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

Michel M. Silva[*], Mario F. M. Campos, Erickson R. Nascimento

[1]Department of Computer Science
Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, Brazil

{michelms, mario, erickson}@dcc.ufmg.br

Abstract. *The availability of low-cost and high-quality wearable cameras combined with the unlimited storage capacity of video-sharing websites have evoked a growing interest in First-Person Videos. 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, it raises the need to provide quick access to the information therein. We propose a Sparse Coding based methodology to fast-forward First-Person Videos adaptively. Experimental evaluations show that the shorter version video resulting from the proposed method is more stable and retain more semantic information than the state-of-the-art. Visual results and graphical explanation of the methodology can be visualized through the link:* https://youtu.be/rTEZurH64ME.

1. Contributions

We list as main contributions of this Ph.D. thesis as: i) a Sparse Sampling-based adaptive frame selection approach to address the problem related to the feature dimensionality scalability in a time-efficient manner; ii) a Machine Intelligence method to learning the user's preference from visual data and their statistics, and the corresponding train and test dataset; and iii) a new multimodal (Depth, IMU, and GPS) dataset comprising 80-hour unconstrained Egocentric Videos with labels regarding the frames, videos, and recorders. The code, dataset, and annotations are publicly available on the project website[1].

2. Awards and Publications

This thesis was awarded as the Best Ph.D. Thesis in the Workshop of Thesis and Dissertation at SIBGRAPI 2019, and as the Highlighted Doctoral Dissertation at 2018 DCC UFMG Day of PPGCC.

Contributions of this work were published on:

- The IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 2020. Impact factor: 17.730 (2018). According to the IEEE Computer Society, TPAMI has the highest impact factor of all Computer Society publications[2];

[*]This work relates to a Ph.D. thesis. Authors order: Ph.D. Candidate, Co-advisor, and Advisor.

[1]https://www.verlab.dcc.ufmg.br/semantic-hyperlapse/

[2]https://www.computer.org/publications/tech-news/insider-membership-news/computer-society-2018-impact-factors

- IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2018. Top 1 Computer Science conference according to Google Scholar[3].
- Journal of Visual Communications and Image Representation (JVCI) 2018. Impact factor: 1.836;
- International Workshop on Egocentric Perception, Interaction and Computing at European Conference on Computer Vision (EPIC@ECCV16) 2016;
- Conference on Graphics, Patterns and Images Tutorials (SIBGRAPI-T) 2019.

3. Methodology

3.1. Introduction

Internet users are contributing to the exponential growth in the amount of multimedia data. They are not only watching, but also recording themselves to share their day-to-day routine. Statics about Internet usage announced that videos represented 70% of global traffic, and studies predict that this number will strike 80% by 2022 [Traffic-Inquiries 2018]. 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 jogging or 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 [del Molino et al. 2017], making significant moments to be lost.

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. Hyperlapse techniques address the shaking effects of fast-forwarding FPVs by performing an adaptive frame selection [Kopf et al. 2014, Joshi et al. 2015, Poleg et al. 2015]. The drawback of these approaches is assuming all frames equally relevant, *e.g.*, in a lengthy stream of daily activity, some portions 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 [Ramos et al. 2016].

Aiming to address both objectives, these methods formulate an optimization problem using the combination of features extracted from the video frames and their transitions. Consequently, the computation time is impacted by the number of features used, once the search space grows exponentially. 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 frame sampling addressing the problem related to the scalability in the number of features to describe the frames.

3.2. Related Work

Video processing to resume the story of FPVs has been studied in the past few years, especially the video summarization problem and fast-forward techniques. The goal of

[3] https://scholar.google.com/citations?view_op=top_venues&hl=en&vq=eng

Video Summarization methods is to produce a compact summary capable of presenting the most discriminative or the most enlightening parts of the video [del Molino et al. 2017] in the format of video skimmings or keyframe collection. This thesis differs from Sparse Coding video summarization since we handle both visual instability and temporal constraints while performing the frame sampling.

