Universidade Federal da Bahia
Instituto de Matemática

Programa de Pós-Graduação em Ciência da Computação

FROM MODELING PERCEPTIONS TO EVALUATING VIDEO SUMMARIZERS

Kalyf Abdalla Buzar Lima

DOCTORAL THESIS

Salvador
18 de dezembro de 2020

KALYF ABDALLA BUZAR LIMA

FROM MODELING PERCEPTIONS TO EVALUATING VIDEO SUMMARIZERS

This thesis was presented to the Postgraduate Program in Computer Science of Universidade Federal da Bahia, as a partial requirement to the degree of PhD in Computer Science.

Advisor: Prof. Dr. Luciano Rebouças de Oliveira
Co-advisor: Prof. Dr. Igor Gomes Menezes

Salvador
18 de dezembro de 2020

Ficha catalográfica elaborada pela Biblioteca Universitária de Ciências e Tecnologias
Prof. Omar Catunda, SIBI - UFBA

L732 Lima, Kalyf Abdalla Buzar.

From modeling perceptions to evaluating video summarizers / Kalyf Abdalla Buzar Lima – Salvador, 2020.

50f.

Orientador: Prof. Dr. Luciano Rebouças de Oliveira.

Co-orientador: Prof. Dr. Igor Gomes Menezes.

Tese (Doutorado) – Universidade Federal da Bahia, Instituto de Matemática, 2020.

1. Ciência da Computação. 2. Processamento digital de imagem. 3. Vídeo.. I. Oliveira, Luciano Rebouças de. II. Menezes, Igor Gomes. III. Universidade Federal da Bahia. Instituto de Matemática. IV. Título.

CDU 004.93

APPROVAL TERM

KALYF ABDALLA BUZAR LIMA

FROM MODELING PERCEPTIONS TO EVALUATING VIDEO SUMMARIZERS

This thesis was judged to obtain the Doctor's Degree in Computer Science, and approved in its final form by the Postgraduate Program in Computer Science of the Federal University of Bahia.

Salvador, 18 de dezembro de 2020

Oliveira

Prof. Dr. Luciano Rebouças de Oliveira
Universidade Federal da Bahia

Profa. Dra. Gecynalda Soares da Silva Gomes
Universidade Federal da Bahia

Prof. Dr. Paulo Canas Rodrigues
Universidade Federal da Bahia

João Paulo Papa.

Prof. Dr. Joao Paulo Papa
Universidade Estadual Paulista

Ricardo Torres

Prof. Dr. Ricardo Da Silva Torres
Norwegian University of Science and Technology

I dedicate this study to everyone who supported me, including the voices in my head.

ACKNOWLEDGEMENTS

Since writing is personally challenging for me, I received a lot of support and assistance throughout this thesis writing.

I want first to thank my advisor, Professor Luciano Oliveira, and my co-advisor, Professor Igor Menezes. They guided me in the challenging task of writing papers. I want to thank you both for your patient support. Your insightful feedback that pushed me to improve my writing abilities, bringing them to a higher level.

I want to thank the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) for providing a scholarship to me. The scholarship was essential for carrying out all the research.

I want to acknowledge my doctoral colleagues. Our discussions inspired me to explore other fields by providing stimulating talks. Also, I want to thank my coworkers at the Instituto Federal de Educação, Ciência e Tecnologia da Bahia for supporting the teaching demands when I needed to dedicate myself to writing the thesis.

I want to thank my parents and friends for their patience and their sympathetic ear. You are there for me with happy distractions to rest my mind outside of my research.

I'm a scientist; because I invent, transform, create, and destroy for a living, and when I don't like something about the world, I change it.

—RICK SANCHEZ

RESUMO

Horas de vídeos são enviados para plataformas de *streaming* a cada minuto, com sistemas de recomendação sugerindo vídeos populares e relevantes para ajudar economizar o tempo dos usuários no processo de busca. Sumarizadores de vídeo foram então desenvolvidos para detectar as partes mais relevantes e automaticamente condensá-las em um vídeo curto. Atualmente, avaliar esse tipo de método é desafiador uma vez que as métricas não avaliam a subjetividade dos usuários, como a concisão das anotações. Para lidar com o critério de concisão, nós propomos uma nova métrica que avalia sumarizadores de vídeo em múltiplas taxas de compressão. Nossa métrica, chamada *Compression Level of USer Annotation* (CLUSA), mensura o desempenho dos sumarizadores de vídeo diretamente a partir dos escores de relevância preditos. Para isso, a CLUSA gera sumários de vídeo descartando gradualmente segmentos de vídeo de acordo com os escores de relevância anotados pelos usuários. Depois de agrupar os sumários de vídeo pelas taxas de compressão, a CLUSA os compara com os escores de relevância preditos. Para preservar informações relevantes em resumos de vídeo concisos, CLUSA então pondera o desempenho dos sumarizadores de vídeo em cada faixa de compressão e, por fim, calcula uma pontuação geral de desempenho. Considerando que a CLUSA pondera todas as faixas de compressão, mesmo aquelas que não foram abrangidas pelas anotações dos usuários, o desempenho de base muda com cada conjunto de dados. Consequentemente, a interpretação do escore de desempenho para os sumarizadores de vídeo não é tão direta quanto em outras métricas. Em nossos experimentos, comparamos a CLUSA com outras métricas de avaliação para sumarização de vídeo. Nossas descobertas sugerem que todas as métricas analisadas avaliam adequadamente sumarizadores de vídeo usando anotações binárias. Para as anotações multivaloradas, a CLUSA mostrou-se mais adequada, preservando as informações de vídeo mais relevantes no processo de avaliação.

Palavras-chave: Sumarização de vídeo. Sumarizadores de vídeo. Avaliação. Métrica. Taxa de compressão.

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. Video summarizers have been developed to detect the video's most relevant parts, automatically condensing them into a shorter video. Currently, evaluating this type of method is challenging since the metrics do not assess user annotations' subjective criteria, such as conciseness. To address the conciseness criterion, we propose a novel metric to evaluate video summarizers at multiple compression rates. Our metric, called Compression Level of USer Annotation (CLUSA), assesses the video summarizers' performance by matching the predicted relevance scores directly. To do so, CLUSA generates video summaries by gradually discarding video segments from the relevance scores annotated by users. After grouping the generated video summaries by the compression rates, CLUSA matches them to the predicted relevance scores. To preserve relevant information in concise video summaries, CLUSA weighs the video summarizers' performance in each compression range to compute an overall performance score. As CLUSA weighs all compression ranges even that user annotations do not span some compression rates, the baseline changes with each video summarization data set. Hence, the interpretation of the video summarizers' performance score is not as straightforward as other metrics. In our experiments, we compared CLUSA with other evaluation metrics for video summarization. Our findings suggest that all analyzed metrics evaluate video summarizers appropriately using binary annotations. For multi-valued ones, CLUSA proved to be more suitable, preserving the most relevant video information in the evaluation process.

Keywords: Video summarization. Video summarizers. Evaluation. Metric. Compression rate

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

CLUSA	Compression Level of USer Annotation	23
GAN	Generative Adversarial Networks	8
LSTM	Long Short-Term Memory	8
DPP	Determinantal Point Processes	8
MLP	Multilayer Perceptron	8
RNN	Recurrent Neural Network	8
CTT	Classical Test Theory	11
IRT	Item Response Theory	11
TP	true positive	14
FP	false positive	14
TN	true negative	14
FN	false negative	14
ROC	Receiver Operating Characteristic	25
PR	Precision-Recall	25
AUC	area under curve	25
RCC	ranked correlation coefficients	23

Chapter 1

INTRODUCTION

Contents

Billions of video hours are watched every day on streaming platforms, such as Youtube, and the total of available videos on those platforms continuously grows (Youtube, 2018). As watching each video takes a long time, users need to select the videos they watch rigorously. Selecting the most relevant videos is an arduous task for users; thus, streaming platforms automate the search for relevant videos according to users' preferences. However, just searching and recommending **entire videos** is no longer enough and users demand for **video summaries** with the most important video information, as illustrated in Fig. 1.1. Determining which video information is relevant to users is challenging as it is affected by users' subjective factors (HUTZ; BANDEIRA; TRENTINI, 2015; PASQUALI, 2017). Consequently, **video summarizers** must not only mimic the way humans understand and judge the relevance of information in videos but also tailor the video summary to users' interests (TRUONG; VENKATESH, 2007).

Earlier studies in the video summarization have focused on identifying how humans judge video information's relevance. He et al. (1999) suggested that users instinctively follow four different but complementary criteria for judging the relevance of video information: Conciseness, coverage, context, and coherence. Conciseness is how much users have shortened the entire video (*i.e.*, the video compression), while coverage is the amount of video information users have summarized indeed. Context and coherence are intrinsically related to video segments' ordering and how video summaries told the story. While users follow all of these criteria to summarize an input video, video summarization studies focus only on coverage and conciseness criteria, as current video summarizers do not address the ordering or generation of video information.

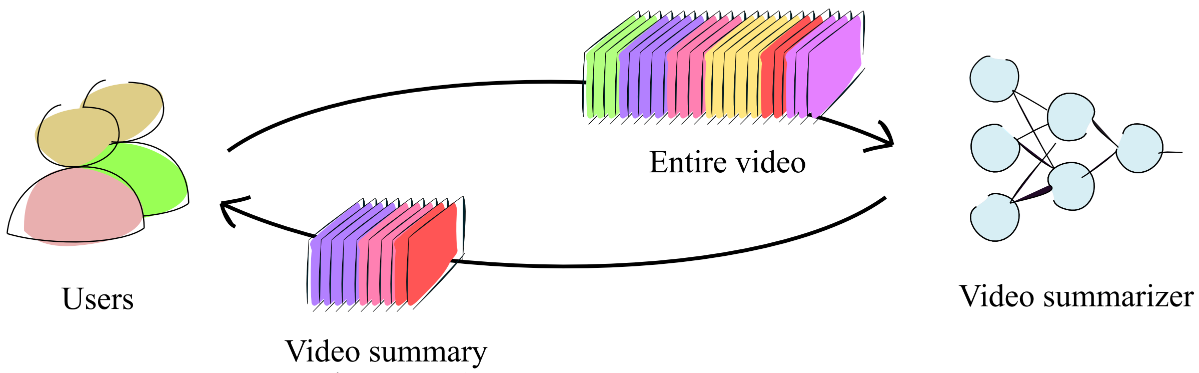


Figure 1.1 Instead of users watching the entire video, users watch a video summary with the most important video information selected by a video summarizer.

Since there were no relevance scores annotated by users for training or testing video summarizers, earlier studies addressed the task of summarizing videos by using ad-hoc summarization heuristics. Each study specified visual elements and events that are supposed to be of interest to users (CHANG; HAN; GONG, 2002; XIONG; RADHAKRISHNAN; DIVAKARAN, 2003; Chong-Wah Ngo; Yu-Fei Ma; Hong-Jiang Zhang, 2003; ZHAO; XING, 2014; SUN; FARHADI; SEITZ, 2014; WU et al., 2016; YAO; MEI; RUI, 2016). For instance, Chang, Han and Gong (2002) formulated a summarization heuristic for baseball game videos based on the detection of the key events: Home run, catch, hit, and infield play. Heuristic-based video summarizers are accurate when users search for known events; otherwise, it is unfeasible to formulate a unique summarization heuristic that ultimately matches human judgments. Therefore, heuristic-based video summarizers can no longer be evaluated in a generic video domain, and some studies merely described video summarizers' results, pointing out advantages and disadvantages (Xiao-Dong Yu et al., 2004).

To properly evaluate video summarizers in a generic video domain, video summarization studies targeted the evaluation at users (SUNDARAM; CHANG, 2001; LIU; ZHANG; QI, 2003; AGNIHOTRI; DIMITROVA; KENDER, 2004; TASKIRAN, 2006; GYGLI et al., 2014; CHU; Yale Song; JAIMES, 2015; SONG et al., 2015), who are genuinely able to determine which video information is relevant or not via user annotations. These ones are feedback collected from **users** who judge the relevance of each **video segment** in a collection of videos, as illustrated in Fig. 1.2.

As there is no consensus on collecting user annotation, different guidelines emerge and change how studies devise new video summarizers. Currently, these summarize videos by discarding or preserving video segments (ZHANG et al., 2016; FAJTL et al., 2019; ZHOU; QIAO; XIANG, 2017; ROCHAN; YE; WANG, 2018; OTANI et al., 2017). In other words, video summarizers act as binary classifiers whose label 0 (zero) represents the discarded video segments and label 1 (one), the preserved ones. Conversely, users find it challenging to discriminate some video segments' relevance from the least relevant to

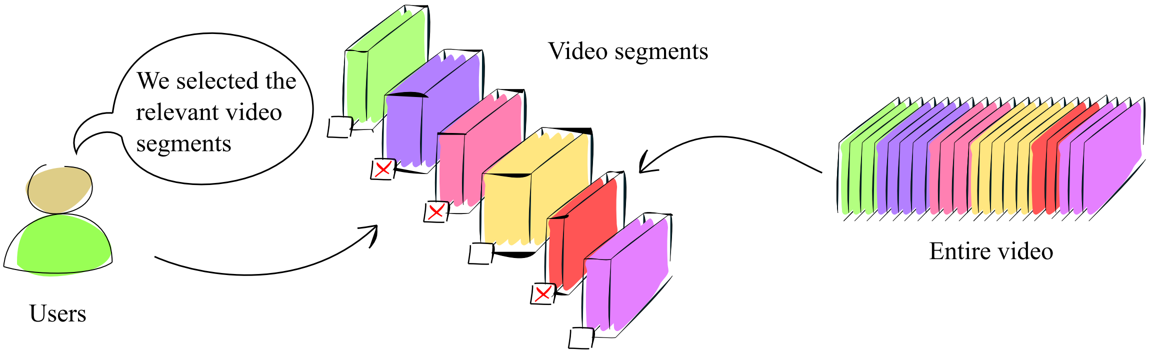


Figure 1.2 Users judge the relevance of video segments, viz., consecutive video frames grouped and that describe the same video information.

the most relevant using a binary scale. Considering that, Song et al. (2015) collected user annotations using an assessment scale with five relevance scores, making it possible to generate video summaries of different lengths. However, video summaries generated from user annotations and video summarizers must have similar length for the evaluation to be accurate using the F_β, a metric commonly used in evaluating classifiers. Therefore, video summarization studies opt to discard the same proportion of video segments (*i.e.*, the compression rate) in user annotations (ZHANG et al., 2016; FAJTL et al., 2019; ZHOU; QIAO; XIANG, 2017; ROCHAN; YE; WANG, 2018; OTANI et al., 2017). In doing so, video summarization studies discard conciseness information of user annotations.

