Modelling perceptions on the evaluation of video summarization

Kalyf Abdalla[a,b], Igor Menezes[c], Luciano Oliveira[a]

[a] *Intelligent Vision Research Lab*
Federal University of Bahia
[b] *Federal Institute of Bahia*
[c] *University of Hull*

Abstract

Hours of video are uploaded to streaming platforms every minute, with recommender systems suggesting popular and relevant videos that can help users save time in the searching process. Recommender systems regularly require video summarization as an expert system to automatically identify suitable video entities and events. Since there is no well-established methodology to evaluate the relevance of summarized videos, some studies have made use of user annotations to gather evidence about the effectiveness of summarization methods. Aimed at modelling the user's perceptions, which ultimately form the basis for testing video summarization systems, this paper seeks to propose: (i) A guideline to collect unrestricted user annotations, (ii) a novel metric called compression level of user annotation (CLUSA) to gauge the performance of video summarization methods, and (iii) a study on the quality of annotated video summaries collected from different assessment scales. These contributions lead to benchmarking video summarization methods with no constraints, even if user annotations are collected from different assessment scales for each method. Our experiments showed that CLUSA is less susceptible to unbalanced compression data sets in comparison to other metrics, hence achieving higher reliability estimates. CLUSA also allows to compare results from different video summarizing

*Corresponding author: Luciano Oliveira.
Email address: kalyfabdalla@gmail.com, igorgmenezes@gmail.com, lrebouca@ufba.br (Luciano Oliveira)

approaches.

Keywords: video summarization, subjective evaluation, evaluation metric

1. Introduction

Streaming services can often suggest videos by popularity and supposedly related to user preferences, with the goal of saving the users' time in the searching process. It is known, for instance, that Youtube users spend one billion hours watching videos daily (Youtube, 2018). With a growing number of videos being made available on a daily basis, recommender systems are an important method to help users choose suitable videos.

Recommender systems usually rely on summarization techniques in order to extract useful information in videos. By analysing video content and their patterns of interaction, video summarization can be considered as a type of expert system able to retrieve relevant information from an input video by means of a relevance score estimation (Gygli et al., 2015; Demir & Bozma, 2015; Wang et al., 2011), as illustrated in Fig. 1.

Video summarization techniques create automatic video summaries by meeting three requirements: The presence of relevant video entities and events, elimination of redundant information, and generation of as much useful information as possible (Truong & Venkatesh, 2007). Truong & Venkatesh (2007) describe some video summarization applications such as **browsing and retrieval**, which is responsible for assisting users on searching and browsing tasks (Awad et al., 2017b; Arman et al., 1994; Zhang et al., 1997; Haojin Yang & Meinel, 2014), **computational reduction and content analysis**, used on semantic abstraction of information to reduce the computational complexity (Plummer et al., 2017), **story navigation and video editing**, which help users on navigating through a video (Nguyen et al., 2012), and **highlighting**, targeted on detection of important events in videos (Yao et al., 2016; Gygli et al., 2014; Xiong et al., 2003). On each of these applications, video summarization techniques try to mimic the ways humans comprehend the most important parts of a video.

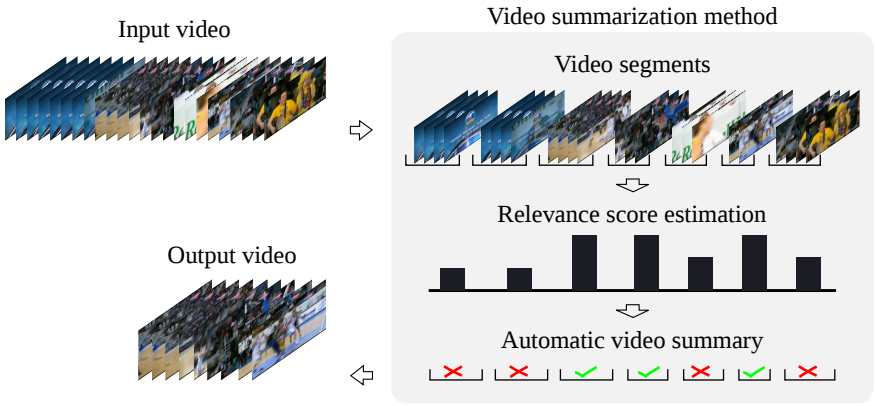


Figure 1: Generic video summarization pipeline. Video segments are generated from the input video. A specific metric scores the relevance of each segment, choosing the video summary, as the output.

According to Roberson (2013), relevance is the perception of what something is interesting and worth knowing, and depends on various individual and cultural aspects. As video summarization is commonly targeted at users, its usefulness is affected by the users' perception of what is either relevant or not in a video. For instance, on sport matches, player substitution information is useful for some users, while useless for others. Thus, the goals that a summarizer is expected to reach change according to the actual convenience (Awad et al., 2017a).

He et al. (1999) suggest that users instinctively follow four separate, but complementary criteria to judge relevant information in videos: Conciseness, coverage, context and coherence. While **conciseness** is related to the length of video summaries, **coverage** has to do with the abstraction level of information, and **context** and **coherence** are inherently related to the flow of information and how a story is told. Although users commonly apply all criteria when judging information, some of them prevail on specific applications of video summarization. For instance, coherence is not as important to story navigation as it is to content analysis. As a consequence, the evaluation of video summarization techniques becomes application-dependent, resulting on several different evaluation methodologies, which are usually applied in three different ways (Truong

& Venkatesh, 2007): **Result description** analyzes the behavior and advantages of automatic video summarization methods, **objective metrics** compare methods being evaluated with a heuristic summarization, and **user studies** measure the extent to which methods are consistent with summaries annotated by humans. Each of these evaluation perspectives has their own limitations that may impinge on a proper application of current techniques. As a result, studies on video summarization typically choose a methodology that is best suited for their research purposes in the field. In this work, important contributions to the task of measuring the performance of video summarization systems are introduced, as discussed henceforth.

1.1. Related work

Earlier works had no annotated data sets to measure the performance of summarization techniques (Liu et al., 2003; Wang et al., 2011). Accordingly, the way found to evaluate proposal methods was limited to describe the advantages and weakness of each one (Xiao-Dong Yu et al., 2004). The result description became inadequate to evaluate video summaries (Truong & Venkatesh, 2007), whether because there were no experimental arguments to enhance the reliability of results, or owing to very much subjective descriptions. As those descriptions were a viewpoint of authors about their own results, evaluation could be biased toward some methods rather than others. Because of that, each video summarization work sought alternative ways to evaluate novel methods (Huang et al., 2004; Taskiran, 2006; Gygli et al., 2014; Sharghi et al., 2017). For some video summarization applications (Truong & Venkatesh, 2007), a solution found to overcome the limitation of result description was to measure the quality of storyboard video summaries directly. In practice, evaluation methodologies compare the estimation of automatic techniques with an objective function (Tiecheng Liu & Kender, 2002), which ultimately matches a known summarization heuristic. Presuming that an heuristic describes optimal video summaries adequately, the evaluation of summarization techniques is limited to the heuristic used at that moment. In general, these heuristics are determined by the

occurrence of appealing objects or events for a target application (Wu et al., 2015; Peng Chang et al., 2002; Xiong et al., 2003) meant to be similar to the goals of specific tasks. For instance, on a highlighting task (Yao et al., 2016; Chong-Wah Ngo et al., 2003; Zhao & Xing, 2014; Sun et al., 2014; Peng Chang et al., 2002), video summarization techniques can focus on specific events, such as players scoring a goal or making a move to summarize sport matches. On browsing applications (Ziyou Xiong et al., 2006), continuous representations of frame dissimilarities can be used to determine how to cluster frames and choose representative ones. As there are several possible applications to use this evaluation approach, objective metrics have become an effective way to assess video summarization performance. On general cases, unfortunately there is no guarantee that any heuristic used to summarize a video matches human judgments properly (Truong & Venkatesh, 2007), which are ultimately affected by several factors such as the video content domain.

