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

Diognei de Matos, Erickson R. Nascimento
Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, Minas Gerais, Brazil
{diogneimatos, erickson}@dcc.ufmg.br

***Abstract*—With the advance in technology and social media usage, first-person recording videos has become a common 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 do not consider inserting background music in the videos, which could make them more enjoyable. This thesis[1] presents a new method that creates accelerated videos and includes the background music keeping the same emotion induced by visual and acoustic modalities. Our approach 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 maintaining the visual quality of the output video compared with other literature methods. Visual results can be seen through the link: https://youtu.be/9ykQa9zhcz8.**

## I. Introduction

In the last years, we have noticed a significant increase in the volume of audio-visual data on the Internet due to the ease in people's access and usage of new digital technologies. Many people started recording diverse first-person videos of their daily activities, referred to as egocentric videos. These videos are usually very long and tiring to watch, containing redundant segments, consequently requiring post-edition. Therefore, a great interest arose in the computer vision community in reducing the length of these videos to make them more pleasant to watch.

Several works in the literature aim to accelerate first-person videos using different strategies and restrictions to make users' experience more pleasant when watching these videos [1]–[8], creating accelerated videos commonly called hyperlapse. Creating a hyperlapse is a technique in time-lapse photography that allows the creation of motion shots, where the goal is to optimize the output number of frames and the visual smoothness [5]. An important extension of the classic hyperlapse, called semantic hyperlapse, includes semantic relevance for each frame, giving lower speedup rates to the most relevant frames in the output video [2], [3], [5]–[7], [9].

Despite the advances, these works did not give attention to the background song that the user wants to insert in the accelerated video. Both visual and sound streams play a significant role in the video-watching experience. In this way, we are interested in combining the song and video content, which is not trivial since they are contents of different natures. The challenge is to combine these contents considering the emotions induced by both to produce a hyperlapse that maintains the video and song's emotional similarity over time.

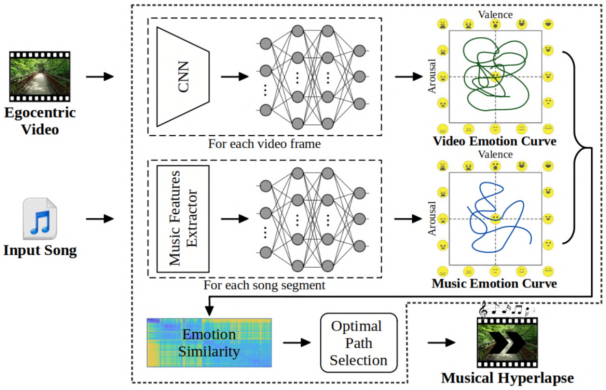

Fig. 1. **Music-driven video acceleration.** After computing emotion curves in the valence-arousal plane, our method accelerates the video according to an optimization algorithm that seeks the best matches between video and song.

We introduce in this thesis a novel problem called *Musical Hyperlapse*, in which the goal is to accelerate the video to the size of a song combining the emotions induced by visual and acoustic contents continuously in time. To solve this problem, we present a new multi-modal approach for creating hyperlapse videos that match the emotion curves produced from an input egocentric video and an input song. As illustrated in Figure 1, our method seeks to find the best subset of video frames that maximizes the similarity of the emotion curves produced from the video and the song, maximizing the video's visual quality.

The contributions of this thesis can be summarized as follows: *i)* New models for automatic music and image emotion recognition; *ii)* a novel optimization algorithm which creates a hyperlapse combining the video and music emotion contents; and *iii)* a new dataset which comprises diverse first-person videos and songs with diverse features.

## II. Related Work

### A. Hyperlapse

Over the past decade, hyperlapse methods have been proposed to reduce the length of long egocentric videos. These works' evolution focuses on improving the quality of the output video by keeping it as pleasant as possible.

[1] This work relates to an M.Sc. dissertation.

One of the first fast-forwarding methods was presented by Kopf *et al.* [10]. They used techniques based on image rendering, such as projecting, stitching, and blending after computing the optimal trajectory of the camera poses. One major drawback of their method is its complexity and high computational cost. Poleg *et al.* [11] presented a method to create classical hyperlapse videos using a graph to model the frame selection, in which the nodes represent the frames and the edge weights represent the cost of including a pair of frames sequentially in the output video. Joshi *et al.* [1] presented a real-time hyperlapse creation algorithm, which 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) [12].

