Semantic Hyperlapse: a Sparse Coding-based and Multi-Importance Approach for First-Person Videos

Michel M. Silva, Mario F. M. Campos, Erickson R. Nascimento
Department of Computer Science
Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, Brazil
E-mails: {michelms, mario, erickson}@dcc.ufmg.br

***Abstract*—The availability of low-cost, high-quality personal wearable cameras combined with the unlimited storage capacity of video-sharing websites has evoked a growing interest in First-Person Videos (FPVs). Such videos are usually composed of long-running unedited streams captured by a device attached to the user body, which makes them tedious and visually unpleasant to watch. Consequently, there is a rise in the need to provide quick access to the information therein. To address this need, efforts have been applied to the development of techniques such as Hyperlapse and Semantic Hyperlapse, which aims to create visually pleasant shorter videos and emphasize semantic portions of the video, respectively. The state-of-the-art Semantic Hyperlapse method SSFF, negligees the level of importance of the relevant information, by only evaluating if it is significant or not. Other limitations of SSFF are the number of input parameters, the scalability in the number of visual features to describe the frames, and the abrupt change in the speed-up rate of consecutive video segments. In this dissertation, we propose a parameter-free Sparse Coding based methodology to adaptively fast-forward First-Person Videos, that emphasize the semantic portions applying a multi-importance approach. Experimental evaluations show that the proposed method creates shorter version video retaining more semantic information, with fewer abrupt transitions of speed-up rates, and more stable final videos than the output of SSFF. Visual results and graphical explanation of the methodology can be visualized through the link: https://youtu.be/8uStih8P5-Y.**

I. Introduction

Statics about Internet usage in 2017 announce that online videos represented 70% of global traffic. Studies predict that this number will strike 80% by 2022 [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. Wearable devices are one of the big players contributing to the rise in the amount of video data. These devices introduced the concept of free-hand recording, allowing the user to perform any activity in the meantime. Due to this feature, wearable cameras are being used to capture many hours of unedited videos from the most memorable events to monotonous and repetitive daily tasks, such as walking, jogging, cooking, driving, and working shift.

Long-running and boring videos decrease the propensity of future viewers to watch the footage, even the recorders could not pay attention to the majority of recordings [2], making significant moments to be lost along with activities that do not merit recording. Thus, a central challenge is to provide quick access to the meaningful parts of the videos without losing the whole message that the user would like to convey. To accelerate the video is one alternative to provide quick access to the information while keeping the context. However, First-Person Videos (FPVs) incorporate the natural body movements of the recorder, since they are recorded with the camera attached to the body. Accelerating these videos naïvely amplifies the movement frequency turning the video unwatchable [3]. Consequently, fast-forward egocentric video had attracted the attention of researchers.

Hyperlapse techniques address the shaking effects of fast-forwarding FPVs by performing an adaptive frame selection [3]–[7]. The drawback of these approaches is assuming every frame equally relevant, *e.g.*, in a lengthy stream of daily activity, some portions of the videos are undoubtedly more relevant than others. Recently, Semantic Hyperlapse techniques have emerged as a solution for fast-forwarding videos emphasizing the relevant content, dealing with visual smoothness and semantic highlighting of FPVs [8], [9].

Aiming to address both objectives, visual smoothness and semantic highlight, Semantic hyperlapse methods use features to describe the video frames and their transitions, then formulate an optimization problem using the combination of these features. Consequently, the number of features used impacts the computation time and memory usage, since the search space grows exponentially. Therefore, such Hyperlapse methods are not scalable regarding the number of features.

The problem addressed by this thesis is the selection of frames with constraints regarding visual smoothness, temporal continuity, and the semantic load of the original video. We tackle this problem by creating a Semantic Hyperlapse technique using sparse coding formulation to perform the adaptive frame sampling addressing the problem related to the scalability of the sampling optimization regarding the number of features to describe the frames.

Contributions. We list the main contributions as:

i) a Sparse Sampling-based adaptive frame selection approach to address the problem related to the scalability of the feature dimensionality in a time-efficient manner.
ii) a Machine Intelligence method to learning the user's preference from visual data and their statistics.
iii) the publicly available 80-hour unconstrained Dataset of Multimodal (Depth, IMU, and GPS) Egocentric Videos with labels regarding the frames, videos, and recorders.

This work relates to an Ph.D. thesis.

II. Related Work

Video processing to resume the story of First-Person Videos has been extensively studied in the past few years, especially the video summarization problem and fast-forward techniques. The fundamental difference between these two techniques is that Hyperlapse methods are focused on creating a visually smooth and temporally continuous shorter version of the input video, *i.e.*, the video is sped up entirely not removing any clips, unless there are stationary camera moments. Video summarization methods, on the other hand, are focused on creating compact visual summaries capable of presenting the most discriminative and/or the most enlightening parts of the video [2]. These summaries are usually presented in the format of video skims or key-frame collection of the relevant moments, not preserving the footage context [10].

