A Multimodal Frame Sampling Algorithm for Semantic Hyperlapses with Musical Alignment

Raphael Nepomuceno, Luísa Ferreira, Michel Silva
Department of Informatics, Universidade Federal de Viçosa - UFV, Viçosa, Brazil
{raphael.nepomuceno,luisa.ferreira,michel.m.silva}@ufv.br

***Abstract*—Producing visually engaging and semantically meaningful hyperlapses presents unique challenges, particularly when integrating an audio track to enhance the watching experience. This paper introduces a novel multimodal algorithm to create hyperlapses that optimize semantic content retention, visual stability, and the alignment of playback speed to the liveliness of an accompanying song. We use object detection to estimate the semantic importance of each frame and analyze the song's perceptual loudness to determine its liveliness. Then, we align the most important segments of the video—where the hyperlapse slows down—with the quieter parts of the song, signaling a shift in attention from the music to the video. Our experiments show that our approach outperforms existing methods in semantic retention and loudness–speed correlation, while maintaining comparable performance in camera stability and temporal continuity.**

I. Introduction

Over the past two decades, recording daily activities has been made accessible with the advent of smartphones, wearable devices, and personal action cameras, such as GoPro™. Sharing photos and videos through social media services has also become commonplace, leading to an ever-growing accumulation of visual data competing for our attention. Hands-free recordings of daily activities often contain repetitive or irrelevant content because the wearer is focused on the activity itself rather than managing the camera, which can make the video unpleasant to watch. Egocentric video summarization aims to infer the intent of the wearer, reduce irrelevant content, and produce a summary that is pleasant to watch [1]. In particular, dynamic fast-forward methods assign semantic importance scores to the video according to domain-specific criteria, such as route guidance [2] or presence of people [3], which are used to lower the playback speed during important segments or raise it in unimportant segments, producing a representative summary video that has no gaps between scenes.

Videos recorded with head-mounted, body-mounted, or handheld cameras are often shaky due to body movements. In the task of dynamic fast-forward, when accelerated, the motion in these videos is exaggerated, making the resulting video unpleasant to watch at best, or even to cause nausea at worst [4]. Research on creating accelerated egocentric videos with stable camera motion, referred to as hyperlapses, includes methods for selecting video frames that minimize shakiness [5] and applying video stabilization with fine-tuning for such videos [4]. An extension called *semantic hyperlapse* [3] combined the semantic criteria from video summarization with the original hyperlapse concept to create egocentric video summaries that minimize unpleasant camera motion.

Although sound plays a major role in the video watching experience, most hyperlapse methods leave it up to the user to insert an audio track themselves. Matching an accelerated video with a background music track presents an ill-posed problem, as the criteria for what constitutes a good match are subjective. Matos *et al*. [6] propose speeding up the video to match the emotions induced by the video with those induced by the song. However, this approach relies on a subjective measure rather than an objective semantic element of the video to determine acceleration, making it difficult to understand why the video speeds up or slows down.

In this work, we introduce a multimodal hyperlapse algorithm that selects the most important frames of a video while matching the playback speed to the momentary loudness of an input song, maintaining a correlation between audio and video, *i.e*., the video slows down when the song is quiet and speeds up when it is loud. For example, in a recording of a morning walk, interactions between the recorder and other people are considered important; therefore, the playback speed in these segments is lowered and aligned with quiet parts of the chosen song. On the other hand, non-interactive moments will be played faster and aligned with loud portions of the song. Compared to previous works, our method achieves significantly better results on both semantic sampling and loudness–playback speed correlation when compared to state-of-the-art hyperlapse methods.

Our code is available at https://github.com/MaVILab-UFV/SemanticMusicalHyperlapse_SIBGRAPI_2024.