1.1 MOTIVATION

Evaluating video summarizers has always proved to be an obstacle in video summarization studies that look for ways to address evaluation issues (GYGLI et al., 2014; OTANI et al., 2019; SHARGHI; LAUREL; GONG, 2017). For example, a recent advance was in changing the assessment scale used to collect annotations from users. While multi-valued assessment scales have made it possible to assess the conciseness criterion properly, F_β can not deal with video summaries at multiple video compression rates. As F_β does not distinguish video summaries' compression rate, video summarizers' performance reduces when high-compressed video summaries (shorter length video summaries) are matched to low-compressed ones (longer length video summaries). Accordingly, video summarization studies opt to limit the conciseness criterion, and the evaluation of video summarizers remains stuck in a single preset compression rate. Here, we emphasize that addressing the conciseness criterion on video summarizers' evaluation is crucial to advance the video summarization.

1.2 GOALS

Video summarization studies collect multi-valued user annotations using different users and collecting guidelines. For instance, Gygli et al. (2014) and Song et al. (2015) split collection of videos into video segments using different techniques for video shot segmentation. Because of these differences in collecting guidelines, we can not directly compare user annotations from current data sets to **investigate the assessment scales' impacts in the user annotations' quality**. We seek to accomplish this by **collecting user annotations in a standard scenario with the same videos and users**.

Assuming that multi-valued assessment scales are suitable for collecting user annotations and F_β may not assess video summarizers' performance at multiple compression rates, we also seek to devise a **novel evaluation metric** to handle multi-valued user annotations and multiple compression rates properly.

1.3 KEY CONTRIBUTIONS

Since Gygli et al. (2014) collected user annotations (the SumMe data set) constraining their compression rate, F_β can use them to evaluate video summarizers, in contrast to the multi-valued annotations collected by Song et al. (2015) (the TVSum50 data set). As our study demonstrated that multi-valued assessment scales deliver higher annotations' quality than binary scales, we devised Compression Level of USer Annotation (CLUSA) metric to overcome the limitations of the F_β when applied in multi-valued scale annotations. This study has already been published in Elsevier's journal Expert Systems with Applications.

While we published the key contributions mentioned above (ABDALLA; MENEZES; OLIVEIRA, 2019), Otani et al. (2019) introduced ranked correlation coefficients (RCC) to assess multi-valued user annotations directly. To investigate the differences between RCC and CLUSA, we ranked five state-of-the-art video summarizers using both metrics in the SumMe and TVSum50 data sets. Although RCC, as a metric for evaluating video summarizers, is arguably more appropriate than F_β, RCC do not perform weighing, and hence, RCC do not target high-compressed video summaries. Conversely, the weighing of compression ranges (*i.e.*, compression rates grouped within a range) is crucial when CLUSA assesses video summarizers' performance, and hence, missing compression ranges can skew CLUSA scores. As long as we evaluate video summarizers with the same compression ranges and user annotations, our study showed that missing compression ranges do not impinge on video summarizers' ranking, but the results' interpretation is challenging using CLUSA.

To sum up, we highlight three key contributions presented in this study: (i) A metric to evaluate video summarizers against user annotations, these latter collected with binary or multi-valued scales, (ii) a study on the quality of user annotations collected from different assessment scales, and (iii) to provide a better understanding of the limitations of RCC and CLUSA while identifying what factors can skew their measurements.

1.4 CHAPTER MAP

Chapter 2 presents a background in the fields of video summarization and psychometric. We detail the task of summarizing videos, the requirements that video summarizers must meet, and a history of how video summarization studies evaluated their proposals. Considering the evaluation of video summarizers targeted at users, we investigated how the collecting process dealt with users' subjectivity and how evaluation metrics assess video summarizers' performance by matching user annotations to video summaries.

Chapter 3 presents how CLUSA intends to overcome the limitations of evaluation metrics commonly used for video summarization. We detail the mathematical formulation of CLUSA and the factors that skew CLUSA. Also, we discuss how missing compression ranges impinge on video summarizers' performance.

Chapter 4 presents the methodology and results of our experimental study. In short, we analyzed: (i) The assessment scales' impact on the user annotations' quality, (ii) the evaluation metrics by exploring the relationship between internal consistency and human consistency, and (iii) how missing compression ranges affect the video summarizers' ranking in the SumMe and TVSum50 data sets.

Chapter 5 presents our discussions about our study's impact on video summarization and how future studies can improve our evaluation approach.

Chapter 2

BACKGROUND

Contents

Although the generation of video summaries must meet three requirements: (i) The presence of visual elements and events relevant to users, (ii) elimination of redundant information, and (iii) generation of useful information as possible from input videos (TRUONG; VENKATESH, 2007), only these are not sufficiently discriminative to rank all video summarization studies. Therefore Truong and Venkatesh (2007) also have grouped the studies on video summarization according to their application goals. They are: **Browsing and retrieval** systems – to assist users on video searching (AWAD et al., 2017; ARMAN et al., 1994; ZHANG et al., 1997; Haojin Yang; MEINEL, 2014), **computational reduction and content analysis** systems – to abstract video information and eliminate redundancies (PLUMMER; BROWN; LAZEBNIK, 2017), **story navigation and video editing** – to help users on video navigation (NGUYEN; NIU; LIU, 2012), and **highlighting** systems to short input videos by selecting relevant video segments or frames (YAO; MEI; RUI, 2016; GYGLI et al., 2014; XIONG; RADHAKRISHNAN; DIVAKARAN, 2003). It is noteworthy that current studies (FAJTL et al., 2019; MAHASSENI; LAM; TODOROVIC, 2017; ZHOU; QIAO; XIANG, 2017) referred to highlighting systems as video summarization methods (*i.e.*, video summarizers). Henceforth, we refer to highlighting systems as video summarizers.

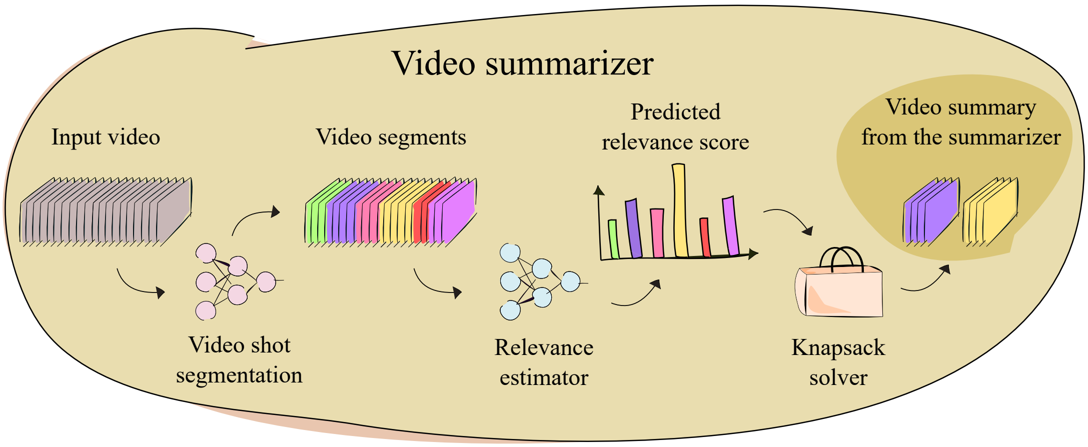


Figure 2.1 In a generic video summarization pipeline, a video shot segmentation split the input video into video segments. After estimating the relevance of each video segment, a knapsack solver selects most relevant segments to generate a video summary.

2.1 PIPELINE OF A VIDEO SUMMARIZER

In general, a **video summarizer** follows a general pipeline (see Fig. 2.1, according to three steps: (a) An **input video** is segmented into **video segments** by grouping consecutive video frames, (b) a relevance score is predicted for each video segment, and (c) the most relevant segments are selected for the video summary by a **knapsack solver**. A video summarizer estimates a non-binary relevance score for each video segment. To comply with video compression constraints when evaluating video summarizers, a knapsack solver generates video summary by selecting video segments whose **predicted relevance scores** are the highest, as depicted in Fig. 2.1. Since video summarizers commonly carry out the **video shot segmentation** by using the same technique used on the collecting process, we cover this topic in more detail in Section 2.4 when discussing the collecting guidelines in the currently available data sets.

After segmenting a video shot, the video summarizer estimates each video segment's relevance. Chu, Yale Song and Jaimes (2015) accomplish this by modeling the summarization as bipartite graphs. The video segments, which are represented by graph nodes, are connected according to the visual similarity. Finally, the video summarizer selects the video segments by ordering the weights of the graph edges. Mahasseni, Lam and Todorovic (2017) use Long Short-Term Memory (LSTM) layers to select some video segments to reconstruct the input video by means of a Generative Adversarial Networks (GAN), with the video summary being a set of video segments whose reconstructed video was similar to the input one. Zhang et al. (2016) use bidirectional LSTM layers and an Multilayer Perceptron (MLP) to compute the probability of each video frame belongs to the final video summary, and a Determinantal Point Processes (DPP) to eliminate redundant video frames. The summarization model is trained with previously collected user annotations. Fajtl et al. (2019) applied attention mechanisms in LSTM layers to select visual elements that are supposed to be relevant in the video. In a similar way, Zhou, Qiao and Xiang (2017) also encode temporal dependencies but using Recurrent

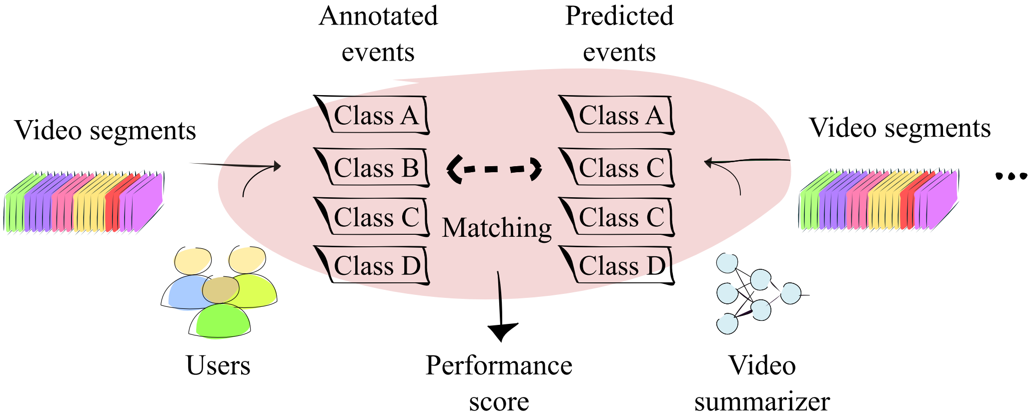


Figure 2.2 In some content domains, video summarizers can focus on detecting key events that are previously defined. In these cases, the evaluation of video summarizers becomes similar to the evaluation of event detection task.

Neural Network (RNN) layers instead.

2.2 A BRIEF HISTORY OF THE EVALUATION OF VIDEO SUMMARIZERS

Currently, video summarization studies evaluate video summarizers using annotated relevance scores. However earlier studies did not take into account users' annotation to measure the video summarizers' performance (LIU; ZHANG; QI, 2003; WANG; CHEN; ZHU, 2011; TRUONG; VENKATESH, 2007). So the evaluation of video summarizers was limited to descriptive analysis (TRUONG; VENKATESH, 2007). The resulting summaries were obtained from the authors' perspective in specific situations, and certainly with biased conclusions. Since there were no experimental arguments to support the descriptive analysis of the advantages and weaknesses of video summarizers, the results were considered inadequate (Xiao-Dong Yu et al., 2004; CO-INVESTIGATOR, 2013; TASKIRAN, 2006; GYGLI et al., 2014; SHARGHI; LAUREL; GONG, 2017).

A solution found to overcome the limitations of descriptive analysis was to change the evaluation of video summarizers according to each content domain (TRUONG; VENKATESH, 2007), *e.g.*, sports, news, and documentaries. For example, in a soccer match, specific events, such as goals, fouls, and penalties, attract users' attention more than others. In other words, video summarization aims to detect events set by users, and the event detection nails the video summarizers' performance (YAO; MEI; RUI, 2016; Chong-Wah Ngo; Yu-Fei Ma; Hong-Jiang Zhang, 2003; ZHAO; XING, 2014; SUN; FARHADI; SEITZ, 2014; CHANG; HAN; GONG, 2002). As depicted in Fig. 2.2, the **performance score** is computed by **matching** the events predicted by a **video summarizer** to events annotated by **users**. As users cannot judge video information from a few preset events for the generic domain's videos, there is no guarantee that any heuristics used to summarize videos will properly match human judgments (TRUONG; VENKATESH, 2007). As such, video summarization studies have looked for other evaluation approaches to assess video summaries from user annotations (Yong Jae Lee; GHOSH; GRAUMAN, 2012; LIU et al., 2015; GYGLI et al., 2014; SONG et al., 2015; CHU; Yale Song; JAIMES, 2015;

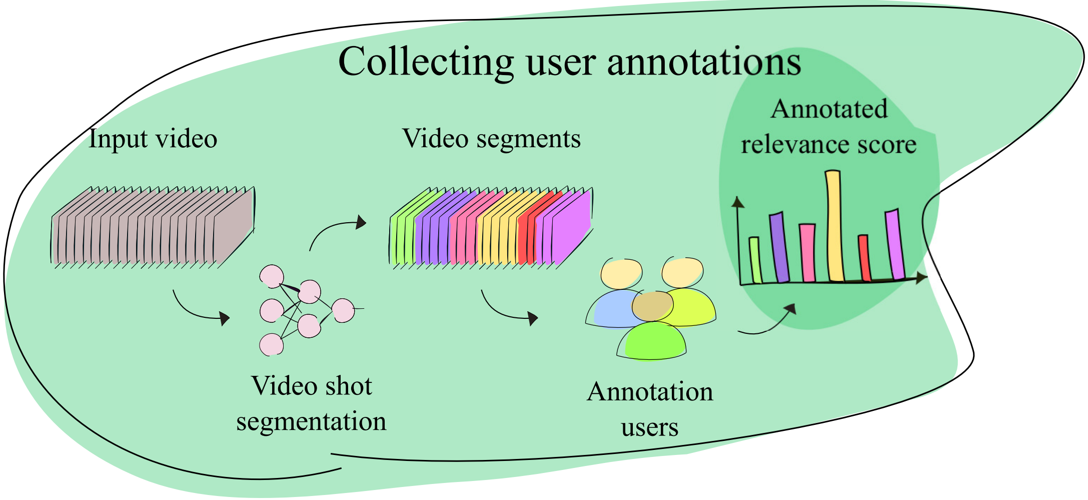


Figure 2.3 A video shot segmentation split the input video into video segments. Users then annotate the video segments' relevance using an assessment scale.

