A Weighted Sparse Sampling and Smoothing Frame Transition Approach for Semantic Fast-Forward First-Person Videos

Michel Silva  Washington Ramos  Joao Ferreira  Felipe Chamone  Mario Campos  Erickson R. Nascimento
Universidade Federal de Minas Gerais (UFMG), Brazil
{michelms, washington.ramos, joaoklock, cadar, mario, erickson}@dcc.ufmg.br

Abstract

Thanks to the advances in the technology of low-cost digital cameras and the popularity of the self-recording culture, the amount of visual data on the Internet is going to the opposite side of the available time and patience of the users. Thus, most of the uploaded videos are doomed to be forgotten and unwatched in a computer folder or website. In this work, we address the problem of creating smooth fast-forward videos without losing the relevant content. We present a new adaptive frame selection formulated as a weighted minimum reconstruction problem, which combined with a smoothing frame transition method accelerates first-person videos emphasizing the relevant segments and avoids visual discontinuities. The experiments show that our method is able to fast-forward videos to retain as much relevant information and smoothness as the state-of-the-art techniques in less time. We also present a new 80-hour multimodal (RGB-D, IMU, and GPS) dataset of first-person videos with annotations for recorder profile, frame scene, activities, interaction, and attention.1

1. Introduction

By 2019, the online video might be responsible for more than 80% of global Internet traffic [1]. Not only are internet users watching more online video, but they are also recording themselves and producing a growing number of videos for sharing their day-to-day life routine. The ubiquity of inexpensive shoot video devices and the lower costs of producing and storing videos are giving unprecedented freedom to the people to create increasingly long-running first-person videos. On the other hand, such freedom might lead the user to create a final long-running and boring video, once most everyday activities do not merit recording.

A central challenge is to selective highlight the meaningful parts of the videos without losing the whole message that the video should convey. Although video summarization techniques [4, 14] provide quick access to videos’ information, they only return segmented clips or single images of the relevant moments. By not including the very last and the following frames of a clip, the summarization might lose the clip context [20]. Hyperlapse techniques yield quick access to the meaningful parts and also preserve the whole video context by fast-forwarding the videos applying an adaptive frame selection [9, 7, 22]. Despite the Hyperlapse techniques being able to address the shake effects of fast-forwarding first-person videos, handling every frame equally important is a major weakness of these techniques. In a lengthy stream recorded using the always-on mode, some portions of the videos are undoubtedly more relevant than others.

Most recently, methods on fast-forward videos emphasizing relevant content have emerged as promising and effective approaches to deal with the tasks of visual smoothness and semantic highlighting of first-person videos. The relevant information is emphasized by playing faster the

1https://www.verlab.dcc.ufmg.br/semantic-hyperlapse/cvpr2018/
non-semantic segments and applying a smaller speed-up rate in the semantic ones [23, 27, 10, 26] or even playing them in slow-motion [30]. To reach both objectives, visual smoothness and semantic highlight, these techniques describe the video frames and their transitions by features, and then formulate an optimization problem using the combination of these features. Consequently, the computation time and memory usage are impacted by the number of features used, once the search space grows exponentially. Therefore, the current Hyperlapse methods are not scalable regarding the number of features.

In this work, we present a new semantic fast-forward method that solves the adaptive frame sampling by modeling the frame selection as a Minimum Sparse Reconstruction (MSR) problem (Figure 1). The video is represented as a dictionary, where each column describes a video frame. The frames selection is defined by the activation vector, and the fast-forwarding effect is reached by the sparsity nature of the problem. In other words, we look for the smallest set of frames that provides the reconstruction of the original video with small error. Additionally, to attenuate abrupt camera movements in the final video, we apply a weighted version of the MSR problem, where frames related to camera movement are more likely to be sampled.

In the proposed modeling, the scalability of features is not a problem anymore, because using a high dimensional descriptor leads to a balance of the dictionary dimensions, which is recommended to solve the MSR problem, and do not substantially affect the computational cost and memory usage. We experimentally demonstrate that our approach creates videos composed of more relevant information than the state-of-the-art Semantic Fast-Forwarding method and as smooth as the state-of-the-art Hyperlapse techniques.

The contributions of our work are: i) a set of methods capable of handling larger feature vectors to better describe the frames and the video transitions, addressing the abrupt camera motions while not increasing the computational processing time; ii) a new labeled 80-hour multimodal (3D Inertial Movement Unit, GPS, and RGB-D camera) dataset of first-person videos covering a wide range of activities such as video actions, party, beach, tourism, and academic life. Each frame is labeled with respect to the activity, scene, recorder ID, interaction, and attention.