The main disadvantage of all these works is that they do not consider the content in the video scene, which is an essential element to ensure a good experience when watching the accelerated video.

### B. Semantic Hyperlapse

There are also approaches in the literature that considers the visual semantics in the optimization process, referred to as semantic hyperlapse. The objective is to accelerate the input video, optimizing stability, speed-up rate, and semantics.

Okamoto *et al.* [13] proposed a method to summarize egocentric moving videos, generating a walking route guidance video. They analyze the video by detecting pedestrian crosswalks and ego-motion classification, estimating importance scores for each video session based on the contents. Ramos *et al.* [2] presented an adaptive frame sampling method that balances the semantics and the traditional hyperlapse objectives by using energy cost minimization. Their method assigns relevance scores for each video frame and give lower speedup rates to the most relevant frames. Silva *et al.* [14] modeled the adaptive frame sampling as a weighted minimum sparse reconstruction problem. They split the video using frame-wise levels of relevance. Each segment is represented as a dictionary from which the output video frames are sparsely selected, reducing abrupt camera motions. Furlan *et al.* [7] proposed considering both visual and acoustic content from the input video. The original video's soundtrack is segmented, and the Psychoacoustic Annoyance (PA) is computed for each segment. The PA values are used to guide the semantic hyperlapse creation since they are used as semantic scores.

Despite the advancement of these works, they did not consider the background music the user wants to insert in the videos, which could make them more enjoyable. Our goal is to create a hyperlapse with background music where visual and acoustic signals induce similar emotions during exhibition.

### C. Emotion Recognition

Significant progress has been made in the field of music emotion recognition. In general, emotions are represented using psychological models, such as Russells's valence-arousal plane [15] and the EmojiGrid [16], in which the $x$-axis is the valence, and the $y$-axis is the arousal.

Some works focus on investigating musical features related to induced emotions. Lu *et al.* [17] performed a thorough study on mood models and features, concluding that music features such as melody, pitch, rhythm, and timbre play a significant role in human psychological functions. Panda *et al.* [18] reviewed the existing audio features and their relationships with musical concepts to improve the classification performance. The authors rely on clues like melodic lines, notes, intervals, and scores to access higher-level musical concepts such as harmony, melody, articulation, and texture.

Diverse works aim to predict emotions using machine learning techniques. Yang *et al.* [19] formulated the musical emotion recognition as a regression problem to predict the emotion labels. They extract features and use two regressors to predict the labels, one for valence and one for arousal. Chowdhury *et al.* [20] proposed a VGG-style deep neural network to obtain emotional contents from a music piece through mid-level perceptual features, using the audio spectrogram as input. However, these works focus on getting the emotion for an entire song instead of creating continuous emotion curves. Thammasan *et al.* [21] proposed a continuous music emotion recognition approach based on brainwave signals. They predict valence and arousal only on two levels (low and high). Dong *et al.* [22] proposed a method for continuous music emotion recognition using segments of $0.5$ seconds. They converted the regression problem into a weighted combination of multiple binary classification problems. Their approach is quite complex compared to other literature methods.

There has also been significant progress in image emotion recognition in recent years. Some works seek to recognize emotions in scenes and facial expressions [23]. In addition, researchers created several datasets to classify image emotions, such as GAPED [24] and MVSO [25]. There is a great interest in automatically retrieving emotional content from an image, which motivates the computer vision community. Human annotations are used on several images to create the datasets. Then, machine learning based on high-level or low-level features is used to predict the emotions from any image.

Joshi *et al.* [26] and Zhao *et al.* [27] explored the use of psychology and art-theory knowledge to determine which emotions may be evoked by a picture. Jia *et al.* [28] demonstrated the effectiveness of using high-level features, such as social network data, than basic low-level features, such as colors. Descriptive data also play an important role in several solutions to recognizing the emotion induced by images. For instance, Borth *et al.* [29] use pairs of adjectives and nouns to classify each picture. They use each of the 24 emotions defined in Plutchik's theory [30] to derive search keywords and retrieve images and videos from Flickr and Youtube. Mittal *et al.* [31] takes a broader range of objects in the scene to sort the most important ones regarding the induced emotion.