II. Related Work

Video summarization techniques, which aim to shorten the length of videos, can be broadly categorized into three classes based on the type of summary produced: *storyboards*, which select a set of representative static frames; *video skimming*, which retains a discontinuous set of the most relevant segments; and *fast-forwarding*, where the video is accelerated at a constant or variable playback speed [1]. Among these, only fast-forwarding produces continuous summaries. Discontinuities in videos can cause viewers to lose context or miss the path taken by the recorder [7]; therefore, our focus is on fast-forward methods.

a) Hyperlapse: Methods to address the exaggerated camera motion in fast-forward egocentric videos, known as hyper-

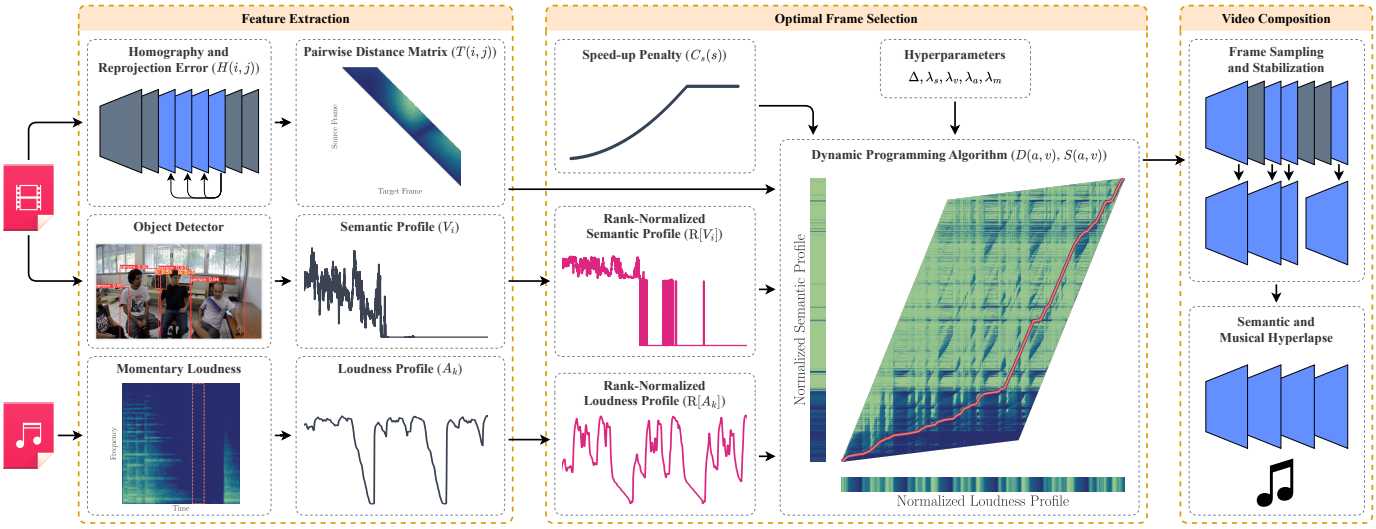

Fig. 1. **Overview of our method**, divided into *feature extraction*, *optimal frame selection*, and *video composition*. Our method extracts semantic and homography features from a video and the loudness curve from a song, using these as inputs in an optimization problem to select an optimal subsequence of frames that: (i) is representative of the semantic content of the video; (ii) aligns quiet moments in the song with lower speed-ups, and vice-versa; (iii) minimizes camera shakiness; and (iv) has no gaps between scenes. The selected frames are then composed into a stabilized video, to which the song is added.

lapses, were pioneered by Kopf *et al*. [8], who reconstructed scenes in 3D and rendered accelerated videos through smooth virtual camera trajectories. While remarkable, the approach has high computational cost and requires sufficient camera motion to perform the reconstruction.

Karpenko [9] created a hyperlapse using gyroscope metadata to estimate the camera orientations for stabilizing the video. Although it requires metadata, this method is notable for being the first hyperlapse method meant for real-time usage.

Poleg *et al*. [10] formulated the hyperlapse as a graph problem, modeling video frames as nodes and frame-to-frame transitions as weighted edges. The frame sampling is solved as a shortest path problem, and the frames along the path are included in the output hyperlapse.

Joshi *et al*. [5] developed a frame sampling method inspired by dynamic time warping. Their algorithm scores frame-to-frame transitions in order to achieve smooth visual transitions while avoiding large deviations from the desired speed-up and abrupt changes in speed-ups, building a sparse dynamic programming matrix to find the optimal frame selection.

b) Semantic fast-forwarding: While the aforementioned methods produce visually stable accelerated videos, they often overlook important content by fast-forwarding through it, causing key information to go practically unnoticed.