2. Related Work

Works on selective highlighting of the meaningful parts of first-person videos have been extensively studied in the past few years. We can broadly classify them into Video Summarization and Hyperlapse approaches.

Video Summarization. The goal of video summarization is to produce a compact visual summary containing the most discriminative and informative parts of the original video. Techniques typically use features that range from low-level such as motion and color [32, 5] to high-level (e.g., important objects, user preferences) [8, 30, 25]. Lee et al. [11] exploit interaction level, gaze, and object detection frequency as egocentric properties to create a storyboard of keyframes with important people and objects. Lu and Grauman [13] present video skims as summaries instead of static keyframes. After splitting the video into subshots, they compute the mutual influence of objects and estimate the subshots importance to select the optimal chain of subshots.

Recent approaches are based on highlight detection [12, 2, 30] and vision-language models [25, 20, 18]. Bettadapura et al. [2] propose an approach for identifying picturesque highlights. They use composition, symmetry and color vibrancy as scoring metrics and leverage GPS data to filter frames by the popularity of the location. Plummer et al. [20] present a semantically-aware video summarization. They optimize a linear combination of visual, i.e., representativeness, uniformity, interestingness, and vision-language objectives to select the best subset of video segments.

Sparse Coding has been successfully applied to many varieties of vision tasks [29, 31, 3, 32, 17, 15, 16]. In video summarization, Cong et al. [3] formulate the problem of video summarization as a dictionary selection problem. They propose a novel model to either extract keyframes or generate video skims using sparsity consistency. Zhao et al. [32] propose a method based on online dictionary learning that generates summaries on-the-fly. They use sparse coding to eliminate repetitive events and create a representative short version of the original video. The main benefit of using sparse coding for frame selection is that selecting a different number of frames does not incur an additional computational cost. This work differs from sparse coding video summarization since it handles the shakiness in the transitions via a weighted sparse frame sampling solution. Also, it is capable of dealing with the temporal gap caused by discontinuous skims.

Hyperlapse. A pioneering work in creating hyperlapse from casual first-person videos was conducted by Kopf et al. [9]. The output video comes from the use of image-based rendering techniques such as projecting, stitching and blending after computing the optimal trajectory of the camera poses. Despite their remarkable results, the method has a high computational cost and requires camera motion and parallax to compute the 3D model of the scene.

Recent strategies focus on selecting frames using different adaptive approaches to adjust the density of frame selection according to the cognitive load. Poleg et al. [22] model the frame selection as a shortest path in a graph. The nodes of this graph represent the frames of the original video and, the edges weights between pairs of frames are proportional to the cost of including the pair sequentially in the out-
Semantic Hyperlapse. Unlike traditional hyperlapse techniques, where the goal is to optimize the output number of frames and the visual smoothness, the semantic hyperlapse techniques also include the semantic relevance for each frame. Ramos et al. [23] introduced an adaptive frame sampling process embedding semantic information within. The methodology assigns scores to frames based on the detection of predefined objects that may be relevant to the recorder. The rate of dropped frames is a function of the relative semantic load and the visual smoothness. Later, Silva et al. [27] extended the Ramos et al.'s method using a better semantic temporal segmentation and an egocentric video stabilization process in the fast-forward output. The drawbacks of these works include abrupt changes in the acceleration and shaky exhibition at every large lateral swing in the camera.

Most recently, two new hyperlapse methods for first-person videos were proposed: the Lai et al.'s system [10] and the Multi-Importance Fast-Forward (MIFF) [26] method. Lai et al.'s system converts 360° videos into normal field-of-view hyperlapse videos. They extract semantics through regions of interest using spatial-temporal saliency and semantic segmentation to guide camera path planning. Low rates of acceleration are assigned to interesting regions to emphasize them in the hyperlapse output. In the MIFF method, the authors applied a learning approach to infer the users' preference and determine the relevance of a given frame. The MIFF calculates different speed-up rates for segments of the video, which are extracted using an iterative temporal segmentation process according to the semantic content.

