electronics

IMPACT FACTOR 2.690

CITESCORE 3.7

Article

# Summarization of Videos with the Signature Transform

J. de Curtò, I. de Zarzà, Gemma Roig and Carlos T. Calafate

Special Issue

Advanced Technologies for Image/Video Quality Assessment

Edited by

Dr. Shaode Yu and Dr. Dingquan Li

MDPI

https://doi.org/10.3390/electronics12071735

*electronics*

MDPI

*Article*

# Summarization of Videos with the Signature Transform

**J. de Curtò [1,2,3,4,*], I. de Zarzà [1,2,3,4], Gemma Roig [3] and Carlos T. Calafate [2]**

[1] Centre for Intelligent Multidimensional Data Analysis, HK Science Park, Shatin, Hong Kong; dezarza@em.uni-frankfurt.de
[2] Departamento de Informática de Sistemas y Computadores, Universitat Politècnica de València, 46022 València, Spain; calafate@disca.upv.es
[3] Informatik und Mathematik, GOETHE-University Frankfurt am Main, 60323 Frankfurt am Main, Germany; roig@cs.uni-frankfurt.de
[4] Estudis d'Informàtica, Multimèdia i Telecomunicació, Universitat Oberta de Catalunya, 08018 Barcelona, Spain
* Correspondence: decurto@em.uni-frankfurt.de

**Abstract:** This manuscript presents a new benchmark for assessing the quality of visual summaries without the need for human annotators. It is based on the Signature Transform, specifically focusing on the RMSE and the MAE Signature and Log-Signature metrics, and builds upon the assumption that uniform random sampling can offer accurate summarization capabilities. We provide a new dataset comprising videos from Youtube and their corresponding automatic audio transcriptions. Firstly, we introduce a preliminary baseline for automatic video summarization, which has at its core a Vision Transformer, an image–text model pre-trained with Contrastive Language–Image Pre-training (CLIP), as well as a module of object detection. Following that, we propose an accurate technique grounded in the harmonic components captured by the Signature Transform, which delivers compelling accuracy. The analytical measures are extensively evaluated, and we conclude that they strongly correlate with the notion of a good summary.

**Keywords:** video summarization; large language models; visual language models; CLIP; signature transform



**Citation:** de Curtò, J.; de Zarzà, I.; Roig, G.; Calafate, C.T. Summarization of Videos with the Signature Transform. *Electronics* **2023**, *12*, 1735. https://doi.org/10.3390/electronics12071735

Academic Editor: Stefanos Kollias

Received: 16 March 2023
Revised: 30 March 2023
Accepted: 2 April 2023
Published: 5 April 2023



## 1. Introduction and Problem Statement

Video data have become ubiquitous, from content creation to the animation industry. The ability to summarize the information present in large quantities of data is a central problem in many applications, particularly when there is a need to reduce the amount of information transmitted and to swiftly assimilate visual contents. Video summarization [1–7] has been extensively studied in Computer Vision, using both handcrafted methods [8] and learning techniques [9,10]. These approaches traditionally use feature extraction on keyframes to formulate an adequate summary.

Recent advances in Deep Neural Networks (DNN) [11–13] have spurred progress across various scientific fields [14–21]. In the realm of video summarization, two prominent approaches have emerged: LSTM- and RNN-based models [22–24]. These models have demonstrated considerable success in developing effective systems for video summarization. Additionally, numerous other learning techniques have been employed to address this challenge [25–28].

In this study, we introduce a novel concatenation of models for video summarization, capitalizing on advancements in Visual Language Models (VLM) [29,30]. Our approach combines zero-shot text-conditioned object detection with automatic text video annotations, resulting in an initial summarization method that captures the most critical information within the visual sequence.

Metrics to assess the performance of such techniques have usually relied on a human in the loop, using services such as Amazon Mechanical Turk (AMT) to provide annotated

summaries for comparison. There have been attempts to introduce quantitative measures to address this problem, the most common being the F1-score, but these measures need human annotators and have shown that many state-of-the-art methodologies perform worse than mere uniform random sampling [31].

However, in this work, we go beyond the current state of the art and introduce a set of metrics based on the Signature Transform [32,33], a rough equivalent to the Fourier Transform that takes order and area into account and that contrasts the spectrum of the original video with the spectrum of the generated summary to provide a measurable score. We then propose an accurate state-of-the-art baseline based on the Signature Transform to accomplish the task. Thorough evaluations are provided, where we can see that the methodologies provide accurate video summaries, and that the technique based on the Signature Transform achieves summarization capabilities superior to the state of the art. Indeed, the temporal content present in a video timeline makes the Signature Transform an ideal candidate to assess the quality of generated summaries where a video stream is treated as a path.

Section 2 gives a primer on the Signature Transform to bring forth in Section 2.1 a set of metrics to assess the quality of visual summaries by considering the harmonic components of the signal. The metrics are then used to put forward an accurate baseline for video summarization in Section 2.2. In the following section, we introduce the concept of Foundation Models, which serves to propose a preliminary technique for the summarization of videos. Thorough experiments are conducted in Section 4, with emphasis on the newly introduced dataset and the set of measures. Section 4.1 gives an assessment of the metrics in comparison to human annotators, whereas Section 4.2 evaluates the performance of the baselines based on the Signature Transform against another technique. Finally, Section 5 delivers conclusions, addresses the limitations of the methodology, and discusses further work.

## 2. Signature Transform

The Signature Transform [34–38] is roughly equivalent to the Fourier Transform; instead of extracting information concerning frequency, it extracts information about the order and area. However, the Signature Transform differs from the Fourier Transform in that it utilizes the space of functions of paths, a more general case than the basis of the space of paths found in the Fourier Transform.

Following the work in [34], the truncated signature of order $N$ of the path $\mathbf{x}$ is defined as a collection of coordinate iterated integrals

$$\mathrm{S}^N(\mathbf{x}) = \left( \left( \int \cdots \int_{0<t_1<\cdots<t_a<1} \prod_{c=1}^{a} \frac{\mathrm{d}f_{z_c}}{\mathrm{d}t}(t_c)\mathrm{d}t_1 \cdots \mathrm{d}t_a \right)_{1\le z_1,\ldots,z_a\le d} \right)_{1\le a\le N}. \tag{1}$$

Here, $\mathbf{x} = (x_1, \ldots, x_n)$, where $x_z \in \mathbb{R}^d$. Let $f = (f_1, \ldots, f_d) \colon [0,1] \to \mathbb{R}^d$ be continuous, such that $f(\frac{z-1}{n-1}) = x_z$, and linear in the intervals in between.

### *2.1. RMSE and MAE Signature and Log-Signature*

The F1-score between a summary and the ground truth of annotated data has been the widely accepted measure of choice for the task of video summarization. However, recent approaches highlighted the need to come up with metrics that can capture the underlying nature of the information present in the video [31].

In this work, we leverage tools from harmonic analysis by the use of the Signature Transform to introduce a set of measures, namely, Signature and Log-Signature Root Mean Squared Error (denoted from now on as RMSE Signature and Log-Signature), that can shed light on what a good summary is and serve as powerful tools to analytically quantize the information present in the selected frames.

As introduced in [32] in the context of GAN convergence assessment, the RMSE and MAE Signature and Log-Signature can be defined as follows, particularized for the application under study:

**Definition 1.** *Given n components of the element-wise mean of the signatures* $\{\tilde{y}^{(c)}\}_{c=1}^{n} \subseteq T(\mathbb{R}^d)$ *from the target summary to the score, and the same number of components of the element-wise mean of the signatures* $\{\tilde{x}^{(c)}\}_{c=1}^{n} \subseteq T(\mathbb{R}^d)$ *from the original video subsampled at a given frame rate and uniformly chosen, we define the Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) as*

$$\mathrm{RMSE}\left(\left\{\tilde{x}^{(c)}\right\}_{c=1}^{n}, \left\{\tilde{y}^{(c)}\right\}_{c=1}^{n}\right) = \sqrt{\frac{1}{n}\sum_{c=1}^{n}\left(\tilde{y}^{(c)} - \tilde{x}^{(c)}\right)^2}, \tag{2}$$

*and*

$$\mathrm{MAE}\left(\left\{\tilde{x}^{(c)}\right\}_{c=1}^{n}, \left\{\tilde{y}^{(c)}\right\}_{c=1}^{n}\right) = \frac{1}{n}\sum_{c=1}^{n}\left|\tilde{y}^{(c)} - \tilde{x}^{(c)}\right|, \tag{3}$$

*respectively, where* $T(\mathbb{R}^d) = \prod_{c=0}^{\infty}\left(\mathbb{R}^d\right)^{\otimes c}$.

The case for Log-Signature is analogous.

For the task of video summarization, two approaches are given. In the case where the user has annotated summaries available, RMSE $(\bar{S}, \bar{S}_{target})$ is computed between an element-wise mean of the annotated summaries and the target summary to the score. If annotations are not available, a comparison against mean random uniform samples is performed, $\bar{S}$, and mean score and standard deviation are provided. Given the properties of the Signature Transform, the measure takes into consideration the harmonic components that are intrinsic to the video under study and that should be preserved once the video is shortened to produce a summary. As a matter of fact, both approaches should lead to the same conclusions, as the harmonic components present in the annotated summaries and the ones present in average in the random uniform samples should also agree. A confidence interval of the scores can be provided for a given measure by analyzing the distances in the RMSEs of annotated summaries or random uniform samples, RMSE $(\bar{S}_a, \bar{S}_c)$.

When comparing against random uniform samples, the underlying assumption is as follows: we assume that good visual summaries capturing all or most of the harmonic components present in the visual cues will achieve a lower standard deviation. In contrast, summaries that lack support for the most important components will yield higher values. For a qualitative example, see Figure 1. With these ideas in mind, we can discern techniques that likely generate consistent summaries from those that fail to convey the most critical information. Moreover, the study of random sample intervals provides a set of tolerances for considering a given summary adequate for the task, meaning it is comparable to or better than uniform sampling of the interval at capturing harmonic components. Consequently, the proposed measures allow for a percentage score representing the number of times a given methodology outperforms random sampling by containing the same or more harmonic components present in the spectrum.

### *2.2. Summarization of Videos with RMSE Signature*

Proposing a methodology based on the Signature Transform to select proper frames for a visual summary can be effectuated as follows: Given a uniform random sample of the video to summarize, we can compare it against subsequent random summaries using RMSE $(\bar{S}, \bar{S}_*)$. We can repeat this procedure $n$ times and choose, as a good candidate, the minimum according to the standard deviation. Using this methodology, we can also repeat the procedure for a range of selected summary lengths, which will give us a set of good candidates, among which we will choose the candidate with the minimum standard deviation. This will provide us with an estimate of the most suitable length. It is important to note that this baseline is completely unsupervised in the sense that no annotations are used, only the metrics based on the Signature Transform. We rely on the fact that,

in general, uniform random samples provide relatively accurate summaries, and among those, we choose the ones that are best according to std(RMSE $(\bar{S}, \bar{S}_*)$), which we denote as RMSE $(\bar{S}, \bar{S}_{u_{min}})|_n$. This will grant us competitive uniform random summaries according to the given measures to use as a baseline for comparison against other methodologies, and with which we can estimate an appropriate summary length to use in those cases.