Hyperlapse methods [Kopf et al. 2014, Joshi et al. 2015, Poleg et al. 2015] 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. Although these solutions have succeeded in creating short versions of long FPVs, they neglect the semantic load of the videos. Unlike traditional hyperlapse techniques, where the goal is to optimize the output number of frames and the visual smoothness, semantic hyperlapse techniques also deal with the semantic load of the frames. Semantic hyperlapse techniques assign emphasis on video portions that are relevant, given a semantic definition, by applying a lower acceleration rate [Ramos et al. 2016, Lan et al. 2018] or zooming effects [Lai et al. 2018].

In this thesis, we aim to create a novel methodology to sample frames adaptively addressing the issues related to the existing works, such as the semantic definition in an *ad hoc* manner, and the unscalability of the frame sampling process regarding the dimension of the feature vector due to its formulation as an optimization problem. 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 Semantic Hyperlapse.

3.3. Method

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.

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 proposed a method using a Convolutional Neural Network (CNN) to learn the users' preference from visual data of Youtube video frames and their number of views and likes. The semantic profile is used for segmenting the video into portions and computing speed-up rates such that it slows down according to their semantic load. The output is a set of segments that feeds the next steps, which process each one separately. In the Weighted sparse frame sampling step, we modeled the adaptive sampling as a weighted Locality-constrained Linear Coding (LLC) [Wang et al. 2010] problem. The dictionary basis are defined as the descriptors of each video frame. The goal is to find a sparse subset of frames that reconstruct the video story, as defined in Eq. 1:

$$\boldsymbol{\alpha}^{\star} = \underset{\boldsymbol{\alpha} \in \mathbb{R}^n}{\arg\min} \left\| \mathbf{v} - D\,\boldsymbol{\alpha} \right\|^2 + \lambda_{\alpha} \left\| W\,\mathbf{g} \odot \boldsymbol{\alpha} \right\|^2, \tag{1}$$

where D is the dictionary of a segment of the original video, $\mathbf{v}$ is the video story defined as the sum of the frame features of the whole segment, *i.e.*, $\mathbf{v} = \sum_{i=1}^{n} \mathbf{d}_i$, and $\mathbf{d}_i$ stands for the feature vector of the i-th video frame. $\mathbf{g}$ 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 weighting vector $\mathbf{w} \in \mathbb{R}^n$, *i.e.*, $W \triangleq \text{diag}(\mathbf{w})$.

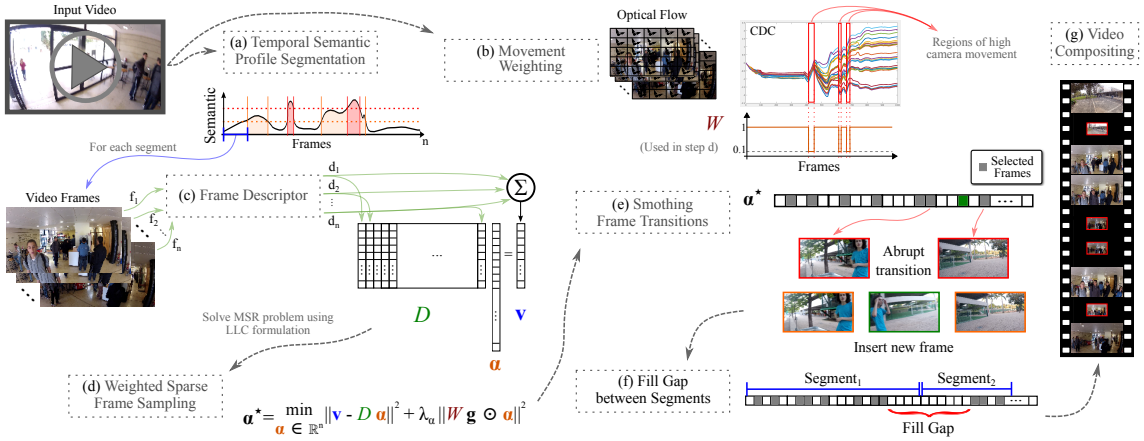

Figure 1. Main steps of our proposed semantic fast-forward framework for FPVs.