A fast-forwarding algorithm based on semantic rules was proposed by Cheng *et al*. [11], designed around the metaphor of a car driver that slows down near areas of interest and using feedback from the viewer to learn what is important to them.

Okamoto and Yanai [2] proposed a semantic fast-forwarding method tailored for route guidance in egocentric videos, where temporal discontinuities can confuse the viewer about their current location. In their approach, frame scores depend on the recorder's actions—turning, stopping, or moving forward—and the presence of crosswalks and turns.

c) Semantic hyperlapse: Ramos *et al*. [3] expanded upon the work of Poleg *et al*. [10], proposing the first semantic hyperlapse. Their method assigns a semantic score to each frame based on the prominence of faces, then fast-forwards the video with low speed for important segments and high speed otherwise. As an extension, Silva *et al*. [4] introduced a stabilization algorithm tailored for semantic hyperlapses and a dataset of videos with varying semantic scores. Further contributions by Silva *et al*. [12] proposed a weighted sparse sampling approach to selecting a set of frames that minimizes the reconstruction error of the original video given the required speed-up in each video segment, and subsequently aimed to solve the problem of abrupt jumps between segments [7].

Most hyperlapses focus on visual features, without input from sound streams. To the best of our knowledge, only two works have presented audio-driven semantic hyperlapses.

Furlan *et al*. [13] proposed the first audio-driven semantic hyperlapse: the audio stream of the video is used to compute the video frames semantic scores, assigning higher speed-ups to unpleasant segments, such as those featuring noisy crowds or streets. However, the resulting hyperlapse has no sound.

Matos *et al*. [6] proposed a hyperlapse in which the emotions of the video and the song are aligned in time. The video and the song are processed frame-by-frame by a neural network to estimate emotion curves based on valence and arousal. The method uses a dynamic programming algorithm to fast-forward the video while aligning the emotion curves. Matos *et al*. [14] extended their work with a method to select the best matching song from a library.

We heavily draw inspiration from the optimization methods proposed by Joshi *et al*. [5] and Matos *et al*. [6], and model the video fast-forward problem as the construction and traversal of a dynamic programming matrix. Unlike the two methods, we account for the semantic features of the video during the frame

selection. Our method differs from Joshi *et al*. [5] in being able to perfectly achieve the requested video length, which is needed to include a song in the output hyperlapse. Regarding Matos *et al*. [6], our method differs in how we match the video to a song: instead of aligning video and music based on emotion as they do, we align hyperlapse playback speed to song loudness, which are objectively measurable metrics.

III. Methodology

In this section, we describe our method to create semantic hyperlapses with musical alignment. As shown in Figure 1, the process can be divided into three major steps:

a) Feature extraction: As proposed by Ramos *et al*. [3], we use an object detection algorithm to identify a set K_i of objects of interest for each frame i of the video, and then calculate each V_i in the **semantic profile** as:

$$V_i = \sum_{k \in K_i} P(k) \cdot G_\sigma(k) \cdot A(k), \tag{1}$$

where k represents a detected object, $P(k)$ denotes the confidence of the detector about k, $G_\sigma(k)$ measures the centrality of the object in the frame, and $A(k)$ is the area of its bounding box. Intuitively, objects with high detection confidence, near the center of the frame and with large bounding boxes (*i.e*., those closer in proximity) are more likely to draw the attention of the viewer and therefore are assigned higher scores.

We also calculate the **pairwise distance matrix** to measure shakiness as the motion of the central pixel of the frames:

$$T(i,j) = \frac{||H(i,j) \cdot \mathbf{M} - \mathbf{M}||_2}{||\mathbf{M}||_2}, \tag{2}$$

where $H(i,j)$ is the homography matrix from frame i to j, $\mathbf{M}$ is the coordinate vector of the central pixel, and $||\dots||_2$ denotes the Euclidean norm. If the homography estimation fails or the reprojection error exceeds $0.2 \cdot ||\mathbf{M}||_2$, we set $T(i,j) = 1$. This is equivalent to the *frame matching cost* defined by Joshi *et al*. [5] and the *shaking ratio* metric defined by Ramos *et al*. [15]. Because our algorithm is guaranteed to never exceed the specified maximum speed-up, only the homographies between frames within Δ indices of each other are calculated, greatly reducing the computational cost and memory usage of T.

