaluation Metrics, and Baselines

We organized a dataset composed of 8 videos presenting different scenes of nature, cities, parks, buildings, cars, people, animals, *etc.*; and 5 songs with varied styles and emotions. Table I shows the list of videos and songs used in the experiments. We collected the videos from various sources, including the YouTube platform, other works in the literature, and self-acquisition. The specific source of the video and the song authors are indicated right after the video and song names, respectively. We resampled all videos to the exact resolution of 640×480. The dataset is available on the project's webpage[1].

To assess the hyperlapse methods, we need to quantify the emotion induced by the video and audio streams, whether the target speed-up rate was achieved, and visual continuity and the stability of the final video. We quantify the emotion in the output video using the Emotion Similarity metric defined as

$$E_{sim} = \frac{1}{S} \sum_{k=1}^{S} \left(1 - \frac{\sqrt{(\hat{x}'_{vk} - y_{vk})^2 + (\hat{x}'_{ak} - y_{ak})^2}}{d_0} \right), \tag{5}$$

where $\hat{x}'_{vk}$ and $\hat{x}'_{ak}$ are the discrete valence and arousal values of the accelerated video $\hat{V}$.

[1] https://github.com/verlab/MusicalHyperlapse_SIBGRAPI_2021

Given a target speed-up rate $Sp^{\star}$, to verify whether the target speed-up rate was achieved, we use the Speed-up Ratio metric, calculated as $Sp_r = \max(Sp^{\star}, \hat{Sp})/\min(Sp^{\star}, \hat{Sp})$, where $\hat{Sp} = \hat{F}/S$ is the speed-up rate achieved by the hyperlapse method.

We also measure if the output visual content is similar to the input and if it is stable. To calculate the similarity, we use the Fréchet Inception Distance (FID) [36], which gives the similarity between two sets of images. We apply this metric to determine the similarity between the original and accelerated videos with respect to visual content. The lower the FID value, the more similar are the input and output videos. To compute the instability of the output frame transitions, we use the Shaking Ratio [10]. The Shaking Ratio uses homography transformations to calculate the average motion of the central pixel between pairs of frames' transitions.

We compare our methods against two hyperlapse baselines: the Microsoft Hyperlapse (MSH) [1], and the extended version of the Sparse Adaptive Sampling (SASv2) [9].

B. Implementation Details

We used a fully-connected network with 4 layers of $1{,}000$ neurons in the image and music emotion classifiers. The classification layer in the image emotion classifier comprises 4 neurons that represent each of the valence-arousal quadrants. In the music emotion classifiers, the classification layer consists of 8 neurons. The cost terms' weights were empirically set to $\lambda_e = 1.00$, $\lambda_i = 0.01$, and $\lambda_s = 0.01$. For our method and all its variants, we set the maximum allowed skip to $w = 2Sp^{\star}$ whose value is bounded to the interval $4 \leq w \leq 16$. We used the *essentia* Python library to extract the $d = 48$ music features used in the classifiers. Our method was fully implemented in Python. For the MSH baseline, we used the desktop version. For the SASv2, we set the hyperparameters as recommended by the authors.

C. Ablation Study

We evaluate the use of two simple path optimization approaches in the ablation study:

- **Greedy Approach:** This method greedily selects the next video frame with the maximum similarity for every song segment until it reaches the last segment. Given the emotion curves X and Y, for each $y_k, k \in \{1, 2, \ldots, S\}$ the method seeks the next frame, v_l, to store in the path by computing $l = \arg\min_{i=l}^{l+w} x_i$, where l stores the frame index of the last selected frame, initially set to $l = 1$.
- **Dynamic Time Warping (DTW):** An algorithm for measuring and aligning similarity between two temporal sequences [37]. To maximize the similarity of the input curves, the original DTW version may repeat video frames. Since this is not allowed in a hyperlapse, we adapted the method to our problem by adding a constraint that forces the algorithm to never repeat frames. We feed the algorithm with the X and Y curves.

Table II shows that our method achieved the best results across all metrics when using the optimal path selection algorithm. The greedy approach maximizes the emotion similarities locally, leading to a significant error in the achieved speed-up, which might remove important frames from the original video, resulting in a high FID. The DTW, in its turn, seeks to find the best alignment globally, which creates many gaps between segments reducing the representability of the accelerated video regarding the original one. Although DTW tries to match the curves, the need to prevent it from repeating frames makes it obtain emotion similarities close to those obtained by the greedy approach. Our method manages to maximize the emotion similarities without repeating frames, reaching the optimal speed-up ratio by taking the exact amount of frames required by the song, also maintaining a balance between frame transitions by using the speedup and inter-frame similarity cost matrices, guaranteeing a lower FID.