Although not focused on the creation of hyperlapses, Yao et al. [30] present a highlight-driven summarization approach that generates skimming and timelapse videos as summaries from first-person videos. They assign scores to the video segments by using late fusion of spatial and temporal deep convolution neural networks (DCNNs). The segments with higher scores are selected as video highlights. For the video timelapse, they calculate proper speedup rates such that the summary is compressed in the non-highlight segments and expanded in highlight segments. It is noteworthy that timelapse videos do not handle the suavity constraint that is a mandatory requirement for hyperlapse videos. Differently from the aforementioned work, our approach optimizes semantic, length and smoothness to create semantic hyperlapses. Most importantly, it keeps the path taken by the recorder avoiding to lose the flow of the story and thus, conveying the full message from the original video in a shorter and smoother version.

3. Methodology

In this section, we describe a new method for creating smooth fast-forward videos that retains most of the semantic content of the original video in a reduced processing time. Our method consists of four primary steps: i) Creation and temporal segmentation of a semantic profile of the input video; ii) Weighted sparse frame sampling; iii) Smoothing Frame Transitions (SFT), and iv) Video compositing.

3.1. Temporal Semantic Profile Segmentation

The first step of a semantic fast-forward method is the creation of a semantic profile of the input video. Once we set a semantic (e.g., faces, type of objects of interest, scene, etc.), a video score profile is created by extracting the relevant information and assigning a semantic score for each frame of the video (Figure 2-a). The confidence of the classifier combined with the locality and size of the regions of interest score are used as the semantic score [23, 27].

The set of scores defines a profile curve, which is used for segmenting the input video into semantic and non-semantic sequences. Following, a refinement process is executed in the semantic segments, creating levels of importance regarding the defined semantic. Finally, speed-up rates are calculated based on the length and level of relevance of each segment. The rates are calculated such that the semantic segments are played slower than the non-semantic ones, and the whole video achieves the desired speed-up. We refer the reader to [26] for a more detailed description of the multi-importance semantic segmentation.

The output of this step is a set of segments that are used to feed the following steps that process each one separately.

3.2. Weighted Sparse Frame Sampling

In general, hyperlapse techniques solve the adaptive frame selection problem searching the optimal configuration (e.g., shortest path in a graph or dynamic programming) in a space of representation where different types of features are combined to represent frames or transitions between frames. A large number of features can be used for improving the representation of a frame or transitions, but such solution leads to a high-dimensional representation space increasing the computation time and memory usage.
We address this problem of representation using a sparse frame sampling approach, Figure 2-d.

Let \( D = [d_1, d_2, d_3, \ldots, d_n] \in \mathbb{R}^{f \times n} \) be a segment of the original video with \( n \) frames represented in our feature space. Each entry \( d_i \in \mathbb{R}^f \) stands for the feature vector of the \( i \)-th frame. Let the video story \( v \in \mathbb{R}^f \) be defined as the sum of the frame features of the whole segment, i.e., \( v = \sum_{i=1}^{n} d_i \). The goal is to find an optimal subset \( S = [d_{s_1}, d_{s_2}, d_{s_3}, \ldots, d_{s_m}] \in \mathbb{R}^{f \times m} \), where \( m \ll n \) and \( \{s_1, s_2, s_3, \ldots, s_m\} \) belongs to the set of frames in the segment.

Let the vector \( \alpha \in \mathbb{R}^n \) be an activation vector indicating whether \( d_i \) is in the set \( S \) or not. The problem of finding the values for \( \alpha \) that lead to a small reconstruction error of \( v \), can be formulated as a Locality-constrained Linear Coding (LLC) [28] problem as follow:

\[
\arg \min_{\alpha \in \mathbb{R}^n} \|v - D \alpha\|^2 + \lambda_\alpha \|W \odot \alpha\|^2, \tag{1}
\]

where \( g \) is the Euclidean distance of each dictionary entry \( d_i \) to the segment representation \( v \), and \( \odot \) is an element-wise multiplication operator. The \( \lambda_\alpha \) is the regularization term of the locality of the vector \( \alpha \).

The benefit of using LLC formulation instead of the traditional Sparse Coding (SC) model is twofold. The LLC provides local smooth sparsity and can be solved by an analytical solution, which results in a smoother final fast-forward video in a lower computational cost.

**Weighted Sampling.** Abrupt camera motions are challenging issues for fast-forwarding video techniques. They might lead to the creation of shaky and nauseating videos. To tackle this issue, we used a weighted Locality-constrained Linear Coding formulation, where each dictionary entry has a weight assigned to it:

\[
\alpha^* = \arg \min_{\alpha \in \mathbb{R}^n} \|v - D \alpha\|^2 + \lambda_\alpha \|W \odot \alpha\|^2, \tag{2}
\]

where \( W \) is a diagonal matrix built from the weight vector \( w \in \mathbb{R}^n \), i.e., \( W \triangleq \text{diag}(w) \).