KIM; SIGAL; XING, 2014; SUNDARAM; CHANG, 2001; AGNIHOTRI; DIMITROVA; KENDER, 2004).

Assuming that only users are truly able to determine which video segments are relevant in videos, two evaluation approaches emerged as user studies: (a) Requesting users to assess the quality of video summaries (LIU; ZHANG; QI, 2003; TASKIRAN, 2006; CHU; Yale Song; JAIMES, 2015), and (b) requesting users to annotate the relevance of video segments (GYGLI et al., 2014; SONG et al., 2015). Since (a) does not guarantee the same conditions when evaluating video summarizers, (b) is currently the most common type of evaluation procedure in the state of the art, nowadays. To clarify this evaluation approach, the **annotated relevance scores** (in Fig 2.3) are used to generate **users' video summaries** with which all video summarizers are evaluated in the same way. Although the evaluation based on relevance scores annotated by users is currently the most straightforward way to assess video summarizers' performance, the collecting process has led to some challenges. Since annotations can be skewed as the users' perception of relevance changes constantly, video summarization studies often apply psychological testing to mitigate bias and improve the quality of the user annotations.

2.3 COLLECTING VIDEO INFORMATION USERS FIND RELEVANT

Psychological testing is part of a complex process aimed at diagnosing individuals' performance on specific tasks. These tests aim to evaluate, measure, or estimate a latent factor in user behavior (URBINA, 2014; HUTZ; BANDEIRA; TRENTINI, 2015; PASQUALI, 2017). In video summarization, psychological tests attempt to measure the users' perception of relevance, which is the intended **latent factor**. Since user's perception is a **psychological phenomenon**, the relevance of a **video segment** can not be measured directly. Instead, it is measured from user feedback via **annotation**, as illustrated in Fig. 2.4.

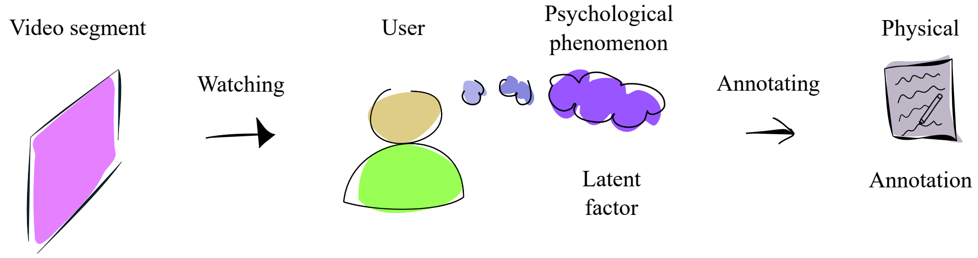


Figure 2.4 After watching a video segment, user assesses the relevance of the video segment. However, user's perception is a psychological phenomenon that can only be measured indirectly through user feedback, in this case, annotations.

In psychometric studies, there are two different epistemological approaches to measure the users' perceptions: Classical Test Theory (CTT), which focuses exclusively on the evaluation of the user annotations and their measurement error, and Item Response Theory (IRT), which focuses on each test item and its influence on the measurement of the latent factor (HUTZ; BANDEIRA; TRENTINI, 2015; PASQUALI, 2017). To the best of our knowledge, there are no video summarization studies, which apply IRT, while Gygli et al. (2014) and Song et al. (2015) applied CTT in the collecting process of the SumMe and TVSum50 data sets, respectively. Given that, our study focuses on CTT theory; however, psychometric studies investigate the advantages of using IRT over CTT (JABRAYILOV; EMONS; SIJTSMA, 2016).

In CTT, video summarization studies (GYGLI et al., 2014; SONG et al., 2015) model the users' perception as

$$t = D - E\,, \tag{2.1}$$

where t is the true relevance of video segments. t is supposed to be the latent factor if it is measured directly. Unfortunately, it is unlikely to control all environmental and psychological conditions when users annotate video segments' relevance. As a result, the measured value is different from t, which leads t to be decomposed into two variables: The D annotated relevance scores and an E random error.

2.4 HOW CURRENT VIDEO SUMMARIZATION STUDIES COPE WITH RANDOM ERROR

Pasquali (2017) and Kline (2013) listed several situations and factors that boost E, shifting D away from t. Among this list, Gygli et al. (2014) and Song et al. (2015) coped with: How much time users spend on completing the collecting process and how much users can distinguish the relevance of video segments on levels.

The time spent by users taking videos is not only determined by the number of videos, but also how many segments they are split into. Before users start annotating the relevance of each video segment, a video shot segmentation split the input videos (YUAN et al., 2007; PAL et al., 2015; HANJALIC, 2002), as illustrated in Fig. 2.3. Gygli et al. (2014) accomplished this by grouping video frames into five-second video segments, stretching them out according to the visual features of consecutive video frames to avoid fragmentation of relevant video information. In turn, Song et al. (2015) split the input videos into two-second video segments. As the length of video segments affects the total of items in the psychological test, there is a limit on how many video segments each user can consistently annotate before getting tired (KLINE, 2013). To the best of our knowledge, the studies with the biggest amount of annotated video segments were carried out by Song et al. (2015), who collected annotations for 6,291 video segments from 50 videos, and Gygli et al. (2014), who collected 820 video segments from 25 videos.

In the annotation process, users judge video segments by selecting the relevance scores that best match users' perception within an assessment scale (SONG et al., 2015). Among the four types of assessment scales, Hutz, Bandeira and Trentini (2015) argue only two are suitable for psychometric studies: **Ordinal scales**, for situations in which it is only possible to discriminate order, and **interval scales**, for when users are also able to discriminate magnitude. Hutz, Bandeira and Trentini (2015) also argue that if we could observe a latent factor directly, there would be an interval scale with infinite values. For that to occur in video summarization, users are required to discriminate a unique relevance score for each video segment, although users are unable to do this (BORSBOOM, 2005). Therefore, each video summarization study presets how many and which relevance scores users can pick on an assessment scale. Gygli et al. (2014), Chu, Yale Song and Jaimes (2015) and Song et al. (2015) used assessment scales with two, three and five relevance scores, respectively. However, choosing the number of relevance scores for the assessment scale is not a trivial decision. As the study carried out by Simms et al. (2019) shows, users tend to disagree when there are not enough relevance scores to discriminate a latent factor appropriately. Users are also unable to make fine-grained distinctions when there are many scores to decide. In short, the assessment scale depends on users' ability to distinguish the relevance of video segments.

To encourage users to be more discerning about which video segments are relevant, Gygli et al. (2014) and Song et al. (2015) impose constraints on the distribution of relevance scores. Gygli et al. (2014) preset a limit of 15% of the length of video that users can annotate as relevant. Similarly, Song et al. (2015) preset a distribution of relevance scores on a five-point assessment scale. Gygli et al. (2014) and Song et al. (2015) argue this tight control over how users annotate the relevance of video segments is necessary to generate high-quality video summaries from user annotations.

2.5 ENSURING THE QUALITY OF USER ANNOTATIONS

Collecting guidelines aim to improve user annotations' quality by coping with biasing factors. Urbina (2014), Pasquali (2017), and Hutz, Bandeira and Trentini (2015) point two quality indicators often used in psychometric studies: Test validity – to verify whether the psychological test and its items measure the psychological phenomenon they intend to measure (in particular, the relevance of video segments), and test reliability – to investigate the internal consistency of scores annotated by users (being specific, how much users agree with the relevance of video segments). To the best of our knowledge, video summarization studies (GYGLI et al., 2014; SONG et al., 2015) only investigate test reliability.

To explain the basis of test reliability, let us take the relevance perception modeling in Eq. 2.1, assuming that t is constant over time for a single user and E is estimated by periodically testing users (PASQUALI, 2017; URBINA, 2014). The smaller the difference between measurements, the less susceptible is the collecting guidelines to factors that boost E. Therefore, users must repeatedly annotate each video segment's relevance several times. Repeating annotation is not feasible in video summarization, as users are often rewarded in cash on Amazon Mechanical Turk (SONG et al., 2015) to perform this. Hence Gygli et al. (2014) and Song et al. (2015) collected user annotations from a cross-sectional perspective in a single test, calculating the internal consistency between users.

Anastasi (2000) points out three ways to measure the internal consistency of user annotations collected from a single psychological test: **Split-half method** – that measures the consistency for a sample of user annotations, **Kuder-Richardson coefficient** – for user annotations collected with binary ordinal scales, and **Cronbach's alpha** – for user annotations collected with generic ordinal scales. Cronbach (1951) derived Cronbach's alpha equation from Kuder-Richardson, and thus, both coefficients are directly comparable on the same scale. Currently, Cronbach's alpha is the most used internal consistency estimator, being applied to user annotations collected by Gygli et al. (2014) and Song et al. (2015).

Formally, Cronbach's alpha coefficient, α, measures the internal consistency of K video segments according to

$$\alpha = \frac{K}{K-1}\left(1 - \frac{\sum_{j=1}^{K} \sigma_{D_j}^2}{\sigma_D^2}\right), \tag{2.2}$$

where the variance of the j-th video segment, $\sigma_{D_j}^2$, is divided by the user annotation variance, σ_D^2. Cronbach's alpha is a direct measure of user disagreement, so the greater the variance of users' responses, the lower the value of Cronbach's alpha. Cronbach's alpha values range from 0 (when user annotations totally differ from each other) and 1 (when user annotation values are all equal). Since these values are constant for any user annotations, the reference values showed in Table 2.1 are used in the literature to assess the reliability of psychometric tests (GEORGE; MALLERY, 2010; HUTZ; BANDEIRA; TRENTINI, 2015). As a rule, user annotations must reach at least the internal consistency value of 0.7.

In the studies conducted by Gygli et al. (2014) and Song et al. (2015), the average qualities calculated by Cronbach's alpha for all videos were 0.74 and 0.81, respectively. However, not all user annotations collected by Gygli et al. (2014) have an acceptable quality. Notably, Gygli et al. (2014) found that 9 out of the 20 videos annotated in the SumMe data set are of unacceptable quality score. For example, the quality of user annotations for the video titled "Saving dolphins" was 0.21. Song et al. (2015) did not report the quality values for each video, as was done by Gygli et al. (2014); hence, it is not possible to assert whether users disagreed on any specific video. Overall, the average quality of the TVSum50's annotations is above 0.8, and therefore, "good" according to Table 2.1.

Table 2.1 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

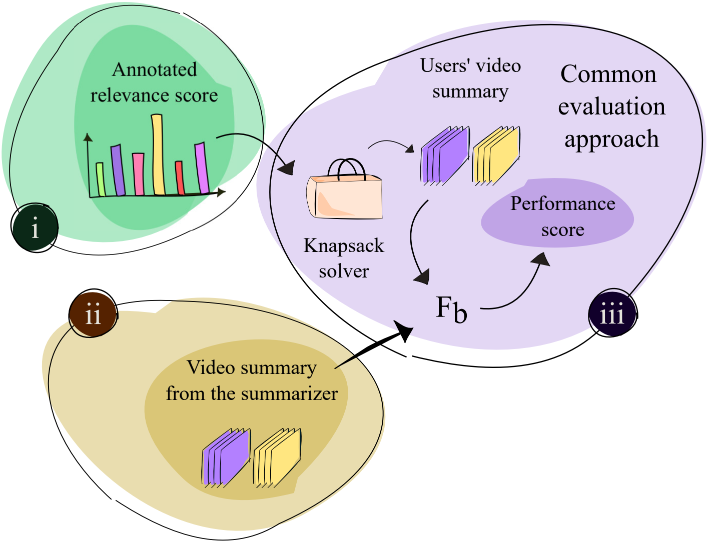

Annotated relevance score
i
Users' video summary
Common evaluation approach
Knapsack solver
Performance score
Fb
iii
ii
Video summary from the summarizer

Figure 2.5 The state-of-the-art approach to evaluating video summarizers is to apply knapsack solver to user annotations to generate a video summary that matches automatic video summaries.

2.6 EVALUATING VIDEO SUMMARIZERS FROM USER ANNOTATIONS

While video summarizers attempt to match the relevance scores collected with multi-valued ordinal scales (*e.g.*, Likert scales), video summarization is binary. Therefore, the **annotated relevance scores** (in the balloon (i) in Fig. 2.5) are mapped to binary with a similar compression rate of the video summary generated by summarizers (in the balloon (ii)). An evaluation metric then measures the similarity of the predicted video summary compared to the annotated one (in the balloon (iii)), with the $F_{\beta=1}$ being the most common in current video summarization studies (GYGLI et al., 2014; SONG et al., 2015). Derived from the confusion matrix, F_β metric is a convenient way to fully describe the performance of a predictive model by matching **expected** values to **predicted** ones (MOSLEY, 2013).

For a binary predictive model, Table 2.2 summarizes the value matching in four categories: **true positive (TP)**, **true negative (TN)**, **false positive (FP)**, and **false negative (FN)**, which correspond to the number of hits and misses for each class label.

Table 2.2 Confusion matrix organizes the matching of expected and predicted values in a classification model in a structured way, allowing detailed analysis of hits and misses.