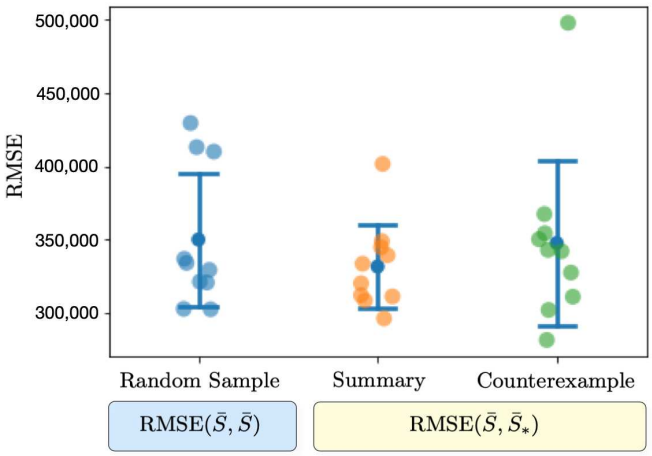



**Figure 1.** Conceptual plot with RMSE $(\bar{S}, \bar{S})$ and RMSE $(\bar{S}, \bar{S}_*)$ standard deviation and mean for two given summaries (our method and a counterexample) of 12 frames using a randomly picked video from Youtube to illustrate how to select a proper summary according to the proposed metric.

Below, we provide a description of the entities involved in the computation of the metrics and the proposed baselines based on the Signature Transform:

- $\bar{S}_*$: Element-wise mean Signature Transform of the target summary to the score of the corresponding video;
- $\bar{S}$: Element-wise mean Signature Transform of a uniform random sample of the corresponding video;
- RMSE $(\bar{S}, \bar{S}_*)$: Root mean squared error between the spectra of $\bar{S}$ and $\bar{S}_*$ with the same summary length. For the computation of standard deviation and mean, this value is calculated ten times, changing $\bar{S}$;
- RMSE $(\bar{S}, \bar{S})$: Root mean squared error between the spectra of $\bar{S}$ and $\bar{S}$ with the same summary length. For computation of standard deviation and mean, this value is calculated ten times, changing both $\bar{S}$ each time;
- RMSE $(\bar{S}, \bar{S}_{u_{min}})|_n$: Baseline based on the Signature Transform. It corresponds to RMSE $(\bar{S}, \bar{S}_*)$, where $\bar{S}_*$ is, in this case, a fixed uniform random sample denoted as $\bar{S}_u$. We repeat this procedure $n$ times and choose the minimum candidate according to standard deviation, $\bar{S}_{u_{min}}$, to propose as a summary;
- $std()$: Standard deviation.

## 3. Summarization of Videos via Text-Conditioned Object Detection

Large Language Models (LLM) [39–42] and VLMs [43] have emerged as indispensable resources for characterizing complex tasks and bestowing intelligent systems with the capacity to interact with humans in unprecedented ways. These models, also called Foundation Models [44–46], excel in a wide variety of tasks, such as robotics manipulation [47–49], and can be integrated with other modules to perform robustly in highly complex situations such as navigation and guidance [50,51]. One fundamental module is the Vision Transformer [52].

We introduce a simple yet effective technique aimed at generating video summaries that accurately describe the information contained within video streams, while also proposing new measures for the task of the summarization of videos. These measures will prove

useful not only when text transcriptions are available, but also in more general cases in which we seek to describe the quality of a video summary.

Building on the text-conditioned object detection using Vision Transformers, as recently proposed in [53], we enhance the summarization task by leveraging the automated text transcriptions found in video platforms. We utilize a module of noun extraction employing NLP techniques [54], which is subsequently processed to account for the most frequent nouns. These nouns serve as input queries for text-conditioned object searches in frames. Frames containing the queries are selected for the video summary; see Figure 2 for a detailed depiction of the methodology.

In this manuscript, we initially present a baseline leveraging text-conditioned object detection, specifically Contrastive Language–Image Pre-training (CLIP) [43]. To assess this approach, we employ a recently introduced metric based on the Signature Transform, which accurately gauges summary quality compared to a uniform random sample. Our preliminary baseline effectively demonstrates the competitiveness of uniform random sampling [31]. Consequently, we introduce a technique utilizing prior knowledge of the Signature, specifically the element-wise mean comparison of the spectrum, to generate highly accurate random uniform samples for summarization. The Signature Transform allows for a design featuring an inherent link between the methodology, metric, and baseline. We first present a method for evaluation, followed by a set of metrics for assessment, and ultimately, we propose a state-of-the-art baseline that can function as an independent technique.

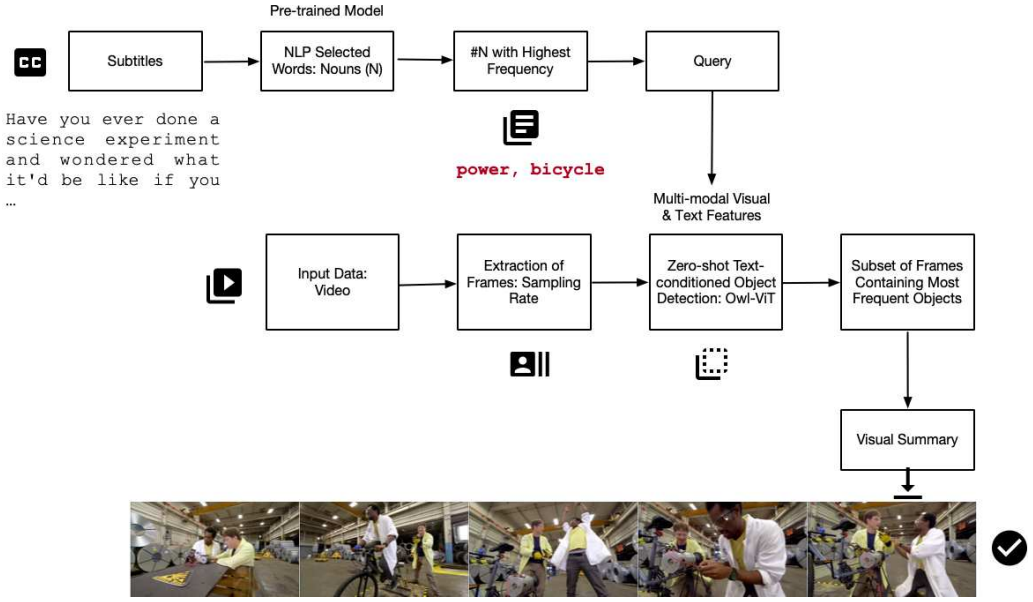



**Figure 2.** Video Summarization via Zero-shot Text-conditioned Object Detection.

## 4. Experiments: Dataset and Metrics

A dataset consisting of 28 videos about science experiments was sourced from Youtube, along with their automatic audio transcriptions, to evaluate the methodology and the proposed metrics. Table 1 provides a detailed description of the collected data and computed metrics, Figure 3 shows the distribution of selected frames using text-conditioned object detection over a subset of videos and the baselines based on the Signature Transform, Figure 4 depicts a visual comparison between methodologies, and Figures 5 and 6 visually elucidate the RMSE distribution for each video with mean and standard deviation.

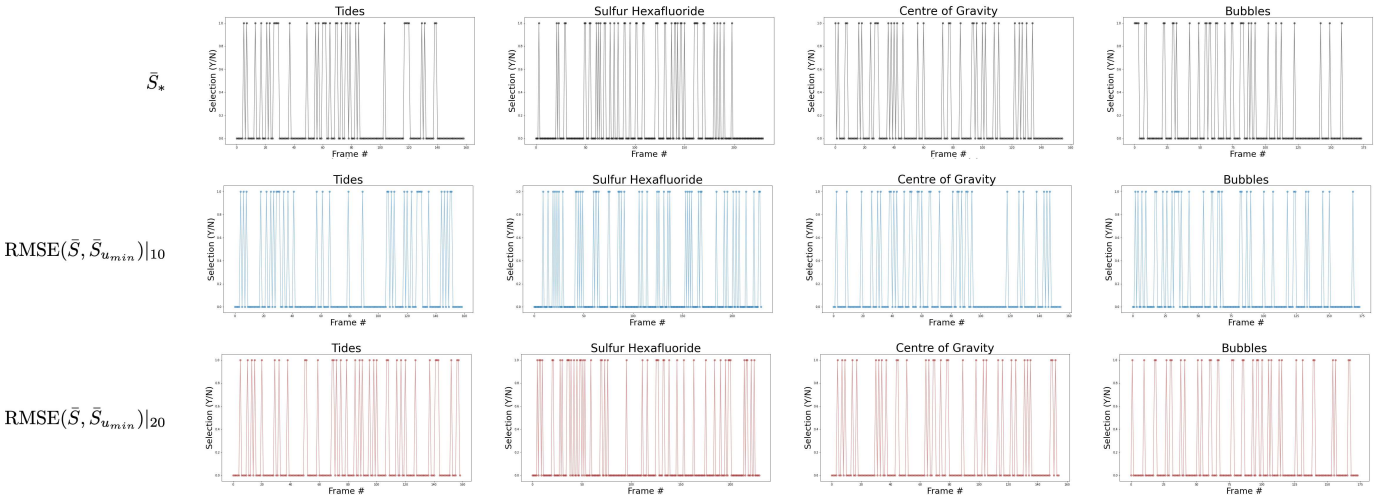



**Figure 3.** Comparison of distribution of selected frames for a subset of videos (Tides, Sulfur Hexafluoride, Centre of Gravity and Bubbles) using the method based on text-conditioned object detection and the baselines using the Signature Transform.

The dataset consists of science videos covering a wide range of experiments on several topics of interest; it has an average number of 264 frames per video (sampling rate $\frac{1}{4}$ s) and an average duration of 17 min 30 s.

Figure 3 depicts the selected frames when using our methodology for a subset of videos in the dataset. The selection coincides with the trigger of the zero-shot text-conditioned object detector by the 20 most frequent word code-phrase queries, which chooses a subset of the methodology that best explains the main factors of the argument. A comparison with the baselines based on the Signature Transform with 10 and 20 points is delivered.

In all experiments that involve the computation of the Signature Transform, we use the parameters proposed in [32] that were originally used to assess synthetic distributions generated with GANs; specifically, we employ truncated signatures of order 3 with a resized image size of $64 \times 64$ in grayscale.

RMSE $(\bar{S}, \bar{S}_*)$ computes the element-wise mean of the signatures of both the target summary to the score and a random uniform sample with the same number of frames, comparing their spectra with the use of the RMSE. Likewise, RMSE $(\bar{S}, \bar{S})$ computes the same measure between two random uniform samples with the same number of frames. The standard deviation of both results is compared to assess the quality of the summarized video concerning the present harmonic components. The preliminary technique based on text-conditioned object detection (see Table 1) achieves a zero-shot of 50% positive cases when compared against std (RMSE $(\bar{S}, \bar{S})$). The number of frames selected by the methodology is consistent, and it automatically selects on average 20% of the total number of frames.

In this paragraph, we discuss the baseline based on the Signature Transform (see Table 1) in terms of the RMSE $(\bar{S}, \bar{S}_{u_{min}})|_{10}$ and RMSE $(\bar{S}, \bar{S}_{u_{min}})|_{20}$. These techniques select a uniform random sample with minimum standard deviation in a set of 10 points and 20 points, respectively, and achieve 100% positive cases when compared to RMSE $(\bar{S}, \bar{S})$. Under the assumption that the summary can be approximated well by a random uniform sample, which holds true in many cases, the methodology finds a set of frames that maximizes the harmonic components relative to those present in the original video.