This weighting formulation provides a flexible solution, where we create different weights for frames based on the camera movement and thus, we can change the contribution for the reconstruction without increasing the sparsity/locality term significantly.

Let \( C \in \mathbb{R}^{c \times n} \) be the Cumulative Displacement Curves [21], i.e., the cumulative sum of the Optical Flow magnitudes, computed over the horizontal displacements in \( 5 \times 5 \) grid windows of the video frames (Figure 2-c). Let \( C' \in \mathbb{R}^{c \times n} \) be the derivative of each curve \( C \) w.r.t. time. We assume frame \( i \) to be within an interval of abrupt camera motion if all curves \( C' \) present the same sign (positive/negative) at the point \( i \), which represents a head-turning movement [21]. We assign a lower weight for these motion intervals to enforce them to be composed of a larger number of frames. We empirically set the weights to \( w_i = 0.1 \) and \( w_i = 1.0 \) for the frame features inside and outside the interval, respectively.

**Speed-up Selection.** All frames related to the activated positions of the vector \( \alpha^* \) will be selected to compose the final video. Since \( \lambda_\alpha \) controls the sparsity, it also manages the speed-up rate of the created video. The zero-value \( \lambda_\alpha \) enables the activation of all frames leading to a complete reconstruction. To achieve the desired speed-up, we perform an iterative search starting from zero, as depicted in Algorithm 1. The function \( \text{NumberOfFrames}(\lambda) \) (Line 4) solves Equation 2 using \( \lambda \) as the value of \( \lambda_\alpha \) and returns the number of activations in \( \alpha^* \).
and 19 can be solved by exhaustive search, since the Equation (3) can be solved by exhaustive search. We identify a shakier transition using the speed-up deviation term. This term calculates how far the desired number frames. Then, we iteratively insert frames into the shakier transitions (Figure 3.3. Smoothing Frame Transitions

A solution \( \alpha^* \) does not ensure a final smooth fast-forward video. Occasionally, the solution might provide a low reconstruction error of small and highly detailed segments of the video. Thus, by creating a better reconstruction with a limited number of frames, \( \alpha^* \) might ignore stationary moments or visually likely views and create videos similar to results of summarization methods.

We address this issue by dividing the frame sampling into two steps. First, we run the weighted sparse sampling to reconstruct the video using a speed-up multiplied by a factor \( SpF \). The resulting video contains 1/\( SpF \) of the desired number frames. Then, we iteratively insert frames into the shakier transitions (Figure 2-e) until the video achieves the exact number of frames.

Let \( I(F_x, F_y) \) be the instability function defined by

\[
I(F_x, F_y) = AC(F_x, F_y) \ast (d_y - d_x - speedup).
\]

The function \( AC(F_x, F_y) \) calculates the Earth Mover's Distance [19] between the color histograms of the frames \( F_x \) and \( F_y \). The second term of the instability function is the speed-up deviation term. This term calculates how far the distance between frames \( F_x \) and \( F_y \), i.e., \( d_y - d_x \), are from the desired speedup. We identify a shakier transition using the Equation 3:

\[
\alpha^* = \arg \max_{i \in \mathbb{R}^m} I(F_{s_i}, F_{s_{i+1}}).
\]

3.3. Smoothing Frame Transitions

Frame Description. The feature vector of the \( i \)-th frame \( d_i \in \mathbb{R}^{446} \) (Figure 2-b) is composed of the concatenation of the following terms. The \( hof_{ni} \in \mathbb{R}^{50} \) and \( hof_{ni} \in \mathbb{R}^{72} \) are histogram of optical flow magnitudes and orientations of the \( i \)-th frame, respectively. The appearance descriptor, \( a \in \mathbb{R}^{144} \), is composed of the mean, standard deviation, and skewness values of HSV color channels of the windows in a \( 4 \times 4 \) grid of the frame \( i \). To define the content descriptor, \( c \in \mathbb{R}^{80} \), we first use the YOLO [24] to detect the objects in the frame \( i \); then, we create a histogram with these objects over the 80 classes of the YOLO architecture. Finally, the sequence descriptor, \( s \in \mathbb{R}^{100} \), is an one hot vector, with the mod\((i, 100)\)-th feature activated.