		Annotated	
		Positive	**Negative**
Predicted	**Positive**	TP	FP
	Negative	FN	TN

From these four categories' values, some rates (*e.g.*, accuracy, precision, recall, and fall-out) emerged to assess the classifier's performance aiming at specific statistical analysis.

Gygli et al. (2014) propose to evaluate video summarizers using metrics based on precision and recall rates, arguing that the evaluation should aim at the selected video segments (*i.e* relevant video segments). While precision describes the hits of relevant video segments concerning all video segments **predicted** to be relevant as

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} = P(\text{annotated} = \text{relevant} \,|\, \text{predicted} = \text{relevant})\,, \tag{2.3}$$

recall describes the hits of relevant video segments in relation to all video segments **annotated** as

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} = P(\text{predicted} = \text{relevant} \,|\, \text{annotated} = \text{relevant})\,. \tag{2.4}$$

Precision and recall rates measure the most relevant video segments from complementary perspectives. Thus, video summarization studies pursue a trade-off between both rates via harmonic mean known as F_β score (KELLEHER; NAMEE; D'ARCY, 2015), is given by

$$F_\beta = (1 + \beta^2)\ \frac{precision \cdot recall}{\beta^2 \cdot precision + recall}\,, \tag{2.5}$$

with $\beta = 1$ being the weight parameter used in current studies. The F_β metric ranges from 0 to 1 which represent the worst and optimal performance value, respectively. Within this range, the value 0.5 represents the random classification method's performance.

The presented F_β assesses the classifiers' binary outputs as categorical targets, being necessary to calculate the arithmetic mean of F_β values for each class label when classifiers' outputs are non-binary. Finally, this average value is the performance of a multi-label classifier, which is not used in video summarization to the best of our knowledge.

2.6.1 Other way to assess classification performance

Evaluation metrics, such as Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves, assess the classifiers' prediction scores (KELLEHER; NAMEE; D'ARCY, 2015). Consequently, performance analysis is not limited to an ad-hoc binarization of the relevance scores. Instead, ROC and PR curves threshold the relevance scores with multiple values, generating C_i confusion matrices for each video summary.

In short, the difference between PR and ROC curves is the rates calculated from C_i confusion matrices. PR curve is computed as (precision$_i$, recall$_i$) pairs, which are summarized in a single performance score by calculating the area under curve (AUC) as

$$\text{AUC-PR} = \sum_{i=1}^{I-1} \frac{\text{precision}_i + \text{precision}_{i+1}}{2} (\text{recall}_{i+1} - \text{recall}_i)\,, \tag{2.6}$$

i being the index for each class label. Conversely, the ROC curve is computed from (recall$_i$, fall-out$_i$) pairs, with the fall-out rate being calculated as

$$\text{fall-out} = \frac{\text{FP}}{\text{FP} + \text{TN}}\,. \tag{2.7}$$

Likewise PR curve, (recall$_i$, fall-out$_i$) pairs of ROC are summarized by

$$\text{AUC-ROC} = \sum_{i=1}^{I-1} \frac{\text{recall}_i + \text{recall}_{i+1}}{2} (\text{fall-out}_{i+1} - \text{fall-out}_i)\,. \tag{2.8}$$

It is worth mentioning that video summarization studies did not evaluate video summarizers using ROC and PR curves to the best of our knowledge.

2.7 RETHINKING THE EVALUATION OF VIDEO SUMMARIZERS

Current studies model video summarizers as classification tasks, but Otani et al. (2019) proposed to evaluate them differently. Assuming that the annotated and predicted relevance scores are equivalent in terms of the order of relevance scores (non-linear monotonic relationship), Otani et al. (2019) assess how high this relationship is by using ranked correlation coefficients (RCC). Otani et al. (2019) applied two RCC (Kendall (KENDALL, 1945) and Spearman (SPEARMAN, 1904)) to match **annotated relevance scores** (in the balloon (i) of Fig. 2.6) to the relevance scores predicted by video summarizers (in the balloon (ii) of Fig. 2.6).

Formally, the τ Kendall coefficient calculates the total agreement and disagreement pairs for ranked scores from predicted and annotated relevance scores. Admittedly, we found three versions of the Kendall coefficient in the literate (KENDALL, 1938; KENDALL, 1945; STUART, 1953), the τ_B being the version used by Otani et al. (2019), and described as

$$\tau_B = \frac{k_c - k_d}{\sqrt{(k_0 - k_1)(k_0 - k_2)}}\,, \tag{2.9}$$

$$k_0 = k(k-1)/2\,, \tag{2.10}$$

$$k_1 = \sum_i t_i(t_i - 1)/2\,, \tag{2.11}$$

$$k_2 = \sum_j u_j(u_j - 1)/2\,, \tag{2.12}$$

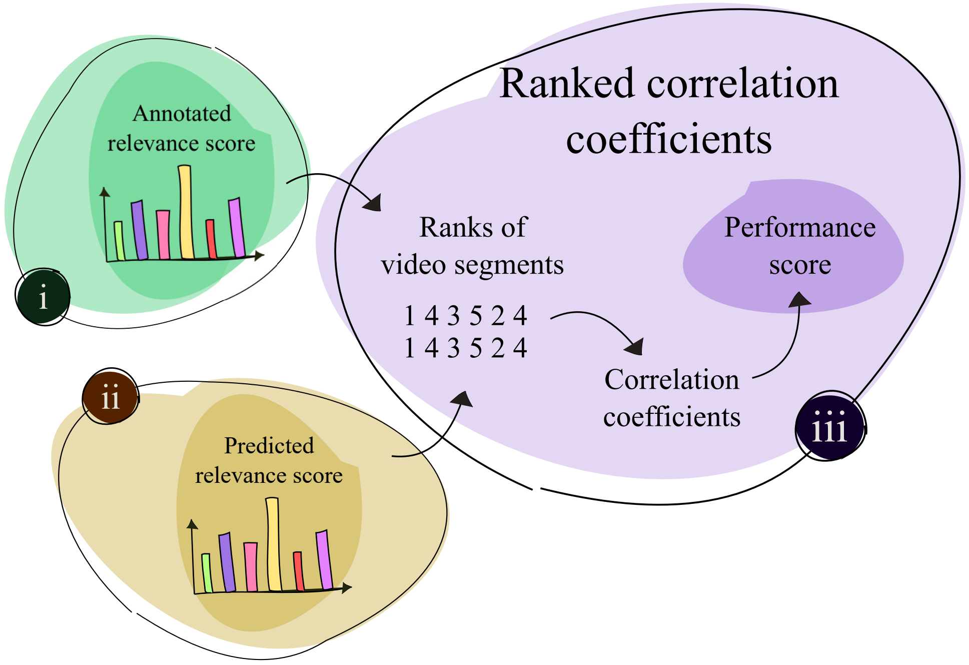


Figure 2.6 The simplified model of evaluation approach using RCC. The relevance scores that come out of the (i) collecting of user annotations and (ii) generic video summarizers are ranked in order of importance and then compared. The internal stages in (i) and (ii) were omitted.

where k_c, k_d, t_i, u_j, k are the total of concordant and discordant pairs, the ties in the first and second group, and the number of video segments, respectively. Similarly, r_s Spearman coefficient matches X ranked scores generated from video summarizers' prediction to M annotated relevance scores, according to

$$r_s = \rho_{\mathrm{r}_X,\mathrm{r}_M} = \frac{\mathrm{cov}(\mathrm{r}_X, \mathrm{r}_M)}{\sigma_{\mathrm{r}_X}\sigma_{\mathrm{r}_M}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(m_i - \bar{m})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(m_i - \bar{m})^2}} \,. \tag{2.13}$$

where r_s is the usual Pearson correlation (PEARSON, 1904), the $\mathbf{x}$ and $\mathbf{y}$ vectors being the annotated and predicted relevance values, respectively, for each i video segment of a input video.

After the computation of both RCC, p-values (*i.e.*, statistical significance) indicate whether the RCC values were not found by chance. Although statistical analyzes can consider several values of significance, three values are often used: 0.05 (5%), 0.01 (1%), and 0.001 (0.01%). If the probability is lower than 0.05 (5%), the association between variables is not incidental and can be deemed statistically significant. Moreover, lower p-values such as 0.01 and 0.001 show an even higher level of significance.

As a rule, the Kendall and Spearman values lie in the range $[-1, +1]$ for any data association. Values below zero (a negative correlation) represents an inverse relationship; that is, the high relevance scores misestimated by the video summarizers match to the low scores annotated by users, and vice-versa. This situation is illustrated in Fig. 2.7(a). The predicted relevance is high for the green video segment (the first bar in the plots),

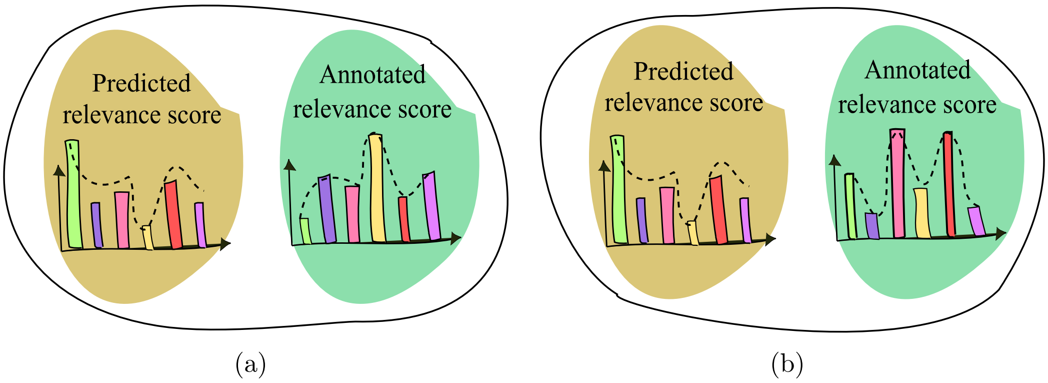


Figure 2.7 Relevance scores whose correlation is negative show inverted values in relation to the $y - axis = 0.5$, whereas the same does not occur when the correlation is null.

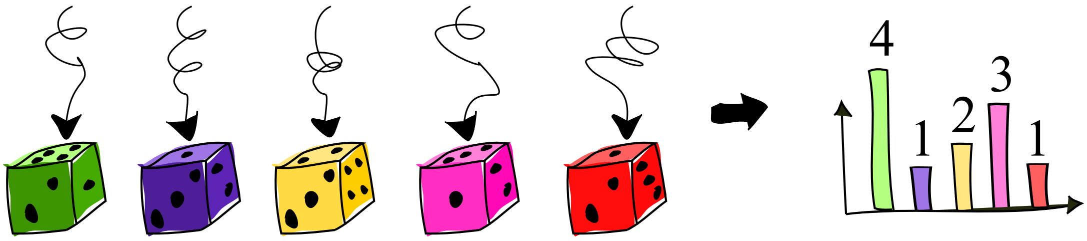


Figure 2.8 For each video segment, represented by different colors, an integer number is randomized (top face of the rolled dice) at a preset range. The set of random integer numbers forms a randomly generated annotation.

but low for users annotation. This oppositional relationship also occurs in other video segments in the plots (such as the yellow one), rendering a negative correlation value. So, the more dissimilar the ranked scores, the closer to -1. In contrast, the more similar the ranked scores, the more positive the correlation coefficient is, peaking at $+1$. When there are no significant associations between relevance scores, the correlation coefficient is null, and hence, 0 (zero). To sum up, video summarizers should pursue positive RCC values, with 0 (zero) being the expected value for the random classification method's performance.

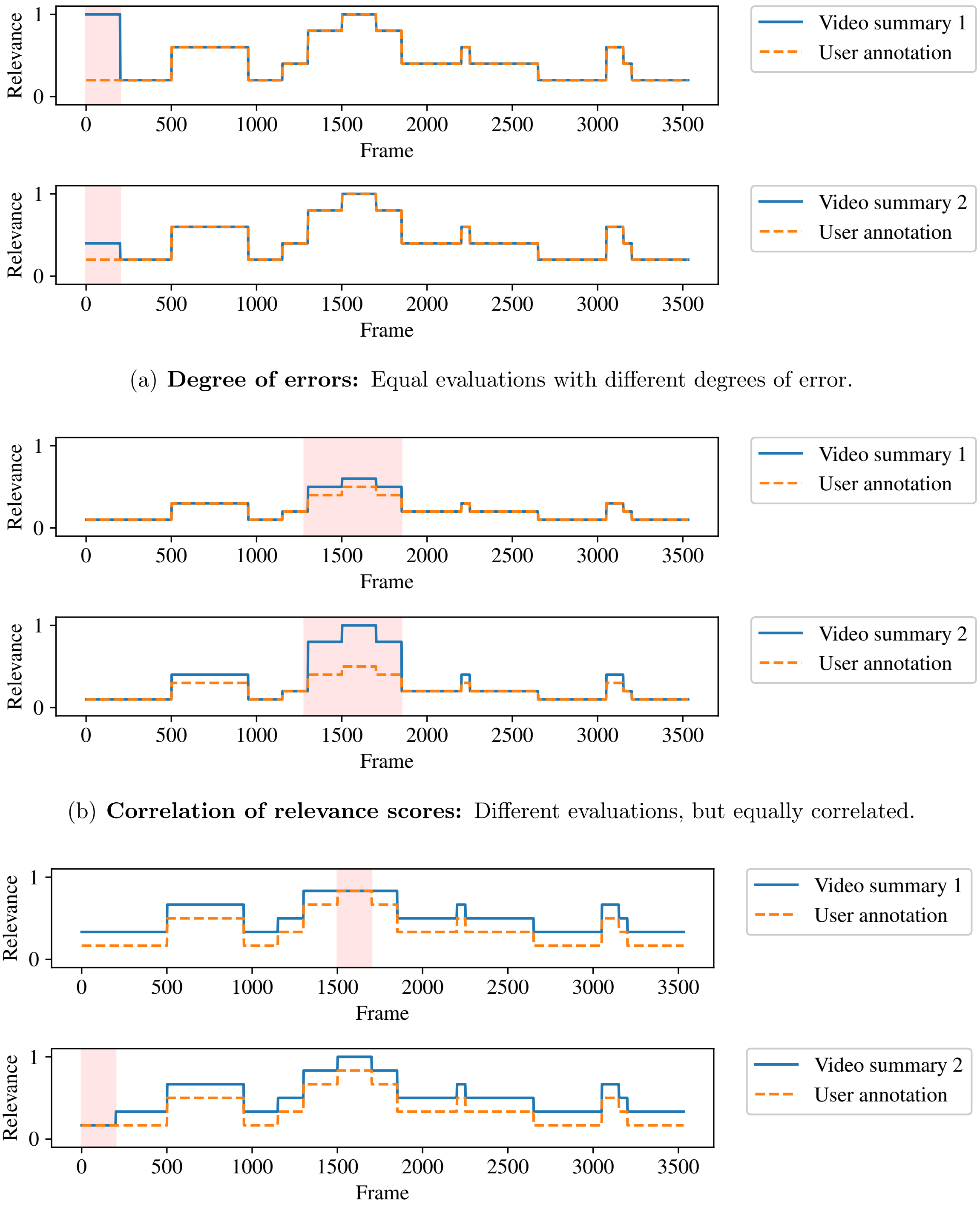
The performance score does not provide information on the quality of the automatic video summaries. For this purpose, at least one video summarizer should be used as a reference for performance, being the random classification method the most common one. In effect, this method generates relevance scores at random for each video segment such as rolling a dice (SONG et al., 2015) (see Fig. 2.8). The top face of the dice simulating a video summarizer that predicts relevance scores by chance. Hence, state-of-the-art video summarizers should pursue a performance score greater than the one achieved by the random classification method, with 0 (zero) being the random classification method's performance in the RCC.