TABLE II
Ablation study. COMPARISON BETWEEN THE DIFFERENT OPTIMIZATION METHODS FOR FRAME SAMPLING (BEST IN BOLD).

Video	**Emotion Sim. ↑**			**Speedup Ratio ↓**			**FID-Score ↓**		
	Greedy	DTW	Ours	Greedy	DTW	Ours	Greedy	DTW	Ours
Berkeley1	0.74	0.74	**0.77**	1.23	1.03	**1.00**	22.06	22.14	**3.30**
Berkeley2	0.72	0.73	**0.77**	1.17	1.02	**1.00**	26.75	27.86	**5.40**
Bike3	0.72	0.72	**0.76**	1.16	1.02	**1.00**	16.75	18.10	**5.04**
CityWalk1	0.70	0.70	**0.71**	1.09	1.02	**1.00**	12.64	13.01	**1.75**
MontOldCity1	0.74	0.75	**0.77**	1.08	1.04	**1.00**	15.47	15.58	**3.10**
NatureWalk1	0.71	0.71	**0.73**	1.08	1.03	**1.00**	15.57	15.55	**2.73**
StockHolm1	0.71	0.71	**0.73**	1.36	1.01	**1.00**	37.58	36.07	**4.21**
Walking4	0.73	0.73	**0.76**	1.08	1.02	**1.00**	14.40	15.32	**2.74**
Mean	0.72	0.72	**0.75**	1.16	1.02	**1.00**	20.15	20.45	**3.53**

D. Results

Table III presents the results for the comparison with the baselines. The columns show the Emotion Similarity, Speed-up Ratio, FID-Score, and Shaking Ratio values for each video in the dataset averaged over the five songs from Table I. Our approach presents the best Emotion Similarity and Speed-up Ratio values while it is on par with the other methods in the Shaking Ratio. We accredit these results to our optimization algorithm that seeks to create a path that is visually stable, temporally continuous, and with high-quality emotion matching. Because our approach samples exact S frames from the input video, it also presents the best Speed-up Ratio values in all cases. MSH, on the flip side, presents the worst values. The reason is that it favors optimizing the stability of the frame transitions over achieving the target speed-up rate.

Although MSH generally presents the best Shaking Ratio values, since the MSH algorithm neglects the video content and only optimizes the frame transition, their FID-Score values are worse than the other approaches by a significant margin. Also, the MSH algorithm includes image warping in its path smoothing and rendering step. This step may crop the image borders; therefore, increasing the FID-Score. In comparison to the MSH, our method presents FID-Score values closer to the SASv2 method, which is, by design, a content-based approach.

Regarding the trained classifiers, the test accuracy obtained with the image classifier was 71% in the MVSO dataset, while for the audio classifiers it was 92% in the DEAM dataset.

TABLE III

Comparison with baselines. COMPARISON OF OUR METHOD AND TWO LITERATURE BASELINES.

Video	**Emotion Similarity ↑**			**Speedup Ratio ↓**			**FID-Score ↓**			**Shaking Ratio ↓**		
	MSH	SASv2	Ours	MSH	SASv2	Ours	MSH	SASv2	Ours	MSH	SASv2	Ours
Berkeley1	0.73	0.72	**0.79**	1.19	1.01	**1.00**	28.90	**4.30**	6.82	**0.02**	**0.02**	**0.02**
Berkeley2	0.72	0.71	**0.77**	1.25	1.01	**1.00**	34.03	**3.74**	7.44	**0.02**	**0.02**	**0.02**
Bike3	0.71	0.71	**0.77**	1.02	1.01	**1.00**	28.31	**3.02**	6.21	**0.03**	0.05	0.05
CityWalk1	**0.72**	0.70	**0.72**	1.57	**1.00**	**1.00**	32.52	**1.09**	2.55	**0.02**	**0.02**	0.03
MontOldCity1	0.74	0.73	**0.77**	1.31	1.02	**1.00**	41.09	**2.09**	4.46	**0.01**	**0.01**	**0.01**
NatureWalk1	0.72	0.71	**0.74**	1.47	1.03	**1.00**	48.43	7.28	**3.63**	**0.01**	**0.01**	**0.01**
StockHolm1	0.71	0.70	**0.74**	1.13	1.16	**1.00**	23.99	7.66	**5.13**	0.02	**0.01**	0.02
Walking4	0.73	0.73	**0.77**	1.12	**1.00**	**1.00**	37.62	**1.40**	3.34	**0.02**	0.03	0.03
Mean	0.72	0.71	**0.76**	1.26	1.03	**1.00**	34.36	**3.82**	4.95	**0.02**	**0.02**	**0.02**