Video Summarization. The goal of video summarization is to produce a compact visual summary containing the most informative parts of the original video. Lee *et al*. [11] exploited interaction level, gaze, and object detection frequency as egocentric properties to create a storyboard of keyframes with important people and objects. Sparse coding theory also has been applied to this task, as the work of Cong *et al*. [16] that formulated the summarization as a dictionary selection problem, and extract keyframes using sparsity consistency. Zhao *et al*. [17] proposed a method based on online dictionary learning that generates keyframes collection summaries on-the-fly using Sparse Coding to eliminate repetitive events. Sparse Coding has also been successfully applied to a variety of vision tasks [16]–[22]. This thesis differs from Sparse Coding video summarization since we handle both visual instability and temporal constraints while performing the frame sampling.

Hyperlapse. Kopf *et al*. [4] proposed the first Hyperlapse method addressing the visual instability through an adaptive frame sampling that reconstructs the 3D scene geometry and creates the final video using a virtual camera traveling on an optimal path. Poleg *et al*. [3] perform the frame sampling by performing the shortest path in a graph modeled as the frames are the nodes, edges are the frame transitions, and edge weights are a linear combination of the shakiness, speed of motion, and appearance between pairs of frames compositing the transitions. Halperin *et al*. [6] extended the work of Poleg with an expansion of the field of view of the output video by using the mosaicking technique on frames from multiples videos and stabilizing the final video by a moving cropping area. Microsoft Hyperlapse [5] is the state-of-the-art Hyperlpase method as far as visual smoothness is concerned. The authors modeled the frame sampling problem using dynamic-time-warping formulation. Wang *et al*. [7] created a Hyperlapse method based on multiple spatially-overlapping sources to synthesize virtual routes created from paths traveled by distinct cameras. Recently, Hyperlapse methods have been extended to omnidirectional videos [23]–[25].

Although these solutions have succeeded in creating short and watchable versions of long first-person shots, they neglect the semantic load of the videos.

Semantic Fast-Forward Methods. To the best of our knowledge, Okamoto and Yanai [26] proposed the pioneering semantic technique by fast-forwarding a guidance video with the emphasis on parts of the route containing street corners and pedestrian crosswalks. The authors applied a lower speed-up rated on these semantic segments regarding the non-semantic ones, and then frames were uniformly sampled. Despite the shaky results of applying uniform sampling in FPVs, other works [27], [28] on this category have also applied naïve sampling after calculating speed-up rates to emphasize the semantic segments.

Ramos *et al*. [8] designed the first Semantic Hyperlapse addressing both visual smooth and semantic load while fast-forwarding FPVs. The proposed method calculates the semantic load of the frames, segments the video into relevant and non-relevant portions, estimates speed-up rates in a manner that the semantic portions are played slower than the non-semantic ones, and performs a graph-based adaptive frame sampling. The drawbacks of this method are the feature scalability, since it is based on graph modeling, and shaky transitions in the non-semantic segments.

Silva *et al*. [9], in the context of this thesis, extended the work of Ramos *et al*. improving the visual smoothness by introducing a methodology combined with a stabilization process specially designed to fast-forwarded videos (Stabilized and Semantic Fast-Forward video - SSFF), which is based on weighted homography transformations and image stitching using frames dropped during the sampling process. Finally, the authors proposed a semantically controlled and labeled dataset to evaluate fast-forward videos regarding the semantic load. The failing cases of this and the previous methodologies are to treat the semantic information as a binary problem disregarding its level of relevance, and the *ad hoc* semantic definition (only faces or pedestrians).

In this thesis, we aim to create a novel methodology to sample frames adaptively addressing issues related to the existing works such as, treat the semantic analysis as a binary problem, *ad hoc* semantic definition, and the scalability regarding the number of frames and dimension of the feature vectors used to describe the frames. We model the frame sampling step as a Minimum Sparse Reconstruction problem. To the best of our knowledge, it is the first Sparse Coding-based Hyperlapse.

III. Methodology

Our method consists of five 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); iv) Filling gaps between segments, and v) Video compositing.