**Table 1.** Descriptive statistics with RMSE $(\bar{S}, \bar{S}_*)$ (target summary against random uniform sample) and RMSE $(\bar{S}, \bar{S})$ (random uniform sample against random uniform sample). RMSE $(\bar{S}, \bar{S}_{u_{min}})|_{10}$ and RMSE $(\bar{S}, \bar{S}_{u_{min}})|_{20}$ correspond to the baselines based on the Signature Transform using 10 and 20 random samples, respectively. Highlighted results in blue/brown correspond to values better than std (RMSE $(\bar{S}, \bar{S})$). Yellow values indicate when std (RMSE $(\bar{S}, \bar{S})$) is lower than std (RMSE $(\bar{S}, \bar{S}_*)$).

| Descriptive Statistics | | | Summary | RMSE $(\bar{S}, \bar{S}_*)$ | | RMSE $(\bar{S}, \bar{S})$ | | RMSE $(\bar{S}, \bar{S}_{u_{min}})|_{10}$ | | RMSE $(\bar{S}, \bar{S}_{u_{min}})|_{20}$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Video** | **# Frames** | **Length** | **# Frames (%)** | **Std** | **Mean** | **Std** | **Mean** | **Std** | **Mean** | **Std** | **Mean** |
| Tides | 159 | 10 m 29 s | 35 (22%) | 13,663 | 202,388 | 14,838 | 155,986 | 8859 | 157,455 | 7312 | 167,480 |
| Sulfur Hexafluoride | 230 | 15 m 12 s | 47 (20%) | 22,727 | 217,935 | 22,607 | 179,409 | 7194 | 161,995 | 7722 | 173,490 |
| Centre of Gravity | 155 | 10 m 14 s | 33 (21%) | 12,333 | 181,460 | 16,404 | 168,824 | 8481 | 160,779 | 12,416 | 175,971 |
| Bubbles | 174 | 11 m 30 s | 35 (20%) | 23,127 | 201,553 | 16,806 | 185,702 | 7461 | 194,993 | 5711 | 175,176 |
| Airplanes | 158 | 10 m 24 s | 22 (14%) | 19,964 | 215,688 | 23,591 | 231,539 | 8417 | 227,391 | 10,235 | 233,020 |
| Protons | 174 | 11 m 30 s | 25 (14%) | 29,853 | 252,224 | 20,186 | 262,434 | 12,835 | 251,907 | 11,542 | 250,512 |
| Hydrophobic | 168 | 11 m 06 s | 29 (17%) | 15,016 | 251,671 | 25,835 | 248,548 | 11,973 | 250,131 | 13,917 | 245,761 |
| States of Matter | 332 | 22 m 03 s | 78 (23%) | 16,249 | 156,408 | 9709 | 130,064 | 6630 | 115,454 | 5340 | 121,028 |
| Spool Racer | 332 | 22 m 02 s | 90 (27%) | 15,903 | 142,520 | 11,883 | 136,147 | 7054 | 137,621 | 8112 | 151,888 |
| Paper Airplane | 332 | 22 m 03 s | 29 (9%) | 20,642 | 235,639 | 11,829 | 221,220 | 5400 | 224,718 | 9385 | 177,448 |
| Loudest Sound | 332 | 22 m 01 s | 93 (28%) | 16,898 | 179,963 | 8304 | 148,885 | 7884 | 138,561 | 4355 | 147,016 |
| Lightning | 332 | 22 m 01 s | 70 (21%) | 15,237 | 169,338 | 21,862 | 162,849 | 9300 | 177,008 | 7494 | 153,797 |
| Light Challenge | 332 | 22 m 02 s | 82 (25%) | 12,566 | 152,488 | 10,546 | 126,117 | 5490 | 139,700 | 4874 | 129,044 |
| Hot Air Balloon | 332 | 22 m 01 s | 98 (30%) | 8620 | 150,366 | 5417 | 144,634 | 3516 | 137,141 | 4165 | 138,453 |
| Hoop Glider | 332 | 22 m 01 s | 82 (25%) | 6419 | 148,065 | 6752 | 132,544 | 4051 | 133,897 | 4966 | 133,894 |
| Drag Race | 332 | 22 m 03 s | 73 (22%) | 9384 | 135,228 | 8931 | 125,264 | 4375 | 122,615 | 4645 | 129,851 |
| All about Balance | 332 | 22 m 03 s | 59 (18%) | 14,023 | 182,063 | 14,238 | 182,179 | 7801 | 176,219 | 6914 | 167,727 |
| Air Pressure | 332 | 22 m 03 s | 65 (20%) | 10,123 | 166,342 | 18,314 | 151,664 | 6386 | 145,897 | 4602 | 148,232 |
| Friction and Momentum | 162 | 10 m 42 s | 28 (17%) | 18,754 | 217,403 | 22,443 | 218,203 | 13,348 | 202,288 | 12,238 | 205,680 |
| Electricity | 162 | 10 m 41 s | 30 (19%) | 24,376 | 298,238 | 22,885 | 279,820 | 16,889 | 268,932 | 10,263 | 270,619 |
| Catapult | 169 | 11 m 11 s | 27 (16%) | 26,413 | 271,643 | 31,265 | 214,727 | 15,158 | 203,290 | 10,222 | 188,008 |
| Carbonation and More | 165 | 10 m 53 s | 40 (24%) | 18,977 | 237,142 | 18,107 | 226,044 | 12,130 | 234,278 | 11,884 | 214,149 |
| Carbon Dioxide | 162 | 10 m 41 s | 38 (23%) | 25,862 | 245,415 | 18,806 | 217,270 | 13,838 | 207,828 | 7760 | 211,504 |
| Bridge | 164 | 10 m 51 s | 21 (13%) | 25,839 | 269,412 | 26,038 | 271,551 | 10,761 | 263,747 | 13,038 | 264,532 |
| Bread Experiment | 337 | 22 m 22 s | 59 (18%) | 15,099 | 189,086 | 8575 | 146,771 | 5542 | 153,224 | 5691 | 156,230 |
| Balloon Power | 337 | 22 m 22 s | 53 (16%) | 14,075 | 157,542 | 29,415 | 147,710 | 7741 | 128,920 | 7351 | 134,545 |
| Attraction and Forces | 654 | 43 m 30 s | 81 (12%) | 5955 | 107,097 | 7486 | 102,965 | 3701 | 96,266 | 2093 | 99,271 |
| Puzzles | 209 | 13 m 48 s | 46 (22%) | 11,258 | 185,502 | 19,012 | 196,762 | 14,620 | 199,556 | 14,622 | 197,064 |
| **Average** | 264 | 17 m 30 s | 52 (20%) | 14/28 (50%) | | | | 28/28 (100%) | | 28/28 (100%) | |

Figure 4 displays examples of summaries using the baseline based on the Signature Transform compared to the summaries using text-conditioned object detection. The figure allows for a visual comparison of the results obtained using RMSE $(\bar{S}, \bar{S}_{u_{min}})|_{10}$, RMSE $(\bar{S}, \bar{S}_{u_{min}})|_{20}$ and $\bar{S}_*$. The best summary among the three baselines according to the metric is highlighted (Table 1).

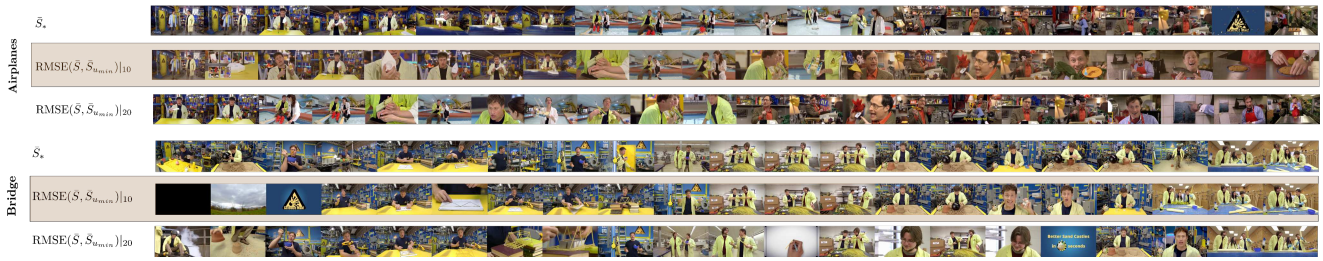



**Figure 4.** Summarization of videos using the baseline based on the Signature Transform in comparison to the summarization using text-conditioned object detection. RMSE $(\bar{S}, \bar{S}_{u_{min}})|_{10}$, RMSE $(\bar{S}, \bar{S}_{u_{min}})|_{20}$ and $\bar{S}_*$ summaries for two videos of the introduced dataset. The best summary among the three, according to the metric, is highlighted.

The selected frames are consistent and provide a good overall description of the original videos. Moreover, the metric based on the Signature Transform aligns well with our expectations of a high-quality summary, with better scores being assigned to summaries that effectively convey the content present in the original video.

Table 2 presents a qualitative analysis of the baseline based on the Signature Transform using 10 points, RMSE $(\bar{S}, \bar{S}_{u_{min}})|_{10}$ and RMSE $(\bar{S}, \bar{S})$ with a varying number of frames per summary. We observe that RMSE $(\bar{S}, \bar{S})$ reflects the variability of the harmonic compo-

nents present; that is, it is preferable to work with lengths for which the variability among summaries is low, according to the standard deviation. RMSE $(\bar{S}, \bar{S}_{u_{min}})|_{10}$ indicates the minimum standard deviation achieved in a set of 10 points, meaning that given a computational budget allowing us to select up to a specific number of frames, a good choice is to pick the length that yields the minimum RMSE $(\bar{S}, \bar{S}_{u_{min}})|_{10}$ with low variability, as per RMSE $(\bar{S}, \bar{S})$.

RMSE $(\bar{S}, \bar{S}_*)$ (Figure 5) and RMSE $(\bar{S}, \bar{S})$ (Figure 6) show the respective distribution of RMSE values (10 points) with the mean and standard deviation. Low standard deviations, in comparison with the random uniform sample counterparts, indicate good summarization capabilities.