The **loudness profile** is the momentary loudness of an input song, described in EBU Tech 3341 [16], which is designed around how the human ear perceives sounds. This measurement is computed using a 400 ms window centered at each timestep with the same frequency as the frames per second (FPS) of the paired video and saved as $A = \langle A_1, A_2, \dots, A_n \rangle$.

b) Optimal frame selection: We model the optimal frame selection as a dynamic programming algorithm, where a recurrence relation $D(a,v)$ represents the minimum cumulative cost of choosing frames and speed-up rates from the start of the sequence up to the audio timestep a and video frame v:

$$D(a,v) = \min_{s \in [1,\Delta]} C(a,v,s) + D(a-1, v-s), \tag{3}$$

where s is a candidate speed-up, Δ is the maximum momentary speed-up, and $D(1,1) = 0$. The traceback matrix $S(a,v)$ stores which choices were made in each $D(a,v)$, and is used to determine what is the set of frames with the lowest cumulative cost that forms a path through the whole video:

$$S(a,v) = \arg\min_{s \in [1,\Delta]} C(a,v,s) + D(a-1, v-s). \tag{4}$$

The best path can be retrieved through a backward traversal of S, storing the values of v while recursively moving from $\langle a, v \rangle$ to $\langle a-1, v-S(a,v) \rangle$ until $\langle a, v \rangle = \langle 1, 1 \rangle$. To ensure that the length of the resulting video matches the song exactly, preventing either of them from being cut short, we traverse from $a = a_{\max}$ and $v = v_{\max}$, *i.e*., the end of the song and video, respectively.

The cost of choosing the video frame v and speed-up s at the timestep a is determined by the weighted combination of the *speed-up cost*, *video semantic cost*, *audio alignment cost* and *frame matching cost*, respectively:

$$C(a,v,s) = \lambda_s C_s(s) + \lambda_v C_v(v) + \lambda_a C_a(a,s) + \lambda_m C_m(v,s). \tag{5}$$

In the following paragraphs, we describe the specifics of each term of the cost function. Similarly to Joshi *et al*. [5] and Matos *et al*. [6], we found that truncating the costs with $\tau = 200$ produces good results, ensuring similar magnitudes for the functions and simplifying the selection of the λ hyperparameters across videos and songs.

The **speed-up cost** (C_s) is a regularization to control how often and how much the playback speed changes, and is defined as:

$$C_s(s) = \min\left\{(s-1)^2, \tau\right\}, \tag{6}$$

which heavily penalizes high speed-up rates.

The **video semantic cost** (C_v) aims to select as many important frames as possible while omitting or speeding through unimportant frames. To achieve this, we define:

$$C_v(v) = -\tau \cdot \mathrm{R}[V_v], \tag{7}$$

such that it decreases based on the relative importance of the semantic score V_v, starting at $C_v = 0$ for the least important frames and down to $C_v = -\tau$ for the most important frame.

The semantic profile is susceptible to outliers; for example, the appearance of a crowd for a few frames can cause the value of V_v to spike, disproportionately lowering the scores of other important frames with fewer people. To avoid this, we compute $\mathrm{R}[V_v]$ using rank-based normalization [17]: each value in V is replaced by its position (rank) in an ordered array, and then divided by the array length to make the result bounded in $[0, 1]$. Ties are handled with competition ranking. Because the transformation depends only on the ordering, it is robust to outliers and requires no knowledge of the distribution of the values of V.

The **audio alignment cost** (C_a) controls the correlation between the playback speed of the hyperlapse and the loudness of the song, slowing down during quiet segments or speeding

up during loud segments. For this, we first calculate the expected speed-up $E(a)$ at each song timestep a as:

$$E(a) = 1 + (\Delta - 1) \cdot \mathrm{R}[A_a]. \tag{8}$$

Similarly to Equation 7, we use rank normalization to obtain values in the $[0, 1]$ range without making assumptions about outliers or the distribution of A. Then, we perform linear interpolation over the normalized loudness values to convert them into proportional speed values. Then, we define a cost function to achieve the desired speed-up behavior:

$$C_a(a, s) = \min \left\{ (s - E(a))^2, \tau \right\}. \tag{9}$$

The alignment between loudness and semantic importance is a product of the joint minimization of C_a and C_v: the video must slow down to highlight important frames, but doing so during loud audio segments would raise C_a. Consequently, the optimal strategy is to line up important segments of the video with periods of quiet audio, and vice-versa.