A. Temporal Semantic Profile Segmentation

In the first step, we create a semantic profile of the input video, by extracting the relevant information and assigning a score for each frame of the video (Fig. 1-a). We use the classifier-based *ad hoc* definition of semantic to perform the experimental evaluation as proposed in the work of Ramos *et*

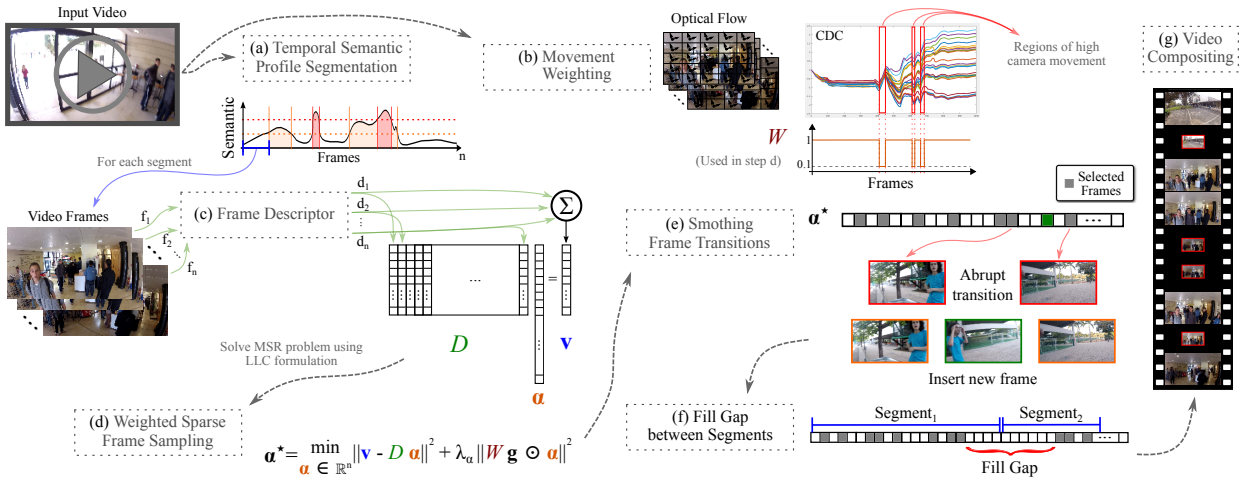

Fig. 1. Main steps of our semantic video fast-forward. For each segment created in the temporal semantic profile segmentation (a), weights based on the camera movement are computed (b) and the frames are described (c). Frames are sampled by minimizing local-constrained and reconstruction problem (d). The smoothing step is applied to tackle the abrupt transitions of the selected frames inside segments (e). Fill processing is applied to handle visual gaps between segments (f). Frames selected in previous steps are used to composite the final fast-forward video (g).

al. [8]. In this work, we proposed the *CoolNet*, a Convolutional Neural Network that learns the preference of the user from visual data of frames of YouTube videos in the YouTube8M dataset [29] and their statistics (number of views, likes, and dislikes). The readers is referred to our work [30] to details about the dataset creation, training routines, and model accuracy. The created semantic profile is used for segmenting the input video into sequences of different levels of semantic, and to compute speed-up rates such that it slows down the video portions according with their semantic load. We refer the reader to our work [30] for a more detailed description of the multi-importance semantic segmentation and speed-up rate assignment. The output is a set of segments that feeds the steps described in Sections III-B and III-C which process each one separately.

B. Weighted Sparse Frame Sampling

Hyperlapse techniques sample frames adaptively by searching the optimal configuration (*e.g.*, shortest path in a graph or dynamic programming) in a representation space where different features are combined to represent frames or frame transitions. Although recent works achieved better results applying a large number of features to represent the data [31]–[33], it increases both the computation time and memory usage since it leads to a high-dimensional space in optimization problems. We address this representation problem using a sparse frame sampling approach as depicted in Fig. 1-d.

Let $D = [\mathbf{d}_1, \mathbf{d}_2, \cdots, \mathbf{d}_n] \in \mathbb{R}^{f \times n}$ be a segment of the original video with n frames represented in our feature space. Each entry $\mathbf{d}_i \in \mathbb{R}^f$ stands for the feature vector of the i-th frame. Let the video story $\mathbf{v} \in \mathbb{R}^f$ be defined as the sum of the frame features of the whole segment, *i.e.*, $\mathbf{v} = \sum_{i=1}^{n} \mathbf{d}_i$. The goal is to find an optimal subset $S = [\mathbf{d}_{s_1}, \mathbf{d}_{s_2}, \cdots, \mathbf{d}_{s_m}] \in \mathbb{R}^{f \times m}$, where $m \ll n$ and $\{s_1, s_2, \cdots, s_m\}$ belongs to the set of frames in the segment.

Let the vector $\boldsymbol{\alpha} \in \mathbb{R}^n$ be an activation vector indicating whether $\mathbf{d}_i$ is in the set S or not. The problem of finding the values for $\boldsymbol{\alpha}$ that lead to a small reconstruction error of $\mathbf{v}$, can be formulated as a weighted Locality-constrained Linear Coding (LLC) [34] problem as follow:

$$\boldsymbol{\alpha}^\star = \arg\min_{\boldsymbol{\alpha} \in \mathbb{R}^n} \|\mathbf{v} - D\,\boldsymbol{\alpha}\|^2 + \lambda_\alpha \|W\,\mathbf{g} \odot \boldsymbol{\alpha}\|^2, \quad (1)$$

where $\mathbf{g}$ is the Euclidean distance between each dictionary entry $\mathbf{d}_i$ and the segment representation $\mathbf{v}$, $\odot$ is an element-wise multiplication operator, λ_α is the regularization term of the locality of the vector $\boldsymbol{\alpha}$, and W is a diagonal matrix built from the weight vector $\mathbf{w} \in \mathbb{R}^n$, *i.e.*, $W \triangleq \text{diag}(\mathbf{w})$.