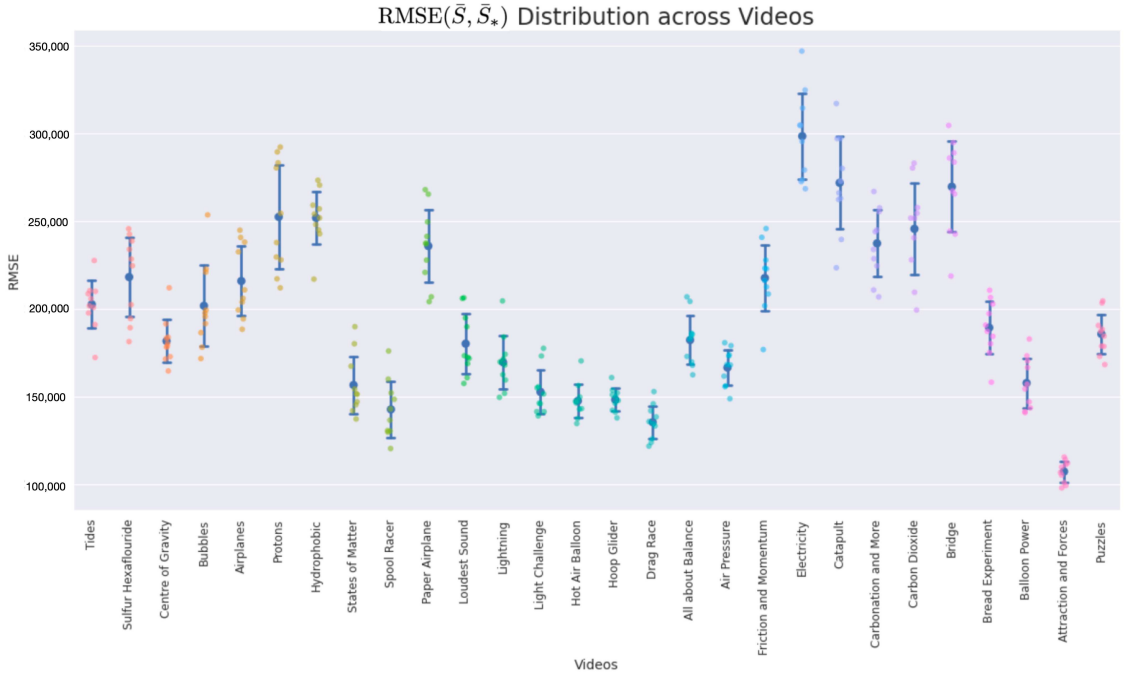



**Figure 5.** Plot with RMSE $(\bar{S}, \bar{S}_*)$ standard deviation and mean.

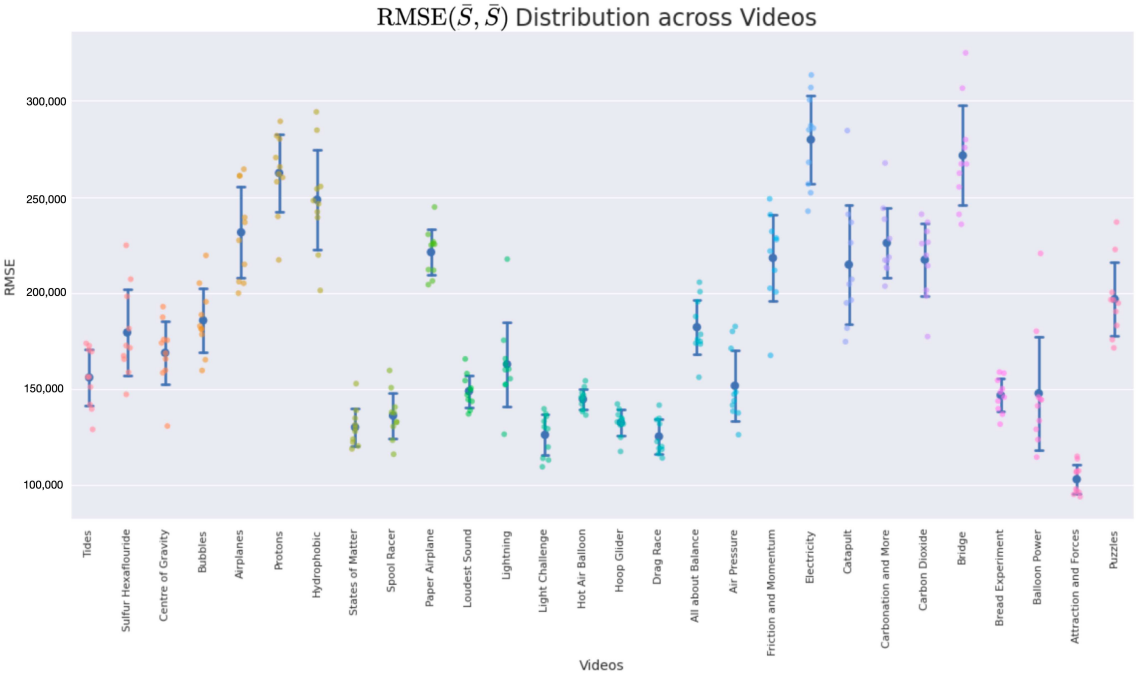
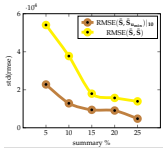
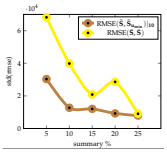
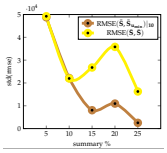
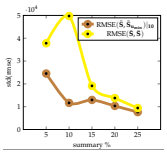



**Figure 6.** Plot with RMSE $(\bar{S}, \bar{S})$ standard deviation and mean.

**Table 2.** Descriptive statistics for a set of videos with varying numbers of frames per summary with RMSE $(\bar{S}, \bar{S}_{u_{min}})|_{10}$ (brown) and RMSE $(\bar{S}, \bar{S})$ (yellow).

| Dataset | | | RMSE $(\bar{S}, \bar{S}_{u_{min}})\|_{10}$ | | RMSE $(\bar{S}, \bar{S})$ | | Visualization |
|---|---|---|---|---|---|---|---|
| **Video** | **# Frames** | **Summary (%)** | **Std** | **Mean** | **Std** | **Mean** | **Plot (Std,Std)** |
| Tides | 159 | 8 (5%) | 22,786 | 422,026 | 54,067 | 390,483 |  |
| | | 16 (10%) | 12,851 | 254,984 | 37,713 | 263,881 | |
| | | 24 (15%) | 9423 | 202,925 | 17,935 | 224,797 | |
| | | 32 (20%) | 9074 | 183,933 | 15,700 | 186,621 | |
| | | 40 (25%) | 4782 | 158,183 | 13,903 | 159,452 | |
| Sulfur Hexafluoride | 230 | 12 (5%) | 30,325 | 452,134 | 68,212 | 362,061 |  |
| | | 23 (10%) | 12,701 | 281,425 | 39,872 | 246,967 | |
| | | 35 (15%) | 12,034 | 228,530 | 20,846 | 201,740 | |
| | | 46 (20%) | 9241 | 190,985 | 28,621 | 175,440 | |
| | | 58 (25%) | 7914 | 161,618 | 9021 | 152,310 | |
| Centre of Gravity | 155 | 8 (5%) | 48,787 | 406,502 | 49,234 | 369,648 |  |
| | | 16 (10%) | 22,163 | 252,841 | 21,974 | 276,366 | |
| | | 24 (15%) | 8050 | 212,893 | 26,776 | 229,959 | |
| | | 31 (20%) | 10,963 | 180,953 | 35,813 | 184,437 | |
| | | 39 (25%) | 2528 | 164,666 | 16,259 | 163,007 | |
| Bubbles | 174 | 9 (5%) | 24,538 | 401,406 | 37,816 | 397,470 |  |
| | | 18 (10%) | 11,669 | 272,430 | 49,740 | 276,152 | |
| | | 27 (15%) | 12,965 | 213,336 | 19,125 | 215,961 | |
| | | 35 (20%) | 10,331 | 190,639 | 13,792 | 183,984 | |
| | | 44 (25%) | 7625 | 173,009 | 9427 | 162,091 | |

### *4.1. Assessment of the Metrics*

The metrics have been rigorously evaluated using the dataset in [1], which consists of short videos sourced from Youtube, and includes 5 annotated summaries per video for a total of 20. Tables 3 and 4 report the results, using a one-frame-per-second sampling rate. In this case, the average number of times that the human annotator outperforms uniform random sampling according to the proposed metric, std (RMSE $(\bar{S}, \bar{S})$), is 87%. Several observations emerge from these findings:

- The proposed metrics demonstrate that human evaluators can perform above average during the task, effectively capturing the dominant harmonic frequencies present in the video.
- Another crucial aspect to emphasize is that the metrics are able to evaluate human annotators with fair criteria and identify which subjects are creating competitive summaries.
- Moreover, the observations from this study indicate that the metrics serve as a reliable proxy for evaluating summaries without the need for annotated data, as they correlate strongly with human annotations.

Figure 7 shows the mean and standard deviation for each human-annotated summary (user 1 to user 5) for the subset of 20 videos from [1], using a sampling rate of 1 frame per second. For each video, a visual inspection of the error plot bar for each annotated summary provides an accurate estimate of the quality of the annotation compared to other users. Specifically:

- Annotations with lower standard deviations offer a better harmonic representation of the overall video;
- Annotations with higher standard deviations suggest that important harmonic components are missing from the given summary;
- The metrics make it simple to identify annotated summaries that may need to be relabeled for improved accuracy.

Furthermore, these metrics remain consistent when applied to various sampling rates.

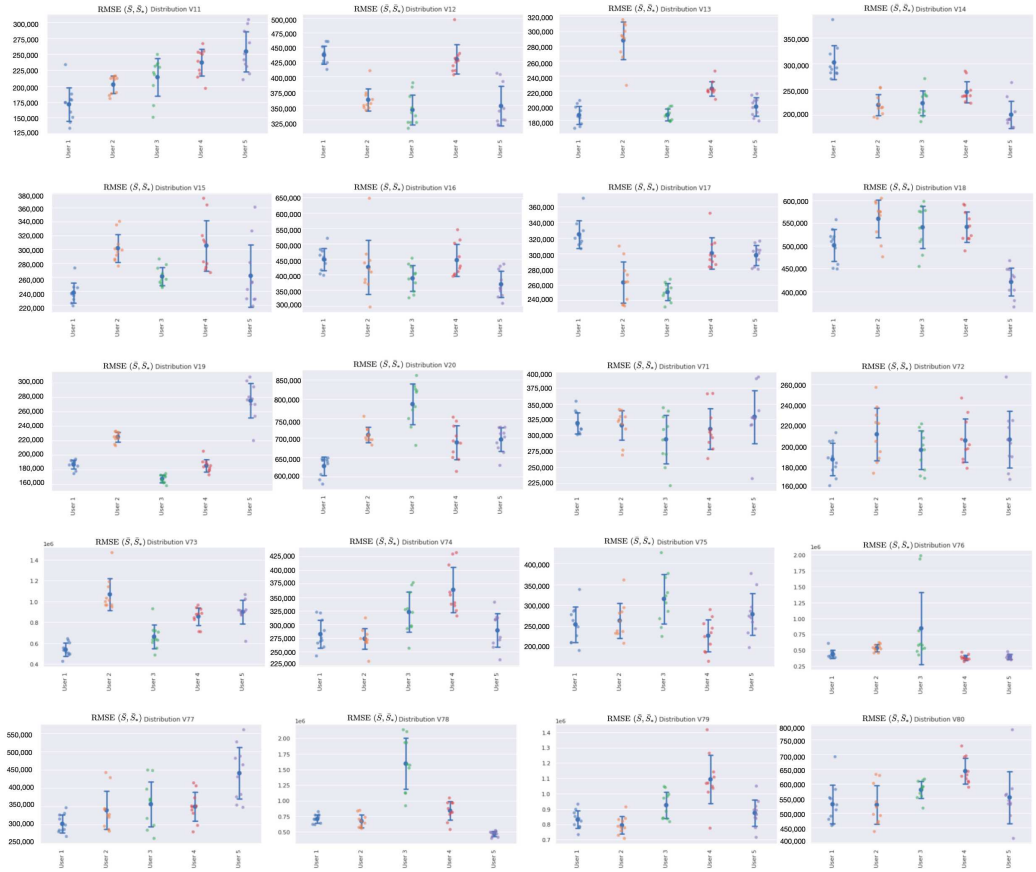


**Figure 7.** Error bar plot with mean and standard deviation for each human-annotated summary of the subset of 20 videos from [1]. Sampling rate: 1 frame per second.