Working on the assumption that video summarization is targeted at users, who are actually able to determine what is relevant or not in a video summary, a third way to evaluate video summarization methods is by investigating the user's perceptions (Sundaram & Chang, 2001; Agnihotri et al., 2004). The first studies under this approach were performed by asking users to judge the results of each video summarization method (Liu et al., 2003; Taskiran, 2006; Chu et al., 2015). Recent works have opted to collect annotated video summaries from users, comparing those against automatic video summaries. As such, the evaluation is carried out from the users' judgments, which are currently the most pragmatic way to evaluate video summarization methods (Truong & Venkatesh, 2007). Hence, a growing number of summarization studies (Yong Jae Lee et al., 2012; Liu et al., 2015; Gygli et al., 2014; Song et al., 2015; Chu et al., 2015; Kim et al., 2014) have collected user annotations with the intent to improve the evaluation of video summarization methods, approach to which this study is particularly concerned.

There are some ways to collect user's annotations on relevant videos, being browsing logs, text annotation and relevance scores the most common ap-

proaches. Browsing logs are mostly used for specific video summarization applications such as story navigation (Wang et al., 2011); they provide less information about the relevance of video elements but are not suitable for highlighting applications. A faster and cheaper way to collect video summaries is through annotating textual video summaries. In this case, users annotate what information is relevant using natural language, or keyword tags. The compilation of these tags compose a textual vocabulary previously built for the annotation process (Sharghi et al., 2017). As a consequence, the performance of automatic methods is limited to the detection of the elements in textual vocabulary. The last common way to collect user annotations is the relevance scores, which focus on a dense annotation of the whole video. The main limitation of this approach, however, is the difficulty in collecting all of the user's annotations about each video frame.

In this study, the main goal is the annotation of relevance scores for **video skim** techniques, which ultimately deal with video segments (sequential frames grouped by contextual similarity). In video skimming, users are expected to watch and judge the relevance of a few video segments, instead of each frame in a video, reducing the amount of data to be annotated. Despite being a seemingly simple task, building up a large data set from video frame annotations is very costly and time-consuming. As a result, current video summarization data sets are usually limited to a small number of videos. For instance, Song et al. (2015) collected 20 user annotations for 50 videos, whereas Gygli et al. (2014) only collected 15 user annotations. Table 1 summarizes the main characteristics of the current the most recent studies in this field; from left to right, the columns represent: (i) The amount of search queries used to define the video domain, (ii) the amount of samples on each knowledge domain, (iii) the amount of participants in the user studies, (iv) what instruments were used to comprehend the user's subjectivity, (v) the annotation constraints used to control the annotation process, and (vi) what evaluation metric was used to match automatic methods to the collected user annotations. Table 1 highlights the three most commonly used benchmark data sets in the field of video summarization: SumMe (Gygli

Table 1: Characteristics of benchmarking data sets on the video summarization literature.

Work	(i) #Queries	(ii) #Videos	(iii) #Users	(iv) Annotation instrument	(v) Annotation constraints	(vi) Evaluation metric
(Gygli et al., 2014)	25	25	16	dichotomous scale	$5\% \leq L \leq 15\%$	F_β
(Song et al., 2015)	10	50	20	5-degree polytomous scale	$1\% \leq p(5) \leq 5\%$ $5\% \leq p(4) \leq 10\%$ $10\% \leq p(3) \leq 20\%$ $20\% \leq p(2) \leq 40\%$	F_β
(Chu et al., 2015)	10	51	3	3-degree polytomous scale	pooling segments with at least two judges	F_β
Ours	2	4	15	dichotomous, 3-degree and 5-degree polytomous scales	No conciseness constraints	CLUSA

et al., 2014), TVSum50 (Song et al., 2015) and CoSum (Chu et al., 2015) data sets, respectively. For comparison reasons, the characteristics of our data set is included in the table.

In spite of the fact that the benchmark data sets in Table 1 have similar goals, each sets different ways to collect user annotations and control the annotation process, as can be observed in column (v). SumMe, TVSum50 and CoSum data sets were gathered by limiting the length of annotated video summaries, by defining probability constraints for the assessment values, and by pooling user annotations from several users in a single one ground truth, respectively. The assessment scales used on each data set is also different, according to column (iv): Annotations on SumMe data set were collected via dichotomous scale, complying with users that ultimately define which video segments should be in the video summary. In contrast, TVSum50 and CoSum annotations use a degree scale, providing more freedom for users to annotate the subjective relevance of video segments.

1.2. Contributions

In addition to the limited number of annotations on current summarization data sets, there is no consensus as to how to collect user's perceptions in this field. Different studies typically deploy distinct techniques to deal with annotation and evaluation, making the achievement of a consolidated benchmark somewhat elusive. With regard to the annotation process, the methods that

have been proposed ignore how mind state, mood, tiredness and personal biases affect annotations, focusing their attention primarily on the user's responses. Considering that summarization aims to reduce the complexity of information, and optimal summaries should be as short as possible, the way found to control subjectivity has been to restrict the conciseness of annotated video summaries. In video skimming works, this is performed by setting the frequency of each relevance on the assessment scale. For instance, annotations on SumMe data set (Gygli et al., 2014) restricted the percentage of relevant video shots to 15% of the video length on a dichotomous assessment scale - a binary scale for annotation. On the other hand, annotations on TVSum50 data set (Song et al., 2015) were limited to a certain frequency by using prearranged *ad hoc* values on an assessment scale with five degrees of relevance levels. With respect to CoSum data set (Chu et al., 2015), the authors collected just annotations from 3 users, which turns this data sets unfeasible to be evaluated as it is. Regardless the assessment scales, the evaluation of video summarizers is performed by matching their generated summaries to user annotations, in general with F_β metric (Song et al., 2015; Gygli et al., 2014). This is done by taking into account absolute errors for each relevance level of user annotations. Although some errors have a common pattern, the relevance values may change from person to person, and as a consequence, F_β scores tend to be low, even in consistent user annotations. To the best of our knowledge, there are no studies in video summarization that considers such a user behavior, particularly considering the use of a polytomous assessment scale in video summarization tasks.

To cope with the aforementioned limitations, our work brings three contributions: (i) A guideline to collect unrestricted user annotations in order to diversify the conciseness of video summaries (Section 2), (ii) a metric to evaluate automatic video summaries against user annotations (Section 3), these latter collected with different assessment scales – our propose metric, named compression level from user annotation (CLUSA), is able to handle with unrestricted conciseness of video summarization task, in contrast with other works (Gygli et al., 2014; Song et al., 2015), and (iii) a study on the quality of an-

notated video summaries collected from different assessment scales, including suggestions regarding the diversification of the conciseness of video summaries in order to improve the evaluation of our proposed metric (Section 4); (ii) and (iii) use SumMe, TVSum50 and our data sets to assess the performance of CLUSA, discarding CoSum due to the small number of users that annotated this data set. A novel evaluation methodology for video summarization metrics that considers the user subjectivity, and annotates the relevance of video segments, is introduced in Section 5.

Even if future works opt not to follow our suggested guideline to collect user annotations, our proposed evaluation metric (CLUSA) provides a benchmark for automatic video summarization methods against user annotations collected with both dichotomous and polytomous assessment scales. A thorough discussion on these and other topics can be found in Section 6.

2. Guideline for annotation process and subjective measurement

The quality of user annotations is inherently related to the method deployed for data collection. Hence, any bias in this collecting process may hinder the reliability of annotations, and therefore the evaluation of the assessment scales. To circumvent this problem, we describe the entire process that involves the collection of unrestricted user annotations, identifying what issues on the annotation process of video segments are likely to occur.

2.1. Determining what users annotate

To determine what video segments are to be annotated by the users, the target videos have to be processed by a baseline boundary video shot detector (Gygli et al., 2014), or by a uniform sampling (Song et al., 2015). Since different users have to annotate the same video segments, the detection of these boundaries is performed off-line following some heuristic, such as motion, object detection and/or frame similarity (Yuan et al., 2007; Pal et al., 2015; Hanjalic, 2002).

There is no a ultimate heuristic to detect the boundaries of any video, thus the evaluation of video summarization is limited to the shot boundaries used in the annotation process. In other words, the user annotations are collected with shot boundaries different from that used in the automatic video summarization methods. The solution adopted to match automatic video summaries to user annotations with different video segments is to use small-length video segments, and perform a segment-to-frame mapping.