Despite the progress of emotion recognition, these works did not apply the correlation of image and music in the hyperlapse. Our work correlates visual and acoustic contents considering emotions in the fast-forwarding process to create an accelerated video combined with the music.

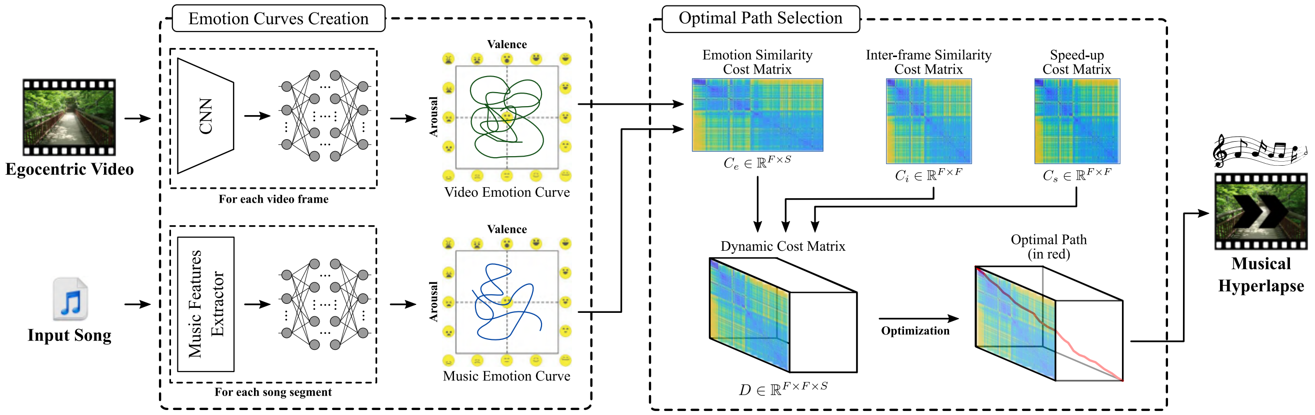

Fig. 2. **Methodology Overview.** In the first step, we extract features from each video frame and each song segment and classify them to obtain their induced emotion. Using the classification results, we create continuous two-dimensional emotion curves in the valence-arousal plane. In the second step, we calculate inter-frame and cross-modal cost matrices to create a three-dimensional dynamic cost matrix 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.

## III. Methodology

Our method accelerates a video to the size of a song by maximizing the emotion curves similarity generated for each one. Given a long egocentric 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, where $S < F$, our goal is to create a shorter video $\hat{V} = [\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_S]$ in which the similarity between the video emotion curve $X \in \mathbb{R}^{F \times 2}$ and the music emotion curve $Y \in \mathbb{R}^{S \times 2}$ is maximized. We show in Fig. 2 an overview of our methodology, divided into two main steps: *i)* Emotion Curves Creation and *ii)* Optimal Path Selection. We detail the methodology in the next sections.

### A. Emotion Curves Creation

Our method creates two emotion curves in the first step, one for the video stream and another for the audio stream. The values in these curves reflect the induced emotion at each instant in time. Based on image and audio feature extraction, classifiers are used to estimate each emotion value, as illustrated in Figure 2-left.

*1) Video Emotion Curve:* To create the video emotion curve, we first feed an image emotion classifier $X' = \phi(V)$ with the frames of the video stream $V$. The output of the $\phi$ classifier is a discrete two-dimensional curve $X' = [x'_1, x'_2, \ldots, x'_F]^T \in \{0,1\}^{F \times 2}$. Then, we decompose the curve into separated values of valence $X'_v = [x'_{v1}, x'_{v2}, \ldots, x'_{vF}]^T \in \{0,1\}^F$ and arousal $X'_a = [x'_{a1}, x'_{a2}, \ldots, x'_{aF}]^T \in \{0,1\}^F$. By this way, the emotion labels of each video frame $v_i$ is represented with the coordinates $x'_{vi}$ and $x'_{ai}$ in the valence-arousal plane. Each 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. To approximate the function $\phi$, we use a pretrained 2D-CNN (ResNet-50 [32]) as a backbone network topped with a fully-connected network.