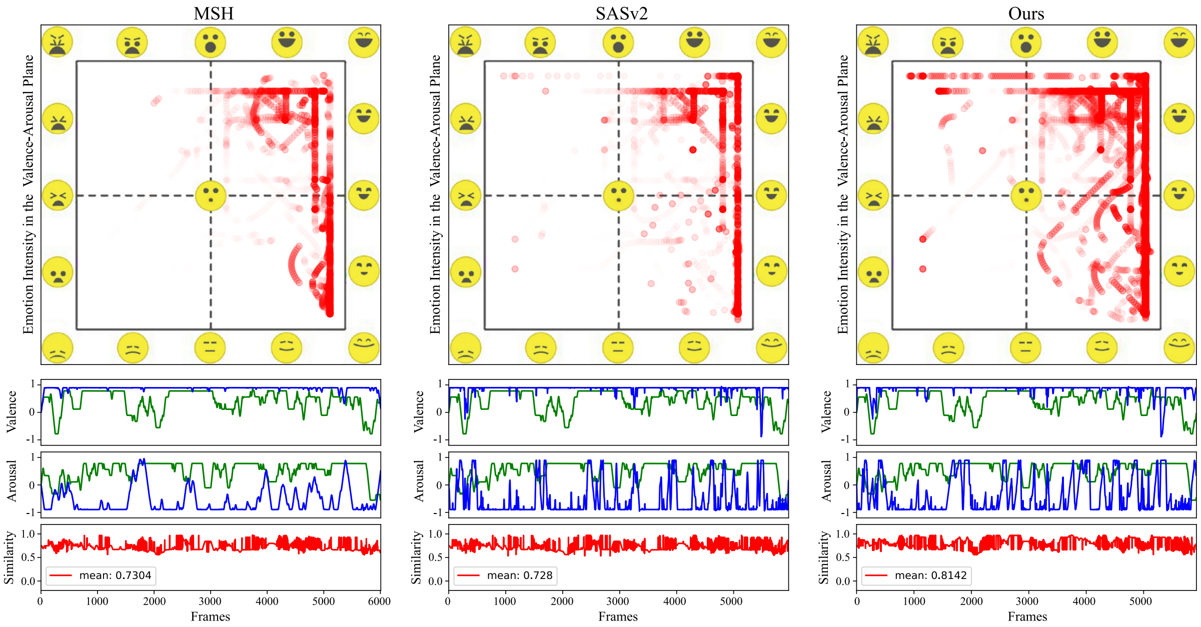

Fig. 4. **Qualitative comparison with baselines.** Each column represents the results of a method. At the top is the EmojiGrid with the similarities of emotions in the regions achieved by video and music emotion curves. The greater the red intensity, the greater the similarity. At the bottom we show the separate curves of valence and arousal throughout the video (blue) and music (green), and the similarity curve (red).

Figure 4 shows the qualitative results for the Emotion Similarity of the *musical hyperlapse* video generated from ‘Bike3’ with the song ‘In The End’. On top, we illustrate the distribution of emotion over the output video in the valence-arousal plane. Higher similarities in emotion curves depicted below the plane produce higher intensities in the plane location. The blue curve represents the video, and the green one the song. The curves similarity is represented by the red curve, at the bottom. Our method presents a distribution with higher intensities in the valence-arousal plane, indicating a higher matching in the induced emotion for the hyperlapse video. MSH and SASv2, on the other hand, have a sparse concentration of correct matching. Additional qualitative results are available on the supplementary material.

1) Limitations: Our method may fail when the video and music emotion curves X and Y are too different, making it challenging to yield a good match. An example case is depicted in Figure 5, which shows the Emotion Similarity of the *musical hyperlapse* video generated from ‘CityWalk1’ with the song ‘Last To Know’. Note that from the 3,000th frame, the song (green curve) increases the induced arousal and reduces the valence while the video (blue curve) induces the opposite.