That being said, there are several standard measures that are commonly used for video summarization, such as F1 score, precision, recall, and Mean Opinion Score (MOS). Each of these measures has its own strengths and weaknesses. Compared to these standard measures, the proposed benchmark based on the Signature Transform has several potential advantages. Here are a few reasons for this:

- Content based: the Signature Transform is a content-based approach that captures the salient features of the video data. This means that the proposed measure is not reliant on manual annotations or subjective human ratings, which can be time consuming and prone to biases.
- Robustness: the Signature Transform is a robust feature extraction technique that can handle different types of data, including videos with varying frame rates, resolutions, and durations. This means that the proposed measure can be applied to a wide range of video datasets without the need for pre-processing or normalization.
- Efficiency: the Signature Transform is a computationally efficient approach that can be applied to large-scale datasets. This means that the proposed measure can be used to evaluate the effectiveness of visual summaries quickly and accurately.
- Flexibility: the Signature Transform can be applied to different types of visual summaries, including keyframe-based and shot-based summaries. This means that the proposed measure can be used to evaluate different types of visual summaries and compare their effectiveness.

Overall, the proposed measure based on the Signature Transform has the potential to provide a more accurate and comprehensive assessment of the standard of visual summaries compared to the preceding measures used in video summarization.

**Table 3.** Descriptive statistics with RMSE $(\bar{S}, \bar{S}_*)$ (target summary against random uniform sample) and RMSE $(\bar{S}, \bar{S})$ (random uniform sample against random uniform sample). Lower is better. Sampling rate: 1 frame per second. Dataset in [1], videos from V11 to V20. Highlighted results in blue/yellow correspond to the lowest values, either std (RMSE $(\bar{S}, \bar{S}_*)$) or std (RMSE $(\bar{S}, \bar{S})$), respectively.

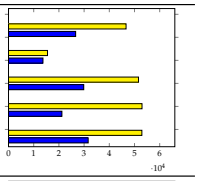
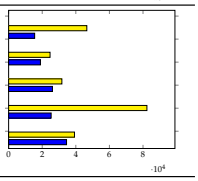
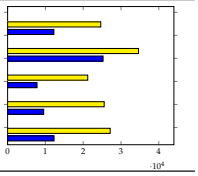
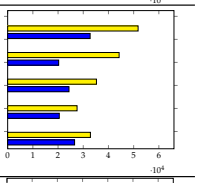
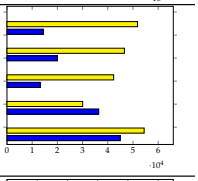
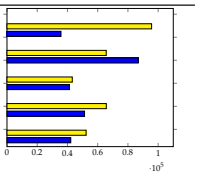
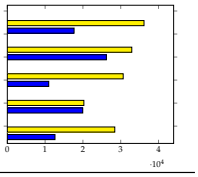
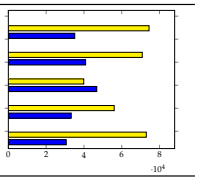
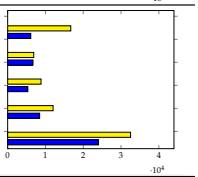
| Youtube, Dataset | | | | RMSE $(\bar{S}, \bar{S}_*)$ | | RMSE $(\bar{S}, \bar{S})$ | | Visualization |
|---|---|---|---|---|---|---|---|---|
| **Video** | **# Frames** | **User** | **# Frames User** | **Std** | **Mean** | **Std** | **Mean** | **Plot (Std,Std)** |
| V11 | 48 | 1 | 10 | 26,644 | 171,106 | 46,655 | 151,483 |  |
| | | 2 | 12 | 13,673 | 202,172 | 15,479 | 155,481 | |
| | | 3 | 10 | 29,857 | 213,880 | 51,590 | 182,327 | |
| | | 4 | 9 | 21,192 | 236,959 | 52,982 | 196,303 | |
| | | 5 | 8 | 31,627 | 254,336 | 52,925 | 193,520 | |
| V12 | 59 | 1 | 11 | 15,497 | 436,723 | 46,551 | 252,142 |  |
| | | 2 | 17 | 18,927 | 359,562 | 24,665 | 177,286 | |
| | | 3 | 15 | 26,071 | 342,161 | 31,703 | 180,066 | |
| | | 4 | 11 | 25,330 | 429,272 | 82,323 | 242,627 | |
| | | 5 | 14 | 34,479 | 348,834 | 39,199 | 188,417 | |
| V13 | 59 | 1 | 19 | 12,238 | 187,001 | 24,649 | 114,155 |  |
| | | 2 | 9 | 25,267 | 287,479 | 34,635 | 166,495 | |
| | | 3 | 18 | 7790 | 187,346 | 21,203 | 126,432 | |
| | | 4 | 14 | 9544 | 222,496 | 25,553 | 140,508 | |
| | | 5 | 18 | 12,298 | 198,349 | 27,138 | 124,386 | |
| V14 | 59 | 1 | 9 | 32,739 | 302,118 | 51,770 | 183,978 |  |
| | | 2 | 16 | 20,249 | 219,068 | 44,235 | 141,927 | |
| | | 3 | 17 | 24,345 | 222,559 | 35,235 | 113,806 | |
| | | 4 | 10 | 20,498 | 244,509 | 27,548 | 155,515 | |
| | | 5 | 16 | 26,561 | 200,139 | 32,840 | 143,384 | |
| V15 | 57 | 1 | 12 | 14,454 | 237,551 | 51,812 | 207,845 |  |
| | | 2 | 11 | 20,018 | 301,650 | 46,590 | 209,491 | |
| | | 3 | 13 | 13,192 | 261,014 | 42,337 | 171,810 | |
| | | 4 | 13 | 36,408 | 305,376 | 30,041 | 179,442 | |
| | | 5 | 14 | 44,931 | 261,859 | 54,428 | 180,145 | |
| V16 | 70 | 1 | 9 | 35,722 | 449,758 | 95,662 | 376,411 |  |
| | | 2 | 9 | 86,863 | 425,107 | 65,626 | 328,563 | |
| | | 3 | 12 | 41,260 | 388,869 | 43,186 | 340,133 | |
| | | 4 | 9 | 51,299 | 447,523 | 65,698 | 375,162 | |
| | | 5 | 13 | 42,200 | 369,517 | 52,316 | 302,677 | |
| V17 | 59 | 1 | 12 | 17,668 | 324,562 | 36,166 | 242,235 |  |
| | | 2 | 13 | 26,203 | 262,895 | 32,930 | 243,366 | |
| | | 3 | 18 | 10,957 | 250,543 | 30,660 | 177,779 | |
| | | 4 | 12 | 19,956 | 300,390 | 20,252 | 223,791 | |
| | | 5 | 16 | 12,611 | 297,707 | 28,433 | 207,258 | |
| V18 | 50 | 1 | 13 | 35,152 | 501,230 | 74,454 | 260,574 |  |
| | | 2 | 14 | 40,896 | 559,244 | 70,863 | 274,572 | |
| | | 3 | 14 | 46,791 | 540,747 | 39,899 | 246,964 | |
| | | 4 | 10 | 33,309 | 541,490 | 56,012 | 329,343 | |
| | | 5 | 14 | 30,663 | 420,924 | 72,998 | 308,756 | |
| V19 | 65 | 1 | 15 | 6114 | 186,893 | 16,695 | 119,136 |  |
| | | 2 | 20 | 6701 | 225,075 | 6899 | 103,517 | |
| | | 3 | 20 | 5339 | 167,085 | 8834 | 103,752 | |
| | | 4 | 13 | 8462 | 185,452 | 12,020 | 129,608 | |
| | | 5 | 6 | 23,992 | 275,155 | 32,512 | 208,629 | |

**Table 3.** *Cont.*

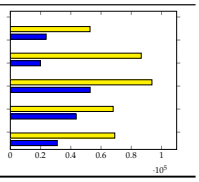
| Youtube, Dataset | | | | RMSE $(\bar{S}, \bar{S}_*)$ | | RMSE $(\bar{S}, \bar{S})$ | | Visualization |
|---|---|---|---|---|---|---|---|---|
| **Video** | **# Frames** | **User** | **# Frames User** | **Std** | **Mean** | **Std** | **Mean** | **Plot (Std,Std)** |
| V20 | 61 | 1 | 15 | 23,716 | 627,121 | 52,711 | 540,857 |  |
| | | 2 | 12 | 19,933 | 707,823 | 86,586 | 609,589 | |
| | | 3 | 9 | 52,818 | 787,188 | 93,656 | 747,199 | |
| | | 4 | 11 | 43,598 | 688,065 | 68,016 | 617,091 | |
| | | 5 | 11 | 31,058 | 695,905 | 69,077 | 618,156 | |

**Table 4.** Descriptive statistics with RMSE $(\bar{S}, \bar{S}_*)$ (target summary against random uniform sample) and RMSE $(\bar{S}, \bar{S})$ (random uniform sample against random uniform sample). Lower is better. Sampling rate: 1 frame per second. Dataset in [1], videos from V71 to V80. Highlighted values correspond to the lowest standard deviation.