2.8 CURRENT ISSUES ON THE EVALUATION OF VIDEO SUMMARIZERS

F_β and RCC have limitations that make it difficult to assess the performance of video summarizers accurately. The evaluation using F_β addresses the video summarization task as a binary classification of video segments from the input video. As video summarizers typically predict non-binary relevance scores for each video segment, a knapsack solver maps the relevance scores to a video summary at a preset compression rate. RCC can overcome this limitation by matching annotated and predicted relevance scores directly. Hence, RCC do not aim to preserve the most relevant video segments as they are not weighted. By carefully investigating how F_β and RCC assess the performance of video summarizers, we identified three issues: (a) The **degree of error**, when F_β matches the expected and predicted relevance scores directly; (b) the **correlation of relevance scores** that are equivalent in rank order; and (c) **weighing** of the most compressed video summaries. These issues are illustrated in Fig. 2.9, and further detailed.

Keeping in mind that user annotations are collected using ordinal scales with several degrees of relevance. The intuitive solution would be to expand the binary classification model to a non-binary model (multi-label classification) using assessment scales with higher representativeness. In this way, each relevance score annotated by users corresponds to a specific class of the multi-label model. For example, on a three-point Likert scale, the relevance scores 3, 2, and 1 correspond to the "relevant", "neutral", and "irrelevant" class labels, respectively. The evaluation of video summarizers is then performed by matching the annotated and predicted relevance scores directly. In practice, multi-label classification conflicts with psychometric studies on the normative reference as there is no guarantee that users will understand the relevance of information in the same way. Users who are aware of the input video's content may judge video information as belonging to the "irrelevant" class; in contrast, other users may judge this same video information as belonging to the "relevant" class. This situation is illustrated in Fig. 2.9(a). Both video summaries accurately predicted the annotated relevance scores, except in the highlighted magenta area. As evaluation metrics for classification ignores the distance between expected and annotated values, all differences are treated equally as an error; in other words, the error of high-compressed video summaries is equal to low-compressed ones. The evaluation by classification approach also ignores relevance scores' rank, as illustrated in Fig. 2.9(b). In the magenta bars, the annotated and predicted relevance scores are different. However, both generate the same video summaries, as the relevance scores follow the same relevance order. To sum up, the multi-label approach is not suitable for evaluating video summarizers.

Estimating the distance and order of relevance scores is useful, as it increases the discrimination of video summarizers' performance. RCC are more suited for comparing relevance scores as RCC assess the correlation between two variables. However, RCC goes against video summarization as RCC does not aim to preserve the video segments with the highest relevance scores. In Fig. 2.9(c), one video summarizer predicted higher relevance scores than the other. Hence, the video summaries generated from them have different compression rates. The video summarizer that discriminates the relevant video information more accurately should have the highest performance score. Nonetheless, the

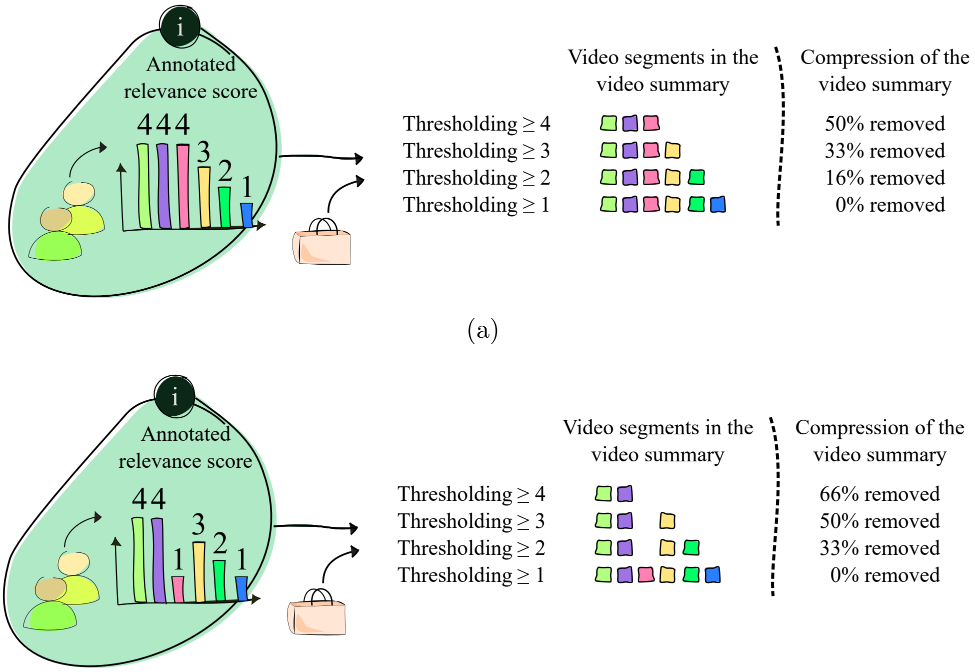


(a) **Degree of errors:** Equal evaluations with different degrees of error.

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

(c) **Weighing of video summaries:** Video summaries are evaluated equally although the compression rates are different, as they hit different relevance levels.

Figure 2.9 Evaluation issues identified in current video summarization metrics.



(a)

(b)

Figure 2.10 The compression rates of videos summaries generated by the thresholding of two user annotations.

performance scores of both video summarizers are equal using RCC. Indeed, there are ways for RCC to weight each relevance score on the ordinal scale. However, even with these approaches, video summaries are not weighted according to the compression rate. Consider Fig. 2.10 that illustrates two users who have annotated six video segments. Except for the magenta video segment, both users annotated the same values for each video segment, as shown in Figs. 2.10(a) and 2.10(b). The user in Fig. 2.10(a) judged the magenta video segment as highly relevant, choosing the relevance score '4'. On the other hand, the user in Fig. 2.10(b) considered the same video segment as of little importance, giving just a '1'. Since both users have the relevance score '4' as the highest order of importance possible for a video segment, this particular relevance score guides the video summary to different compression rates. For instance, selecting the video segments whose relevance score is equal to or above '4' in Fig. 2.10(a), a video summarizer removes 50% of the video segments. Conversely, the compression rate is 66% for the user in Fig. 2.10(b) by selecting video segments with the same relevance score.

2.9 CLOSURE

Current studies evaluate video summarizers using F_β, this being the most widely used (MAHASSENI; LAM; TODOROVIC, 2017; ZHANG et al., 2016; FAJTL et al., 2019; ZHOU; QIAO; XIANG, 2017; OTANI et al., 2017; ROCHAN; YE; WANG, 2018; GYGLI et al., 2014; SONG et al., 2015; CHU; Yale Song; JAIMES, 2015). More recently, (OTANI et al., 2019) introduced RCC with aim at coping with the limitations of F_β . However, both metrics present issues when applied to video summarization. Next chapter, we propose a novel metric to overcome these issues.

Chapter

3

COMPRESSION LEVEL OF USER ANNOTATION (CLUSA)

Contents

To overcome the issues of F_β and ranked correlation coefficients (RCC) in video summarization, we conceived a novel metric named Compression Level of USer Annotation (CLUSA) by assigning relevance scores to the order of importance. While current studies (FAJTL et al., 2019; ZHOU; QIAO; XIANG, 2017; ROCHAN; YE; WANG, 2018; OTANI et al., 2017; ZHANG et al., 2016) evaluate video summarizers by mapping multi-valued relevance scores to video summaries at a single compression rate, CLUSA evaluates the relevance scores predicted by video summarizers (see the balloon (ii) in Fig. 3.1) at multiple compression rates. To do so, CLUSA generates all possible video summaries from the **annotated relevance scores** (see the balloon (i) in Fig. 3.1) by gradually discarding video segments according to their relevance (the thresholding process in the balloon (iii) of Fig. 3.1). For example, CLUSA extracted three video summaries with different compression rates from annotated relevance scores in the balloon (i) by assigning value 1 to the selected video segments (the bars in $\mathbf{X}$) and value 0, otherwise. After grouping these video summaries, each video summary is matched to relevance scores predicted by video summarizers in the balloon (iii). CLUSA then computes an overall performance score by weighing the mean matching scores.

3.1 FORMAL DEFINITION OF CLUSA

Let $\mathbf{m} = (m_j) \in \mathbb{R}^K$ be a vector containing relevance scores predicted by a video summarizer for K video segments. To properly evaluate $\mathbf{m}$, CLUSA requires a data set

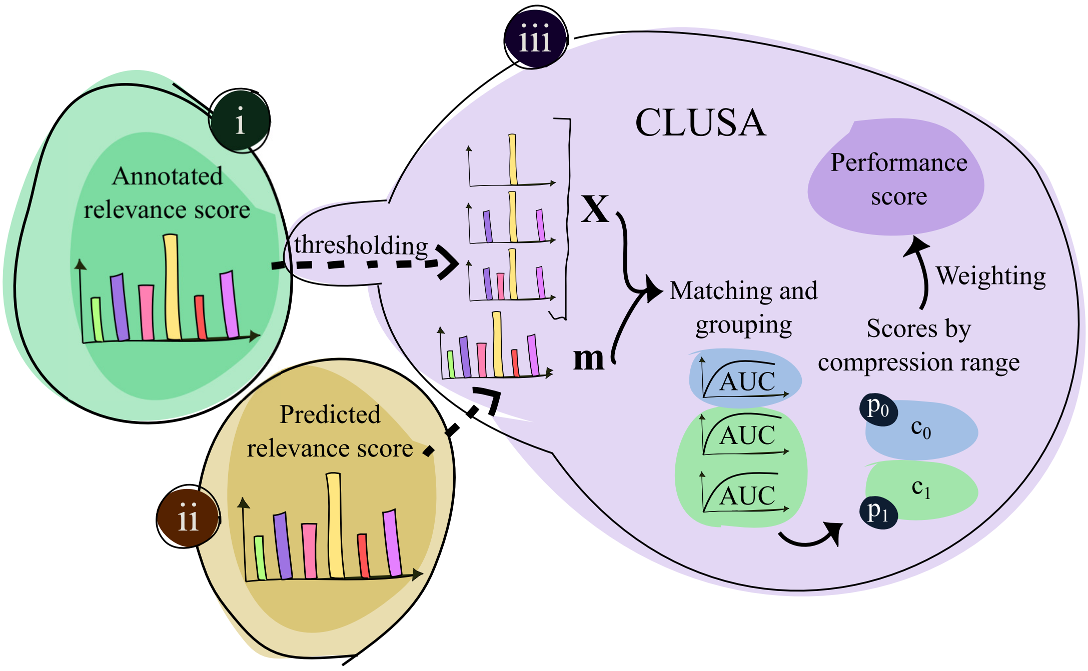


Figure 3.1 The simplified model of CLUSA's evaluation approach. After mapping the relevance scores that come out of (i) collecting user annotations, the video summaries are matched with the output from the (ii) generic video summarization method, being later grouped and weighed by compression range. The internal stages in (i) and (ii) were omitted.

$\mathbf{D} = (d_{i,j}) \in \mathbb{R}^{U\times K}$ annotated by U users. As $\mathbf{D}$ values can be non-binary, they are mapped to $\mathbf{O}_i$ video summaries considering the unique relevance scores in each row, $\mathbf{u}_i$. Here, we express this operation by defining a set of distinct relevance scores as

$$\mathbf{u}_i = \{d_{i,j} : \forall j, 1 \leq i \leq U, 1 \leq j \leq K\}\,. \tag{3.1}$$

The top-down example in Fig. 3.2 illustrates this mapping process. Starting on a single row-vector of $\mathbf{D}$ matrix, CLUSA applies thresholds of 0.2 and 0.6 to generate two concatenated video summaries, $\mathbf{O}_i$. $\mathbf{O}_i$ is given by

$$\mathbf{O}_i = ([d_{i,j} \geq u_{i,k}] : 1 \leq k \leq |\mathbf{u}_i|) - 1 \in \mathbb{R}^{|\mathbf{u}_i|-1\times K}\,, \tag{3.2}$$

$|\mathbf{u}_i|$ being the cardinality of $\mathbf{u}_i$, this is the number of elements. CLUSA thresholds the row-vectors $\mathbf{d_i}$ using each $u_{i,k}$ value. Together, the resulting row-vectors make up the $\mathbf{O}$ matrices.

As the highest values in $\mathbf{u}_i$ leads all values in $\mathbf{O}_i$ to zero, CLUSA discards them and concatenates $\mathbf{O}_i$ video summaries into a single matrix, $\mathbf{X} = (x_{i,j}) \in \mathbb{R}^{(\sum(|\mathbf{u}_i|-1)\times K} : (\mathbf{O}_1^T|\mathbf{O}_2^T|...|\mathbf{O}_i^T)^T$. The steps described above built a set of video summaries (as shown in each row of Fig. 3.3(a)), $\mathbf{X}$, from the user annotations, $\mathbf{D}$ (illustrated in each row of Fig. 3.3(b)).