From the user's viewpoint, the length of video segments is associated with the time required to complete the annotation process. The longer the annotation process, the fewer complete annotations are collected from the users. This situation occurs because users tend not to complete long tests. For instance, considering a two-second uniform sampling such as the one used in (Song et al., 2015), users had to annotate more than one hundred video segments for videos with approximately five minutes. Time to collect annotations in video segments must be reduced in order to avoid user withdrawal. The easiest solution to this problem is to increase the length of video segments, searching for a trade-off between brevity and reliability. If the goal is to build more reliable user annotations, authors can shorten the video segments to produce more annotated video segments. Alternatively, brevity makes it easier to collect more complete annotations at the expense of more reliable annotated video summaries.

2.2. Preparing users to annotate the video segment

Users are expected to express their opinions about the relevance of each video segment as sincerely as possible, however hard it may be to fully guarantee them. Psychometrics suggests that task instructions are one of the important elements that may affect the reliability of user annotations (Rosner & Cronbach, 1960). When users do not comprehend what they are supposed to do, user annotations provide reduced information regarding the evaluation of video summaries or any other psychological characteristic. In face of that, our study introduced clear and unequivocal instructions to the respondents in order to control for potential biases stemming from random responses. Hence, the goal of video

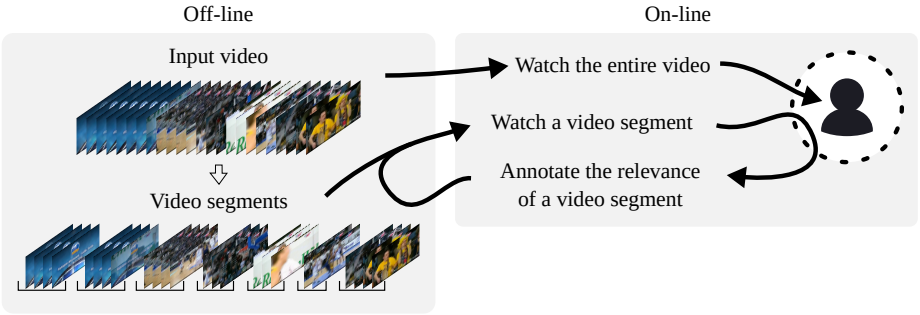


Figure 2: Annotation process flow to collect user annotations.

summarization was clearly explained at the very beginning of the session, and so how users should proceed in each step of the annotation process. As some respondents might feel tempted to skip the instructions, we set up a system to prevent this behavior from happening. This constraint was therefore done by our annotation tool[1].

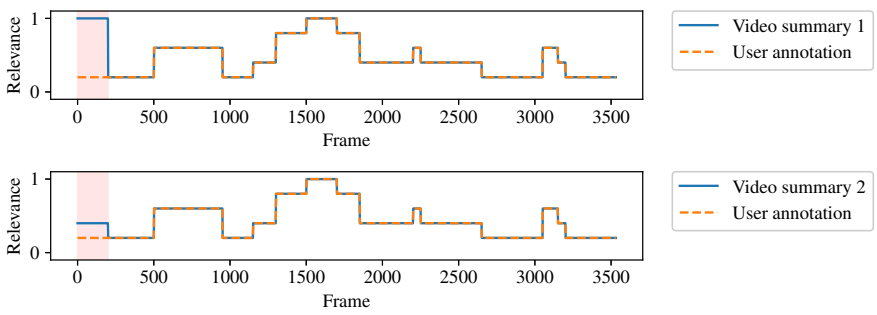
By watching entire videos, users are able to abstract and make sense of their context. However, there is no guarantee that user annotations are driven solely by the context of each video as previous knowledge and opinions about the content of the video can ultimately influence the user's perception. In order to tackle this problem, users were compelled to watch the entire video, and then judge the relevance of each segment. Since the sequence that video segments are presented to users also affects the data annotation (Song et al., 2015), and users tend to annotate higher relevance scores to the video segments that appear earlier, we presented the video segments randomly. Then users annotated the video segments continuously until having the entire video completed, according to the flow illustrated in Fig. 2. Also the video segments were muted in order to allow the users to focus their attention only on the visual stimuli.

[1]Paper is currently under review. The tool will be made publicly available once accepted.

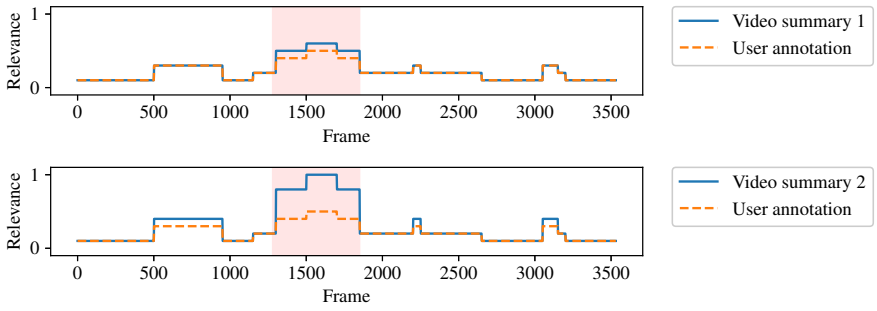
3. Evaluating automatic video summarization methods

Once user annotations are collected, they are used to assess the performance of automatic video summaries using an evaluation metric. Since the very first goal of video summarization tasks is to determine which video segments are relevant to users, state-of-the-art works treat the evaluation of automatic video summaries as a classification problem, either binary or not. For non-binary classification (*i.e.*, multi-label), an evaluation metric matches the expected relevance label of video segments to the users' annotated label. Here we identified three issues on the current approaches to evaluate video summarization tasks: **Degree of error**, which measures how far the estimated relevance is from the expected one for each video segment (see Fig. 3(a)), **correlation of relevance scores**, since different relevance estimations could produce exact video summaries (see Fig. 3(b)), and **relevance weighing**, determining which relevance levels suit individual video summarization the best (see Fig. 3(c)). For multi-label classification, the evaluation metrics consider the relevance scores as labels, but users do not perceive the relevance as such. As a consequence, the user annotations keep changing, harming the label matching and the evaluation of automatic video summaries. For evaluation metrics based on classification, any label different from the expected is treated as an error, therefore ignoring the degree of abstraction's relevance, which is useful to rank video summarization methods.

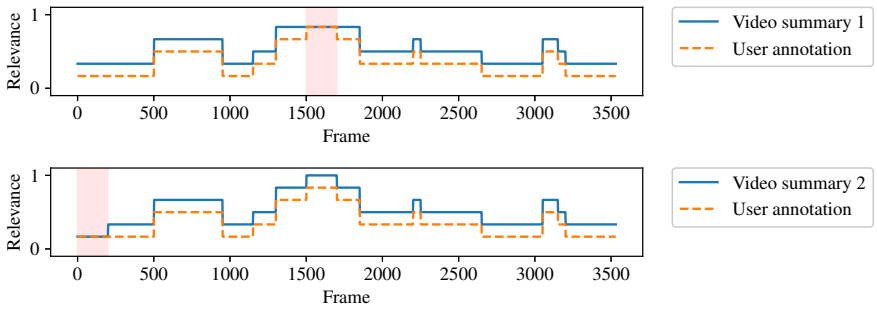
Let us take the examples shown in Fig. 3. Current metrics considers that both video summaries are equal, even if the first video summary exhibits more relevance than the second one (see Fig. 3(a)), as can be seen in the pink areas in sub-figures. An easier way to deal with the degree of error is considering video summarization tasks as regression problems instead of classification, like the reconstruction error metric, so the evaluation metrics measure dissimilarity distances from the user annotations. However, those video summaries that seem to be dissimilar produce exact binary automatic video summaries by retrieving the most relevant video segments, as depicted in Fig. 3(b). This occurs because



(a) **Degree of errors:** Equal evaluations with different degrees of error.



(b) **Correlation of relevance scores:** Different evaluations, but equally correlated.



(c) **Relevance weighting:** Equal evaluations hitting different relevance levels.

Figure 3: Evaluation issues identified in current video summarization metrics.

there is a direct monotonic relationship between both relevance scores. Besides, regression metrics are not able to weigh degrees of relevance, since video summarization tasks prioritize higher relevant video segments than lower ones, as

illustrated in Fig. 3(c). All things considered, here a novel metric is proposed to overcome these limitations of the previous ones.