We perform a fine-tunning in the video emotion classifier using the MVSO dataset [25]. The MVSO dataset comprises about 7 million images and their respective concepts represented by adjective-noun pairs, such as *crying-baby*, *colorful-clouds*, *old-books*, *sad-dogs*, and others. The dataset also comprises labels associating each adjective-noun pair with a distribution over the 24 emotion categories from Plutchik's Wheel of Emotions [30] (*e.g.*, joy, anger, sadness). For each image, we select the emotion with the highest score, convert it to a point in the valence-arousal plane, and use its position as a label, associating the image with the quadrant in which the label is positioned. We perform the fine-tuning by randomly splitting the converted set into training, validation, and test sets in the proportion 70:15:15, using cross-entropy loss.

We convert the discrete video emotion curve into a continuous emotion curve $X = f(X') \in \mathbb{R}^{F \times 2}$, represented as the blue curve in the EmojiGrid in Fig. 2, where $f : \{0,1\} \rightarrow \mathbb{R}$ is a smoothing function that applies a quadratic interpolation to the sequential values. The curve is rescaled into a range of $[-1, +1]$, the range displayed on the EmojiGrid.

*2) Music Emotion Curve:* To create the music emotion curve, we feed a music emotion classifier $Y' = \psi(M)$ with the segments of the audio stream $M$. The output of the $\psi$ classifier is a discrete two-dimensional 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 levels. 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 positive valence and higher $y'_{ak}$ values indicate a higher arousal. The music emotion classifier $\psi$ is composed of a feature extractor topped with two fully-connected networks, one for valence and another for arousal. We use a sliding window of size $\alpha = 6$ and stride of $\delta = 0.5$ seconds over the audio stream to extract the segment-wise features. Then, following Panda *et al.* [33], we extract from each segment a $d$-dimensional feature vector

$\hat{\mathbf{m}}_k \in \mathbb{R}^d$ dedicated to the song. We feed each $\hat{\mathbf{m}}_k$ to the classifiers to obtain the discrete curves $Y'_v$ and $Y'_a$.

We train the music emotion classifier using the DEAM dataset [34]. This 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. 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. We used training, validation, and test splits in the same proportion used in our image classifier.

We apply a linear interpolation in the valence and arousal values, which have only 2 samples per second, to match the video's sampling rate of 30 frames per second. Then we apply a smoothing function to create the continuous curve $Y = g(Y') \in \mathbb{R}^{S \times 2}$, represented as the green curve in the EmojiGrid in Fig. 2. Smoothing is performed to avoid abrupt transitions in the emotion curve.

### B. Optimal Path Selection

Once we have created the video and music emotion curves, the next step is to find the subset of frames $\hat{V}$ that maximizes the similarity of the emotion curves and the video's visual quality. To attend to both objectives, we model the frame sampling problem inspired by Joshi *et al.*'s formulation [1], which considers only inter-frame transitions and visual modality. However, we extend their formulation also to include audio-visual correlation.

We create an Inter-frame Similarity Cost Matrix $C_i \in \mathbb{R}^{F \times F}$ to model the visual continuity term for the optimization, aiming to avoid visual discontinuity. In this matrix, each element is computed as

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

where $i, j \in \{1, \ldots, F\}$ are indices of frames in the video and $\text{SSIM}(\cdot, \cdot)$ is the structural similarity index [35]. High SSIM value indicates high inter-frame similarities.

We also create a Speedup Cost Matrix $C_s \in \mathbb{R}^{F \times F}$ to avoid skips too distant from the target speed-up rate 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 the Emotion Similarity Cost Matrix $C_e \in \mathbb{R}^{F \times S}$ to determine the cost of skipping relevant frames regarding the video and music emotion similarity, given by

$$C_e(i, k) = \frac{1}{d_0} \sqrt{(x_{vi} - y_{vk})^2 + (x_{ai} - y_{ak})^2}, \tag{3}$$

where $k \in \{1, 2, \ldots, S\}$ 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, used as a normalization factor.

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 [10] | 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 |

We normalize all the cost matrices $C_i$, $C_s$, and $C_e$ to $[0, 1]$ and use them to create the 3D Dynamic Cost Matrix $D \in \mathbb{R}^{F \times F \times S}$, where 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 $k$. 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 $\lambda_e$, $\lambda_s$ and $\lambda_i$ are the weights associated with each cost term and $w$ is the maximum skip between adjacent frames in the path. We also concurrently populate the traceback matrix as $T(i, j, k) = \arg\min_{1 \leq h \leq w} D(i - h, i, k - 1)$.