The **frame matching cost** (C_m) reduces camera shakiness by penalizing motion with:

$$C_m(v, s) = \tau \cdot T(v - s, v), \tag{10}$$

where T is the pairwise distance matrix from Equation 2.

c) Video composition: The output from the previous step is a sequence of frame indices that should be included in the hyperlapse. However, to be useful for humans, this sequence must be composed into a video.

In the final step, we compose the hyperlapse using the stabilization algorithm proposed by Silva *et al*. [4] for egocentric fast-forward videos. The camera movements between frames sampled by our algorithm are smoothed using weighted homographies, while the discarded frames are used to cover blank spaces produced by the homographies. Images corrupted by the smoothing step are replaced with discarded frames that have a high semantic score. Then, the chosen song is added as the audio stream of the hyperlapse to complete the output of our method.

IV. Experiments

In this section, we describe the *datasets*, *evaluation criteria* and *hyperparameters* used in our experimental evaluation. Then, we present the *results and discussion* in the comparisons made to existing hyperlapse methods and in the ablation study.

a) Datasets: We evaluate our method on videos from the *Annotated Semantic Dataset* [4], which consists of 11 videos of daily activities, each 4 to 10 minutes long, annotated for the presence of people. The "0p" videos have few or no people, while the "75p" videos feature many. The videos also vary in camera movements, being recorded while walking, biking, or driving. We used the songs from the *Database for Emotional Analysis of Music* [18], which contains 1,744 excerpts, each about 45 seconds long. Instead of using the provided emotion annotations, we used the Essentia library to extract perceptual loudness curves. We conducted experiments using all possible combinations of videos and songs.

b) Experimental evaluation: Our quantitative analysis of the output fast-forward video is based on four aspects: (i) **semantic** score measures how much information was retained out of the maximum possible within the same amount of frames [12]; (ii) **correlation** score measures how good is the match between the output speed-up curve and the momentary song loudness using the Spearman correlation coefficient [19]; (iii) **instability** score measures the visual smoothness as the average standard deviation of pixels within a sliding window of seven neighbor frames [20]; and (iv) **discontinuity** score is the root-mean-square error (RMSE) between the speed-up from each frame and the overall target speed-up, to measure how abrupt and large changes between the frames are [7].

We compare the performance of our method to three other representative methods, based on their primary goals: (i) EgoSampling (ES) [10] for hyperlapses; (ii) expanded Sparse Adaptive Sampling (SAS2) [7] for semantic-driven hyperlapses; and (iii) Musical Hyperlapse (MH) [6] for music-driven hyperlapses. Because *Microsoft Hyperlapse* [5] does not report the frame indices or momentary speed-ups needed to calculate the metrics, we omitted it from the evaluation.

c) Hyperparameters: Our method requires five parameters to be determined: the maximum speed-up Δ and the four λ weights, which control the following: semantic importance (λ_v), camera movement (λ_m), audio-speed correlation (λ_a), and speed-up penalization (λ_s).

We found that using $\langle \lambda_v, \lambda_m, \lambda_a, \lambda_s \rangle = \langle 2, 2, 1, 1 \rangle$ provides a good balance between our criteria while giving higher priority to video semantics and stability. In practice, our method is robust to changes in these parameters, and this choice ultimately depends on user preferences.

Increasing Δ reduces the time spent on unimportant frames but worsens the temporal continuity of the video. Larger jumps make it harder to obtain a smooth camera transition due to either a greater distance moved between frames or a complete lack of matching keypoints. We found that $\Delta = 20$ presents a good balance between these factors.

A. Results and discussion

In this section, we present and discuss the results summarized in Table I.

Regarding the **semantic** score, we attribute the higher performance of our method over SAS2 to a difference in optimization criteria. Unlike SAS2, which explicitly assigns speeds to whole segments based on their importance, our method attempts to maximize the total semantic content of the fast-forward video with frame-level granularity. Consequently, our method performs better at selecting semantic frames in videos with low semantic content (*e.g.*, 0p) or in non-relevant segments with some important frames.

