An Open Source Quantitative Evaluation Framework for Automatic Video Summarization Algorithms

Leandro Balmaceda[1], Ariel I. Diaz[1], Adrián Rostagno[1], Santiago L. Aggio[1], Anibal M. Blanco[2], and Javier Iparraguirre[1]

[1] Universidad Tecnológica Nacional, Facultad Regional Bahía Blanca, Argentina
[2] PLAPIQUI, Universidad Nacional del Sur, CONICET, Bahía Blanca, Argentina
j.iparraguirre@computer.org

Abstract. The creation, consumption, and manipulation of video play a central role in everyday life as the amount of video data is growing at an exponential rate. Video summarization consists on producing a condensed output from a video that allows humans to rapidly understand and browse the content of the original source. Although there are several evaluation approaches proposed in the literature, multiple challenges make the quantitative evaluation of a summarization a complex process. In this paper we present a completely open video summarization evaluation framework that is compatible with existing datasets and published results. Standard metrics are considered and a new metric that captures unbalanced-class video summarization evaluation is proposed. Two legacy datasets are integrated in a standard format. Finally, new quantitative results based on already published algorithms are presented.

1 Introduction

The creation, consumption, and manipulation of video play a central role in everyday life. The amount of video data is growing at an exponential rate. This reality makes almost impossible to manually process a video collection. As a consequence, humans need tools that help to understand and process video. Video summarization (VS) consists on producing a condensed output from a video that allows humans to understand and browse the content of the original source. A successful summarization should present the user all relevant details available in the original stream in a intelligible and practical way. There are multiple ways to classify summarization algorithms. Truong and Venkatesh [7] cataloged the methods by the produced output. There are VS methods that produce keyframes (static summarization) and others that produce a video skim (dynamic summarization). Some authors propose a classification based on the classification scheme [8, 9]. Another possible classification depends on the online operation of the method [4].

Although a diverse collection of VS methods are present in the literature [1, 3, 8], multiple barriers arise making the quantitative evaluation a complex process. Some of the challenges are: i) datasets are not available in a consistent format, ii) different evaluation methods are focused on different relevant aspects iii) the use of proprietary software complicates the comparison of different approaches. As a consequence, it is a

challenging task to get consistent quantitative results after any VS algorithm is implemented.

In this paper we present a completely open VS evaluation framework. Quantitative results of already published VS algorithms can be reproduced in the provided toolkit. Additionally, a new evaluation method is proposed. Common metrics can also be calculated and legacy datasets are integrated in a flexible way. We named the new framework Open Summarization Toolbox (OST) and it will available as an open source project [3]. The spirit behind OST consists on simplifying the quantitative evaluation of VS techniques while keeping the compatibility with previous published results as much as possible. Additionally, the architecture allows the use of new available datasets.

The main contributions of this work are: i) a clear implementation of a VS evaluation methodology that is compatible with existing evaluation proposals, ii) a new VS evaluation methodology that captures time distance between selected relevant data, iii) the integration of metrics used in previous publications and the introduction of a new metric, iv) the integration of two existing datasets in a standard format allowing the evaluation of existing and new summaries, v) new quantitative results based on already published algorithms, vi) a completely open source toolbox that allows the operation over multiple VS outputs.

2 Related Work

The work published by De Avila *et al*. [1] was fundamental in terms of setting the basic concepts in a solid way. In this article a VS method (VSUMM), an evaluation methodology called Comparison of Summaries (CUS), and two datasets are provided. The first dataset contains 50 videos selected from the Open Video Project (OVP)[4]. The second dataset contains 50 videos selected from YouTube[5]. Both datasets provide frames selected by 5 users for each video. Although this work has been intensively cited, there are aspects that may be subject to improvement. In first place, the implementation of CUS is a closed-source binary. Additionally, the provided datasets are available as raw videos and frames, and there is no tool that allows their conversion to other formats such as vector or matrix data structures.

Song *et al*. [6] presented a title-based image search VS algorithm and the TVSumm dataset. The dataset consists of 50 videos from YouTube representing multiple genres such as news, sports, and user generated content. The annotated content was generated using crowd-sourcing and it is classified by shot-level importance. There are two restrictions related to this dataset. The first barrier is that the annotated information is not open, therefore, interested users must request access. Other limitation is that shot-level importance is not a common practice in current VS research. Legacy results must be converted, therefore, to this particular format.

Gygli *et al*. [3] published a summarization method based on superframes, an evaluation framework called the SumMe benchmark, and a dataset that consists on 25

[3] https://github.com/BHI-Research/ost-python
[4] http://www.open-video.org/
[5] http://www.youtube.com/

Table 1: Related datasets published in previous works.

Dataset	# Videos	Description	Annotations format	Open
SumMe [3]	25	User generated videos	Interval-based shots	Yes
TVSumm [6]	50	YouTube videos	Frame-level importance	No
OVP [1]	50	Open Video Project	Keyframes	Yes
YouTube [1]	50	YouTube videos	Keyframes	Yes

videos with annotations of 15 to 18 humans for each video. The work presents a solid statistical approach related to VS evaluation. However, the provided implementation of the benchmark depends on proprietary software, and the human annotations are embedded into the framework. Therefore, SumMe presents limitations for a wide adoption by the research community.