The benefit of using the LLC formulation instead of the traditional L0-pseudo norm or L1-norm Sparse Coding (SC) models is twofold: i) the LLC provides local smooth sparsity, and ii) it can be solved by an analytical solution, which results in a lower computational cost.

This weighting formulation provides a flexible solution, in which we create weights for frames based on the camera motion and thus we can modify the contribution for the reconstruction without increasing the sparsity term. This manner, we oversample frames in region of abrupt camera movement.

Let $C \in \mathbb{R}^{c \times n}$ be the Cumulative Displacement Curves [35] (Fig. 1-b), and $C' \in \mathbb{R}^{c \times n}$ 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 turning movement [35]. We empirically assign $\mathbf{w}_i = 0.1$ for frames in these intervals to enforce a denser sampling in these intervals, and $\mathbf{w}_i = 1.0$ for the remaining ones.

1) Speed-up Control: All frames related to the activated positions of the vector $\boldsymbol{\alpha}^\star$ will compose the final video. Since λ_α controls the sparsity, it also controls the speed-up rate of the created video. Therefore, we perform an iterative adjust in the λ_α value to achieve the desired number of frames.

2) Frame Description: The feature vector of the i-th frame $\mathbf{d}_i \in \mathbb{R}^{446}$ (Fig. 1-c) is composed of the concatenation of the following terms. The $\mathbf{hof_m} \in \mathbb{R}^{50}$ and $\mathbf{hof_o} \in \mathbb{R}^{72}$ are the histogram of the optical flow magnitudes and the orientations of the i-th frame, respectively. The appearance descriptor $\mathbf{a} \in \mathbb{R}^{144}$ contains the mean, standard deviation, and skewness values of the HSV color channels of the windows in a 4×4 grid of the frame i. To define the content descriptor $\mathbf{c} \in \mathbb{R}^{80}$, we use the YOLO [36] 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 $\mathbf{s} \in \mathbb{R}^{100}$ is an one-hot vector, with the $\text{mod}(i, 100)$-th feature activated indicating the video portion where the frame is located.

C. Smoothing Frame Transitions

A solution $\boldsymbol{\alpha}^\star$ does not ensure a final continuous fast-forward video. 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, $\boldsymbol{\alpha}^\star$ may ignore stationary moments or visually similar views and create videos akin to results of summarization methods.

We address this problem 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 (Fig. 1-e) until the video achieves the exact number of frames.

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

$$I(F_x, F_y) = AC(F_x, F_y) * (d_y - d_x - speedup). \quad (2)$$

The function $AC(F_x, F_y)$ calculates the Earth Mover's Distance [37] 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:

$$i^\star = \underset{i \in \mathbb{R}^m}{\arg\max}\, I(F_{s_i}, F_{s_{i+1}}). \quad (3)$$

The transition composed of $F_{s_{i^\star}}$ and $F_{s_{i^\star+1}}$, *i.e.*, solution of Eq. 3, has visually dissimilar frames with a distance between them larger than the required speed-up.

After identifying the shakier transition from the subset with frames ranging from $F_{s_{i^\star}}$ to $F_{s_{i^\star+1}}$, we choose the frame $F_{j^\star}$ that minimizes the instability of the frame transition as follows:

$$j^\star = \underset{j \in \mathbb{R}^n}{\arg\min}\, I(F_{s_{i^\star}}, F_j)^2 + I(F_j, F_{s_{i^\star+1}})^2. \quad (4)$$

Since the interval is small, Eq. 3 and 4 can be solved by exhaustive search (we use $SpF = 2$ in the experiments). Larger values increase the search interval, also increasing the time for solving Eq. 4.

D. Fill Gap between segments

Temporal discontinuities between some video segments may occur due to the frame selection being performed for each segment at a time while neglecting the remaining ones. If the last selected frame of one segment is far from the first selected frame of the following video segment, it creates a visual gap in the final video. Section III-C provides a valid solution by inserting frames and tackling the visual discontinuities created within the segments. However, it has no effect on frame transitions between segments.

Abrupt speed-up differences between video segments are additional issues present in most semantic fast-forward methods in the literature. These abrupt differences are caused by the selection of speed-up rates assigned to video segments. Generally, they occur when one segment containing a significant amount of semantic information is followed by, or follows, a non-semantic segment. This would create abrupt differences among speed-up rates assigned to each segment. For instance, in the experiment "Biking_50p" a $2\times$ speed-up semantic segment follows a non-semantic segment with speed-up $14\times$. In this section, we propose to address both the visual gap and the abrupt speed-up difference issues.