A downside of our approach is that speed-up decisions are made implicitly based on semantics, assuming neighboring frames have similar semantic scores and that it is optimal to slow down for them. However, this assumption may not hold if the semantic score fluctuates due to camera motion,

TABLE I
AVERAGE PERFORMANCE OF METHODS ACROSS VIDEO AND SONG PAIRINGS.

Video	Semantic (%, ↑)				Correlation (↑)				Instability (↓)				Discontinuity (↓)			
	ES	SAS2	MH	Ours	ES	SAS2	MH	Ours	ES	SAS2	MH	Ours	ES	SAS2	MH	Ours
Biking 0p	9.1	15.3	14.9	**91.1**	0.017	-0.001	0.007	**0.775**	26.1	**21.7**	24.1	22.5	11.2	5.6	**0.4**	4.2
Biking 25p	7.8	28.6	11.9	**33.4**	0.002	-0.009	0.009	**0.507**	50.7	**45.0**	50.2	47.5	7.7	15.1	**0.5**	7.2
Biking 50p	18.7	27.1	19.3	**50.0**	0.009	0.002	0.006	**0.762**	33.8	**29.2**	32.7	30.4	6.3	10.1	**0.5**	7.5
Biking 50p 2	17.3	31.9	22.3	**79.9**	-0.013	-0.002	0.008	**0.446**	25.2	24.0	24.8	**21.9**	11.7	6.8	**0.9**	7.0
Driving 0p	7.0	34.4	12.5	**93.9**	-0.015	-0.013	-0.002	**0.801**	44.7	**37.6**	42.8	39.3	11.7	6.8	**0.6**	4.5
Driving 25p	9.0	28.6	16.4	**92.7**	-0.020	-0.002	0.006	**0.394**	37.6	**31.4**	36.5	31.6	8.5	5.9	**0.5**	5.2
Driving 50p	9.2	18.9	12.4	**73.0**	0.002	-0.001	0.010	**0.452**	41.3	33.9	38.7	**32.6**	13.3	7.3	**0.4**	7.8
Walking 0p	10.1	13.0	17.9	**95.5**	-0.009	0.009	-0.005	**0.747**	27.5	28.2	28.6	**26.6**	11.9	4.2	**0.9**	4.1
Walking 25p	3.3	32.7	19.6	**57.2**	-0.005	0.014	0.007	**0.387**	31.4	30.1	32.7	**28.9**	13.3	7.2	**0.7**	7.3
Walking 50p	5.5	29.3	23.5	**60.5**	0.003	-0.015	0.017	**0.569**	35.1	33.3	34.5	**31.6**	22.4	12.7	**0.5**	7.2
Walking 75p	12.6	47.5	39.1	**58.5**	-0.038	0.005	0.010	**0.708**	40.3	**33.5**	38.0	35.2	29.2	10.0	**0.3**	7.1
Overall	10.0	27.9	18.1	**71.4**	-0.006	-0.001	0.007	**0.595**	35.8	**31.6**	34.9	**31.6**	13.4	8.3	**0.6**	6.3

in which case our algorithm will select the most important frames without necessarily slowing down.

Our method is the only one to achieve a positive **correlation** between the song loudness and video playback speed. In particular, the poor performance of MH in this metric is expected, since their method aims to match the induced emotions between video and music, a goal unrelated to this correlation metric. Similarly, we expect our method to perform poorly on their emotion matching metric.

Our method fails to achieve positive correlation when the loud segments of the song fully coincide with the important segments of the video. For example, in Figure 2, the beginning of the video is highly important, necessitating a lower playback speed. This makes the song in (iii) a poor fit, as its loud segments also occur at the start. In such cases, increasing the correlation would require significant drops in the semantic score, which is undesirable.

The primary objectives of semantic and musical methods also affect **instability** performance. By lowering the playback speed, these methods select similar frames in sequence, which improves stability.