Each row-vector, $\mathbf{x}_i \in \mathbf{X}$, denotes a binary form obtained from user annotation, so CLUSA computes a matching score vector, $\mathbf{z}_i$, given by

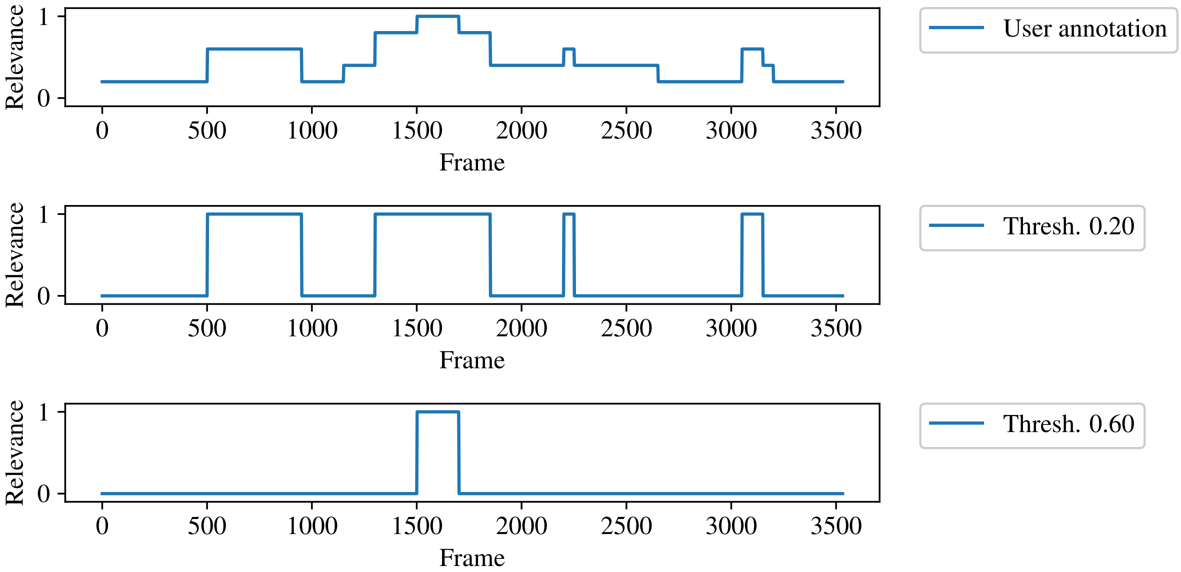
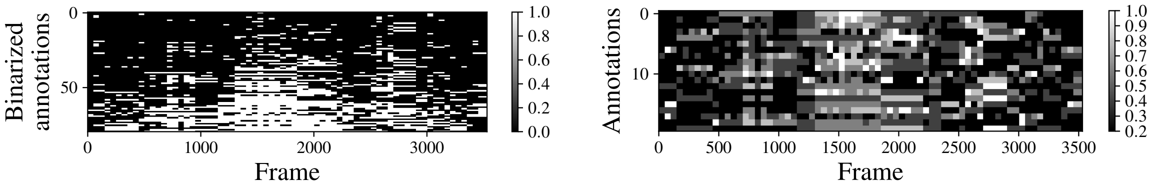


Figure 3.2 Thresholding a user annotation into several relevance levels.



(a) Binarized user annotations.

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

Figure 3.3 (a) illustrates the **X** result map generated from the user annotations arranged as row-vectors in (b). The row-vectors in (a) were ordered by their compression rate.

$$\mathbf{z}_i = \left(\theta(\mathbf{m}, \mathbf{x}_i) : 1 \leq i \leq \sum |\mathbf{u}_i|\right), \tag{3.3}$$

where θ is a vanilla function, which matches $\mathbf{m}$ with $\mathbf{x}_i$ values, such as the area under Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves (AUC-ROC and AUC-PR). Hence, if $\mathbf{m}$ relates to $\mathbf{x}_i$ on an exact monotonic association, all area under curve (AUC) values reach the maximum area, $z_i = 1$. The matching process is depicted in Fig. 3.4: User annotations are binarized (on the left) and then matched to the relevance scores (on the right), delivering a different ROC curve for each match.

To weight the matching scores properly, CLUSA calculates the ratio between the discarded video segments and the entire video for each row-vector in **X** data. This ratio is here called the compression rate, $\mathbf{w}_i = P(\mathbf{x}_i = 0)$, which represents the proportion of video segments not included in the video summaries. The score vector, $\mathbf{z}_i$, is then grouped into clusters, $\mathbf{c}_i$, according to the video summaries' compression rate, $\mathbf{w}_i$. The

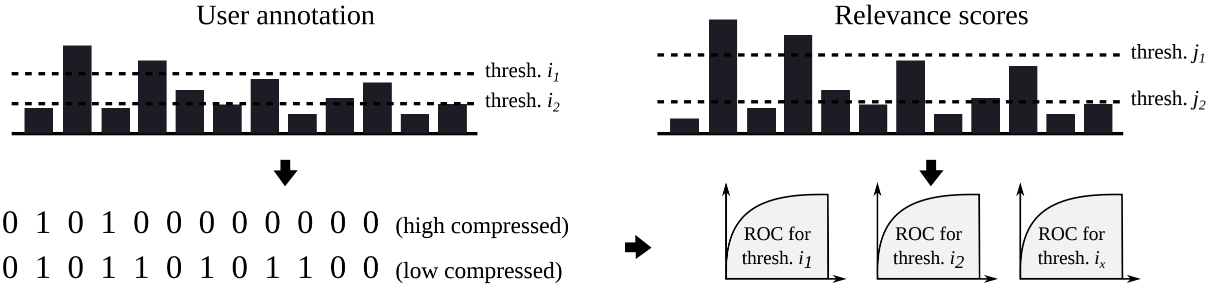


Figure 3.4 To match user annotation and predicted relevance scores, CLUSA metric computes area under ROC curves from the video summaries associated to a compression rate.

$\mathbf{c}_i$ clusters are defined as

$$\begin{gathered}\mathbf{c}_i = \mu(\mathbf{z}_k : \|\mathbf{w}_k - \mathbf{p}_i\|^2 \leq \|\mathbf{w}_k - \mathbf{p}_j\|^2, \forall j, 1 \leq i \leq j \leq B, \\ 1 \leq k \leq \sum |\mathbf{u}_i|)\,,\end{gathered} \tag{3.4}$$

where B represents the number of compression ranges, while $\mathbf{p}_i$ is a median point in each range, given by $\mathbf{p}_i = (2i-1)(2B)^{-1}, 1 \leq i \leq B$. $\mathbf{c}_i$ clusters are suitable to assess video summarizers' performance in a general way; that is to say that we can now compare techniques considering any compression range.

The mean scores of $\mathbf{c}_i$ clusters are weighted using the $\mathbf{p}_i$ values, and at the end, CLUSA is giving by

$$\text{CLUSA}(\mathbf{D},\mathbf{m}) = \mathbf{p}^T\mathbf{c}\,. \tag{3.5}$$

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

3.2 SELECTING A SUITABLE θ FUNCTION FOR CLUSA

In the SumMe (GYGLI et al., 2014) and TVSum50 (SONG et al., 2015) data sets, the total of relevant and non-relevant video segments are unbalanced, while the video segments labeled as "less relevant" are more frequent. As ROC is known for skewed assessments in situations whereby the distribution of labels is unbalanced (FAWCETT, 2006; DAVIS; GOADRICH, 2006), a less bias-sensitive approach is needed to establish guidelines for the general interpretation of CLUSA scores. Therefore, we propose to calculate the $\theta(\mathbf{m}, \mathbf{x}_i)$ **matching** in Fig. 3.1 using the PR curve, instead of ROC curve. PR curves focus on positive labels (*i.e.*, relevant video segments) by calculating pairs of precision and recall rates. Hence, the predominant label does not skew the performance scores when calculating the AUC.

Both AUC-ROC and AUC-PR scores behave differently, as depicted in Fig. 3.5. For the c_i scores with AUC-PR, the random classification method's performance inversely follows the video summaries' compression rate, p_i, because the probability of having a

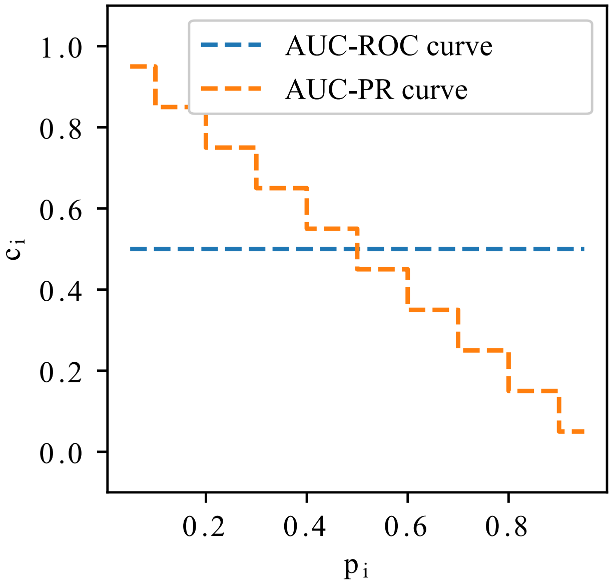


Figure 3.5 The expected mean scores c_i for each p_i compression range using CLUSA with AUC-ROC and AUC-PR curves considering randomly generated relevance scores.

relevant label assigned is higher at low compression (the orange dashed line). On the other hand, c_i should not deviate from a constant value (the blue dashed line) for the AUC-ROC curve. Both curves plotted in Fig. 3.5 illustrate the expected performance for a random classification method if CLUSA evaluates all p_i compression ranges. However, video summarization data sets cover only a few compression rates, changing how we interpret the CLUSA score values.

3.3 INTERPRETING CLUSA SCORE VALUES

By default, CLUSA sets the value of B to 10. CLUSA presets ten equally divided compression ranges, which together cover all compression rates. Although this parameter can be changed, values lower than the default tend to increase empty groups' occurrence. These, namely the compression ranges that do not have annotated video summaries, are not counted toward CLUSA general score, reducing video summarizers' performance.

To illustrate how empty groups impair the evaluation of video summarizers, consider the video summaries produced by the user at the top of Fig. 2.10. Three compression ranges are covered, $]10\%, 20\%]$, $]30\%, 40\%]$, and $]40\%, 50\%]$, as opposed to the user at the bottom whose values include $]30\%, 40\%]$, $]40\%, 50\%]$, and $]60\%, 70\%]$. Even if we concatenate these two users' video summaries, there are no video summaries in the compression range $]70\%, 100\%]$. As CLUSA does not assess video summarizers' performance at these compression ranges, the performance score decreases for all video summarizers, including the random classification method. As it is necessary to identify the video summaries' compression ranges in video summarization data sets, the interpretation of the CLUSA is not as straightforward as for the RCC. Formally, the total number of combinations for the

Cumulative score for random performance	CLUSA w/ AUC-ROC	CLUSA w/ AUC-PR
$\sum_{i=0.9}^{1.0} c_i p_i$	0.09	0.01
$\sum_{i=0.8}^{1.0} c_i p_i$	0.18	0.04
$\sum_{i=0.7}^{1.0} c_i p_i$	0.26	0.07
$\sum_{i=0.6}^{1.0} c_i p_i$	0.32	0.12
$\sum_{i=0.5}^{1.0} c_i p_i$	0.38	0.17
$\sum_{i=0.4}^{1.0} c_i p_i$	0.42	0.22
$\sum_{i=0.3}^{1.0} c_i p_i$	0.45	0.26
$\sum_{i=0.2}^{1.0} c_i p_i$	0.48	0.30
$\sum_{i=0.1}^{1.0} c_i p_i$	0.49	0.33
$\sum_{i=0.0}^{1.0} c_i p_i$	**0.50**	**0.35**

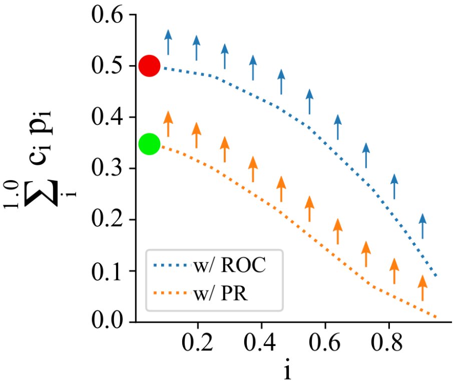


Figure 3.6 Performance scores expected in the evaluation of relevance values generated randomly in specific subsets of compression ranges p_i.

compression ranges, p_i, could be arranged as $\sum_{i=1}^{10}\binom{10}{i}$, which for 10 compression range results in 1023 possible combinations. To simplify the task of interpreting the CLUSA scores, we selected a few combinations by gradually removing the lower compression range to reduce the total number of combinations. The last row of the table in Fig. 3.6 shows how all compression ranges (*i.e.*, $[0\%, 100\%[$ video segments removed) were used for the computation of the performance score. For any video summarizer to be better than a random classification method, it must have a performance greater than 0.50 and 0.35 for CLUSA with AUC-ROC and AUC-PR curves, respectively. Conversely, the first row (*i.e.*, $[90\%, 100\%[$ video segments removed) comprises 10% of the most important video segments of the targeted video that has been used. Thus, if video summarizers can be assessed only in this situation, their performance should be greater than 0.09 and 0.01 for the ROC and PR curves, respectively. All values in Fig. 3.6 are performance scores of the random classification method, but the choice of which score is suitable depends on the statistical distribution of the compression rate for each data set. Therefore, it is necessary to investigate the statistical distribution of $\mathbf{X}$ for each video summarization data set and, from their statistical distribution, determine which compression ranges are missing.

3.4 CLOSURE

Although CLUSA supposedly fulfills all video summarization requirements, issues such as missing compression ranges and unbalanced class labels can skew CLUSA scores. As these particular issues are present in the SumMe and TVSum50 data sets, it is necessary to analyze how CLUSA reacts to both issues. In the next chapter, we detail how we analyze CLUSA when evaluating video summarizers in the SumMe and TVSum50.