To address the visual gap problem, we first calculate the instability index (Eq. 2) between the last frame of a segment A and the first frame of its consecutive segment B. If the instability index is larger than the average instability over all transitions of segment A, then we create a new segment delimited by the last frame of segment A and the first frame of the segment B (Fig. 1-f). This newly created segment is then used to smooth the speed-up transition and fill the visual gap. To solve the abrupt speed-up difference problem, we define the speed-up rate for the new segment as the average value between the speed-ups of A and B. Then, we fill the visual gap by running the Weighted Sparse Frame Sampling and Smoothing Frame Transitions, defined in Sections III-B and III-C respectively, using the smoother calculated speed-up.

E. Video compositing

All selected frames of each segment are concatenated to compose the final video (Fig. 1-g). After the concatenation is done, we run the video stabilization designed to fast-forwarded videos proposed in the context of this work [9]. The stabilizer creates smooth transitions by applying weighted homography transformations. Frames corrupted by the homography transformations are reconstructed using image stitching and blending of the non-selected frames of the original video.

IV. Experiments

Competitors. We compare our method with: i) EgoSampling (ES) [3]; ii) Microsoft Hyperlapse (MSH) [5], the state-of-the-art method in terms of visual smoothness; and iii) Stabilized Semantic Fast-Forward (SSFF) [9], the state-of-the-art method in terms of retained amount of semantics.

Datasets. Two datasets were used for the evaluation process. The first one, Annotated Semantic Dataset (ASD), is composed of small and controlled videos regarding the amount semantic

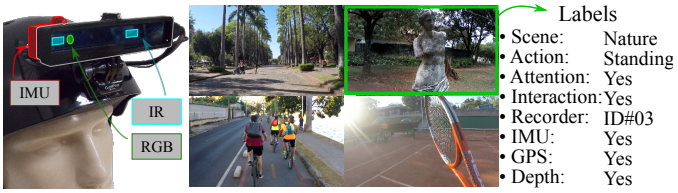

Fig. 2. Left: setup used to record videos with RGB-D camera and IMU. Center: frame samples from DoMSEV. Right: an example of the available labels for the image highlighted in green.

information of each video. We used it for finding the the fast-forward approach that retains the highest semantic load of the original video. Aside from the ASD dataset, we extend the evaluation process on a challenging dataset. Because of the absence of unrestricted and annotated data to work with egocentric tasks, we proposed an 80-hour *Dataset of Multimodal Semantic Egocentric Videos (DoMSEV)* covering a wide range of activities, light and weather conditions, places, camera mounting, device, and recorders. All details mentioned earlier are annotated along with the attention of the user while recording and their personal preferences. The multimodal data contains visual, depth, GPS, sound, and inertial information. A few examples of frames and the some labels are depicted in Fig. 2. DoMSEV, built setup, and video details are publicly available in www.verlab.dcc.ufmg.br/semantic-hyperlapse/cvpr2018-dataset/.

Metrics. The quantitative analysis presented in this work is based on four aspects: temporal discontinuity, visual instability, amount of semantic information retained in the fast-forward video, and processing time.

1) *Discontinuity*: we calculate the Root-Mean-Square Error (RMSE) over the selected frames jumps and the required speed-up rate for that video. Higher values indicate the accelerated video contains long jumps, which creates visual gaps.

2) *Instability*: it is measured by the cumulative sum of the standard deviation of pixels in a sliding window over the video [30]. The lower the value, less shaky is the video, indicating that the frame selection is visually pleasant to watch.

3) *Semantic*: this 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 for the video [8]. We consider the semantic labels defined in the Semantic Dataset.

4) *Processing Time*: we measure the time spent to run the frame sampling process comparing the time performance between the graph-based approach and the proposed sparse coding formulation. The reader is referred to the work [8], [30] for more details about the metrics.

Parameter settings. We used SpF = 2 during the Smoothing Frame Transitions. Half of the frames compositing the final video were sampled to reconstruct well the context of the original video, and the other half to smooth the transitions.

V. Results

We first evaluate quantitatively the *CoolNet*, then we perform a quantitative analysis over the proposed methodology.

A. CoolNet

Since most of the "Cool" images in our Dataset are related to radical sports and beautiful landscapes, the Network classifies with high score frames with nature elements, e.g., forest and gardens. Visually uniform frames, like indoor looking images, walls, and offices, yield to a low rating. Figure 4 depicts network score related to different scenes. In the left image, when the wearer passes through an inside garden, the network assigns an average rating. In the center image, the wearer is walking inside a building hall, which the net considers unattractive. In the right image, the wearer goes to an outside area composed of many trees and gardens, which are highly rated by the *CoolNet*.