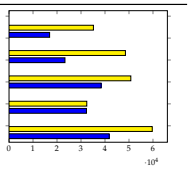
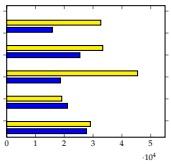
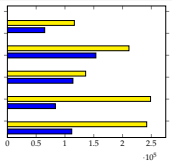
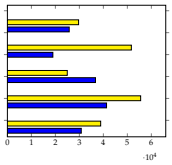
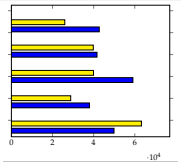
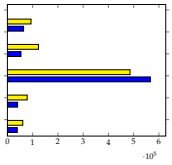
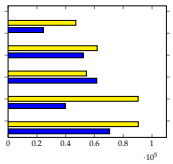
| Youtube, Dataset | | | | RMSE $(\bar{S}, \bar{S}_*)$ | | RMSE $(\bar{S}, \bar{S})$ | | Visualization |
|---|---|---|---|---|---|---|---|---|
| **Video** | **# Frames** | **User** | **# Frames User** | **Std** | **Mean** | **Std** | **Mean** | **Plot (Std,Std)** |
| V71 | 277 | 1 | 18 | 16,916 | 319,975 | 35,173 | 330,114 |  |
| | | 2 | 18 | 23,314 | 315,996 | 48,511 | 339,793 | |
| | | 3 | 20 | 38,384 | 293,853 | 50,766 | 345,021 | |
| | | 4 | 17 | 32,270 | 310,193 | 32,411 | 359,049 | |
| | | 5 | 18 | 41,753 | 329,353 | 59,688 | 334,337 | |
| V72 | 536 | 1 | 18 | 15,842 | 187,019 | 32,676 | 194,820 |  |
| | | 2 | 16 | 25,427 | 211,466 | 33,363 | 202,442 | |
| | | 3 | 16 | 18,684 | 196,149 | 45,453 | 217,699 | |
| | | 4 | 18 | 21,112 | 205,421 | 19,122 | 177,117 | |
| | | 5 | 18 | 27,718 | 206,335 | 29,057 | 205,808 | |
| V73 | 201 | 1 | 11 | 64,802 | 538,239 | 116,284 | 484,970 |  |
| | | 2 | 7 | 153,682 | 106,8305 | 211,124 | 704,655 | |
| | | 3 | 8 | 113,805 | 661,992 | 135,899 | 653,041 | |
| | | 4 | 8 | 83,387 | 856,406 | 248,619 | 689,301 | |
| | | 5 | 7 | 111,767 | 899,150 | 241,947 | 794,828 | |
| V74 | 293 | 1 | 17 | 25,780 | 282,200 | 29,674 | 309,051 |  |
| | | 2 | 16 | 18,954 | 273,776 | 51,670 | 331,322 | |
| | | 3 | 15 | 36,714 | 322,833 | 24,961 | 335,618 | |
| | | 4 | 13 | 41,327 | 363,665 | 55,543 | 369,875 | |
| | | 5 | 16 | 30,798 | 289,135 | 38,881 | 353,928 | |
| V75 | 383 | 1 | 14 | 42,736 | 254,385 | 25,959 | 282,877 |  |
| | | 2 | 13 | 41,632 | 263,431 | 39,826 | 337,124 | |
| | | 3 | 10 | 59,083 | 315,531 | 39,925 | 330,766 | |
| | | 4 | 17 | 37,954 | 227,411 | 28,843 | 250,314 | |
| | | 5 | 12 | 49,908 | 278,966 | 63,236 | 312,366 | |
| V76 | 89 | 1 | 6 | 64,097 | 440,825 | 93,524 | 422,565 |  |
| | | 2 | 4 | 53,727 | 536,138 | 123,009 | 464,922 | |
| | | 3 | 1 | 566,208 | 843,799 | 485,614 | 878,793 | |
| | | 4 | 6 | 40,356 | 382,643 | 78,354 | 424,418 | |
| | | 5 | 6 | 39,194 | 395,906 | 60,916 | 401,751 | |
| V77 | 168 | 1 | 12 | 24,546 | 302,076 | 47,095 | 366,748 |  |
| | | 2 | 9 | 52,176 | 339,285 | 61,880 | 385,056 | |
| | | 3 | 9 | 61,623 | 355,883 | 54,390 | 413,118 | |
| | | 4 | 10 | 39,765 | 349,207 | 90,313 | 400,379 | |
| | | 5 | 7 | 70,562 | 440,656 | 90,468 | 451,833 | |

**Table 4.** *Cont.*

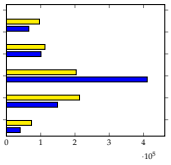
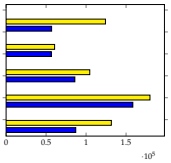
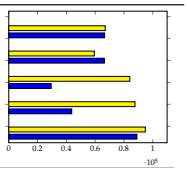
| Youtube, Dataset | | | | RMSE $(\bar{S}, \bar{S}_*)$ | | RMSE $(\bar{S}, \bar{S})$ | | Visualization |
|---|---|---|---|---|---|---|---|---|
| **Video** | **# Frames** | **User** | **# Frames User** | **Std** | **Mean** | **Std** | **Mean** | **Plot (Std,Std)** |
| V78 | 310 | 1 | 13 | 65,238 | 706,978 | 96,368 | 770,000 |  |
| | | 2 | 14 | 100,771 | 672,121 | 112,412 | 807,250 | |
| | | 3 | 3 | 410,792 | 159,3229 | 203,589 | 188,2757 | |
| | | 4 | 9 | 149,063 | 839,743 | 213,286 | 106,1204 | |
| | | 5 | 23 | 40,178 | 466,571 | 73,228 | 614,140 | |
| V79 | 49 | 1 | 7 | 56,918 | 831,057 | 124,249 | 835,575 |  |
| | | 2 | 8 | 56,569 | 793,831 | 60,657 | 859,241 | |
| | | 3 | 6 | 85,973 | 925,025 | 104,621 | 990,479 | |
| | | 4 | 5 | 158,480 | 109,3141 | 179,902 | 109,9105 | |
| | | 5 | 6 | 87,104 | 873,950 | 131,597 | 895,318 | |
| V80 | 159 | 1 | 18 | 66,585 | 529,875 | 67,019 | 572,836 |  |
| | | 2 | 17 | 66,367 | 527,930 | 59,432 | 602,819 | |
| | | 3 | 13 | 29,459 | 579,078 | 84,101 | 726,883 | |
| | | 4 | 12 | 43,740 | 643,016 | 87,688 | 685,117 | |
| | | 5 | 14 | 89,016 | 553,274 | 94,849 | 649,317 | |

Figure 8 shows a summary that is well annotated by all users, demonstrating that the metrics can accurately indicate when human annotators have effectively summarized the information present in the video.

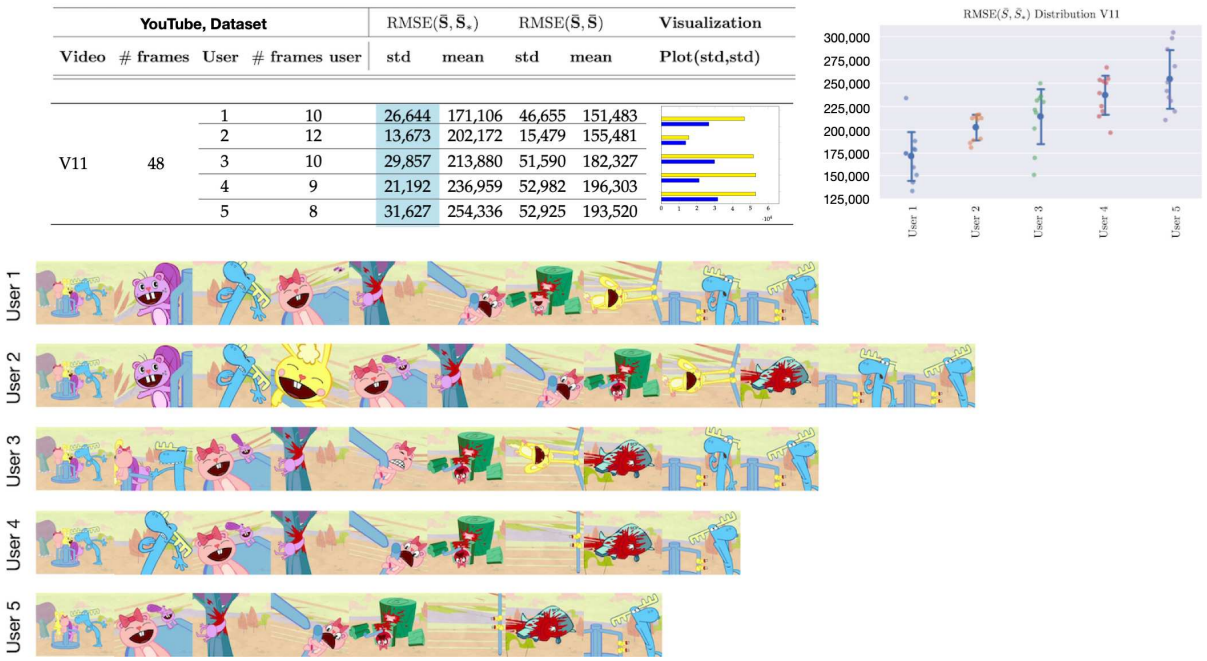



**Figure 8.** Visual depiction of human annotated summaries together with RMSE $(\bar{S}, \bar{S}_*)$ and RMSE $(\bar{S}, \bar{S})$ of video V11, Table 3. Sampling rate: 1 frame per second. Highlighted values on the table correspond to the lowest standard deviation.

To illustrate how these metrics can help improve annotations, Figure 9 displays the metrics along with the annotated summaries of users 1 to 5. We observe that selecting the frames highlighted by users 1–4 can increase the performance if user 5 is asked to relabel its summary.

| YouTube, Dataset | | | | RMSE$(\bar{S}, \bar{S}_*)$ | | RMSE$(\bar{S}, \bar{S})$ | | Visualization |
|---|---|---|---|---|---|---|---|---|
| Video | # frames | User | # frames user | std | mean | std | mean | Plot(std,std) |
| V19 | 65 | 1 | 15 | 6114 | 186,893 | 16,695 | 119,136 | |
| | | 2 | 20 | 6701 | 225,075 | 6899 | 103,517 | |
| | | 3 | 20 | 5339 | 167,085 | 8834 | 103,752 | |
| | | 4 | 13 | 8462 | 185,452 | 12,020 | 129,608 | |
| | | 5 | 6 | 23,992 | 275,155 | 32,512 | 208,629 | |

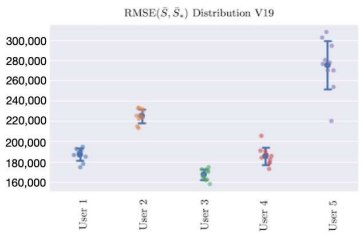




**Figure 9.** Visual depiction of human annotated summaries together with RMSE $(\bar{S}, \bar{S}_*)$ and RMSE $(\bar{S}, \bar{S})$ of video V19, Table 3. Sampling rate: 1 frame per second. Highlighted frames can increase the accuracy of the annotated summary by user 5. Highlighted values on the table correspond to the lowest standard deviation.

Figure 10 showcases an example in which random uniform sampling outperforms the majority of human annotators. This occurs because the visual information is uniformly distributed throughout the video. In this case, user 5 performs the best, scoring slightly higher than std (RMSE $(\bar{S}, \bar{S})$. Highlighted values on the table correspond to the lowest standard deviation.).

| YouTube, Dataset | | | | RMSE$(\bar{S}, \bar{S}_*)$ | | RMSE$(\bar{S}, \bar{S})$ | | Visualization |
|---|---|---|---|---|---|---|---|---|
| Video | # frames | User | # frames user | std | mean | std | mean | Plot(std,std) |
| V75 | 383 | 1 | 14 | 42,736 | 254,385 | 25,959 | 282,877 | |
| | | 2 | 13 | 41,632 | 263,431 | 39,826 | 337,124 | |
| | | 3 | 10 | 59,083 | 315,531 | 39,925 | 330,766 | |
| | | 4 | 17 | 37,954 | 227,411 | 28,843 | 250,314 | |
| | | 5 | 12 | 49,908 | 278,966 | 63,236 | 312,366 | |

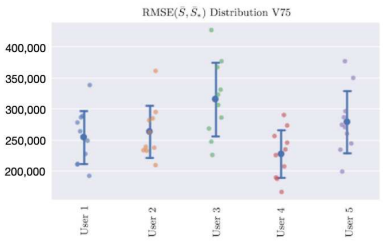




**Figure 10.** Visual depiction of human annotated summaries, together with RMSE $(\bar{S}, \bar{S}_*)$ and RMSE $(\bar{S}, \bar{S})$ of video V75, Table 4. Sampling rate: 1 frame per second. Highlighted values on the table correspond to the lowest standard deviation.