We assign weights for frames based on the camera motion (Fig. 1-b), modifying the contribution for the reconstruction without increasing the sparsity term, which leads to a frame oversampling in a region of abrupt camera movement. The benefit of using the LLC instead of the traditional sparse coding models is twofold, the LLC provides local smooth sparsity, and it can be solved by an analytic solution, which is more efficient. 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 video. Therefore, we perform an iterative adjust in the λ_α value to achieve the desired speed-up. A solution $\boldsymbol{\alpha}^\star$ does not ensure a final continuous fast-forward video since it ignores visually similar frames and creates videos akin to the results of summarization methods. We address this problem in the Smoothing Frame Transition step by running the sparse sampling to reconstruct the video using a speed-up multiplied by a factor SpF. Then, we iteratively identify the shakier transitions and insert a frame from the original video that minimizes the instability (Fig. 1-e) until the video achieves the exact number of frames.

Temporal discontinuities and abrupt speed-up differences between video segments may occur. These issues are produced due to the frame selection and speed-up rate estimation being performed for each segment at a time while neglecting the others. In the Filling gaps between segments steps, 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 created segment is used to fill the visual gap running the Weighted Sparse Frame Sampling and Smoothing Frame Transitions steps, and also to smooth the speed-up rate transition by setting the new acceleration rate as the average value between the speed-ups of A and B. In the last step, all selected frames of each segment are concatenated to compose the final video (Fig. 1-g). Finally, we run a video stabilization based on weighted homography transformations designed to fast-forwarded videos proposed in the context of this work.

3.4. Experiments

Competitors and metrics. We compare our method with: i) EgoSampling (ES) [Poleg et al. 2015]; ii) Microsoft Hyperlapse (MSH) [Joshi et al. 2015], the state-of-the-art method in terms of visual smoothness; and iii) Stabilized Semantic Fast-Forward (SSFF) [Silva et al. 2016], the state-of-the-art method in terms of retained amount of semantics. The experimental evaluation is based on the temporal discontinuity, visual instability and

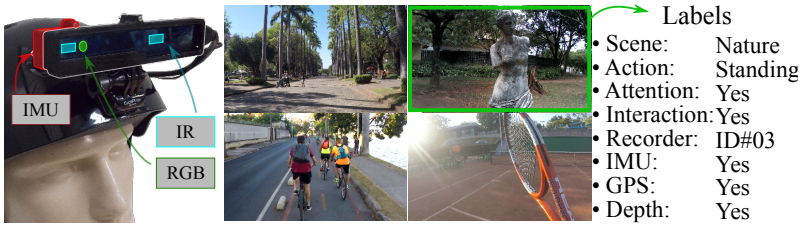

Figure 2. Example of RGB-D sensor and IMU, frame samples and available labels.

amount of semantic information of the output video, and processing time.

Datasets. Two datasets were used for the experimental evaluation, the Annotated Semantic Dataset (ASD) and the *Dataset of Multimodal Semantic Egocentric Videos (DoMSEV)*. This last is a proposed and public available 80-hour dataset with recorded videos covering a wide range of activities, light and weather conditions, places, camera mounting, device, and recorders. The mentioned details are annotated along with the level of attention of the user while recording and their personal preferences (Fig. 2). The multimodal data contains visual, depth, GPS, sound, and inertial information.

Results. Regarding the proposed CNN to assign relevance to the frames, the *CoolNet*, by using the statistics about the training Youtube videos, the Network classifies with high score frames with nature elements, *e.g.*, forest and gardens, once most of the "Cool" images in are related to radical sports and beautiful landscapes. Uniform frames, like indoor looking images, walls, and offices, yield to a low rating. The scores assigned by the CNN were consistent with the human behavior extracted from the statistics of the videos.

Fig. 3 summarizes the results of the output fast-forward videos compared with the competitors. As shown in (a), the Semantic retained is more than the double of the best competitor, SSFF, which is the state-of-the-art in this metric (higher values are better). In (b) are depicted the results for Visual Instability, presented as the mean value over all sequences in the ASD Dataset (lower values are better). The result shows that our methodology created videos smoother than the state-of-the-art method MSH. The chart in (c) depicts the impact of the Filling Gaps step. Our proposed method achieved the lowest value compared to the semantic competitor SSFF, excluding the non-semantic methods.