Fig. 3-a shows the results of the Semantic evaluation performed using the sequences in the ASD Dataset, in which the area under the curves measures the retained semantic content. The area under the curve of our proposed method is more than the double of the area under the curve regarding the best competitor, SSFF, which is the state-of-the-art in this metric. Non-semantic hyperlapse techniques such as MSH and ES achieved at best 19.4% of our result.

The results for the Instability metric are presented as the mean of the instability indexes calculated over all sequences in the ASD Dataset (Fig. 3-b, lower values are better). The black dotted and the cyan dashed lines stand for the mean instability index when using a uniform sampling and for the original video, respectively. Ideally, it is better to yield an instability index as close as possible to the original video. The chart shows that our methodology created videos smoother than the state-of-the-art method MSH.

The Chart in Fig. 3-c depicts the visual gap problem related to the frame selection of Semantic Hyperlapse techniques. Our proposed method achieved a value of 10.1, while the lowest value 5.7 was achieved by MSH. However, it is noteworthy that MSH is a non-semantic hyperlapse method, *i.e.*, all segments are sped-up with at the same rate. The discontinuity value for semantic fast-forward methods is expected to be higher since semantic segments are accelerated in a rate smaller than the required for the whole video. Consequently, the non-semantic segments will have a greater speed-up rate

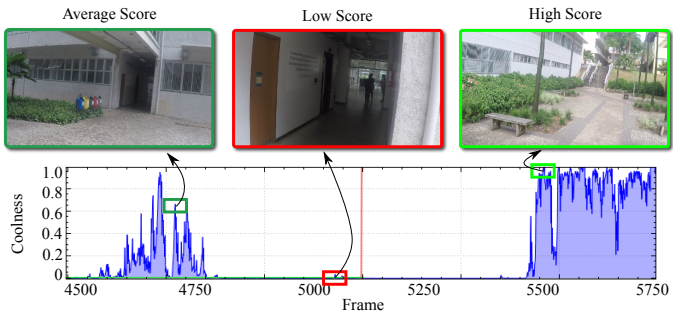

Fig. 4. Semantic Profile curve of the *CoolNet* for a sample video. *Left*: image depicts an inside garden, with its medium score. *Center:* a building hall, that the CoolNet does not consider containing semantic content. *Right:* garden with a outdoor view, for which CoolNet gives the highest scores.

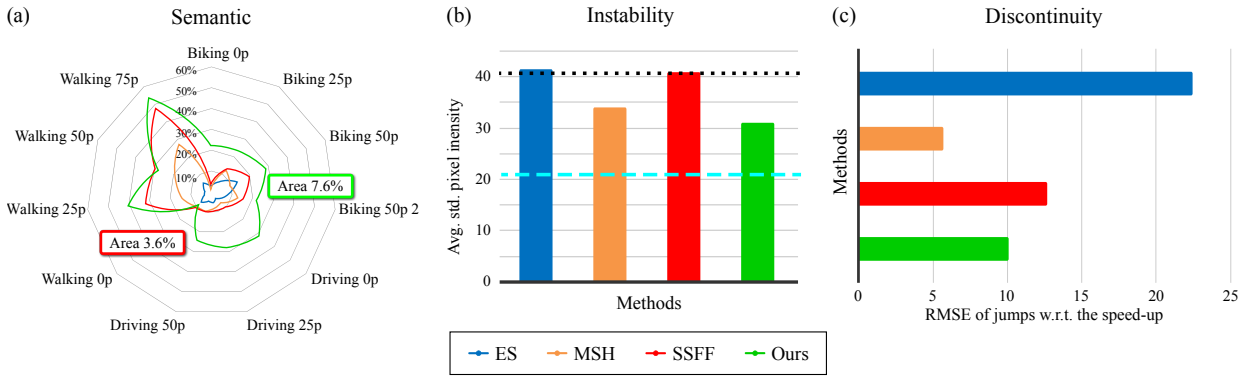

Fig. 3. Evaluation of the proposed Sparse Sampling methodology against the competitors using the Annotated Semantic Dataset. Dashed and doted lines in (b) are related to the mean instability of the input video and the uniform sampling, respectively. Better values are: (a) higher, (b) and (c) lower.

assigned to it.

Fig. 5 shows the time for the frame sampling step of our method and the best semantic competitor SSFF. We run a parameter setup based on Particle Swarm Optimization and the shortest path in the SSFF. Our methodology runs minimum reconstruction, frame transition smoothing, and fill gap between segments steps. The execution time of SSFF grows exponentially while our method was not influenced by the growth in the number of frames in the input video.