The **discontinuity** score measures the RMSE to a constant speed-up, and is a competing goal to the correlation score, which, for most songs, would require the speed-up to change. Despite this, our method performs within an acceptable range, being slightly better than SAS2, indicating that we obey the required speed-up without creating large gaps between scenes. In our tests, we found that MH was able to maximize their emotion alignment with little variation in the speed-up, resulting in low discontinuity. ES allows for skips of up to 100 frames, which produces gaps in the video and is reflected by the metric. Among the four methods, only ours and MH match the required speed-up exactly, which is crucial to prevent the video ending before the song and vice-versa.

a) Ablation study: In Figure 3, we show how each component of the cost function affects our method. Removing the *video semantic cost* (C_v) leads to a sharp decline in the semantic metric and makes the frame selection primarily driven by the song. This change shows that large swings in speed are required to select as many high-semantic frames as possible, which is reflected in the discontinuity metric.

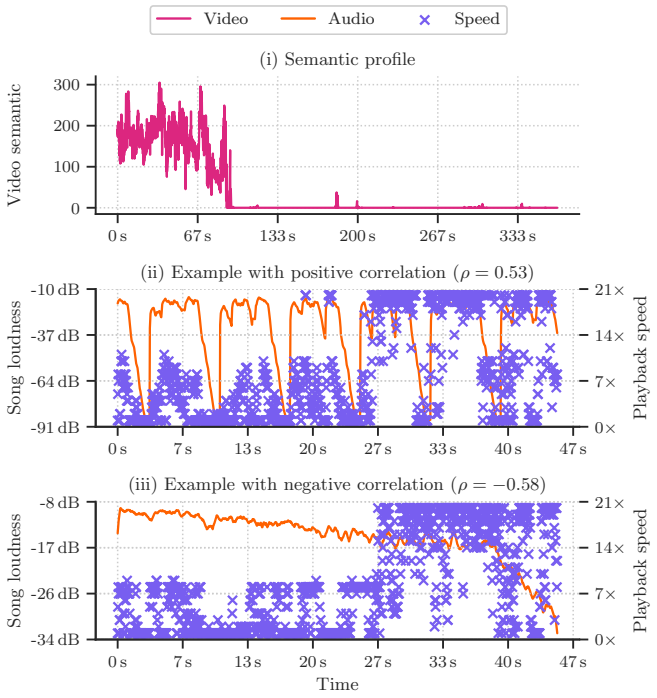

Fig. 2. We present the results of pairing the video in (i), which requires a reduced playback speed in the first half, with two different songs: the song in (ii) produces a speed-up curve that aligns with both the video's semantic content and the song's loudness; in (iii), only the video's semantic content is matched, as the loud segments of the song coincide with the important segments of the video.

Removing the *audio alignment cost* (C_a) causes the total loss of correlation between speed-up and loudness, although the resulting increase in the semantic score is smaller due to the originally smaller importance of this term. This is expected and reflects the results in Table I.

The impact of removing the *frame matching cost* (C_m) is relatively minor across the metrics. In particular and unexpectedly, no changes occurred in the instability metric. Despite this, we chose to maintain this term in the optimization cost

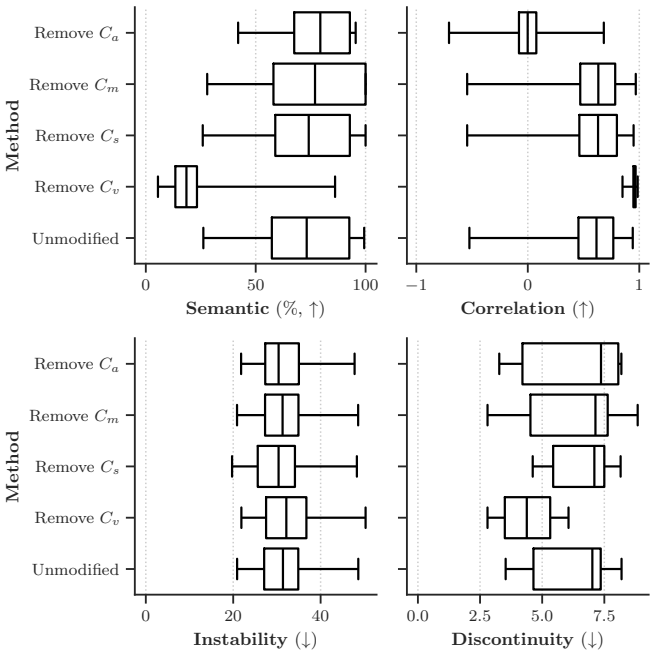

Fig. 3. Results of the ablation study, aggregated over the whole dataset.

to assist in the used homography-based video stabilizer [4].