Regarding the time evaluation, differently from the state-of-the-art method SSFF that runs an optimization method to automatic parameter setup and frame sampling, our

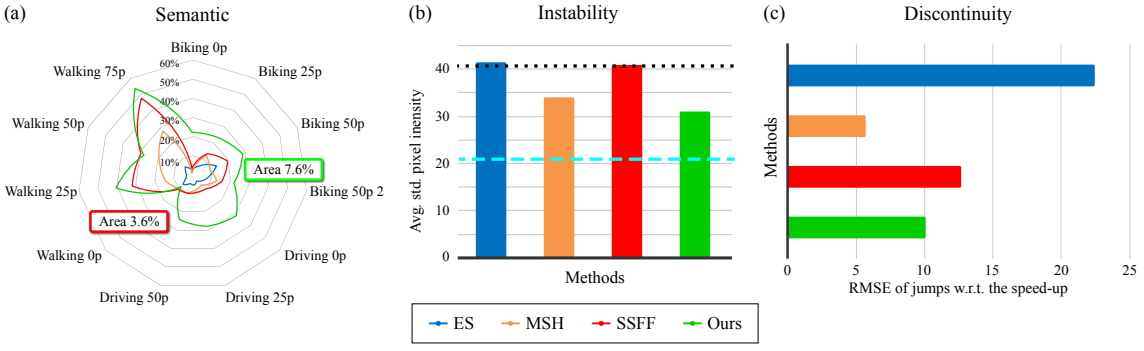

Figure 3. Quantitative analysis of the proposed method against the competitors.

methodology runs minimum reconstruction to frame sampling in an analytical form. The execution time of SSFF grows exponentially with the video length, while the growth in the number of frames in the input video did not influence the efficiency of our method.

We run an ablation study showing the effects of the weighting formulation and the SFT step, to investigate the usage of deep-features to represent the video frames, and to demonstrate the benefits of the LLC over Lasso and Orthogonal Matching Pursuit.

4. Conclusion

In this Ph.D. thesis, we tackled the challenging task of creating Semantic Hyperlapse for an FPV through a sparse coding-based framework composed of the adaptive frame sampling, Smooth Frame Transition, and Fill Gap steps. Experimental evaluation showed that our hyperlapse videos kept the double of semantic information, were smoother, and presented fewer visual discontinuities when compared with the state-of-the-art methods. Moreover, the related improvement did not affect the running time of the frame sampling process.

References

del Molino, A. G., Tan, C., Lim, J. H., and Tan, A. H. (2017). Summarization of Egocentric Videos: A Comprehensive Survey. IEEE Trans. Human-Machine Syst., 47(1):65–76.

Joshi, N., Kienzle, W., Toelle, M., Uyttendaele, M., and Cohen, M. F. (2015). Real-time hyperlapse creation via optimal frame selection. ACM Trans. Graph., 34(4):63:1–63:9.

Kopf, J., Cohen, M. F., and Szeliski, R. (2014). First-person hyper-lapse videos. ACM Trans. Graph., 33(4):78:1–78:10.

Lai, W. S., Huang, Y., Joshi, N., Buehler, C., Yang, M. H., and Kang, S. B. (2018). Semantic-driven generation of hyperlapse from 360° video. IEEE Trans. Visualization and Computer Graphics, 24(9):2610–2621.

Lan, S., Panda, R., Zhu, Q., and Roy-Chowdhury, A. K. (2018). FFNet: Video fast-forwarding via reinforcement learning. In Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pages 6771–6780, Salt Lake City, USA.

Poleg, Y., Halperin, T., Arora, C., and Peleg, S. (2015). Egosampling: Fast-forward and stereo for egocentric videos. In Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pages 4768–4776, Boston, USA.

Ramos, W. L. S., Silva, M. M., Campos, M. F. M., and Nascimento, E. R. (2016). Fast-forward video based on semantic extraction. In Proc. IEEE Int. Conf. Image Process. (ICIP), pages 3334–3338, Phoenix, USA.

Silva, M. M., Ramos, W. L. S., Ferreira, J. P. K., Campos, M. F. M., and Nascimento, E. R. (2016). Towards semantic fast-forward and stabilized egocentric videos. In Proc. Europ. Conf. Comput. Vis. Workshops (ECCVW), pages 557–571, Amsterdam, NLD.

Traffic-Inquiries (2018). Cisco visual networking index: Forecast and methodology, 2017-2022. Technical Report 1543280537836565, CISCO.

Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., and Gong, Y. (2010). Locality-constrained linear coding for image classification. In Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pages 3360–3367, San Francisco, USA.