It is noteworthy that unlike SSFF which requires 14 parameters to be adjusted, our method is parameter-free. Therefore, the average processing time spent per frame to perform the frame sampling step using our methodology was 0.2 ms, while the automatic parameter setup process and the sampling processing of SSFF spent 36 ms per frame, indicating that our method is 170× faster, with no code optimization. The descriptor extraction for each frame ran in 320 ms facing 1,170 ms of SSFF. The experiments were ran in a machine with an i7-6700K CPU @ 4.00GHz and 16 GB of memory.

VI. Conclusion

We tackled the challenging task of creating Semantic Hyperlapse for a First-Person Video through a sparse coding-based framework composed of the adaptive frame sampling, Smooth Frame Transition, and Fill Gap steps. The frame sampler was modeled as a weighted minimum sparse reconstruction problem allowing a denser sampling along the segments with high camera movement. The Smoothing Frame Transitions step addressed visual instability by inserting frames into abrupt transitions, while the Fill Gap step dealt with temporal discontinuities. Contrasting with previous fast-forward methods 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. Experimental evaluation showed that our hyperlapse videos kept the double of semantic information, were smoother, and present fewer temporal discontinuities when compared with the best competitors SSFF and MSH. Moreover, the improvements did not affect the running time of the frame sampling process. An additional contribution is the smoothing of abrupt speed-up transitions, leading to more natural accelerated videos. An ablation study was performed to evaluate the contributions of each step of the methodology, the results can be visualized in the thesis.

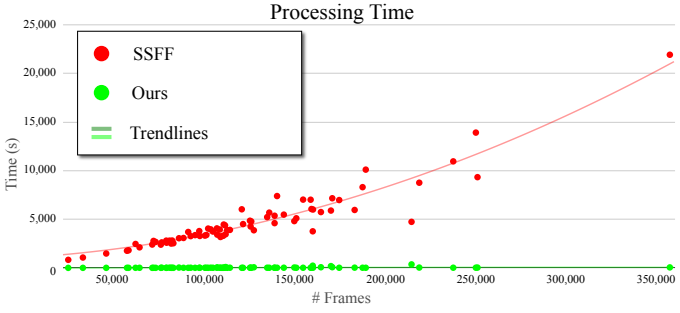

Fig. 5. Processing time regarding to video length. Y-axis is shown in logarithmic scale. Trend-lines follow a second order polynomial curve.

Limitations and Future Work. The main drawback of this work is to model the frame sampling problem regardless of the temporal information of frames, *i.e.*, the transitions information between frames are not encoded. Future steps to continue evolving the result are to address the characterization of frame transition and to perform the Smooth Frame Transition step adding virtual frames shaped by encoding temporal information of dropped frames.

Acknowledgment. We would like to thank the PPGCC-UFMG, CAPES and CNPq for funding this work.

VII. Awards & Publications

Contributions of this work were published on the Workshop on Egocentric Perception, Interaction and Computing at European Conference on Computer Vision (EPIC@ECCVW) 2016, Journal of Visual Communications and Image Representation (JVCI) 2018, IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2018, and an journal extension is under review on the Transactions on Pattern Analysis and Machine Intelligence (TPAMI). The thesis related to this work was awarded with the Highlighted Doctoral Dissertation at the 2018 DCC UFMG Day of Post-Graduate Program, and with an IEEE Society Travel Grants.

REFERENCES

[1] Traffic-Inquiries, "Cisco visual networking index: Forecast and methodology, 2017-2022," CISCO, Tech. Rep. 1543280537836565, November 2018. [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white-paper-c11-741490.html

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

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

[4] J. Kopf, M. F. Cohen, and R. Szeliski, "First-person hyper-lapse videos," ACM Trans. Graph., vol. 33, no. 4, pp. 78:1–78:10, Jul. 2014.

[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, pp. 63:1–63:9, Jul. 2015.

[6] 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. PP, no. 99, pp. 1–1, 2017.

[7] M. Wang, J. Liang, S. Zhang, S. Lu, A. Shamir, and S. Hu, "Hyper-lapse from multiple spatially-overlapping videos," IEEE Transactions on Image Processing (TIP), vol. 27, no. 4, pp. 1735–1747, April 2018.

[8] W. L. S. Ramos, M. M. Silva, M. F. M. Campos, and E. R. Nascimento, "Fast-forward video based on semantic extraction," in Proceedings of the IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, USA, Sept 2016, pp. 3334–3338.

[9] 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 Proceedings of the European Conference on Computer Vision Workshop (ECCVW). Amsterdam, NL: Springer International Publishing, October 2016, pp. 557–571.

[10] B. A. Plummer, M. Brown, and S. Lazebnik, "Enhancing video summarization via vision-language embedding," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, USA, July 2017, pp. 1052–1060.

[11] Y. J. Lee, J. Ghosh, and K. Grauman, "Discovering important people and objects for egocentric video summarization," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2012, pp. 1346–1353.

[12] Z. Lu and K. Grauman, "Story-driven summarization for egocentric video," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2013, pp. 2714–2721.