3.1. CLUSA: Compression Level from USer Annotation

Let be $\mathbf{m} = (m_j) \in \mathbb{R}^K$ a vector containing K relevance scores of the video segments provided by an automatic video summarization method. In order to properly assess the performance of video summaries, a set $\mathbf{D} = (d_{i,j}) \in \mathbb{R}^{U \times K}$ of annotations by U users is required that ultimately represents the relevance for each video segment. $\mathbf{m}$ denotes the scores of a binary classifier, following the assumption that video summary techniques select a set of video segments by relevance. Additionally, user annotations contained in $\mathbf{D}$ can be in polytomous scales, with a preprocess step being required. In this case, the annotations $\mathbf{D}$ are binarized into $\mathbf{O}_i$ summaries, considering the unique relevance levels in each row, $\mathbf{u}_i$, as

$$\mathbf{u}_i = \{d_{i,j} : \forall j, 1 \leq i \leq U, 1 \leq j \leq K\}\,. \tag{1}$$

So, each value in $\mathbf{u}_i$ is used on thresholding the $\mathbf{O}_i$ matrices, as illustrated in the top-down example in Fig. 4, starting on the annotation and applying thresholds of 0.2 and 0.6, respectively. It is noteworthy that the highest values in $\mathbf{u}_i$ are not used since it leads all values in $\mathbf{O}_i$ to zero. $\mathbf{O}_i$ is given by

$$\mathbf{O}_i = ([\mathbf{D}_{i,j} \geq \mathbf{u}_{i,k}] : 1 \leq k \leq |\mathbf{u}_i|) - 1 \in \mathbb{R}^{|\mathbf{u}_i|-1 \times K}\,. \tag{2}$$

Each user annotation is mapped onto $\mathbf{O}_i$ summary matrices, which are concatenated into a single matrix, $\mathbf{X} = (x_{i,j}) \in \mathbb{R}^{(\sum(|\mathbf{u}_i|-1) \times K}$. All this preprocessing step builds a set of binary annotations (as can be seen in each row of Fig. 5(a)), $\mathbf{X}$, from the user annotations, and $\mathbf{D}$ (illustrated in each row of Fig. 5(b)). By proceeding in this way, we are able not only to normalize the relevance scores, but also to address the degree of error and correlation issues on the score matching.

As each row-vector, $\mathbf{x}_i \in \mathbf{X}$, denotes a binary form obtained from user annotation, now we are able to compute a matching score vector, $\mathbf{z}_i$, given by

$$\mathbf{z}_i = (\theta(\mathbf{m}, \mathbf{x}_i) : 1 \leq i \leq \textstyle\sum |\mathbf{u}_i|)\,, \tag{3}$$

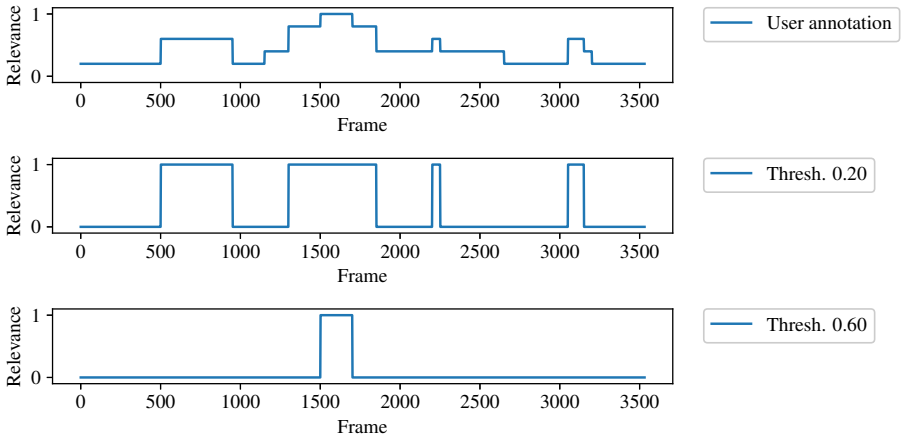
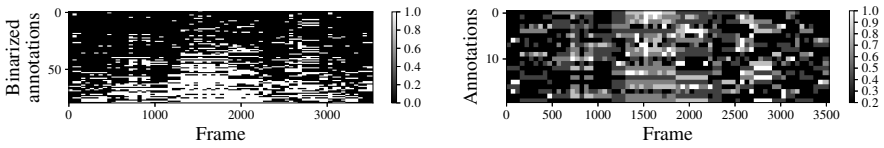


Figure 4: Thresholding an user annotation into several relevance levels.



(a) Binarized user annotations.

(b) Data set of user annotations for a target video.

Figure 5: Illustration of (a) the result map of the preprocessing step from (b) the discretized map of a user annotation.

where θ is a vanilla function, which matches $\mathbf{m}$ with $\mathbf{x}_i$ values. Instead of using F_β, we decided to use the area under curve from a receiver operating characteristic (ROC) curve to measure the **degree of error**. The use of ROC curve allows to evaluate the relevance scores, $\mathbf{m}$, given by a video summarization method on each binary video summary, $\mathbf{x}_i$. Indeed, the ROC curve allows to identify the thresholding values that maximize the matching with the binary video summaries. If there is an exact monotonic association between the annotated and the estimated relevance, all areas under ROC curve reach the maximum area, $z_i = 1$, addressing the issue of the **correlation of relevance scores**. This process is depicted in Fig. 6: the ROC is used to map the $thresh_i$ values from

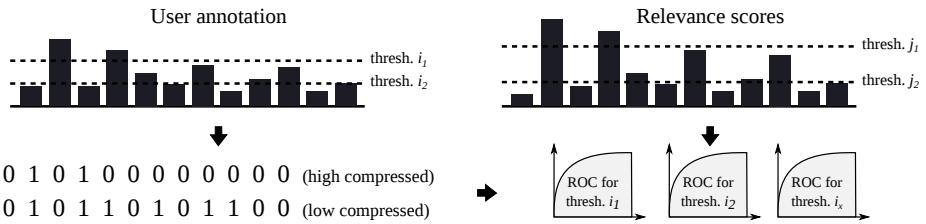


Figure 6: To match user annotation and estimated relevance scores, CLUSA metric computes area under ROC curves from the binary video summaries associated to a compression rate.

user annotations (on the left), to $thresh_j$ values from automatic methods (on the right).

Relevance weighting is not possible to be directly carried out over $\mathbf{X}$ data, since this matrix is result of the binarization of the user relevance values. Hence, we proposed to calculate the ratio between the amount of video summary and the entire video. This relation is here called compression rate, $\mathbf{w}_i = P(\mathbf{x}_i = 0)$, which represents the probability of the video segments not to be included in the video summary by the user annotations. Although compression and relevance are different concepts, in practice they are related: The most relevant video segments are consistent with the highest compressed summaries. Moreover, the compression weighting is expected to be a strong candidate for the evaluation of a video summary. In other words, the compression weighting overcomes the **relevance weighting** issue.

The sets $\mathbf{D}$ can concentrate user annotations in specific compression intervals due to annotation constraints of each video summarization work. This strategy limits the ranking of the video summarization methods on different data sets. In order to circumvent this, the score vector, $\mathbf{z}_i$, is grouped into clusters, $\mathbf{c}_i$, according to the compression rate of each annotated video summary, $\mathbf{w}_i$, and is defined as

$$\begin{aligned} \mathbf{c}_i = \mu(\mathbf{z}_k : \|\mathbf{w}_k - \mathbf{p}_i\|^2 \leq \|\mathbf{w}_k - \mathbf{p}_j\|^2, \forall j, 1 \leq i \leq j \leq B, \\ 1 \leq k \leq \textstyle\sum |\mathbf{u}_i|) \,, \end{aligned} \tag{4}$$

where B represents the number of compression intervals, while $\mathbf{p}_i$ is a median

point in each interval, given by $\mathbf{p}_i = (2i-1)(2B)^{-1}, 1 \leq i \leq B$. The clusters, $\mathbf{c}_i$, are suitable to assess the quality of the scope of each video summarization work. That is to say that we are now able to compare techniques considering the most relevant video segments. $\mathbf{c}_i$ assists in the interpretation of the results of a video summarization method, but it does not allow for a direct single score to evaluate the summarization performance. To obtain this single score, the compression score vector is weighed, $\mathbf{c}$, comprised of $\mathbf{c}_i$ values, by the median points, $\mathbf{p}$, comprised of $\mathbf{p}_i$. At the end, CLUSA metric is defined as

$$\text{CLUSA}(\mathbf{D},\mathbf{m}) = \mathbf{p}^T\mathbf{c}\,. \tag{5}$$

Since CLUSA metric does not require that $\mathbf{D}$ and $\mathbf{m}$ be on the same assessment scales, our proposed metric is expected to set a benchmark for different video summarization methods and data sets.

