[Supplementary Material]
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

In this Supplementary Material, we provide additional information about implementation details, experiments, and the proposed dataset. The document is organized as follows:

- ◇ Section 1: implementation details of the semantic criterion, the estimation of the weights and the Cumulative Displacement Curves (CDC).
- ◇ Section 2: individual results for each experiment performed to create the Figure 4 in the main paper.
- ◇ Section 3: details about the proposed dataset, *i.e.* the type of sensors used to capture the multi-modal data, also annotations and info available for the videos.

1. Implementation details

Semantic criterion. In the main paper, we performed the quantitative analysis measuring the Semantic, Instability, and Speed-up achieved in the created fast-forward video. In this Supplementary Material, we detail the implementation of computing the semantic criterion.

Let $S(f)$ be the semantic content of a frame f, defined by Equation 1:

$$S(f) = \sum_{r \in R_f} c_r \cdot s_r \cdot l_r, \tag{1}$$

where r is a Region of Interest (ROI) in the set of regions R_f returned by the classifier used to indicate the semantic content of the frame f. The term c_r is the classifier confidence about the region r, s_r is the size of the ROI delimiting the semantic content, and l_r is the locality term indicating how close is the semantic to the recorded central view.

Thus, regions r with the highest semantic scores will have a higher confidence assigned by the classifier, a bigger area, and located in the central part of the image. The reader is referred to [2] for more details about the terms.

We calculate the Semantic Value (SV) of a fast-forward video defined in [2] as follows:

$$SV = \frac{\sum_{i=1}^{m_c} S(f_i)}{\sum_{i=1}^{m_r} S(t_i)}, \tag{2}$$

where m_c is the number of frames in the fast-forwarded video, m_r is the number of frames needed to create a fast-forward video with the required speed-up, f_i stands for the i-th frame of the fast-forward video, and t_i stands for the top-i ranked frame of the original video regarding the semantic content.

Weights estimation. To estimate the weights used in the matrix W (Equation 2 in the main paper), we first identify the video segments with sharp camera movements. To identify these segments we estimate the optical flow in a 5×5 grid window between all consecutive frames by applying the sparse optical flow proposed in the work of Poleg *et al.* [1]. The information of the horizontal displacements is used to create the Cumulative Displacement Curves (CDC) [1].

Let $C' \in \mathbb{R}^{25 \times n}$ be the derivate w.r.t. time of the CDC. A frame f_i is considered into a sharp camera motion sequence if $C'_j(i) > 0 \; \forall \; j \in C'$ or $C'_j(i) < 0 \; \forall \; j \in C'$, where j is the window grid index.

2. Individual experiment results

In this Section, we present the individual values for the experimental evaluation performed in the main paper.

Table 1 shows the values used to create the plots in Figure 4 of the main text regarding the comparison between the proposed Weighted Sparse Sampling and the state-of-the-art methods.

Table 1. Results of the comparison between our weighted sparse sampling methodology and the state-of-the-art methods concerning to Semantic Evaluation (a), Instability (b) and Speed-up (c) achieved in the final fast-forward video.

(a) Semantic Evaluation - *Higher is better.*

Videos	ES	MSH	FFSE	MIFF	Ours
Biking 0p	3.6%	2.1%	4.5%	16.4%	**24.6%**
Biking 25p	6.8%	13.1%	13.9%	**26.4%**	20.4%
Biking 50p	13.5%	9.6%	19.5%	25.2%	**26.3%**
Biking 50p 2	6.9%	12.4%	14.9%	**18.7%**	18.1%
Driving 0p	3.0%	6.1%	9.1%	10.7%	**30.0%**
Driving 25p	4.1%	6.6%	7.9%	**31.9%**	24.7%
Driving 50p	3.3%	8.2%	8.3%	**24.9%**	19.0%
Walking 0p	5.7%	8.6%	9.2%	**14.8%**	7.5%
Walking 25p	2.8%	13.2%	30.2%	**43.7%**	36.7%
Walking 50p	3.0%	18.0%	**28.2%**	23.7%	25.2%
Walking 75p	6.1%	27.2%	47.0%	**56.9%**	49.4%
Area	*0.3%*	*1.3%*	*3.6%*	*6.8%*	***6.9%***

(b) Instability - *Lower is better.*

	Baselines		Methods				
Videos	Original	Naïve	ES	MSH	FFSE	MIFF	Ours
Biking 0p	15.9	29.3	31.5	**22.4**	30.8	27.2	23.4
Biking 25p	35.8	54.6	55.4	**47.6**	52.7	50.1	48.9
Biking 50p	21.6	37.1	38.0	30.6	34.9	33.0	**29.0**
Biking 50p 2	19.5	32.0	31.2	26.4	30.5	27.3	**25.8**
Driving 0p	24.5	49.3	50.2	**41.4**	55.0	52.2	43.9
Driving 25p	21.8	44.4	44.2	37.0	47.4	37.4	**34.2**
Driving 50p	23.0	43.7	45.9	**35.7**	45.7	38.5	**35.7**
Walking 0p	16.6	37.0	36.3	**32.7**	39.2	35.1	36.9
Walking 25p	17.5	38.8	38.3	34.2	37.9	**31.1**	33.3
Walking 50p	18.3	39.9	40.6	**31.7**	38.2	33.9	34.7
Walking 75p	19.2	40.4	44.0	34.8	37.9	**32.8**	33.0
Mean	*21.3*	*40.6*	*41.4*	***34.0***	*40.9*	*36.2*	*34.4*