The set of frames from \( F_{s_i} \) to \( F_{s_{i+1}} \), i.e., solution of Equation 3, has visually dissimilar frames with a distance between them higher than the required speed-up.

After identifying the shakier transition, from the subset with frames ranging from \( F_{s_i} \) to \( F_{s_{i+1}} \), we choose the frame \( F_j \), that minimizes the instability of the frame transition as follows:

\[
\alpha^* = \arg \min_{j \in \mathbb{R}^m} I(F_{s_i}, F_j)^2 + I(F_j, F_{s_{i+1}})^2.
\]

Equations 3 and 4 can be solved by exhaustive search, since the interval is small. In this work, we use \( SpF = 2 \) in the experiments. Higher values enlarge the search interval, increasing the time for solving Equation 4.

3.4. Video compositing

All selected frames of each segment are concatenated to compose the final video (Figure 2-f). In this last step, we also run the egocentric video stabilization proposed in the work of Silva et al. [27], which is properly designed to fast-forward egocentric videos. The stabilizer creates smooth transitions by applying weighted homographies. Images corrupted during the smoothing step are reconstructed using the non-selected frames of the original video.

4. Experiments

In this section, we describe the experimental results on the Semantic Dataset [27] and a new multimodal semantic egocentric dataset. After detailing the datasets, we present the results followed by the ablation study on the components and efficiency analysis.

4.1. Datasets and Evaluation criterion

Semantic Dataset. We first test our method using the Semantic Dataset, proposed by Silva et al. [27]. This dataset is composed of 11 annotated videos. Each video is classified having 0%, 25%, 50%, or 75% of semantic content in the semantic portions (a set of frames with high semantic score) on average. For instance, in the Walking 25p video, the recorder is walking and there are an average of 25% of frames with faces and/or pedestrians. It is worth noting that even when video belongs to the class Op, it still contains semantics on its frames. The reason of being classified as Op

![Image](3743x6874 to 4287x7191)

Figure 3. Left: setup used to record videos with RGB-D camera and IMU. Center: frame samples from DoMSEV. Right: an example of the available labels for the image highlighted in green.

![Image](4291x6553 to 4835x6869)

Image 2658x6472 to 3740x7283

![Image](4837x4081)
is mainly because it does not have a minimum number of frames with high semantic score.

Because this dataset has the annotation of the semantic load, we can use it for finding the best semantic fast-forward method, i.e., the fast-forward approach that retains the highest semantic load of the original video.

Multimodal Semantic Egocentric Videos. Aside from the Semantic Dataset, we also evaluated our approach on a new 80-hour dataset. Because of the absence of unrestricted and available multimodal data to work with egocentric tasks, we propose an 80-hour Dataset of Multimodal Semantic Egocentric Videos (DoMSEV). The videos cover a wide range of activities such as shopping, recreation, daily life, attractions, party, beach, tourism, sports, entertainment, and academic life.

The multimodal data was recorded using either a Go-Pro Hero™ camera or a built setup composed of a 3D Inertial Measurement Unit (IMU) attached to the Intel Realsense™ R200 RGB-D camera. Figure 3 shows the setup used and a few of examples of frames from the videos. Different people recorded the videos in varied illumination and weather conditions.

The recorders labeled the videos informing the scene where some segment were taken (e.g., indoor, urban, crowded environment, etc.), the activity performed (walking, standing, browsing, driving, biking, eating, cooking, observing, in conversation, etc.), if something caught their attention and when they interacted with some object. Example of labels are depicted in Figure 3. Also, we create a profile for each recorder representing their preferences over the 80 classes of the YOLO classifier [24] and the 48 visual sentiment concepts defined by Sharghi et al. [25]. To create the recorders’ profile, we asked them to indicate their interest in each class and concepts in a scale from 0 to 10.

Table 1 summarizes in the “Info” and “Videos” columns the diversity of sensors and activities that can be found in the dataset. Due to the lack of space, we chose the videos which best represent the diverse activities, camera models, mounting, and the presence/absence of sensors info. The dataset, source code and the 3D model for printing the built setup are publicly available in https://www.verlab.dcc.ufmg.br/semantic-hyperlapse/cvpr2018.