Chapter

4

EXPERIMENTAL EVALUATION

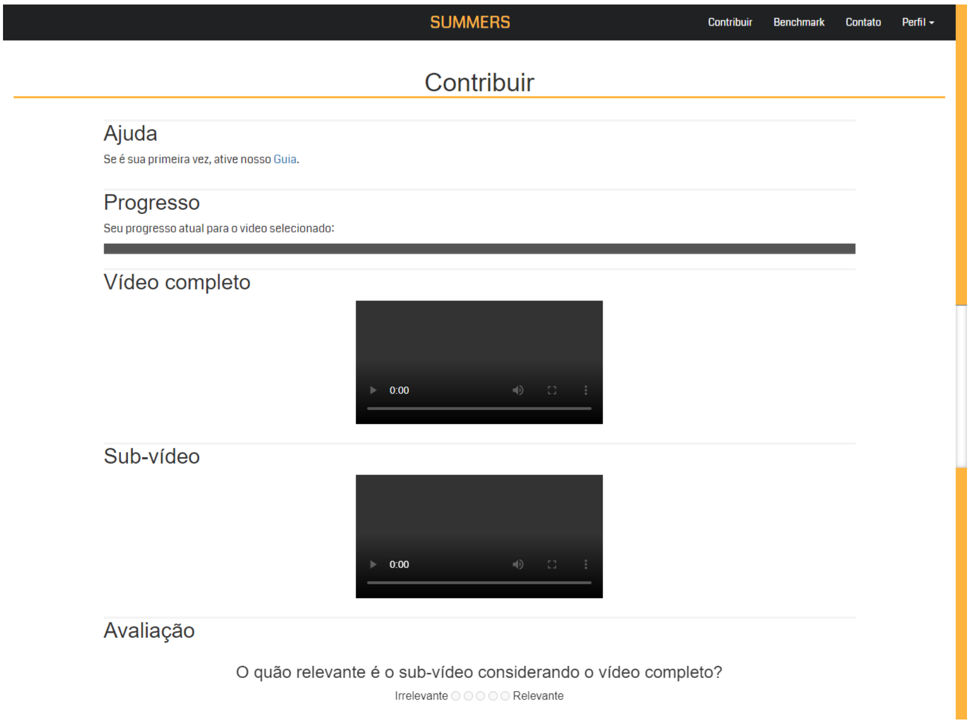
Contents

Our study involves the analysis and discussion around evaluating video summarizers at multiple compression rates. Notably, we investigate three issues in this scope: (a) the quality of user annotations collected with different assessment scales on a standard scenario, (b) the human consistency compatibility with the internal consistency, and (c) an analysis of evaluation metrics while deploying state-of-the-art video summarizers at multiple compression rates. We describe each of these experiments in this chapter.

4.1 METHODOLOGY OF OUR STUDY

4.1.1 Collecting user annotations on a standard scenario

As the quality of user annotations is inherently related to the guideline deployed for collecting data, the assessment scales are also supposed to affect user annotations' quality. SumMe and TVSum50 have collected annotations in different ways, and as a result,



Contribuir

Ajuda

Se é sua primeira vez, ative nosso Guia.

Progresso

Seu progresso atual para o video selecionado:

Vídeo completo

Sub-vídeo

Avaliação

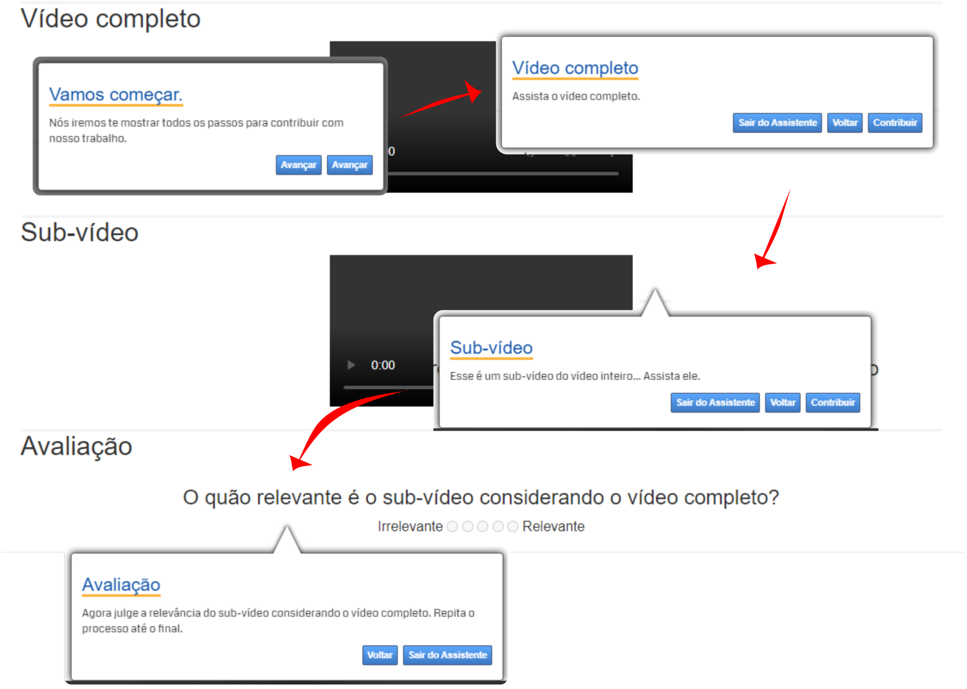
O quão relevante é o sub-vídeo considerando o vídeo completo?

Irrelevante ○ ○ ○ ○ ○ Relevante

Figure 4.1 User interface of Summers.

we cannot compare their collecting process directly. To circumvent this problem, we collected annotations using our web-based annotation tool called Summers on a standardized scenario where the same users annotated the same videos using three types of ordinal assessment scales: Binary, three-point Likert scale, and five-point Likert scale. Figure 4.1 illustrates the Summers' user interface. In top-down order, Summers comprises of: (i) A guide with step-by-step instructions to enable users to use our annotation tool, (ii) a progress bar for the current video, (iii) the input video, (iv) a video segment to be judged, and finally, (v) the assessment scale with the available relevance scores.

Users may not comprehend how to proceed in the collecting process. Thus, we introduced clear and unequivocal instructions to the users at the very beginning of the session, as illustrated in Fig. 4.2. Admittedly, users might feel tempted to skip the instructions. Therefore, we set up a lock mechanism on our annotation tool to prevent this behavior from happening the first time test instructions are shown. Figure 4.2 illustrates four messages displayed to users. First, the tool clarifies that the messages are step-by-step instructions. Second, the annotation tool compels users to watch the entire video. Only after they have watched the entire video, the third message is shown requesting users to watch a video segment. Only after watching the video segment, Summers enables the assessment scale for the user to pick a relevance score. Finally, users are requested to judge relevance on an assessment scale. After guiding users on their first annotation, Summers shuffles the sequence of muted video segments presented to users, as users tend to annotate higher relevance scores to the video segments that appear earlier Song et al. (2015).



Vídeo completo

Vamos começar.

Nós iremos te mostrar todos os passos para contribuir com nosso trabalho.

Vídeo completo

Assista o vídeo completo.

Sub-vídeo

Sub-vídeo

Esse é um sub-vídeo do vídeo inteiro... Assista ele.

Avaliação

O quão relevante é o sub-vídeo considerando o vídeo completo?

Irrelevante ○ ○ ○ ○ ○ Relevante

Avaliação

Agora julge a relevância do sub-vídeo considerando o vídeo completo. Repita o processo até o final.

Figure 4.2 Our annotation tool presents the instructions to users at the beginning of an annotation session. The next instruction is shown to the user only after completing the action required in the previous instruction.

(a)

(b)

Figure 4.3 Samples of used videos from UCF101 data set: (a) Surfing and (b) basketball.

To determine what videos are to be annotated by the users in Summers, we queried two types of actions in the UCF101 collection of videos (SOOMRO; ZAMIR; SHAH, 2012): **Surfing** and **basketball**. Assuming that users are familiar with this kind of content, we selected ten videos whose duration was around three minutes. Some samples of the selected videos are illustrated in Fig. 4.3. Following, the boundary video shot detector proposed by Gygli et al. (2014) split the input videos into video segments to be annotated.

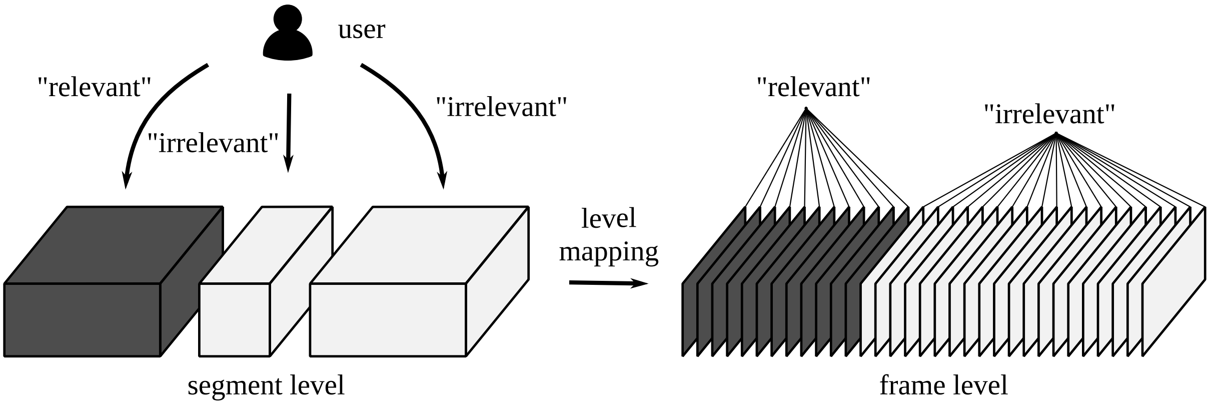


Figure 4.4 Mapping the annotated relevance at video segments to frame level (level mapping).

4.1.2 Assessing the quality of user annotations collected from Summers

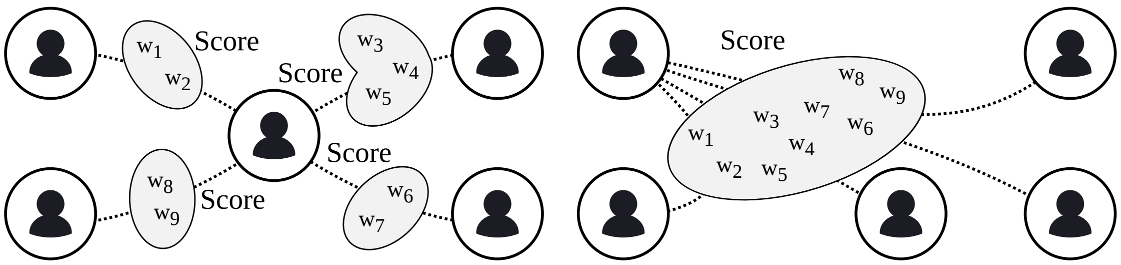
Users annotated the relevance of a specific arrangement of video segments shuffled in the annotation collecting. However, video summarizers are not limited to the segments annotated by the users, but any resulting arrangement of a video shot segmentation. The relevance scores annotated by users were then mapped to the frame level to allow for the evaluation of video summarizers regardless of their video shot segmentation. As illustrated in Fig. 4.4, this segment-to-frame mapping is performed by repeating the video segments' relevance value in the video frames. Conversely, psychometric estimators assess the quality at the level where users annotated the data; in our case, at the segment level. Therefore, we bring the level mapping back to the segment level to calculate Cronbach's alpha values. By comparing Cronbach's alpha values of user annotations collected with different assessment scales, we sought to elucidate the relationship between assessment scales' representativeness and Cronbach's alpha values.

4.1.3 Relating evaluation metrics for video summarization to internal consistency

Gygli et al. (2014) calculated human consistency by averaging the distances between pairs of users using F_β, as shown in Fig. 4.5(a). Similarly, we calculate the human consistency of user annotations using CLUSA. Unlike F_β, CLUSA matches one user annotation, $\mathbf{m}$, to multiple user annotations concatenated in a single matrix, $\mathbf{D}$. As the pair-wise strategy implies that $\mathbf{D}$ contains only one user annotation, the pair-wise strategy impinges CLUSA. Accordingly, we also calculated the human consistency using a leave-one-out strategy. In this strategy, a row-vector of $\mathbf{D}$ is removed and becomes the vector $\mathbf{m}$. In other words, CLUSA evaluates one user from all others, as illustrated in Figure 4.5(b).

(see Fig. 4.5(b))

After we assessed the human consistency with F_β and CLUSA, we related the ordering of their scores from different assessment scales to Cronbach's alpha. Cronbach's alpha and evaluation metrics measure users' agreement; thus, we assumed that both should follow a



(a)

(b)

Figure 4.5 Approaches used to compute human consistency using CLUSA: (a) Pair-wise and (b) leave-one-out.

similar order. By comparing the order relation, we investigated: (a) Whether the F_β and CLUSA deal with more representative assessment scales similar to psychometric quality estimators, and (b) whether F_β and CLUSA can deal with relevance scores directly. When we carried out this stage of our experimental study, Otani et al. (2019) had not yet introduced RCC in the evaluation of video summarizers. For this reason, we did not assess human consistency using Kendall and Spearman coefficients.

4.1.4 Assessing video summarizers' performance using evaluation metrics

Assuming that Cronbach's alpha values are at least acceptable using the values in Table 2.1 as a reference, we used the user annotations to evaluate video summarizers. We selected state-of-the-art video summarizers by focusing primarily on video summarizers whose source codes are available in code repositories, such as GitHub. Our search query on GitHub was performed by date (*e.g.*, recently updated) with the combined words "video summarization". The search results returned six automatic state-of-the-art methods based on deep learning techniques. Aiming at video summarizers whose backbone layers were pre-trained (see Table 4.1), we downloaded from GitHub five video summarizers: dppLSTM (ZHANG et al., 2016), vasnet (FAJTL et al., 2019), vsummRI (ZHOU; QIAO; XIANG, 2017), vsumDSF (OTANI et al., 2017), and VSwFCSN (ROCHAN; YE; WANG, 2018). Accordingly, we discarded the AVS as the pre-trained weights are not available on GitHub. Given that our study aims to evaluate video summarizers at multiple compression rates, we removed the knapsack solver stage in the source codes provided by the authors.