(c) Speed-up - *Better closer to zero.*

Videos	ES	MSH	FFSE	MIFF	Ours
Biking 0p	14.0	**0.1**	7.9	2.6	-0.2
Biking 25p	1.4	-1.6	1.0	**0.0**	-0.6
Biking 50p	4.0	0.7	1.6	**0.0**	-0.2
Biking 50p 2	8.1	-1.6	2.2	**-0.1**	-0.3
Driving 0p	22.2	**0.0**	6.0	3.8	-0.4
Driving 25p	15.9	0.4	4.5	**0.1**	-2.4
Driving 50p	16.1	1.2	6.0	**0.0**	-1.6
Walking 0p	4.2	-2.6	3.2	**0.0**	**0.00**
Walking 25p	3.3	-1.6	1.6	**-0.2**	-0.9
Walking 50p	14.3	-2.3	0.5	-0.1	**-0.0**
Walking 75p	17.2	-0.8	-1.6	-1.7	**-0.0**
Mean Biking	*6.9*	*-0.6*	*3.2*	*0.6*	***-0.3***
Mean Driving	*18.1*	***0.6***	*5.5*	*1.3*	*-1.4*
Mean Walking	*9.8*	*-1.9*	*0.9*	*-0.5*	***-0.3***

3. Multi-modal Dataset

In addition to the new method for fast-forward first-person videos, this work also presents a new 80-hour Dataset of Multimodal Semantic Egocentric Videos (DoMSEV)[1]. Both videos and frames were annotated as presented in Table 3. A group of 8 persons recorded a total of 73 videos in a wide range of cameras, sensors, mounting setups, activities, illumination, weather conditions, and purposes.

To encourage the research about semantic definitions and personalized retrieval, we added the personal information and general preferences of each recorder. The recorder general preferences $p \in \mathbb{R}^{120}$ was defined as being a feature vector, where the entries are the 80 YOLO [3] classifier classes and the 48 concepts defined in the work of Shargi *et al.* [4]. We removed the duplicates due to the intersection between the YOLO classes and Shargi concepts. Each feature p_i has embedded a level of interest from 1 up to 10, indicating the recorder's interest in the respective concept or object.

Also, each video was annotated with respect to the frequency that the recorder used to perform the activity in the video. Additionally, each frame of all videos has a label indicating the type for the scene where the images were taken, the action performed, and the attention of the recorder along with the action. The "Frame Attention" indicates if the user was paying attention or interacting with some component of the scene.

Sensors. The dataset is composed of different types of information: GPS, Inertial Measurement Unit (IMU), depth, temperature, and RGB. For GPS and temperature information, we used the GoPro® Hero™ 5 built-in sensors. For IMU, we used either the inertial information provided by the Hero built-in sensor or the external LORD MicroStrain®3DM-GX3®-25 sensor for videos recorded with depth information. The videos with the similar names in the column "Videos", *e.g.*, Academic_Life_13 and Academic_Life_13_c, refer to videos recorded by cameras with different mounting but recording the same activity. The videos with the suffix "c" were recorded by the Intel® RealSense™ R200 using the head mounting support synchronized with the GoPro 5 mounted in the chest.

For videos with depth information, we built a setup using a 3D printer to attach a computer depth camera with an external IMU sensor (both were mounted in a helmet). The depth sensor used was an Intel® RealSense™ R200. We used a Robotic Operational System (ROS) package[2] to gather the sensor data. The inertial data was measured using the LORD MicroStrain®3DM-GX3®-25 sensor, and

[1] www.verlab.dcc.ufmg.br/semantic-hyperlapse/cvpr2018-dataset/

[2] wiki.ros.org/RealSense

Table 2. Annotated information for videos and frames. The symbol ○ indicates the possible values for the respective annotation.

Video	Camera	Resolution FPS Field of View Stabilization
	Sensors	GPS IMU Depth
	Recorder	ID Height Weight Age Gender Preferences
	Camera Mounting	○ *Head* ○ *Helmet* ○ *Chest* ○ *Shoulder* ○ *Hand*
	Activity Frequency	○ *Every Day* ○ *Often* ○ *Sometimes* ○ *Rarely* ○ *First time*
Frame	Scene	○ *Indoor* ○ *Nature* ○ *Crowed environment* ○ *Urban*
	Action	○ *Walking* ○ *Running* ○ *Standing* ○ *Biking* ○ *Driving* ○ *Playing* ○ *Cooking* ○ *Eating* ○ *Observing* ○ *In conversation* ○ *Browsing* ○ *Shopping*
	Attention	○ *None* ○ *Paying attention* ○ *Interacting*

the data was also gathered using a ROS package[3]. Finally, we used the ROS Bag System to record both sensor data in ROS environment, ensuring time synchronization between RGB, depth, and inertial data.

References

[1] Y. Poleg, C. Arora, and S. Peleg. Temporal segmentation of egocentric videos. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2537–2544, Columbus, USA, June 2014. 1

[2] W. L. S. Ramos, M. M. Silva, M. F. M. Campos, and E. R. Nascimento. Fast-forward video based on semantic extraction. In *The IEEE International Conference on Image Processing (ICIP)*, pages 3334–3338, Phoenix, USA, Sept 2016. 1

[3] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, Las Vegas, USA, June 2016. 2

[4] A. Sharghi, J. S. Laurel, and B. Gong. Query-focused video summarization: Dataset, evaluation, and a memory network based approach. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2127–2136, Honolulu, USA, July 2017. 2

[3] wiki.ros.org/microstrain_3dmgx2_imu