4. Evaluating the quality of assessment scales on video summarization tasks

To evaluate the performance of automatic video summarization methods, first the quality of the collected annotations is needed to be ensured. In psychometrics, two parameters are usually pursued as indicators of the quality of annotations: (i) **Test validity** refers to whether or not the test or any of its items measures the characteristic intended to be measured (in particular, whether different video segments are consistent with the video relevance provided by the users) and (ii) **test reliability**, which seeks to investigate the precision or internal consistency of test scores (specifically, how much users agree on video relevance). Test validity is used to compare different annotation guidelines, while test reliability computes user consistency on a specific guideline. The latter is the most suitable indicator to analyze the quality of assessment scales on video summarization tasks.

For test reliability, there are two main approaches to investigate the agreement of user annotations: (i) **stability over time**, where the aim is to evaluate

annotations collected with a time interval between them, being test-rest with the same users considered the most suitable parameter for video summarization, and (ii) **internal consistency** that evaluates the quality of annotations under cross-sectional perspective, collecting user annotations only once. The main problem with the former is the difficulty to find users engaged to repeat the annotation process. In practice, the main limitation resides in contacting users to guarantee repeatibility. In view of that, internal consistency ends up to be the main approach to measure the user agreement on the relevance of video segments using different assessment scales.

There are several ways to measure internal consistency, being Kuder-Richardson and Cronbach's alpha (Rosner & Cronbach, 1960) the most widely used coefficients. Whereas Kuder-Richardson is used for dichotomous scales, Cronbach's alpha is deployed for polytomous assessment scales. As Cronbach's alpha equation is derived from Kuder-Richardson's, as well as the interpretation of both coefficients is in the same directly comparable scale, then the name Cronbach's alpha was used in this study to refer to both types of internal consistency, whether applied to dichotomous or polytomous scales. Therefore, Cronbach's alpha, α, measures the reliability of K video segments according to

$$\alpha = \frac{K}{K-1}\left(1 - \frac{\sum_{j=1}^{K} \sigma_{D_j}^2}{\sigma_D^2}\right), \tag{6}$$

where the variance of the j-th video segment, $\sigma_{D_j}^2$, is divided by the annotation variance, σ_D^2.

The reference values to evaluate the Cronbach's alpha is shown in Table 2. Since SumMe (Gygli et al., 2014) and TVSum50 (Song et al., 2015) provide a disjoint collection of videos, we are not able to compare directly the reported Cronbach's alpha on the annotation of these two data sets. This comparison could only be accomplished under a standardized scenario, involving the administration of the same video segments annotated by the same individuals. As we are interested in identifying which assessment scale is more suitable to video summarization tasks, we collected user annotations for a common data set,

Table 2: Reference values to evaluate Cronbach's alpha estimations.

Cronbach's alpha	**Internal consistency**
$0.9 \leq \alpha$	Excellent
$0.8 \leq \alpha < 0.9$	Good
$0.7 \leq \alpha < 0.8$	Acceptable
$0.6 \leq \alpha < 0.7$	Questionable
$0.5 \leq \alpha < 0.6$	Poor
$\alpha < 0.5$	Unacceptable

following our guideline described in Section 2, considering three types of assessment scales: Dichotomous, three-point Likert and five-point Likert. Cronbach's alpha was then used as the measure of average quality of each assessment scale.

5. Experimental evaluation

Two experiments were carried out in this study: The first measures the quality of our collected user annotations with different assessment scales, and the second evaluates CLUSA performance regarding its internal consistency in comparison to other metrics. The former experiment was devised to investigate the most adequate assessment scale to collect annotations on video summarization tasks, while the latter aimed at evaluating how CLUSA performs in face of different scales, as well as how compression in video summaries affects the performance of automatic video summarization methods.

5.1. Collecting user annotations

The evaluation of the quality of user annotations is usually done for each annotated video (Gygli et al., 2014; Song et al., 2015). This is because users are not necessarily the same on different videos, and the Cronbach's alpha has to be calculated for each annotated video. Here we propose to collect user annotations on a standardized scenario where the same users annotated the same videos using different assessment scales. With that, we are able to relate the quality of

(a) (b)

Figure 7: Samples of used videos from UCF101 data set: (a) Surfing and (b) basketball.

user annotation comparatively, measured by the average Cronbach's alpha of each assessment scales. Since the annotation collected on a standardized scenario is an arduous workforce for the users, the amount of annotated videos was reduced to avoid the users' withdraw during a long annotation processes (refer to a discussion in Section 2). From UCF101 collection (Soomro et al., 2012), although initially ten action videos were offered to the users to be annotated, just four of them was guaranteed to have annotations of all users on all videos using all assessment scales. Two types of actions were queried in UCF101 data set: **Surfing** and **basketball**, and some samples are illustrated in Fig. 7. We used videos whose duration was around three minutes with well-defined video shot boundaries, splitting the video into segments using a vanilla boundary shot detection, based on motion and frame similarity before the annotation process. In order to automatize this process and make it more reliable, we developed an annotation tool, complying with all the requirements described in our proposed guideline (see Section 2). In this stage, the analysis on the relation of the different assessment scales is not affected by the number of videos, being important just to guarantee that the same users annotated the same videos using different assessment scales.

5.2. Assessing the quality of data sets collected with different assessment scales

Users annotated the relevance of a specific arrangement of video segments, which were previously determined in the annotation task. However, automatic

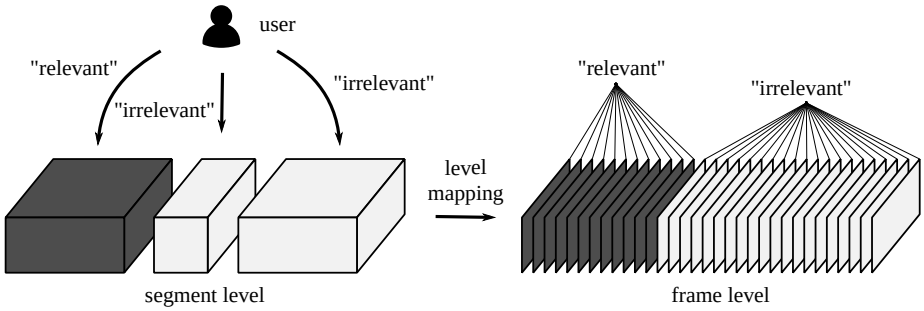


Figure 8: Mapping the annotated relevance at video segments to frame level (level mapping).

video summarization methods are not limited to the segments annotated by the users, but any resulting arrangement of a boundary shot detector. As a result, the relevance of video segments were mapped to the frame level to allow for the evaluation of automatic methods. As illustrated in Fig. 8, this segment-to-frame mapping is performed by repeating the annotated relevance of video segments in the video frames. The automatic video summarization methods are then evaluated regardless of their boundary shot detector.

Since data sets for video summarization typically do not provide the boundaries of the video segments to bring the level mapping back to the segment level, the quality of the user annotations was calculated at the frame level.

Psychometric estimators are computed at the level where users annotated the data, in our case, at the segment level. Therefore the Cronbach's alpha is computed on segment level to verify the difference in Cronbach's alpha values at both levels. Table 3 shows the Cronbach's alphas for our collected user annotations.

On our standardized scenario, with the same annotated videos and users, we are interested in the impact of assessment scales on video summarization tasks. Our user annotations were grouped by the assessment scale used on the annotation process: Dichotomous (dich) and polytomous (Likert-3 and Likert-5). Based on this, we were able to compare the Cronbach's alphas from different assessment scales directly. In Table 3, Cronbach's alpha values for our user annotations are observed to increase proportionally to the degree of the assessment

Table 3: Cronbach's alpha for different assessment scales: dichotomous (dich) and polytomous (Likert-3 and Likert-5).