The *speed-up cost* (C_s) achieves our goals of reducing the occurrence of large skips, as reflected by the discontinuity score, and has no impact in the other factors, which we attribute to lower weight (λ_s) assigned to speed-up cost. Conversely, increasing λ_s would gradually lead the optimizer to behave like a uniform sampler, lowering the discontinuity to the detriment of the other scores.

V. Conclusion

We presented a novel method to create hyperlapses that are representative of the semantic content in the video while maximizing the correlation between song loudness and video playback speed. Our extension of the semantic hyperlapse aims to add music to the video summaries while maintaining a correlation between audio and video: the segments when the playback speed is reduced to show important frames must coincide the quiet parts of the song, drawing the attention to the video; conversely, segments with higher speed-ups must match the louder parts of the song.

Our proposed algorithm achieves superior performance in producing a hyperlapse that is representative of the original video while introducing the novel loudness–speed correlation, and has comparable performance in the stability and temporal continuity to previous methods.

For future work, we aim to develop a method for selecting songs that produce a good match from a music library and to explore hyperparameter optimization to find better default values for the weights of our cost function.

Acknowledgment

The authors would like to thank CAPES, CNPq, FAPEMIG and Finep for supporting this project.

References

[1] A. G. del Molino, C. Tan, J.-H. Lim, and A.-H. Tan, "Summarization of egocentric videos: A comprehensive survey," IEEE Transactions on Human-Machine Systems, vol. 47, no. 1, pp. 65–76, Feb. 2017.

[2] M. Okamoto and K. Yanai, "Summarization of egocentric moving videos for generating walking route guidance," in Image and Video Technology, R. Klette, M. Rivera, and S. Satoh, Eds., 2014, pp. 431–442.

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

[4] 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," in European Conference on Computer Vision Workshops, Oct. 2016, pp. 557–571.

[5] 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.

[6] D. de Matos, W. Ramos, L. Romanhol, and E. R. Nascimento, "Musical hyperlapse: A multimodal approach to accelerate first-person videos," in 34th SIBGRAPI Conference on Graphics, Patterns and Images, Oct. 2021, pp. 184–191.

[7] 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, Apr. 2021.

[8] 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, Aug. 2014.

[9] A. Karpenko, "The technology behind hyperlapse from instagram," http://instagram-engineering.tumblr.com/post/95922900787/hyperlapse, Aug. 2014.

[10] Y. Poleg, T. Halperin, C. Arora, and S. Peleg, "Egosampling: Fast-forward and stereo for egocentric videos," in IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Jun. 2015, pp. 4768–4776.

[11] K.-Y. Cheng, S.-J. Luo, B.-Y. Chen, and H.-H. Chu, "Smartplayer: user-centric video fast-forwarding," in Conference on Human Factors in Computing Systems, 2009, pp. 789–798.

[12] M. Silva, W. Ramos, J. Ferreira, F. Chamone, M. Campos, and E. R. Nascimento, "A weighted sparse sampling and smoothing frame transition approach for semantic fast-forward first-person videos," in IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 2383–2392.

[13] 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, 2018, pp. 2504–2507.

[14] 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, vol. 166, pp. 174–181, 2023.

[15] 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 IEEE Winter Conference on Applications of Computer Vision (WACV), Mar. 2020, pp. 3260–3269.

[16] "Loudness metering: 'EBU Mode' metering to supplement loudness normalisation in accordance with EBU R 128," European Broadcasting Union, Techn. Rep. 3341, Aug. 2011.

[17] A. Tsodikov, A. Szabo, and D. Jones, "Adjustments and measures of differential expression for microarray data," Bioinformatics, vol. 18, no. 2, pp. 251–260, Feb. 2002.

[18] A. Alajanki, Y.-H. Yang, and M. Soleymani, "Benchmarking music emotion recognition systems," PLOS ONE, 2016.

[19] C. Spearman, "The proof and measurement of association between two things," The American Jour. of Psych., vol. 15, no. 1, pp. 72–101, 1904.

[20] 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, pp. 55–64, 2018.