Finally, we trace back the optimal path, starting from position $k = F$, and selecting, at each step, the index stored in $T(i, j, k - 1)$ while $k >= 0$. The sorted order of the frames selected during this step is the final set that composes the hyperlapse video. We add the input audio stream to the composed hyperlapse video to generate the *Musical Hyperlapse* video.

## IV. Experiments

### A. Implementation Details

We used a fully-connected network with 4 layers of 1,000 neurons in the image and music emotion classifiers. In the image emotion classifier, the classification layer comprises 4 neurons representing each valence-arousal quadrant. In the music emotion classifiers, the classification layer comprises 8 neurons, corresponding to the number of discretization levels. We used the *essentia* Python library to extract $d = 48$ music features used in the music classifiers. For the optimal path selection algorithm, we set the cost terms' weights empirically to $\lambda_e = 1.00$, $\lambda_i = 0.01$, and $\lambda_s = 0.01$.

### *B. Experimental Setup*

We created a dataset comprising 8 first-person videos with diverse contents, such as cities, buildings, parks, people, cars, nature, animals, and others; and 5 songs with various styles and emotions. We present the list of videos and songs used in the experiments and their sources in Table I. We resampled all the videos to $640 \times 480$. We compare our methods against two hyperlapse baselines: the Microsoft Hyperlapse (MSH) [1], and the extended version of the Sparse Adaptive Sampling (SASv2) [36]. For both competitors, we used the default values for all parameters.

We measure the similarity of the video and music emotion curves with the Emotion Similarity metric, defined as

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

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

We defined the Speed-up Ratio metric to measure the ratio between the desired and obtained speed-ups, calculated as $Sp_r = \max(Sp^\star, \hat{Sp})/\min(Sp^\star, \hat{Sp})$, where $Sp^\star$ is the desired speed-up and $\hat{Sp} = \hat{F}/S$ is the obtained speed-up. We also measure the similarity between the input and output videos, using the Fréchet Inception Distance (FID) [37] to assess how much useful content is preserved. The lower the FID value, the more similar the input and output videos. We also compute the instability of the frame transitions in the output video, using the Video Shaking Ratio [9].

In our experiments, we run each method for the eight videos, each with the five songs presented in Table I, and present the average values of the five songs for each metric.

### *C. Ablation Study*

We perform an ablation study evaluating the use of two other simple curve alignment algorithms:

- **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 adaptation of the original DTW algorithm, used to measure and align similarities between two temporal [12]. Since the original DTW may repeat frames, which is not allowed in hyperlapses, we added a constraint in the algorithm that forces it to never repeat frames. We feed the algorithm with the $X$ and $Y$ curves, and the output is the sequence of selected frames.

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** |

We show in Table II the ablation study results. Our optimal path selection algorithm achieved the best results across all metrics. The greedy approach maximizes the emotional similarities locally, leading to a significant error in the achieved speedup, which might remove important frames from the original video, resulting in a high FID. The DTW 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 speedup ratio by taking the exact number of frames required by the song and maintaining a balance between frame transitions by using the speedup and inter-frame similarity cost matrices, guaranteeing a lower FID.

### *D. Results*

Table III presents the results for the comparison with the baselines. 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 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, gives 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. Compared 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.