Data set	**Assessment scale**	**Annotations per video (mean)**	**Cronbach's alpha (mean)**	
			Frame-level	**Segment-level**
Ours	dich.	16	0.712	0.718
	Likert-3	16	0.809	0.799
	Likert-5	16	0.842	0.833

scale, suggesting that polytomous scales are more suitable to collect user annotations for video summarization tasks. Five-point Likert turned out to be the most adequate assessment scale for video summarization tasks, considering the increase in the internal consistency. This finding does not rule out the potential use of higher degree assessment scales, though the increasing in user response time and overlapping responses between similar adjacent categories (*e.g.*, somewhat disagree versus slightly disagree) can be regarded as a deterrent to the use of higher degree scales.

5.3. Using CLUSA to obtain the internal consistency for video segments

In addition to the use of Cronbach's alpha to calculate the internal consistency of user annotations, F_β metric is also exploited to assess the internal consistency of user annotation in all state-of-the-art works. It is done by calculating the distances between pairs of users, as shown in Fig. 9(a). In a different way, CLUSA was originally conceived in a leave-one-out strategy, matching one user annotation to a collection of user annotations (see Fig. 9(b)). This leads to computing more compression scores, c_i, and compression rates, w_i, than those provided in a pair-wise strategy.

To evaluate CLUSA's performance, we compared our metric with Cronbach's alpha and F_β. CLUSA was also calculated in a pair-wise fashion. Table 4 summarizes the results. It is worth noting that F_β metric presents the opposite behavior (decreasing as the degree of assessment scales increases) with respect to Cronbach's alpha in our standardized scenario (row "Ours"); F_β indicates that

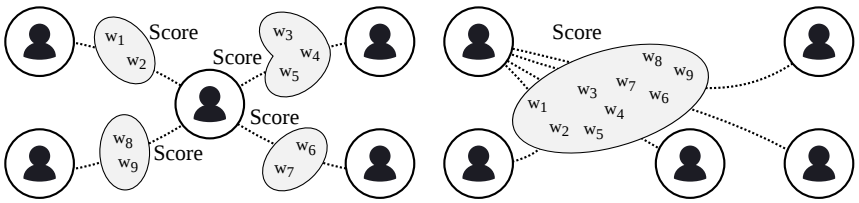


Figure 9: Approaches used to compute internal consistency using CLUSA: (a) Pair-wise and (b) leave-one-out.

dichotomous scale should be more consistent than polytomous. Conversely, the behavior of CLUSA became similar to Cronbach's alpha, suggesting that our proposed metric are closer to the psychometric studies when dealing with subjectivity on video summarization tasks.

Table 4 also summarizes the results of SumMe and TVSum50 according to the characteristics of each data set, rather than different assessment scales. SumMe data set is formed by three types of videos: Egocentric, moving and static, which were determined by the camera and scene motions, whereas TVSum50 collected user annotations for the following video contents: Changing Vehicle Tire (VT), getting Vehicle Unstuck (VU), Grooming an Animal (GA), Making Sandwich (MS), ParKour (PK), PaRade (PR), Flash Mob gathering (FM), BeeKeeping (BK), attempting Bike Tricks (BT), and Dog Show (DS).

CLUSA was also calculated on SumMe and TVSum50 annotations in order to investigate its behaviour in other scenarios. SumMe and TVSum50 annotations have been collected using different guidelines, video contents and users, and hence, we are not able to compare the Cronbach's alphas directly. In the leave-one-out strategy, CLUSA in both SumMe and our dichotomous user annotations approaches to 0.2, while in both TVSum50 and our Likert-5 user annotations, CLUSA is around 0.5, even considering the difference in the guidelines and video contents of the three data sets. This allows us to state that leave-one-out CLUSA is affected by the assessment scales rather than guidelines and video contents.

Table 4: Internal consistency using F_β and CLUSA in their respective assessment scales: Dichotomous (dich) and polytomous (Likert-3 and Likert-5)

Data set	**Assessment scale**	**Internal consistency**			
			Pair-wise		**Leave-One-Out**
		Cronbach's α	**F$_\beta$**	**CLUSA**	**CLUSA**
Ours	dich.	0.712	0.647	0.033	0.271
	Likert-3	0.809	0.516	0.066	0.432
	Likert-5	0.842	0.333	0.151	0.635
SumMe	dich. (ego)	0.766	0.292	0.103	0.212
	dich. (moving)	0.748	0.308	0.104	0.176
	dich. (static)	0.850	0.359	0.110	0.228
TVSum50	Likert-5 (BK)	0.791	0.377	0.338	0.505
	Likert-5 (BT)	0.871	0.385	0.357	0.550
	Likert-5 (DS)	0.760	0.350	0.319	0.494
	Likert-5 (FM)	0.789	0.367	0.323	0.486
	Likert-5 (GA)	0.866	0.394	0.362	0.533
	Likert-5 (MS)	0.826	0.380	0.338	0.529
	Likert-5 (PK)	0.741	0.359	0.308	0.494
	Likert-5 (PR)	0.813	0.378	0.332	0.533
	Likert-5 (VT)	0.875	0.410	0.359	0.540
	Likert-5 (VU)	0.783	0.367	0.332	0.495

Note that CoSum data set (Chu et al., 2015), which is often used as benchmark data set, was annotated by just 3 users. This number of user annotations directly affects the Cronbach's alpha and CLUSA analysis, hence making CoSum data set unsuitable for experimental analysis.

6. Discussion and concluding remarks

6.1. Quality of user annotations

Psychometrics studies suggest reference values for Cronbach's alphas that can be used to evaluate the quality of user annotations. Table 2 shows these references, with 0.7 as being the minimal "acceptable" score. All user annotations collected on our data set are above 0.7, assuring the minimum quality to

properly evaluate the assessment scales and CLUSA. According to the ranges in Table 3, our user annotations collected with dichotomous scale have lower quality in comparison to the ones collected with polytomous, both reaching 0.718 and 0.833, respectively. Since our collecting process is performed on a standardized scenario, with the same videos and users, the increase of quality in polytomous suggested that this assessment scale is more suitable to collect user annotations for video summarization tasks.

User annotations in SumMe (Gygli et al., 2014) and TVSum50 (Song et al., 2015) data sets were collected with different guidelines, videos, users and assessment scales, and hence, we cannot guarantee the impact of changing assessment scales in their experiments. Although the behaviors of the internal consistency on the annotations of SumMe and TVSum50 data sets are not directly comparable, Cronbach's alphas on the two data set annotations behaved similarly to the results of our standardized scenario. The dichotomous scale in SumMe produced user annotations with lower values in comparison with user annotations collected with polytomous in TVSum50 data set, as can be seen in Table 4.

All user annotations in the three data sets (SumMe, TVSum50 and ours) were collected using a specific arrangement of video segments. Notwithstanding, automatic video summarization methods can use different boundary shot detection approaches, which result in different video segments than those annotated by users. To uniform the evaluation of different video summarization methods, user annotations are usually mapped from segment to frame level. The difference between the Cronbach's alpha values at both segment and frame levels is named Cronbach's alpha inflation, which impinges on the qualitative analysis, resulting on erroneous classifications of the annotation quality. In Table 3, row "Ours", the quality of the user annotations collected with three-point Likert scale was reduced from "good" to "acceptable", in frame and segment levels, respectively. This phenomenon occurs because the Cronbach's alphas were close to the quality classification boundary of 0.8. Note that Cronbach's alpha inflation alters the real quality of user annotations, but does not hamper the quantitative comparison of Cronbach's alphas at one level. Cronbach's al-

phas under dichotomous scales are lower than polytomous for both frame and segment levels in Table 3, then it is true to say we are able to quantitatively analyze the results even with the inflation issue.

6.2. Annotation consistency as compression scores

The growing of Cronbach's alpha in comparison to the assessment scale is the pillar of our analysis (see Table 4). Gygli et al. (2014) introduced the concept of human consistency by computing F_β. Considering our user annotations collected with different assessment scales (see Table 4), F_β behaved differently to the Cronbach's alpha: While Cronbach's alphas suggested that the quality of user annotations increased, F_β suggested the opposite. On the other hand, CLUSA coped with annotation consistency in a similar fashion to the Cronbach's alpha for both pair-wise and leave-one-out approaches.