Evaluation criterion. The quantitative analysis presented in this work is based on three aspects: instability, speed-up, and amount of semantic information retained in the fast-forward video. The Instability index is measured by using the cumulative sum over the standard deviation of pixels in a sliding window over the video [26]. The Speed-up metric is given by difference of the achieved speed-up rate from the required value. The speed-up rate is the ratio between the number of frames in the original video and in its fast-forward version. In this work, we used 10 as required speed-up. For the Semantic evaluation, we consider the labels defined in the Semantic Dataset, which characterize the relevant information as pedestrian in the “Biking” and “Driving” sequences, and face in the “Walking”. The semantic index is given by the ratio between the sum of the semantic content in each frame of the final video and the maximum possible semantic value (MPSV). The MPSV is the sum of the semantic scores of the $n$ top-ranked frames of the output video, where $n$ is the expected number of frames in the output video, given the required speed-up.

4.2. Comparison with state-of-the-art methods

In this section, we present the quantitative results of the experimental evaluation of the proposed method. We compare it with the methods: EgoSampling (ES) [22], Stabilized Semantic Fast-Forward (SSFF) [27], Microsoft Hyperlapse (MSH) [7] the state-of-the-art method in terms of visual smoothness, and Multi-Importance Fast-Forward (MIFF) [26] the state-of-the-art method in terms of the amount of semantics retained in the final video.
Table 1. Results and videos details of a sample of the proposed multimodal dataset. Duration is the length of the video before the acceleration. RS in the Camera column stands for RealSense™ by Intel® and Hero is a GoPro® line product.

<table>
<thead>
<tr>
<th>Videos</th>
<th>Ours MIFF</th>
<th>Speed-up</th>
<th>Instability</th>
<th>Time</th>
<th>Duration (hh:mm:ss)</th>
<th>Mount Camera</th>
<th>Info</th>
</tr>
</thead>
<tbody>
<tr>
<td>Academic_Life_09</td>
<td>21.80</td>
<td>24.74</td>
<td>0.01</td>
<td>0.00</td>
<td>47.56 59.38</td>
<td>145.6 3,298.5</td>
<td>01:02:53 helmet RS R200 ✓ ✓</td>
</tr>
<tr>
<td>Academic_Life_10</td>
<td>24.99</td>
<td>25.12</td>
<td>-0.02</td>
<td>1.53</td>
<td>47.47 51.62</td>
<td>282.2 7,654.7</td>
<td>02:04:33 head Hero5 ✓ ✓</td>
</tr>
<tr>
<td>Academic_Life_11</td>
<td>21.03</td>
<td>20.14</td>
<td>0.00</td>
<td>0.20</td>
<td>30.19 42.64</td>
<td>96.6 3,176.9</td>
<td>01:02:04 hand Hero4 ✓ ✓</td>
</tr>
<tr>
<td>Attraction_02</td>
<td>65.04</td>
<td>59.22</td>
<td>0.10</td>
<td>0.00</td>
<td>24.68 25.65</td>
<td>95.0 5,284.6</td>
<td>01:31:10 chest Hero5 ✓ ✓</td>
</tr>
<tr>
<td>Attraction_08</td>
<td>80.29</td>
<td>77.52</td>
<td>0.35</td>
<td>1.72</td>
<td>34.78 37.78</td>
<td>8.7 1,762.0</td>
<td>00:32:41 chest Hero5 ✓ ✓</td>
</tr>
<tr>
<td>Attraction_09</td>
<td>43.83</td>
<td>44.35</td>
<td>-0.18</td>
<td>0.29</td>
<td>51.30 52.42</td>
<td>27.7 3,265.1</td>
<td>00:52:43 helmet RS R200 ✓ ✓</td>
</tr>
<tr>
<td>Attraction_11</td>
<td>27.28</td>
<td>31.55</td>
<td>-0.05</td>
<td>-0.02</td>
<td>31.93 35.79</td>
<td>185.6 4,779.3</td>
<td>01:17:20 helmet RS R200 ✓ ✓</td>
</tr>
<tr>
<td>Daily_Life_01</td>
<td>18.76</td>
<td>20.01</td>
<td>0.04</td>
<td>2.56</td>
<td>47.06 49.05</td>
<td>126.3 5,222.0</td>
<td>01:16:45 head Hero5 ✓ ✓</td>
</tr>
<tr>
<td>Daily_Life_02</td>
<td>25.68</td>
<td>25.51</td>
<td>-0.10</td>
<td>3.48</td>
<td>38.16 46.80</td>
<td>46.4 5,741.3</td>
<td>01:33:39 head Hero5 ✓ ✓</td>
</tr>
<tr>
<td>Entertainment_05</td>
<td>24.63</td>
<td>23.93</td>
<td>0.04</td>
<td>0.01</td>
<td>33.79 39.12</td>
<td>20.8 3,786.1</td>
<td>00:55:25 helmet RS R200 ✓ ✓</td>
</tr>
<tr>
<td>Recreation_03</td>
<td>76.52</td>
<td>72.70</td>
<td>-0.04</td>
<td>0.45</td>
<td>41.69 43.64</td>
<td>37.8 3,518.7</td>
<td>00:57:39 helmet Hero4 ✓ ✓</td>
</tr>
<tr>
<td>Recreation_08</td>
<td>24.20</td>
<td>26.33</td>
<td>-0.05</td>
<td>3.74</td>
<td>34.98 38.44</td>
<td>59.2 5,957.0</td>
<td>01:44:15 shoulder Hero5 ✓ ✓</td>
</tr>
<tr>
<td>Recreation_11</td>
<td>67.94</td>
<td>65.25</td>
<td>0.20</td>
<td>0.02</td>
<td>12.49 12.15</td>
<td>17.9 2,802.9</td>
<td>00:46:04 chest Hero5 ✓ ✓</td>
</tr>
<tr>
<td>Sport_02</td>
<td>13.62</td>
<td>14.85</td>
<td>-0.13</td>
<td>6.25</td>
<td>44.96 52.59</td>
<td>20.0 2,387.6</td>
<td>00:43:20 head Hero5 ✓ ✓</td>
</tr>
<tr>
<td>Tourism_01</td>
<td>64.00</td>
<td>62.90</td>
<td>-0.01</td>
<td>2.15</td>
<td>28.93 31.57</td>
<td>33.6 3,283.4</td>
<td>00:55:35 chest Hero4 ✓ ✓</td>
</tr>
<tr>
<td>Tourism_02</td>
<td>48.24</td>
<td>47.22</td>
<td>-0.23</td>
<td>3.22</td>
<td>52.38 54.27</td>
<td>118.2 9,331.0</td>
<td>02:22:52 head Hero5 ✓ ✓</td>
</tr>
<tr>
<td>Tourism_04</td>
<td>27.20</td>
<td>29.24</td>
<td>0.00</td>
<td>0.10</td>
<td>53.14 56.41</td>
<td>229.4 8,302.5</td>
<td>01:46:38 helmet RS R200 ✓ ✓</td>
</tr>
<tr>
<td>Tourism_07</td>
<td>42.93</td>
<td>42.72</td>
<td>0.09</td>
<td>4.47</td>
<td>39.44 37.08</td>
<td>27.1 3,906.1</td>
<td>01:05:03 head Hero5 ✓ ✓</td>
</tr>
</tbody>
</table>