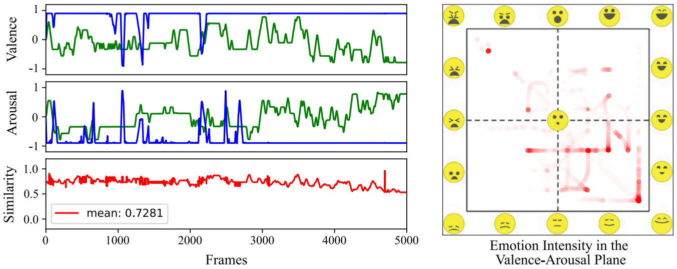

Fig. 5. **Failure case.** A particular case in which our method obtained a lower similaritydue to the low similarity between the song and video.

V. Conclusion

This paper introduced the novel task of accelerating first-person videos and aligning the emotion induced by visual and acoustic signals. As a solution for this task, we proposed a new multimodal method capable of accelerating egocentric videos while taking visual information in the video scenes and audio information in the background music. We also presented a new multimodal dataset comprising different videos and songs. The proposed method achieved superior performance in terms of video representation, required speed-up and emotional alignment for different videos and songs without losing the visual quality of the hyperlapse, as compared to previous methods. The results show that it is possible to create a hyperlapse combining media of distinct nature according to their respective affective semantic. For future work, it is possible to improve the emotion recognition models and perform experiments with more videos.

Acknowledgments

The authors would like to thank CAPES, CNPq, and FAPEMIG for funding this work. We also thank NVIDIA Corporation for the donation of a TITAN Xp GPU.

References

[1] N. Joshi, W. Kienzle, M. Toelle, M. Uyttendaele, and M. F. Cohen, "Real-time hyperlapse creation via optimal frame selection," ACM Trans. Graph., vol. 34, no. 4, Jul. 2015.

[2] W. L. S. Ramos, M. M. Silva, M. F. M. Campos, and E. R. Nascimento, "Fast-forward video based on semantic extraction," in 2016 IEEE International Conference on Image Processing (ICIP), Sep. 2016, pp. 3334–3338.

[3] W.-S. Lai, Y. Huang, N. Joshi, C. Buehler, M.-H. Yang, and S. B. Kang, "Semantic-driven generation of hyperlapse from 360 video," ArXiv, vol. abs/1703.10798, 2017.

[4] T. Halperin, Y. Poleg, C. Arora, and S. Peleg, "Egosampling: Wide view hyperlapse from egocentric videos," IEEE Transactions on Circuits and Systems for Video Technology, vol. 28, no. 5, pp. 1248–1259, 2018.

[5] M. M. Silva, W. L. S. Ramos, F. C. Chamone, J. P. K. Ferreira, M. F. M. Campos, and E. R. Nascimento, "Making a long story short: A multi-importance fast-forwarding egocentric videos with the emphasis on relevant objects," Journal of Visual Communication and Image Representation, vol. 53, p. 55 – 64, 2018.

[6] M. M. Silva, W. L. S. Ramos, J. P. K. Ferreira, F. C. Chamone, M. F. M. Campos, and E. R. Nascimento, "A weighted sparse sampling and smoothing frame transition approach for semantic fast-forward first-person videos," in 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, USA, Jun. 2018, pp. 2383–2392.

[7] V. S. Furlan, R. Bajcsy, and E. R. Nascimento, "Fast forwarding egocentric videos by listening and watching," in In IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshop on Sight and Sound. IEEE Computer Society, 2018, p. 2504–2507.

[8] M. Wang, J.-B. Liang, S.-H. Zhang, S.-P. Lu, A. Shamir, and S.-M. Hu, "Hyper-lapse from multiple spatially-overlapping videos," Trans. Img. Proc., vol. 27, no. 4, p. 1735–1747, Apr. 2018.

[9] M. Silva, W. Ramos, M. Campos, and E. R. Nascimento, "A sparse sampling-based framework for semantic fast-forward of first-person videos," vol. 43, no. 4, pp. 1438–1444, 2021.

[10] W. L. S. Ramos, M. M. Silva, E. R. Araujo, A. C. Neves, and E. R. Nascimento, "Personalizing fast-forward videos based on visual and textual features from social network," in 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 3260–3269.

[11] Y. Yang, Y. Lin, Y. Su, and H. H. Chen, "A regression approach to music emotion recognition," IEEE Transactions on Audio, Speech, and Language Processing, vol. 16, no. 2, pp. 448–457, Feb 2008.

[12] R. Panda, R. M. Malheiro, and R. P. Paiva, "Novel audio features for music emotion recognition," IEEE Transactions on Affective Computing, pp. 1–1, 2018.