CLUSA was conceived based on a leave-one-out strategy, because it uses all user annotations to compute the compression scores. As a consequence, CLUSA's performance with a pair-wise approach is lower than the leave-one-out one, as shown in Table 4. The worst CLUSA score was achieved with pair-wise strategy in the dichotomous scales. Because of the unrestricted scenario, the probability that a pair of users annotates the same compression rates are lower than considering the compression rates provided by all users at the same time. Yet for the dichotomous scale, the compression rates were sparse and focused on low compression, as illustrated in Figs. 10(a) and 10(d), where the compression scores (box plots) are concentrated on the left side of the plots (low compression scores). Since video summarization tasks pursue high compression scores, CLUSA penalizes all user annotations collected with dichotomous in comparison to the other assessment scales, as can be observed in all plots of Fig. 10, where the plots in Fig. 10(b) and 10(e) show the three-point Likert scale results. As illustrated in Figs. 10(c) and 10(f), the box plots occupy the entire x-axis, meaning that our proposed guideline with five-point Likert scale tends to collect user annotations on all available compression rates. Following that, we can state that leave-one-out CLUSA is able to evaluate all the conciseness

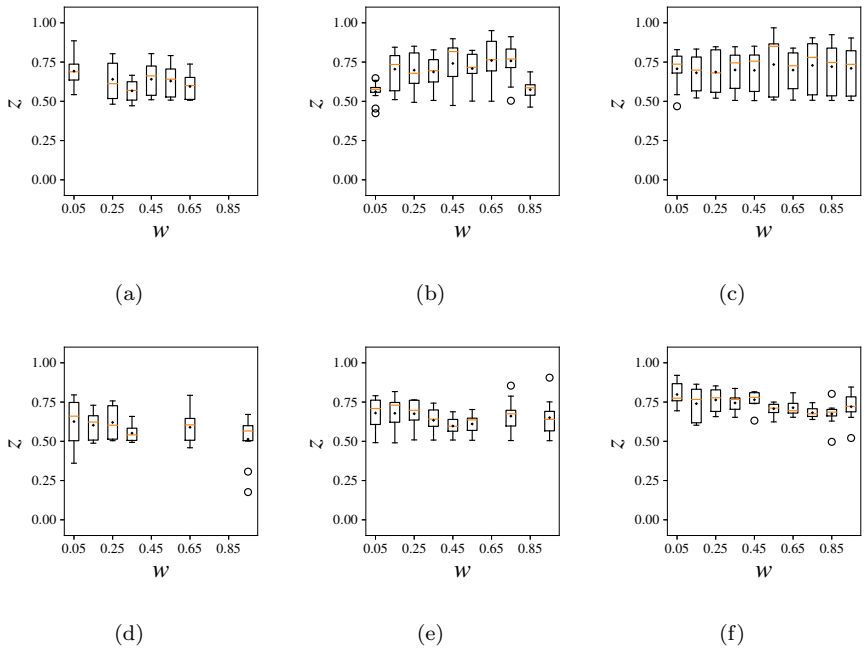


Figure 10: Relating z_i scores to w_i compression rates on collected annotations: (a) and (d) Dichotomous scale, (b) and (e) Polytomous Likert-3 scale, and (c) and (f) Polytomous Likert-5 scale, with leave-one-out approach.

of a video summary, even in an unrestricted scenario.

Considering the compression scores of one video annotation in SumMe and TVSum50 data sets, illustrated in Figs. 11(a) and 11(b), respectively, we can observe that the annotation process of SumMe and TVSum50 focused on high compression rates (blox plots more on the right of the x-axis). Even that TVSum50 occupies all the right side of x-axis, CLUSA does not penalize annotations in this data set. On the other hand, SumMe occupies only a small portion of the right side in x-axis, and hence the compression weight in Eq. 5 penalizes the SumMe results. These two situations are explained for the accumulated of the high compression weights that always corresponds to 75% of the CLUSA score. Because of that, CLUSA considers the high compression rates to be crucial to score the performance of an automatic video summarization method.

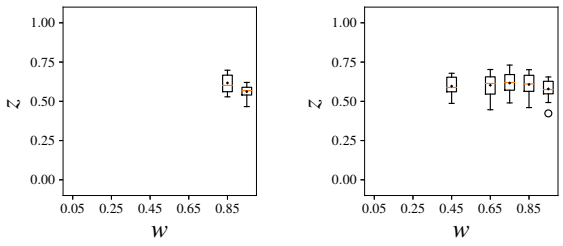


Figure 11: Relating z_i scores to w_i compression rates on user annotations collected on: (a) SumMe and (b) TVSum50 data sets.

6.3. CLUSA limitations

A possible weakness of CLUSA resides in the cluster scores, c_i, that should follow a normal distribution. In Figures 10 and 11, the scales of the box plots represent the variance of user annotations. Users diverge more about some video contents (see Figure 10(c)) than others (see Figs. 10(f) and 11(a)). Supposing that users annotate the video segments by biasing the relevance (a certain group saying that the segments are very relevant, and another group saying that it is very irrelevant), the cluster score would be no longer representative to evaluate the hypothetical user annotations. In that case, we suggest that future works explore non-normal distribution approaches or mixture of normal distributions.

6.4. Future work

Our study, as well as the previous studies introduced herein, assumed that all video segments are annotated from a single relevance perception, measuring the relevance of all objects in the scene together into a single relevance score. However, the relevance could be attached to a collection of visual elements in the video segment. So, in an alternative scenario, users should also describe these representative elements (*e.g.,* objects, places). For instance, regarding a video depicting images of surfing, beaches and surfers could be split between (i) landscape and (ii) bonds among surfers, so that some users could place more emphasis on environment (i), whereas others would consider relationships (ii)

as the most important characteristic of the video. This is already performed by video captioning tasks, and can be incorporated to our guideline on the video segment level to improve the interpretation of the results, improving also relevance laid by users, such as performed in Sharghi et al. (2017). In that case, CLUSA can be improved to perform also text matching, similar to matching metrics in natural language field.

The relation between the Cronbach's alpha and CLUSA was explored with the goal of analyzing the behavior of this novel metric, though there are several data sets not suitable for the computation of the Cronbach's alpha due to an insufficient number of collected user annotations (*e.g.*, (Chu et al., 2015)). A future work could be aimed at analyzing CLUSA's efficiency under such a restricted sampling scenario.

By showing the importance of establishing a research agenda able to surmount the limitations of previous studies conducted on video summarization, this work presented the findings focused on the development of a novel metric less sensitive to user annotations with unbalanced compression. Further investigations ensuing from this study should follow on the compilation of video summarization data sets and methods into benchmark testing, facilitating the evaluation of novel automatic methods in a model comparison perspective.

Acknowledgements

This work was supported by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) which provided scholarship to Kalyf Abdalla. Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) supported Luciano Oliveira under grant 307550/2018-4.

References

Agnihotri, L., Dimitrova, N., & Kender, J. R. (2004). Design and evaluation of a music video summarization system. In *Proceedings of International Conference Multimedia and Expo* (pp. 1943–1946). volume 3.

Arman, F., Depommier, R., Hsu, A., & Chiu, M.-Y. (1994). Content-based browsing of video sequences. In *Proceedings of the second ACM international conference on Multimedia* (pp. 97–103). New York, New York, USA: ACM Press. doi:10.1145/192593.192630.

Awad, G., Butt, A., Fiscus, J., Joy, D., Delgado, A., Michel, M., Smeaton, A. F., Graham, Y., Kraaij, W., Quénot, G., Eskevich, M., Ordelman, R., Jones, G. J. F., & Huet, B. (2017a). TRECVID 2017: Evaluating Ad-hoc and Instance Video Search, Events Detection, Video Captioning and Hyperlinking. In *Proceedings of TRECVID 2017*. NIST, USA.

Awad, G., Kraaij, W., Over, P., & Satoh, S. (2017b). Instance search retrospective with focus on TRECVID. *International Journal of Multimedia Information Retrieval*, *6*, 1–29. doi:10.1007/s13735-017-0121-3.

Chong-Wah Ngo, Yu-Fei Ma, & Hong-Jiang Zhang (2003). Automatic video summarization by graph modeling. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 104–109). IEEE. doi:10.1109/ICCV.2003.1238320.

Chu, W.-S., Yale Song, & Jaimes, A. (2015). Video co-summarization: Video summarization by visual co-occurrence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3584–3592). IEEE. doi:10.1109/CVPR.2015.7298981.

Demir, M., & Bozma, H. I. (2015). Video Summarization via Segments Summary Graphs. In *Proceedings of the IEEE International Conference on Computer Vision Workshop* (pp. 1071–1077). IEEE. doi:10.1109/ICCVW.2015.140.