Similarly, Figure 11 presents an example in which incorporating the highlighted frames improves the accuracy of the annotated summary by user 3, which is currently performing worse than uniform random sampling, according to the metrics.

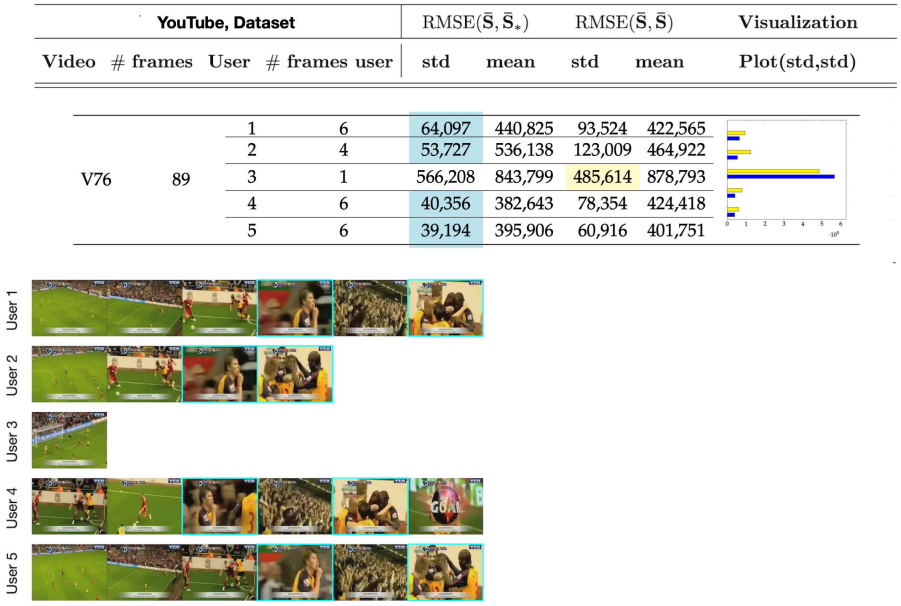



**Figure 11.** Visual depiction of human annotated summaries together with RMSE $(\bar{S}, \bar{S}_*)$ and RMSE $(\bar{S}, \bar{S})$ of video V76, Table 4. Sampling rate: 1 frame per second. Highlighted frames can increase the accuracy of the annotated summary by user 3. Highlighted values on the table correspond to the lowest standard deviation.

### *4.2. Evaluation*

In this section, we evaluate the baselines and metrics compared to VSUMM [1], a methodology based on handcrafted techniques that performs particularly well on this dataset. Table 5 displays the comparison between the standard deviation of RMSE $(\bar{S}, \bar{S}_*)$ and RMSE $(\bar{S}, \bar{S})$, as well as against the baselines based on the Signature Transform, RMSE $(\bar{S}, \bar{S}_{u_{min}})|_{10}$ and RMSE $(\bar{S}, \bar{S}_{u_{min}})|_{20}$, with 10 and 20 points, respectively.

We can observe how the metrics effectively capture the quality of the visual summaries and how the introduced methodology based on the Signature Transform achieves state-of-the-art results with both 10 and 20 points. The advantages of using a technique that operates on the spectrum of the signal, compared to other state-of-the-art systems, is that it can generate visual summaries without fine-tuning the methodology. In other words, there is no need to train on a subset of the target distribution of videos, but rather, compelling summaries can be generated at once for any dataset. Moreover, this approach is highly efficient, as computation is performed on the CPU and consists only of calculating the Signature Transform, element-wise mean, and RMSE. These operations can be further optimized for rapid on-device processing or for deploying in parallel at the tera-scale level.

**Table 5.** VSUMM [1] comparison against baseline based on the Signature Transform for the first 20 videos of the dataset crawled from Youtube. Descriptive statistics with RMSE $(\bar{S}, \bar{S}_*)$ (target summary against random uniform sample) and RMSE $(\bar{S}, \bar{S})$ (random uniform sample against random uniform sample). RMSE $(\bar{S}, \bar{S}_{u_{min}})|_{10}$ and RMSE $(\bar{S}, \bar{S}_{u_{min}})|_{20}$ correspond to the baselines based on the Signature Transform using 10 and 20 random samples, respectively. Highlighted results are better than std (RMSE $(\bar{S}, \bar{S})$). Sampling rate: 1 frame per second. Highlighted results correspond to lowest standard deviation as described in Table 1.

| **Descriptive Statistics** | | **VSUMM** | **RMSE $(\bar{S}, \bar{S}_*)$** | | **RMSE $(\bar{S}, \bar{S})$** | | **RMSE $(\bar{S}, \bar{S}_{u_{min}})|_{10}$** | | **RMSE $(\bar{S}, \bar{S}_{u_{min}})|_{20}$** | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Video** | **# Frames** | **# Frames** | **Std** | **Mean** | **Std** | **Mean** | **Std** | **Mean** | **Std** | **Mean** |
| V11 | 48 | 11 | 25,981 | 185,959 | 37,907 | 175,031 | 16,343 | 148,128 | 18,343 | 159,157 |
| V12 | 59 | 13 | 56,274 | 313,156 | 41,613 | 205,004 | 17,770 | 181,533 | 11,665 | 206,951 |
| V13 | 59 | 19 | 7018 | 184,865 | 15,319 | 120,307 | 10,578 | 110,258 | 6655 | 134,846 |
| V14 | 59 | 8 | 21,415 | 281,969 | 39,412 | 171,935 | 19,069 | 157,531 | 10,104 | 180,199 |
| V15 | 57 | 10 | 20,159 | 271,197 | 46,041 | 219,182 | 27,536 | 192,667 | 27,765 | 218,787 |

**Table 5.** *Cont.*

| Descriptive Statistics | | VSUMM | RMSE $(\bar{S}, \bar{S}_*)$ | | RMSE $(\bar{S}, \bar{S})$ | | RMSE $(\bar{S}, \bar{S}_{u_{min}})\|_{10}$ | | RMSE $(\bar{S}, \bar{S}_{u_{min}})\|_{20}$ | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Video** | **# Frames** | **# Frames** | **Std** | **Mean** | **Std** | **Mean** | **Std** | **Mean** | **Std** | **Mean** |
| V16 | 70 | 9 | 65,997 | 513,440 | 84,667 | 428,025 | 38,088 | 283,324 | 30,235 | 446,068 |
| V17 | 59 | 15 | 10,697 | 255,666 | 41,831 | 197,136 | 17,625 | 197,944 | 19,102 | 227,646 |
| V18 | 50 | 14 | 42,731 | 449,324 | 51,635 | 230,695 | 33,525 | 261,288 | 30,179 | 242,746 |
| V19 | 65 | 16 | 3891 | 235,797 | 5739 | 121,766 | 5883 | 116,245 | 4582 | 111,766 |
| V20 | 61 | 9 | 43,864 | 796,448 | 39,035 | 733,547 | 28,460 | 684,546 | 39,414 | 644,681 |
| V71 | 277 | 17 | 20,840 | 383,945 | 43,176 | 341,779 | 14,908 | 352,365 | 20,657 | 327,732 |
| V72 | 536 | 12 | 61,886 | 233,649 | 48,603 | 252,688 | 17,604 | 276,631 | 18,966 | 248,489 |
| V73 | 201 | 10 | 40,261 | 717,107 | 156,051 | 533,457 | 64,344 | 681,064 | 38,361 | 711,039 |
| V74 | 293 | 17 | 26,274 | 270,374 | 36,674 | 334,265 | 17,622 | 354,621 | 17,486 | 330,606 |
| V75 | 383 | 10 | 37,516 | 272,804 | 38,026 | 366,510 | 23,163 | 339,078 | 21,295 | 360,216 |
| V76 | 89 | 7 | 36,084 | 353,323 | 114,266 | 377,699 | 31,131 | 335,958 | 34,724 | 405,954 |
| V77 | 168 | 9 | 26,653 | 361,516 | 67,134 | 422,612 | 33,214 | 407,085 | 27,562 | 480,795 |
| V78 | 310 | 13 | 95,305 | 831,043 | 127,705 | 823,938 | 33,903 | 980,397 | 36,361 | 951,784 |
| V79 | 49 | 7 | 67,052 | 965,267 | 101,325 | 878,917 | 42,513 | 818,629 | 47,401 | 885,023 |
| V80 | 159 | 15 | 48,115 | 613,702 | 118,428 | 644,529 | 43,411 | 589,256 | 37,487 | 808,984 |
| **Average** | 153 | 12 | 17/20 (85%) | | | | 19/20 (95%) | | 19/20 (95%) | |

## 5. Conclusions and Future Work

In this manuscript, we propose a benchmark based on the Signature Transform to evaluate visual summaries. For this purpose, we introduce a dataset consisting of videos obtained from Youtube related to science experiments with automatic audio transcriptions. A baseline, based on zero-shot text-conditioned object detection, is used as a preliminary technique in the study to evaluate the metrics. Subsequently, we present an accurate baseline built on the prior knowledge that the Signature provides. Furthermore, we conduct rigorous comparison against human-annotated summaries to demonstrate the high correlation between the measures and the human notion of a good summary.

One of the main contributions of this work is that techniques based on the Signature Transform can be integrated with any state-of-the-art method in the form of a gate that activates when the method performs worse than the metric, std $(\text{RMSE}\ (\bar{S}, \bar{S}_*)) >$ std $(\text{RMSE}\ (\bar{S}, \bar{S}))$.

The experiments conducted in this work lead to the following conclusion: if a method for delivering a summarization technique is proposed that involves complex computation (e.g., DNN techniques or Foundation Models), it must provide better summarization capabilities than the baselines based on the Signature Transform, which serve as lower bounds for uniform random samples. If not, there is no need to use a more sophisticated technique that would involve greater computational and memory overhead and possibly require training data. The only exception to this would be when additional constraints are present in the problem, such as when summarization must be performed by leveraging audio transcriptions (as in the technique based on text-conditioned object detection) or any other type of multimodal data.

That being said, the methodology proposed based on the Signature Transform, although accurate and effective, is built on the overall representation of harmonic components of the signal. Videlicet, under certain circumstances, can provide summaries in which frames are selected due to low-level representations of the signal, such as color and image intensity, rather than the storyline. Moreover, it assumes that, in general, uniform random sampling can provide good summarization capabilities, which is supported by the literature. However, this assumption is not fulfilled in all circumstances. Therefore, in subsequent works, it would be desirable to develop techniques that perform exceptionally well according to the metrics while simultaneously bestowing a level of intelligence similar to the methodology based on Foundation Models. This would take into account factors such as the human concept of detected objects, leading to more context-aware and meaningful summarization.

**Author Contributions:** Conceptualization, J.d.C. and I.d.Z.; funding acquisition, C.T.C. and G.R.; investigation, J.d.C. and I.d.Z.; methodology, J.d.C. and I.d.Z.; software, J.d.C. and I.d.Z.; supervision, G.R. and C.T.C.; writing—original draft, J.d.C.; writing—review and editing, C.T.C., G.R., J.d.C. and I.d.Z. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported by the HK Innovation and Technology Commission (InnoHK Project CIMDA). We acknowledge the support of Universitat Politècnica de València; R&D project PID2021-122580NB-I00, funded by MCIN/AEI/10.13039/501100011033 and ERDF. We thank the following funding sources from GOETHE-University Frankfurt am Main; 'DePP—Dezentrale Plannung von Platoons im Straßengüterverkehr mit Hilfe einer KI auf Basis einzelner LKW' and 'Center for Data Science & AI'.