There are also differences between CUS, TVSumm, and SumMe datasets in terms of the format of the ground truth information. CUS compares keyframes, TVSumm assigns a relevance value to each frame, and SumMe focuses on video segments. Zhang *et al*. [8] acknowledged the difficulties associated with the quantitative evaluation and proposed a unified evaluation methodology. Although they provided a public repository that contains the evaluation methodology, the proposal is still based on proprietary software. Additionally, it is hard to decouple the datasets and the source code used in the evaluation. Table 1 summarizes the mentioned datasets, the content covered, and the format used to provide the annotations.

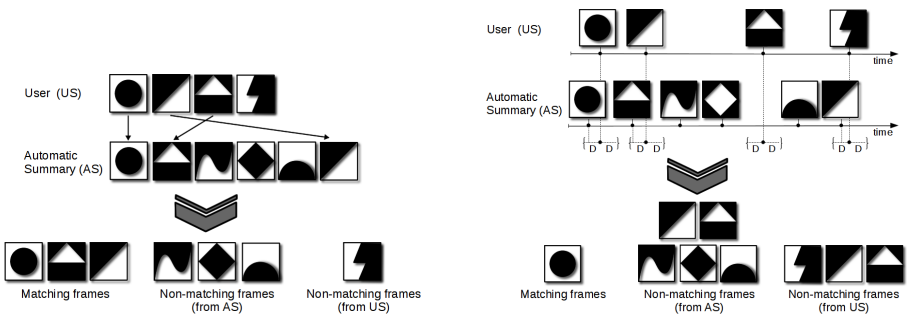
3 Proposed Framework

3.1 Evaluation Methods

Two evaluation methods are implemented in OST: CUS [1], which is very popular in the VS community, and a novel method named BHI. Since the authors of CUS do not provide the source code, we implemented it in OST following the published concepts. All evaluation methods considered in OST take as input a set of keyframes.

CUS As it is shown in Figure 1a, CUS demands as input a set of keyframes selected by the user (User Summary, US) and another set selected by the VS algorithm (Automatic Summary, AS). Two frames from the AS and US sets are matched if the distance of color histograms is below a threshold (e). After the matching process is completed, three groups of frames are obtained: (i) matched frames, (ii) non-matched user selected frames, and (iii) non-matched algorithm selected frames. Afterwards, a set of metrics can be calculated based on the available results.

BHI Although CUS provides a clear way to compare two sets of keyframes, there is no consideration related to the temporal distance of the frames. In some cases, the section of the video sequence in which a keyframe appears may have a particular significance. For example, in surveillance video, the temporal distance between two



(a) CUS (b) BHI

Fig. 1: CUS evaluation method (left) and BHI evaluation method (right).

Table 2: Square contingency table based on the results of any evaluation method.

		Automatic VS		
		Yes	No	Total
User VS	Yes	TP	FN	$TP+FN$
	No	FP	TN	$FP+TN$
	Total	$TP+FP$	$FN+TN$	T

keyframes containing the same person may be critical. In order to address this issue, we decided to create an evaluation method that measures not only the similarity of the frames, but also the temporal distance among them.

Figure 1b shows the mechanism behind the proposed evaluation method. Given a frame selected by the user, a similar frame is searched at a temporal distance D. If a frame is located, color histograms are compared in a similar way as in the case of CUS. This new scheme keeps compatibility in terms of frame similarity and adds the temporal component to the evaluation method. BHI produces as output (i) the set of frames that are matched, (ii) the non-matched frames selected by the user, and (iii) the non-matched frames selected by the automatic summarization. It is relevant to remark that both conditions, similarity and locality, must be satisfied to produce a match.

3.2 Metrics

The result of any of the proposed evaluation methods can be condensed into a contingency table. Matched frames can be considered True Positive , frames selected by the user that were not matched can be labeled as False Negatives , and frames selected by the algorithm that were not matched can be marked as False Positives. The remaining frames in the video fall into the True Negative category. After a contingency table is created (see Table 2), multiple metrics can be calculated such as **CUSa** [1], **CUSe** [1], **Precision**, **Recall**, and **F-Score**.

$$CUS_A = \frac{TP}{TP+FN}, CUS_E = \frac{FP}{TP+FN} \quad (1)$$

Table 3: CUS and BHI evaluation methods using OVP as test dataset.