| Mean              | 39.89     | 39.63      | 0.00        | 1.72  | 38.38 42.08         | 87.7 4,636.6 |

1 Higher is better. 2 Better closer to 0. 3 Lower is better.

Figure 4-a shows the results of the Semantic evaluation performed using the sequences in the Semantic Dataset. We use the area under the curves as a measure of the retained semantic content. Our approach outperformed the other methodologies. The area under the curve of the proposed method was 100.74% of the area under the MIFF curve, which is the state-of-the-art in semantic hyperlapse. The second Semantic Hyperlapse technique evaluated, SSFF, had 52.01% of the area under curve of MIFF. Non-semantic hyperlapse techniques such as MSH and ES achieved at best 19.63% of the MIFF area.

The results for Instability are presented as the mean of the instability indexes calculated over all sequences in the Semantic Dataset (Figure 4-b, lower values are better). The black dotted and the green dashed lines stand for the mean instability index when using an uniform sampling and for the original video, respectively. Ideally, it is better to yield an instability index as closer as possible to the original video. The reader is referred to the Supplementary Material for the individual values. The chart shows that our method created videos as smooth as the state-of-the-art method MSH and smoother than the MIFF.

Figure 4-c shows the speed-up achieved by each method. The bar represent the average difference between the required speed-up and the rate achieved by a respective method for each class of video in the Semantic Dataset. Values closer to zero are desirable. The chart shows that our method provided the best acceleration for “Driving” and “Walking” experiments. In “Biking” experiments MSH held the best speed-up.