**Data Availability Statement:** https://doi.org/10.24433/CO.7648856.v2 (accessed on 1 March 2023).

**Conflicts of Interest:** The authors declare that they have no conflicts of interest. The funders had no role in the design of the study, in the collection, analyses, or interpretation of data, in the writing of the manuscript, or in the decision to publish the results.

## Abbreviations

The following abbreviations are used in this manuscript:

| | |
|---|---|
| DNN | Deep Neural Networks |
| AMT | Amazon Mechanical Turk |
| RMSE | Root Mean Squared Error |
| MAE | Mean Absolute Error |
| VLM | Visual Language Models |
| LLM | Large Language Models |
| GAN | Generative Adversarial Networks |
| CLIP | Contrastive Language–Image Pre-training |
| LSTM | Long Short-Term Memory |
| RNN | Recurrent Neural Network |
| NLP | Natural Language Processing |
| CPU | Central Processing Unit |
| MOS | Mean Opinion Score |

## References

1. de Avila, S.E.F.; Lopes, A.; da Luz, A., Jr.; de Albuquerque Araújo, A. VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method. *Pattern Recognit. Lett.* **2011**, *32*, 56–68. [CrossRef]
2. Gygli, M.; Grabner, H.; Gool, L.V. Video summarization by learning submodular mixtures of objectives. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015.
3. Gygli, M.; Grabner, H.; Riemenschneider, H.; Van Gool, L. (2014). Creating summaries from user videos. In Proceedings of the Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September; Springer: Berlin/Heidelberg, Germany, 2014.
4. Kanehira, A.; Gool, L.V.; Ushiku, Y.; Harada, T. Viewpoint-aware video summarization. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018.
5. Liang, G.; Lv, Y.; Li, S.; Zhang, S.; Zhang, Y. Video summarization with a convolutional attentive adversarial network. *Pattern Recognit.* **2022**, *131*, 108840. [CrossRef]
6. Song, Y.; Vallmitjana, J.; Stent, A.; Jaimes, A. TVSum: Summarizing web videos using titles. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 5179–5187.
7. Zhu, W.; Lu, J.; Han, Y.; Zhou, J. Learning multiscale hierarchical attention for video summarization. *Pattern Recognit.* **2022**, *122*, 108312. [CrossRef]
8. Ngo, C.-W.; Ma, Y.-F.; Zhang, H.-J. Automatic video summarization by graph modeling. In Proceedings of the Ninth IEEE International Conference on Computer Vision, Nice, France, 13–16 October 2003.
9. Fajtl, J.; Sokeh, H.S.; Argyriou, V.; Monekosso, D.; Remagnino, P. Summarizing videos with attention. In Proceedings of the Computer Vision—ACCV 2018 Workshops: 14th Asian Conference on Computer Vision, Perth, Australia, 2–6 December; Springer: Berlin/Heidelberg, Germany, 2018.
10. Zhu, W.; Lu, J.; Li, J.; Zhou, J. DSNet: A flexible detect-to-summarize network for video summarization. *IEEE Trans. Image Process.* **2020**, *30*, 948–962. [CrossRef] [PubMed]
11. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.

12. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
13. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015.
14. de Curtò, J.; de Zarzà, I.; Yan, H.; Calafate, C.T. On the applicability of the hadamard as an input modulator for problems of classification. *Softw. Impacts* **2022**, *13*, 100325. [CrossRef]
15. de Zarzà, I.; de Curtò, J.; Calafate, C.T. Detection of glaucoma using three-stage training with efficientnet. *Intell. Syst. Appl.* **2022**, *16*, 200140. [CrossRef]
16. Dwivedi, K.; Bonner, M.F.; Cichy, R.M.; Roig, G. Unveiling functions of the visual cortex using task-specific deep neural networks. *PLoS Comput. Biol.* **2021**, *17*, e100926. [CrossRef]
17. Dwivedi, K.; Roig, G.; Kembhavi, A.; Mottaghi, R. What do navigation agents learn about their environment? In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 10276–10285.
18. Rakshit, S.; Tamboli, D.; Meshram, P.S.; Banerjee, B.; Roig, G.; Chaudhuri, S. Multi-source open-set deep adversarial domain adaptation. In Proceedings of the Computer Vision—ECCV: 16th European Conference, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 735–750.
19. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI*; Springer: Cham, Switzerland, 2015.
20. Tan, M.; Chen, B.; Pang, R.; Vasudevan, V.; Le, Q.V. Mnasnet: Platform-aware neural architecture search for mobile. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
21. Thao, H.; Balamurali, B.; Herremans, D.; Roig, G. Attendaffectnet: Self-attention based networks for predicting affective responses from movies. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 8719–8726.
22. Mahasseni, B.; Lam, M.; Todorovic, S. Unsupervised video summarization with adversarial lstm networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
23. Zhang, K.; Chao, W.-L.; Sha, F.; Grauman, K. Video summarization with long short-term memory. In Proceedings of the Computer Vision–ECCV: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Cham, Switzerland, 2016.
24. Zhao, B.; Li, X.; Lu, X. Hierarchical recurrent neural network for video summarization. In Proceedings of the 25th ACM International Conference on Multimedia, Mountain View, CA, USA, 23–27 October 2017.
25. Rochan, M.; Ye, L.; Wang, Y. Video summarization using fully convolutional sequence networks. In Proceedings of the Computer Vision–ECCV: 15th European Conference, Munich, Germany, 8–14 September 2018; Springer: Cham, Switzerland, 2018.
26. Yuan, L.; Tay, F.E.; Li, P.; Zhou, L.; Feng, J. Cycle-sum: Cycle-consistent adversarial lstm networks for unsupervised video summarization. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019.
27. Zhang, K.; Grauman, K.; Sha, F. Retrospective encoders for video summarization. In Proceedings of the Computer Vision–ECCV: 15th European Conference, Munich, Germany, 8–14 September 2018; Springer: Cham, Switzerland, 2018.
28. Zhou, K.; Qiao, Y.; Xiang, T. Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward. In Proceedings of the Association for the Advancement of Artificial Intelligence Conference (AAAI), New Orleans, LA, USA, 2–7 February 2018.
29. Narasimhan, M.; Rohrbach, A.; Darrell, T. Clip-it! Language-Guided Video Summarization. *Adv. Neural Inf. Process. Syst.* **2021**; *34*, 13988–14000.
30. Plummer, B.A.; Brown, M.; Lazebnik, S. Enhancing video summarization via vision-language embedding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
31. Otani, M.; Nakashima, Y.; Rahtu, E.; Heikkilä, J. Rethinking the evaluation of video summaries. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
32. de Curtò, J.; de Zarzà, I.; Yan, H.; Calafate, C.T. Signature and Log-signature for the Study of Empirical Distributions Generated with GANs. *arXiv* **2022**, arXiv:2203.03226.
33. Lyons, T. Rough paths, signatures and the modelling of functions on streams. *arXiv* **2014**, arXiv:1405.4537.
34. Bonnier, P.; Kidger, P.; Arribas, I.P.; Salvi, C.; Lyons, T. Deep signature transforms. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December2019; Volume 32.
35. Chevyrev, I.; Kormilitzin, A. A primer on the signature method in machine learning. *arXiv* **2016**, arXiv:1603.03788.
36. Kidger, P.; Lyons, T. Signatory: Differentiable computations of the signature and logsignature transforms, on both CPU and GPU. *arXiv* **2020**, arXiv:2001.00706.
37. Liao, S.; Lyons, T.J.; Yang, W.; Ni, H. Learning stochastic differential equations using RNN with log signature features. *arXiv* **2019**, arXiv:1908.0828.
38. Morrill, J.; Kidger, P.; Salvi, C.; Foster, J.; Lyons, T.J. Neural CDEs for long time series via the log-ode method. *arXiv* **2021**, arXiv:2009.08295.
39. Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. Flamingo: A visual language model for few-shot learning. *arXiv* **2022**, arXiv:2204.14198.
40. Gu, X.; Lin, T.-Y.; Kuo, W.; Cui, Y. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv* **2022**, arXiv:2104.13921.

41. Wei, J.; Tay, Y.; Bommasani, R.; Raffel, C.; Zoph, B.; Borgeaud, S.; Yogatama, D.; Bosma, M.; Zhou, D.; Metzler, D.; et al. Emergent abilities of large language models. *arXiv* **2022**, arXiv:2206.07682.
42. de Curtò, J.; de Zarzà, I.; Calafate, C.T. Semantic scene understanding with large language models on unmanned aerial vehicles. *Drones* **2023**, *7*, 114. [CrossRef]
43. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J. et al. Learning transferable visual models from natural language supervision. *arXiv* **2021**, arXiv:2103.00020.
44. Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; Sutskever, I. Zero-shot text-to-image generation. In Proceedings of the 38th International Conference on Machine Learning, Online, 18–24 July 2021; pp. 8821–8831.
45. Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-resolution image synthesis with latent diffusion models. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022.
46. Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E.; Ghasemipour, S.K.S.; Ayan, B.K.; Mahdavi, S.S.; Lopes, R.G.; et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv* **2022**, arXiv:2205.11487.
47. Cui, Y.; Niekum, S.; Gupta, A.; Kumar, V.; Rajeswaran, A. Can foundation models perform zero-shot task specification for robot manipulation? In Proceedings of the Learning for Dynamics and Control Conference, Palo Alto, CA, USA, 23–24 June 2022 .
48. Nair, S.; Rajeswaran, A.; Kumar, V.; Finn, C.; Gupta, A. R3M: A universal visual representation for robot manipulation. *arXiv* **2022**, arXiv:2203.12601.
49. Zeng, A.; Florence, P.; Tompson, J.; Welker, S.; Chien, J.; Attarian, M.; Armstrong, T.; Krasin, I.; Duong, D.; Wahid, A.; et al. Transporter networks: Rearranging the visual world for robotic manipulation. In Proceedings of the Conference on Robot Learning, Online, 15–18 November 2020.
50. Huang, W.; Abbeel, P.; Pathak, D.; Mordatch, I. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. *arXiv* **2022**, arXiv:2201.07207.
51. Zeng, A.; Attarian, M.; Ichter, B.; Choromanski, K.; Wong, A.; Welker, S.; Tombari, F.; Purohit, A.; Ryoo, M.; Sindhwani, V.; et al. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv* **2022**, arXiv:2204.00598.
52. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16 $\times$16 words: Transformers for image recognition at scale. *arXiv* **2021**, arXiv:2010.11929.
53. Minderer, M.; Gritsenko, A.; Stone, A.; Neumann, M.; Weissenborn, D.; Dosovitskiy, A.; Mahendran, A.; Arnab, A.; Dehghani, M.; Shen, Z.; et al. Simple open-vocabulary object detection with vision transformers. *arXiv* **2022**, arXiv:2205.06230.
54. Bird, S.; Klein, E.; Loper, E. *Natural Language Processing with Python*, 1st ed.; O'Reilly Media, Inc.: Sebastopol, CA, USA, 2009.