VS Algorithm	CUSa		CUSe		Precision		Recall		F-score		κ	
	CUS	BHI	CUS	BHI	CUS	BHI	CUS	BHI	CUS	BHI	CUS	BHI
VSUMM1	0.868	0.703	0.435	0.599	0.753	0.609	0.868	0.703	0.795	0.643	0.786	0.635
VSUMM2	0.716	0.553	0.294	0.457	0.779	0.600	0.717	0.553	0.732	0.563	0.723	0.556
DT	0.552	0.358	0.316	0.510	0.698	0.456	0.552	0.358	0.590	0.381	0.583	0.376
OV	0.726	0.601	0.556	0.670	0.644	0.528	0.726	0.601	0.647	0.533	0.640	0.526
STIMO	0.730	0.529	0.749	0.949	0.602	0.443	0.730	0.529	0.638	0.464	0.630	0.457
FLASH	0.618	0.456	0.810	0.988	0.555	0.399	0.618	0.456	0.547	0.397	0.540	0.391

$$Precision = \frac{TP}{TP+FP}, Recall = \frac{TP}{TP+FN}, \text{F-score} = \frac{2 * Precision * Recall}{Precision + Recall} \tag{2}$$

In the case where there is no matching frame, no true positives cases are detected and therefore F-Score is zero. It is possible to argue that in cases that no matches are found there is a coincidence that both raters agree that most on frames are true positive. In order to capture this agreement, OST introduces an additional metric: the Conhen's kappa (κ) coefficient. This new metric helps to quantify unbalanced classification problems such as VS.

$$\kappa = \frac{P_o - P_e}{1 - P_e}, P_o = \frac{TP+TN}{T}, P_e = \frac{TP+FP}{T} * \frac{TP+FN}{T} + \frac{TN+FN}{T} * \frac{TN+FP}{T} \tag{3}$$

4 Results

In this section, results related to published VS algorithms are presented to illustrate most relevant features of OST. Using the OVP dataset as reference, the following VS algorithms were evaluated: VSUMM1, VSUMM2 [1], DT [5], OV (keyframes provided by OVP), STIMO [2]. Additionally, new results from another online VS algorithm already published (FLASH) [4] are included. An open source version of FLASH is available at a public accessible website [6].

Table 3 shows quantitative results using CUS evaluation method. The first relevant conclusion is that obtained values are consistent with previous publications [1]. In the case of FLASH, lower evaluation results can be observed. The scores are reasonable since FLASH is designed to operate in real time without any knowledge of the incoming video stream.

Results using BHI evaluation method are also shown in Table 3. A distance value (D) of 120 frames is selected. In this case, all VS algorithms receive lower scores. This result is reasonable because BHI introduces a temporal restriction. The number of matched keyframes declines and the scores reflect the new scenario.

[6] https://github.com/javierip/flash-video-summarization

The values provided by κ are consistent with those of F-score. Although there are no practical differences between κ and F-score, we believe that in the cases of videos longer than those provided by OVP, values will differ since the total number of frames is taken into account in the case of κ. Another advantage of κ is that the metric will be greater that 0 in the cases where there are no matches. This topic may be subject to further research.

5 Conclusions

A completely open source VS evaluation framework was presented in this work. The design is compatible with existing proposals available in the literature. Popular datasets, evaluation methods, and metrics were considered. Moreover, a new evaluation method (BHI) that takes into account the temporal distance between keyframes was introduced. Cohen's kappa coefficient was introduced in VS quantitative evaluation as a metric. Multiple barriers that make VS quantitative evaluation difficult to achieve were removed, or at least lowered.

Results show that OST reports unbiased metrics and it is dataset agnostic. The experiments reproduce previous results in a consistent way. Additionally, new quantitative results based on already published FLASH algorithm were also presented. As future work, we expect to integrate new datasets into the OST repository and add a consistent collection of external results. Finally, we invite researchers to use OST and to make progress in the field based on a common reference framework for evaluation.

References

1. De Avila, S.E.F., Lopes, A.P.B., da Luz Jr, A., de Albuquerque Araújo, A.: Vsumm: A mechanism designed to produce static video summaries and a novel evaluation method. Pattern Recognition Letters **32**(1), 56–68 (2011)
2. Furini, M., Geraci, F., Montangero, M., Pellegrini, M.: Stimo: Still and moving video storyboard for the web scenario. Multimedia Tools and Applications **46**(1), 47 (2010)
3. Gygli, M., Grabner, H., Riemenschneider, H., Van Gool, L.: Creating summaries from user videos. In: European conference on computer vision. pp. 505–520. Springer (2014)
4. Iparraguirre, J., Delrieux, C.: Speeded-up video summarization based on local features. In: Multimedia (ISM), 2013 IEEE International Symposium on. pp. 370–373. IEEE (2013)
5. Mundur, P., Rao, Y., Yesha, Y.: Keyframe-based video summarization using delaunay clustering. International Journal on Digital Libraries **6**(2), 219–232 (2006)
6. Song, Y., Vallmitjana, J., Stent, A., Jaimes, A.: Tvsum: Summarizing web videos using titles. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5179–5187 (2015)
7. Truong, B.T., Venkatesh, S.: Video abstraction: A systematic review and classification. ACM transactions on multimedia computing, communications, and applications (TOMM) **3**(1), 3 (2007)
8. Zhang, K., Chao, W.L., Sha, F., Grauman, K.: Video summarization with long short-term memory. In: European conference on computer vision. pp. 766–782. Springer (2016)
9. Zhou, K., Qiao, Y., Xiang, T.: Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)