[13] B. Xiong, G. Kim, and L. Sigal, "Storyline representation of egocentric videos with an applications to story-based search," in Proceedings of the IEEE International Conference on Computer Vision (ICCV), Dec 2015, pp. 4525–4533.

[14] J. A. Yang, C. H. Lee, S. W. Yang, V. S. Somayazulu, Y. K. Chen, and S. Y. Chien, "Wearable social camera: Egocentric video summarization for social interaction," in IEEE International Conference on Multimedia Expo Workshops, July 2016, pp. 1–6.

[15] S. Lan, R. Panda, Q. Zhu, and A. K. Roy-Chowdhury, "Ffnet: Video fast-forwarding via reinforcement learning," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018, pp. 6771–6780.

[16] Y. Cong, J. Yuan, and J. Luo, "Towards scalable summarization of consumer videos via sparse dictionary selection," IEEE Transactions on Multimedia, vol. 14, no. 1, pp. 66–75, Feb 2012.

[17] B. Zhao and E. P. Xing, "Quasi real-time summarization for consumer videos," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, USA, June 2014, pp. 2513–2520.

[18] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 31, no. 2, pp. 210–227, Feb 2009.

[19] B. Zhao, L. Fei-Fei, and E. P. Xing, "Online detection of unusual events in videos via dynamic sparse coding," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Colorado Springs, USA, 2011, pp. 3313–3320.

[20] G. Oliveira, E. Nascimento, A. Vieira, and M. Campos, "Sparse spatial coding: A novel approach to visual recognition," IEEE Transactions on Image Processing (TIP), vol. 23, no. 6, pp. 2719–2731, June 2014.

[21] S. Mei, G. Guan, Z. Wang, M. He, X. S. Hua, and D. D. Feng, "L2,0 constrained sparse dictionary selection for video summarization," in 2014 IEEE International Conference on Multimedia and Expo (ICME), July 2014, pp. 1–6.

[22] S. Mei, G. Guan, Z. Wang, S. Wan, M. He, and D. D. Feng, "Video summarization via minimum sparse reconstruction," Pattern Recognition, vol. 48, no. 2, pp. 522 – 533, 2015.

[23] M. Ogawa, T. Yamasaki, and K. Aizawa, "Hyperlapse generation of omnidirectional videos by adaptive sampling based on 3d camera positions," in Proceedings of the IEEE International Conference on Image Processing (ICIP), Sep. 2017, pp. 2124–2128.

[24] P. Rani, A. Jangid, V. P. Namboodiri, and K. S. Venkatesh, "Visual odometry based omni-directional hyperlapse," in National Conference on Computer Vision, Pattern Recognition, Image Processing, and Graphics, R. Rameshan, C. Arora, and S. Dutta Roy, Eds. Singapore: Springer Singapore, 2018, pp. 3–13.

[25] 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 e-prints, Mar. 2017.

[26] M. Okamoto and K. Yanai, Summarization of Egocentric Moving Videos for Generating Walking Route Guidance. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014, pp. 431–442.

[27] T. Yao, T. Mei, and Y. Rui, "Highlight detection with pairwise deep ranking for first-person video summarization," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016.

[28] K. Higuchi, R. Yonetani, and Y. Sato, "Egoscanning: Quickly scanning first-person videos with egocentric elastic timelines," in Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, ser. CHI '17. New York, NY, USA: ACM, 2017, pp. 6536–6546.

[29] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan, "Youtube-8m: A large-scale video classification benchmark," CoRR, vol. abs/1609.08675, 2016.

[30] 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 (JVCI), vol. 53, pp. 55 – 64, 2018.

[31] M. Otani, Y. Nakashima, E. Rahtu, J. Heikkilä, and N. Yokoya, "Video summarization using deep semantic features," in Proceedings of the Asian Conference on Computer Vision (ACCV). Cham: Springer International Publishing, 2017, pp. 361–377.

[32] S. Lal, S. Duggal, and I. Sreedevi, "Online video summarization: Predicting future to better summarize present," in Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), Hawaii, USA, January 2019, to appear.

[33] T.-J. Fu, S.-H. Tai, and H.-T. Chen, "Attentive and adversarial learning for video summarization," in Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), Hawaii, USA, January 2019, to appear.

[34] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, USA, June 2010, pp. 3360–3367.

[35] Y. Poleg, C. Arora, and S. Peleg, "Temporal segmentation of egocentric videos," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2014, pp. 2537–2544.

[36] J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger," ArXiv e-prints, Dec. 2016.

[37] O. Pele and M. Werman, "Fast and robust earth mover's distances," in Proceedings of the IEEE International Conference on Computer Vision (ICCV), Sept 2009, pp. 460–467.

[38] 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 Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, USA, Jun. 2018, pp. 2383–2392.