Figure 3 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

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** |

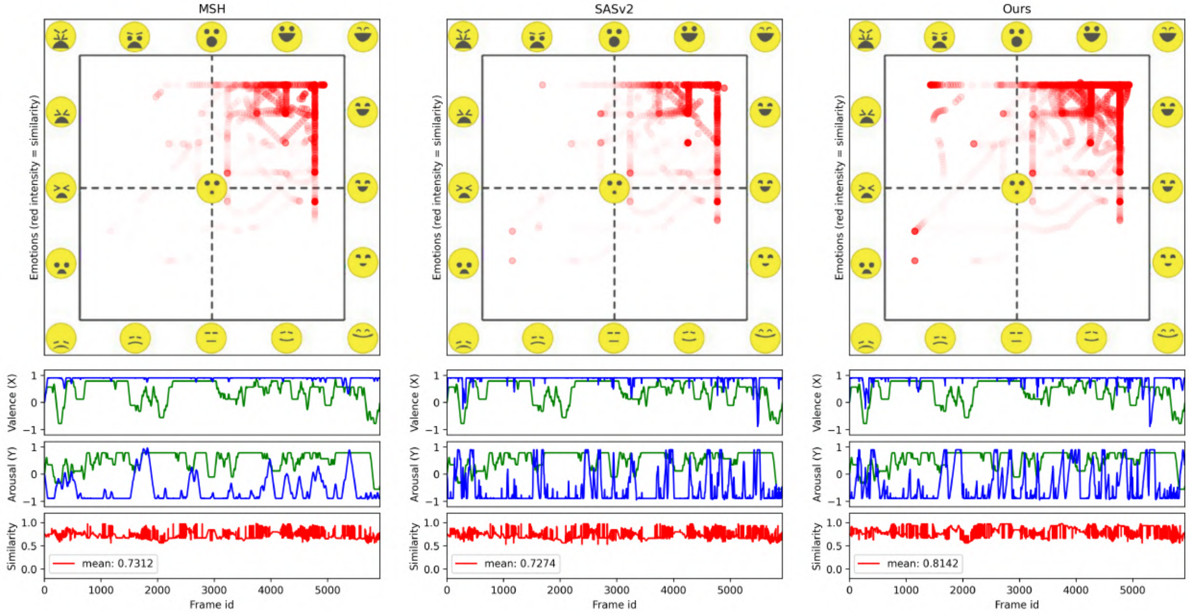

Fig. 3. **Qualitative comparison with baselines.** Each column represents the results of a method. At the top is the EmojiGrid with the emotion similarities in the regions achieved by video and music emotion curves. The greater the red intensity, the greater the similarity. At the bottom, the separate curves of valence and arousal throughout the video (blue) and song (green), and the similarity curve (red).

plane location. The blue curve represents the video, and the green one the song. The red curve represents the curves' similarity 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.

## V. Conclusion

We introduced in this thesis the novel task of creating a Musical Hyperlapse, in which we accelerate first-person videos aligning the emotions induced by visual and acoustic signals. Our method reduces an input egocentric video to the size of a song seeking to align the feelings induced by both over time. We also presented a new multimodal dataset comprising diverse first-person videos and songs of various styles. Our method achieved superior performance in terms of video representation, required speed-up, and emotional alignment for different videos and songs. The results showed the possibility of creating a hyperlapse combining media of distinct nature based on their affective semantics.

For future work, it is possible to improve the emotion recognition models, performing regressions instead of classifications, and using a shared embedding space to measure the similarities between image and audio data. We can also reduce the complexity of the optimal path algorithm and perform the experiments with more videos and songs.

**Acknowledgments:** We would like to thank the PPGCC-UFMG, CAPES, CNPq, and FAPEMIG for funding this work.

## VI. Awards & Publications

The results of this thesis were published at the 34th Conference on Graphics, Patterns, and Images - SIBGRAPI 2021 [38], where it was awarded as the best paper, extended, and published in a special issue of the Pattern Recognition Letters (PRL) journal [39].

## 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] 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.

[10] 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.

[11] Y. Poleg, T. Halperin, C. Arora, and S. Peleg, "Egosampling: Fast-forward and stereo for egocentric videos," 2015, pp. 4768–4776.

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

[13] M. Okamoto and K. Yanai, "Summarization of egocentric moving videos for generating walking route guidance," pp. 431–442, 2014.

[14] 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.

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

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

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

[26] 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.

[27] 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.

[28] 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.

[29] D. Borth, R. Ji, T. Chen, T. Breuel, and S. Chang, "Large-scale visual sentiment ontology and detectors using adjective noun pairs," MM 2013 - Proceedings of the 2013 ACM Multimedia Conference, pp. 223–232, 10 2013.

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

[31] 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.

[32] 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.

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

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

[38] D. de Matos, W. Ramos, L. Romanhol, and E. R. Nascimento, "Musical hyperlapse: A multimodal approach to accelerate first-person videos," in Conference on Graphics, Patterns and Images (SIBGRAPI), 2021, pp. 184–191.

[39] D. de Matos, W. Ramos, M. Silva, L. Romanhol, and E. R. Nascimento, "A multimodal hyperlapse method based on video and songs' emotion alignment," Pattern Recognition Letters, 2022.