[13] S. Chowdhury, A. Vall, V. Haunschmid, and G. Widmer, "Towards explainable music emotion recognition: The route via mid-level features," International Society for Music Information Retrieval Conference, 07 2019.

[14] N. Thammasan, K. Moriyama, K.-i. Fukui, and M. Numao, "Continuous music-emotion recognition based on electroencephalogram," IEICE Transactions on Information and Systems, vol. E99.D, pp. 1234–1241, 04 2016.

[15] Y. Dong, X. Yang, X. Zhao, and J. Li, "Bidirectional convolutional recurrent sparse network (bcrsn): An efficient model for music emotion recognition," IEEE Transactions on Multimedia, pp. 1–1, 2019.

[16] J. Kopf, M. Cohen, and R. Szeliski, "First-person hyperlapse videos," in ACM Transactions on Graphics (Proc. SIGGRAPH 2014), vol. 33. ACM - Association for Computing Machinery, August 2014.

[17] M. M. Silva, W. L. S. Ramos, J. P. K. Ferreira, M. F. M. Campos, and E. R. Nascimento, Towards Semantic Fast-Forward and Stabilized Egocentric Videos, Amsterdam, NL, Oct. 2016, p. 557–571.

[18] E. Zwicker and H. Fastl, Psychoacoustics: Facts and models. Springer Science & Business Media, 2013, vol. 22.

[19] Lie Lu, D. Liu, and Hong-Jiang Zhang, "Automatic mood detection and tracking of music audio signals," IEEE Transactions on Audio, Speech, and Language Processing, vol. 14, no. 1, pp. 5–18, Jan 2006.

[20] A. Alpher, "A circumplex model of affect," Journal of Personality and Social Psychology, vol. 39, no. 6, pp. 1161—-1178, 1980.

[21] R. E. Thayer, "The biopsychology of mood and arousal," 1989, oxford University Press.

[22] A. Toet and J. B. van Erp, "The emojigrid as a tool to assess experienced and perceived emotions," Psych, vol. 1, no. 1, pp. 469–481, 2019.

[23] A. Toet and v. Erp, "Emomadrid: An emotional pictures database for affect research," 12 2019.

[24] S. E. Dan-Glauser and R. K. Scherer, "The geneva affective picture database (gaped): a new 730-picture database focusing on valence and normative significance," Behavior Research Methods, pp. 468–477, 2011.

[25] D. Joshi, R. Datta, E. Fedorovskaya, Q. Luong, J. Z. Wang, J. Li, and J. Luo, "Aesthetics and emotions in images," IEEE Signal Processing Magazine, vol. 28, no. 5, pp. 94–115, 2011.

[26] S. Zhao, Y. Gao, X. Jiang, H. Yao, T.-S. Chua, and X. Sun, "Exploring principles-of-art features for image emotion recognition," MM 2014 - Proceedings of the 2014 ACM Conference on Multimedia, pp. 47–56, 11 2014.

[27] J. Jia, S. Wu, X. Wang, P. Hu, L. Cai, and J. Tang, "Can we understand van gogh's mood? learning to infer affects from images in social networks," ACM International Conference on Multimedia, 10 2012.

[28] D. Borth, R. Ji, T. Chen, T. Breuel, and S. Chang, "Large-scale visual sentiment ontology and detectors using adjective noun pairs," 10 2013, pp. 223–232.

[29] T. Mittal, P. Guhan, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "Emoticon: Context-aware multimodal emotion recognition using frege's principle," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2020.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.

[31] V. Dalmia, H. Liu, and S. Chang, "Columbia mvso image sentiment dataset," ArXiv, vol. abs/1611.04455, 2016.

[32] R. Plutchik, Emotion, a Psychoevolutionary Synthesis. Harper & Row, 1980.

[33] R. Panda, R. M. Malheiro, and R. P. Paiva, "Audio features for music emotion recognition: a survey," IEEE Transactions on Affective Computing, pp. 1–1, 2020.

[34] M. Solymanil, A. Aljanakil, and Y.-H. Yang, "DEAM: Mediaeval database for emotional analysis in music," 2018.

[35] Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," IEEE Transactions on Image Processing, vol. 13, no. 4, pp. 600–612, 2004.

[36] A. Mathiasen and F. Hvilshøj, "Fast fréchet inception distance," 2020, aarhus University.

[37] M. Müller, "Dynamic time warping," Information Retrieval for Music and Motion, vol. 2, pp. 69–84, 01 2007.