As far as the semantic metric is concerned (Figure 4-a), our approach leads followed by MIFF. We ran a more detailed performance assessment comparing our method to MIFF in the multimodal dataset. The results are shown in Table 1. As can be seen, our method outperforms MIFF in all metrics. The column “Time” shows the time for the frame sampling step of each method (MIFF runs a parameter setup and the shortest path, and our method runs minimum reconstruction followed by the smoothing step). Our method was 53× faster than MIFF. It is noteworthy that, unlike MIFF that requires 14 parameters to be adjusted, our method is parameter-free. Therefore, the average processing time spent per frame was 0.5 ms, while the automatic parameter setup process and the sampling processing of MIFF spent 30 ms per frame. The descriptor extraction for each frame ran in 320 ms facing 1,170 ms of MIFF. The experiments were conducted in a machine with i7-6700K CPU @ 4.00GHz and 16 GB of memory.

4.3. Ablation analysis

In this Section, we discuss the gain of applying the steps Weighted Sparse Frame Sampling and Smoothing Frame Transitions to the final fast-forward video. All analysis were conducted in the Semantic Dataset.
Figure 5. The effect of applying the Weighted Sparse Sampling in an abrupt camera movement segment. Black arrows are the frames of the original video, red arrows are the frames selected by non-weighted sparse sampling, and the green arrows represent the frames sampled by the weighted sparse sampling. Each image is related with the respective numerated arrow.

Weighted Sparse Sampling. As stated, we introduce a new model based on weighted sparse sampling to address the problem of abrupt camera motions. In this model, small weights are applied to frames containing abrupt camera motions to increase the probability of these frames being selected and, consequently, to create a smooth sequence.

Considering all sequences of abrupt camera motions present in all videos of the Semantic Dataset, the weighted version manages to sample, in average, three times more frames than the non-weighted version. Figure 5 illustrates the effect of solving the sparse sampling by weighting the activation vector. It can be seen that the weighting strategy helps by using a denser sampling in curves (on the right) than when applying the non-weighted sparse sampling version (on the left). In this particular segment, our approach sampled twice the number of frames, leading to less shaky lateral motions.

Smoothing Frame Transitions. By computing the coefficient of variation (CV), we measured the relative variability of the points representing the appearance cost of the frames (blue and red points in Figure 6). The appearance cost is computed as the Earth Mover’s Distance [19] between the color histogram of frames in a transition.

After applying the proposed smoothing approach we achieved \( CV = 0.97 \), while the simple sampling provided \( CV = 2.39 \). The smaller value for our method indicates a smaller dispersion and consequently fewer visual discontinuities. Figure 6 shows the result when using SFT and non-smoothed sparse sampling. The horizontal axis contains the index of selected frames and the vertical axis represents the appearance cost between the \( i \)-th frame and its following in the final video. The points in the red line represent the oversampling pattern of non-smoothed sparse sampling, in which many frames are sampled in segments hard to reconstruct followed by a big jump.

The abrupt scene changing is depicted by high values of appearance cost. The red-bordered frames in the figure show an example of two images that compose the transition with the highest appearance cost for a fast-forwarded version of the video “Walking 25p” using non-smoothed sparse sampling. After applying the SFT approach, we have a more spread sampling covering all segments, and with less video discontinuities. The blue-bordered images present the frames composing the transition with the highest appearance cost using the sparse sampling with the SFT step. By comparing the red and blue curves, one can clearly see that after using SFT, we achieve smoother transitions, i.e., lower values for the appearance cost.

5. Conclusion

In this work, we presented a new semantic fast-forward parameter-free method for first-person videos. It is based on a weighted sparse coding modeling to address the adaptive frame sampling problem and smoothing frame transitions to tackle abrupt camera movements by using a denser sampling along the segments with high movement. Contrasting with previous fast-forward techniques that are not scalable in the number of features used to describe the frame/transition, our method is not limited by the size of feature vectors.

The experiments showed that our method was superior to state-of-the-art semantic fast-forward methods in terms of semantic, speed-up, stability, and processing time. We also performed an ablation analysis that showed the improvements provided by the weighted modeling and the smoothing step. An additional contribution of this work is a new labeled 80-hour multimodal dataset with several annotations related to the recorder preferences, activity, interaction, attention, and the scene where the video was taken.

Acknowledgments. The authors would like to thank the agencies CAPES, CNPq, FAPEMIG, and Petrobras for funding different parts of this work.
References


and stabilized egocentric videos. In *European Conference on Computer Vision Workshop (ECCVW)*, pages 557–571, Amsterdam, NL, October 2016. 2, 3, 5, 6