Gygli, M., Grabner, H., Riemenschneider, H., & Van Gool, L. (2014). Creating Summaries from User Videos. In *Proceedings of the European Conference on Computer Vision* (pp. 505–520). doi:10.1007/978-3-319-10584-0{_}33.

Gygli, M., Grabner, H., & Van Gool, L. (2015). Video summarization by learning submodular mixtures of objectives. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3090–3098). doi:10.1109/CVPR.2015.7298928.

Hanjalic, A. (2002). Shot-boundary detection: unraveled and resolved? *IEEE Transactions on Circuits and Systems for Video Technology*, *12*, 90–105. doi:10.1109/76.988656.

Haojin Yang, & Meinel, C. (2014). Content Based Lecture Video Retrieval Using Speech and Video Text Information. *IEEE Transactions on Learning Technologies*, *7*, 142–154. doi:10.1109/TLT.2014.2307305.

He, L., Sanocki, E., Gupta, A., & Grudin, J. (1999). Auto-summarization of audio-video presentations. In *Proceedings of the ACM international conference on Multimedia* (pp. 489–498). doi:10.1145/319463.319691.

Huang, M., Mahajan, A. B., & DeMenthon, D. F. (2004). *Automatic performance evaluation for video summarization*. Technical Report Maryland Univ. College Park Inst. for Advanced Computer Studies.

Kim, G., Sigal, L., & Xing, E. P. (2014). Joint Summarization of Large-Scale Collections of Web Images and Videos for Storyline Reconstruction. In *2014 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4225–4232). IEEE. doi:10.1109/CVPR.2014.538.

Liu, T., Zhang, H. J., & Qi, F. (2003). A Novel Video Key-Frame-Extraction Algorithm Based on Perceived Motion Energy Model. *IEEE Transactions on Circuits and Systems for Video Technology*, *13*, 1006–1013. doi:10.1109/TCSVT.2003.816521.

Liu, W., Mei, T., Zhang, Y., Che, C., & Luo, J. (2015). Multi-task deep visual-semantic embedding for video thumbnail selection. In *Proceedings of the Conference on Computer Vision and Pattern Recognition* (pp. 3707–3715). doi:10.1109/CVPR.2015.7298994.

Nguyen, C., Niu, Y., & Liu, F. (2012). Video summagator: an interface for video summarization and navigation. In *Proceedings of the Special Interest Group on Computer-Human Interaction on Conference on Human Factors in Computing Systems* (pp. 647–650).

Pal, G., Rudrapaul, D., Acharjee, S., Ray, R., Chakraborty, S., & Dey, N. (2015). Video Shot Boundary Detection: A Review. In S. C. Satapathy, A. Govardhan, K. S. Raju, & J. K. Mandal (Eds.), *Emerging ICT for Bridging the Future - Proceedings of the Annual Convention of the Computer Society of India* (pp. 119–127). Cham: Springer International Publishing. doi:10.1007/978-3-319-13731-5{_}14.

Peng Chang, Mei Han, & Yihong Gong (2002). Extract highlights from baseball game video with hidden Markov models. *Proceedings of the International Conference on Image Processing*, *1*, I–609–I–612. doi:10.1109/ICIP.2002.1038097.

Plummer, B. A., Brown, M., & Lazebnik, S. (2017). Enhancing video summarization via vision-language embedding. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, *2017-Janua*, 1052–1060. doi:10.1109/CVPR.2017.118.

Roberson, R. (2013). Helping Students Find Relevance. URL: http://www.apa.org/ed/precollege/ptn/2013/09/students-relevance.aspx.

Rosner, B., & Cronbach, L. J. (1960). *Essentials of Psychological Testing* volume 73 of *Essentials of Behavioral Science*. Wiley. doi:10.2307/1419921.

Sharghi, A., Laurel, J. S., & Gong, B. (2017). Query-Focused Video Summarization: Dataset, Evaluation, and a Memory Network Based Approach. In *Proceedings of the Conference on Computer Vision and Pattern Recognition* (pp. 2127–2136). IEEE. doi:10.1109/CVPR.2017.229.

Song, Y., Vallmitjana, J., Stent, A., & Jaimes, A. (2015). TVSum: Summarizing web videos using titles. *Proceedings of the IEEE Computer Society Confer-*

ence on Computer Vision and Pattern Recognition, *07-12-June*, 5179–5187. doi:10.1109/CVPR.2015.7299154.

Soomro, K., Zamir, A. R., & Shah, M. (2012). UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. *CoRR*, *abs/1212.0*.

Sun, M., Farhadi, A., & Seitz, S. (2014). Ranking domain-specific highlights by analyzing edited videos. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *8689 LNCS*, 787–802. doi:10.1007/978-3-319-10590-1{_}51.

Sundaram, H., & Chang, S.-F. (2001). Condensing computable scenes using visual complexity and film syntax analysis. In *Proceedings of International Conference Multimedia and Expo*.

Taskiran, C. M. (2006). Evaluation of automatic video summarization systems. In *Multimedia Content Analysis, Management, and Retrieval 2006* (p. 60730K). International Society for Optics and Photonics volume 6073.

Tiecheng Liu, & Kender, J. (2002). An efficient error-minimizing algorithm for variable-rate temporal video sampling. In *Proceedings of the IEEE International Conference on Multimedia and Expo* (pp. 413–416). doi:10.1109/ICME.2002.1035806.

Truong, B. T., & Venkatesh, S. (2007). Video abstraction. *ACM Transactions on Multimedia Computing, Communications, and Applications*, *3*, 3–es. doi:10.1145/1198302.1198305.

Wang, X., Chen, J., & Zhu, C. (2011). User-Specific Video Summarization. In *Proceedings of the International Conference on Multimedia and Signal Processing* (pp. 213–219). doi:10.1109/CMSP.2011.51.

Wu, T., Gurram, P., Rao, R. M., & Bajwa, W. U. (2015). Hierarchical Union-of-Subspaces Model for Human Activity Summarization. In *Proceedings of the International Conference on Computer Vision Workshop* (pp. 1053–1061). volume 2016-Febru. doi:10.1109/ICCVW.2015.138.

Xiao-Dong Yu, Lei Wang, Qi Tian, & Ping Xue (2004). Multilevel video representation with application to keyframe extraction. In *Proceedings of the IEEE International Multimedia Modelling Conference* (pp. 117–123). doi:10.1109/MULMM.2004.1264975.

Xiong, Z., Regunathan Radhakrishnan, & Ajay Divakaran (2003). Generation of sports highlights using motion activity in combination with a common audio feature extraction framework. In *Proceedings of the International Conference on Image Processing* (pp. I–5). volume 1. doi:10.1109/ICIP.2003.1246884.

Yao, T., Mei, T., & Rui, Y. (2016). Highlight Detection with Pairwise Deep Ranking for First-Person Video Summarization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 982–990). doi:10.1109/CVPR.2016.112.

Yong Jae Lee, Ghosh, J., & Grauman, K. (2012). Discovering important people and objects for egocentric video summarization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1346–1353). doi:10.1109/CVPR.2012.6247820.

Youtube (2018). Youtube for Press. URL: https://www.youtube.com/intl/en-US/yt/about/press/.

Yuan, J., Wang, H., Xiao, L., Zheng, W., Li, J., Lin, F., & Zhang, B. (2007). A Formal Study of Shot Boundary Detection. *IEEE Transactions on Circuits and Systems for Video Technology*, *17*, 168–186. doi:10.1109/TCSVT.2006.888023.

Zhang, H. J., Wu, J., Zhong, D., & Smoliar, S. W. (1997). An integrated system for content-based video retrieval and browsing. *Pattern Recognition*, *30*, 643–658. doi:10.1016/S0031-3203(96)00109-4.

Zhao, B., & Xing, E. P. (2014). Quasi Real-Time Summarization for Consumer Videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2513–2520). doi:10.1109/CVPR.2014.322.

Ziyou Xiong, Xiang Sean Zhou, Qi Tian, Yong Rui, & Huangm TS (2006). Semantic retrieval of video - review of research on video retrieval in meetings, movies and broadcast news, and sports. *IEEE Signal Processing Magazine*, *23*, 18–27. doi:10.1109/MSP.2006.1621445.