Briefly, the results of human consistency (presented and discussed later in this chapter) showed that the F_β is not suitable for evaluating video summaries directly from the relevance scores. Therefore, we did not analyze F_β in this second experimental study. Instead, we analyze Kendall and Spearman correlation coefficients and CLUSA with ROC or PR curves, using the performance of a random classification method as a reference.

To determine the performance score expected for a random classification method, we simulated one by generating integer numbers arbitrarily within the range $[1, 5]$ for each annotated video in the SumMe (GYGLI et al., 2014) and TVSum50 (SONG et al., 2015). Then, we compared the average performance of 500 video summaries generated by the

Table 4.1 State-of-the-art video summarizers retrieved from GitHub repositories.

Method	**Pre-trained in data set**	
	SumMe	TVSum50
AVS (MAHASSENI; LAM; TODOROVIC, 2017)		
dppLSTM (ZHANG et al., 2016)	features	features
vasnet (FAJTL et al., 2019)	features	features
vsummRI (ZHOU; QIAO; XIANG, 2017)	features	features
vsumDSF (OTANI et al., 2017)	features	
VSwFCSN (ROCHAN; YE; WANG, 2018)		features

random classification method with state-of-the-art video summarizers.

As missing compression ranges are supposed to affect CLUSA, we analyzed the frequency of compression ranges in each data set to compare the performance of the random classification method against the performance scores depicted in Fig. 3.6.

4.2 EXPERIMENTAL RESULTS

4.2.1 The quality of user annotations collected with different assessment scales

Initially, we asked users to annotate ten action videos. However, just four were annotated by sixteen users (on average) using the Binary, Likert-3, and Likert-5 assessment scales. Therefore we reduce the total of videos to avoid the users' withdraw during the annotation process.

In our standardized scenario, with the same annotated videos and users, we are interested in the impact of assessment scales on video summarization tasks. Therefore, we calculated Cronbach's alphas for user annotations grouped by the assessment scale in Table 4.2. The Cronbach's alpha values for our user annotations increase proportionally to the assessment scale's degree, suggesting that multi-valued assessment scales are more suitable to collect user annotations for video summarization tasks. The five-point Likert scale turned out to be the most suitable assessment scale for video summarization tasks, considering human consistency growth. This finding does not rule out the potential use of assessment scales higher than 5 points; however, the increase 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 more scale points.

Cronbach's alpha values in Table 4.2 are different at frame- and segment- levels. The repeated relevance scores at the frame level skewed the Cronbach's alpha values, and the quality of users annotations collected with Likert-3 were reduced from "good" (at the frame level) to "acceptable" (at the segment level) by taking Table 2.1 as a reference.

4.2.2 Using CLUSA to calculate the human consistency

We compared the ordering of CLUSA scores with the Cronbach's alpha and F_β scores. Table 4.3 summarizes the results, and the arrows in the row "Ours" highlight the in-

Table 4.2 Cronbach's alpha for different assessment scales: Binary and multi-valued (Likert-3 and Likert-5).

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

Table 4.3 Human consistency using F_β and CLUSA in their respective assessment scales: Binary and multi-valued (Likert-3 and Likert-5). The green, red and blue arrows highlight the growth of the values (from smallest to largest).

Data set	**Assessment scale**	**Internal consistency**			
			Pair-wise		**Leave-One-Out**
		Cronbach's α	**F$_\beta$**	**CLUSA**	**CLUSA**
Ours	binary	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	binary (ego)	0.766	0.292	0.103	0.212
	binary (moving)	0.748	0.308	0.104	0.176
	binary (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

ternal consistency growth on the Cronbach's alpha, F_β, and CLUSA by changing the assessment scale. The order of F_β scores presents the opposite behavior (decreasing as the degree of assessment scales increases) concerning Cronbach's alpha in our standardized scenario (row "Ours"); in other words, F_β indicated that binary assessment scale should be more consistent than multi-valued ones. Conversely, the growth of human consistency in CLUSA became similar to Cronbach's alpha, suggesting that both deal with user annotations similarly.

Table 4.3 also summarizes the results of SumMe and TVSum50 according to the characteristics of each data set. 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). As SumMe and TVSum50 annotations were collected using different guidelines, video contents, and users, we cannot directly compare the Cronbach's alpha values. Therefore, we analyzed each row in isolation. By comparing CLUSA values using pair-wise and leave-one-out strategies, the human consistency of annotations collected with binary assessment scales (in the rows 'SumMe' and 'Ours') came closer than the annotations collected with a Five-point Likert scale (in the rows 'TVSum50' and 'Ours'). This increase of CLUSA values in the five-point Likert scale suggests that the leave-one-out strategy is less impaired by the issue of missing compression ranges.

4.2.3 Ranking video summarizers by their performance scores

Using the SumMe's and TVSum50's annotations, we assessed the performance scores of the earlier listed state-of-the-art video summarizers and the random classification method. Table 4.4 summarizes the performance scores computed using the Spearman and Kendall correlation coefficients, and using CLUSA with AUC-ROC and AUC-PR curves. We highlighted the highest performance score on each row.

Table 4.4 Performance scores of state-of-the-art video summarization methods using Kendall and Spearman correlation coefficients, and using CLUSA metrics with AUC-ROC and AUC-PR curves.

Data set	Evaluation metric	Automatic method					
		dppLSTM	**vasnet**	**vsummRI**	**VSwFCSN**	**vsumDSF**	**random**
SumMe	Kendall	-0.074***	**0.057***	-0.074		-0.049*	0.000
	Spearman	-0.097**	**0.074***	-0.101		-0.055*	0.000
	CLUSA AUC-ROC	0.314	**0.405**	0.300		0.345	0.157
	CLUSA AUC-PR	0.176	**0.228**	0.172		0.204	0.063
TVSum50	Kendall	**0.043***	-0.074**	-0.012*	-0.004		0.000
	Spearman	**0.055***	-0.101*	-0.013*	-0.005		0.000
	CLUSA AUC-ROC	0.448	**0.692**	0.486	0.499		0.423
	CLUSA AUC-PR	0.262	**0.480**	0.347	0.349		0.285

Statistical significance. * $p < 0.05$. ** $p < 0.01$. *** $p < 0.005$.

For each evaluation metric in Table 4.4, we ranked the video summarizers by decreasing order of performance in the SumMe and TVSum50 data sets. Table 4.5 summarizes the rank position of each video summarizer. The performance scores computed using AUC-ROC and AUC-PR curves delivered the same rank order for video summarizers in the SumMe and TVSum50 so that we merged the results of both curves in the column "CLUSA" and the results of the Spearman and Kendall coefficients in the column "RCC". It is worth noting that there was a divergence in the Kendall and Spearman coefficients in the TVSum50: The vsummRI rank position (the value indicated by the magenta arrow) tied for 3rd using the Kendall coefficient, but not using the Spearman coefficient.

Whereas vasnet is ranked first using CLUSA, it switches to the last position using RCC (as indicated by the blue arrow in Table 4.5). The same trend, but in the opposite direction, was observed for dppLSTM, whose position was 4th using CLUSA, moving up to 1st using RCC (as indicated by the green arrow).

Table 4.5 Rank of state-of-the-art video summarization methods using RCC and CLUSA metric. The subcategories of each evaluation metric were omitted since the ranks are the same across metrics, with the exception of the method vsummRI using RCC.

Data set	Automatic method	Metrics (Ranking order) CLUSA	 RCC
SumMe	vasnet	1st	1st
	vsumDSF	2nd	2nd
	dppLSTM	3rd	3rd
	vsummRI	4th	3rd/4th ←
TVSum50	vasnet	1st	4th
	VSwFCSN	2nd	2nd
	vsummRI	3rd	3rd
	dppLSTM	4th	1st

Table 4.6 The compression rate distribution of video summaries generated from user annotations in each video summarization data set.

Data set	p_i 0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
SumMe	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.874	0.126
TVSum50	0.000	0.000	0.001	0.004	0.189	0.056	0.077	0.159	0.192	0.321

4.2.4 The compression ranges in which video summarizers are accurate

The reason why RCC and CLUSA ranked video summarizers differently is in the assessed compression ranges. Therefore, we investigated the compression rate distribution on SumMe and TVSum50. Gygli et al. (2014) prioritized the annotation of highly compressed video summaries in the SumMe, as shown in Table 4.6. So the compression rate distribution lied in the ranges [0.8.0.9[and $[0.9, 1.0[$. Since [0.8.0.9[contains 87% of the total annotated summaries, CLUSA weighted almost all video summaries equally, and hence, the ranking of video summarizers using CLUSA and RCC came equally in the SumMe.

The compression rate distribution in the SumMe reveals why CLUSA and RCC behave similarly, but to explain the contrasting pattern of both metrics in TVSum50. In contrast to SumMe, the compression rates in TVSum50 are more spread out, covering more compression ranges. As shown in Table 4.6, for all compression ranges above 20%, there is at least one video summary.

Since the weighing introduced by CLUSA plays a crucial role in the evaluation of video summarizers, we investigate in which compression ranges the video summarizers are accurate on the TVSum50. In Fig. 4.6, we combined the mean scores of all 50 videos that comprise TVSum50 by assigning zero to the missing compression ranges. Computing both AUC-ROC (the left plot) and AUC-PR (the right plot) curves, dppLSTM performed

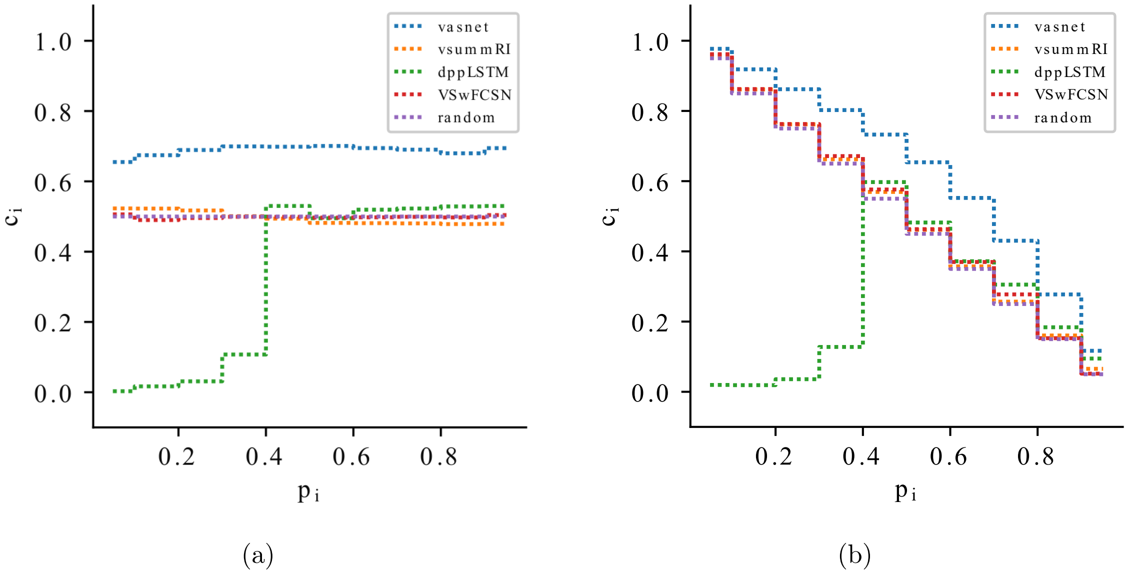


Figure 4.6 Mean performance scores on each compression range using CLUSA with: (a) Area under ROC curve and (b) area under PR curve in the TVSum50 data set.

poorly in compression ranges below 40%, albeit slightly better than other video summarizers for higher ranges. Although compression ranges above 40% are more weighed, the higher values' performance was insufficient to compensate because dppLSTM did not predict the video segments' relevance in low compression rates. For this reason, dppLSTM ranked 4^{th} in CLUSA.

The performance evaluation of video summarizers in Fig. 4.6 includes two situations of label imbalance: Preserving or removing as many video segments as possible. AUC-ROC (the left plot) and AUC-PR (the right plot) curves respond differently to the label imbalance. At very low ($]0.0, 0.1[$) and very high ($[0.9, 1.0[$) compression ranges, the ROC curve is more sensitive. Hence, AUC-ROC boosted the mean scores as the compression weights are higher at the very high compression range than the others. In contrast, the mean scores at very low and at very high compression ranges tend to be more steady for the PR curve, causing the AUC-PR curve to come closer to the random classification at both compression ranges. To sum up, how the ROC and PR curves deal with label imbalance justifies the substantial difference in the performance scores found between both curves.

4.2.5 Comparing video summarizers with a baseline

The performance of the random classification method sets a baseline for the quantitative evaluation of video summarizers. For the Spearman and Kendall correlation coefficients, this baseline performance score reached the null correlation (the value 0 in Table 4.4) for both SumMe and TVSum50 data sets. However, this does not apply to CLUSA. As Table 4.6 indicates that it is only possible to generate video summaries by discarding at least 20% of the videos in the TVSum50 data set, the random classification method's performance using the PR curve is supposed to be 0.30 accordingly to the reference values in Fig. 3.6. Instead, the performance score achieved for the random classification method was 0.28 in the experimental study. Similarly, the performance score using AUC-ROC is supposed to be 0.48 in the TVSum50, while the performance score reached was 0.42 in the experimental study.

To clarify why there was a substantial difference between the expected and obtained performance scores for the random classification method, we must turn to the compression rate distribution in Table 4.6. Only 0.1 % of all video summaries generated from TVSum50 are at the compression range $[0.2, 0.3[$. Consequently, $p_i = 0.2$ is not the minimum p_i value for all videos, only for a few. Since defining a minimum compression range for all videos in a data set is not trivial, the performance score achieved experimentally is more accurate.

4.3 CLOSURE

In the next chapter, we discuss the consequences of the results presented here, the limitations of F_β, RCC and CLUSA in the evaluation of video summarizers, and how CLUSA advances the field of video summarization.

Chapter

5

DISCUSSION AND CONCLUSIONS

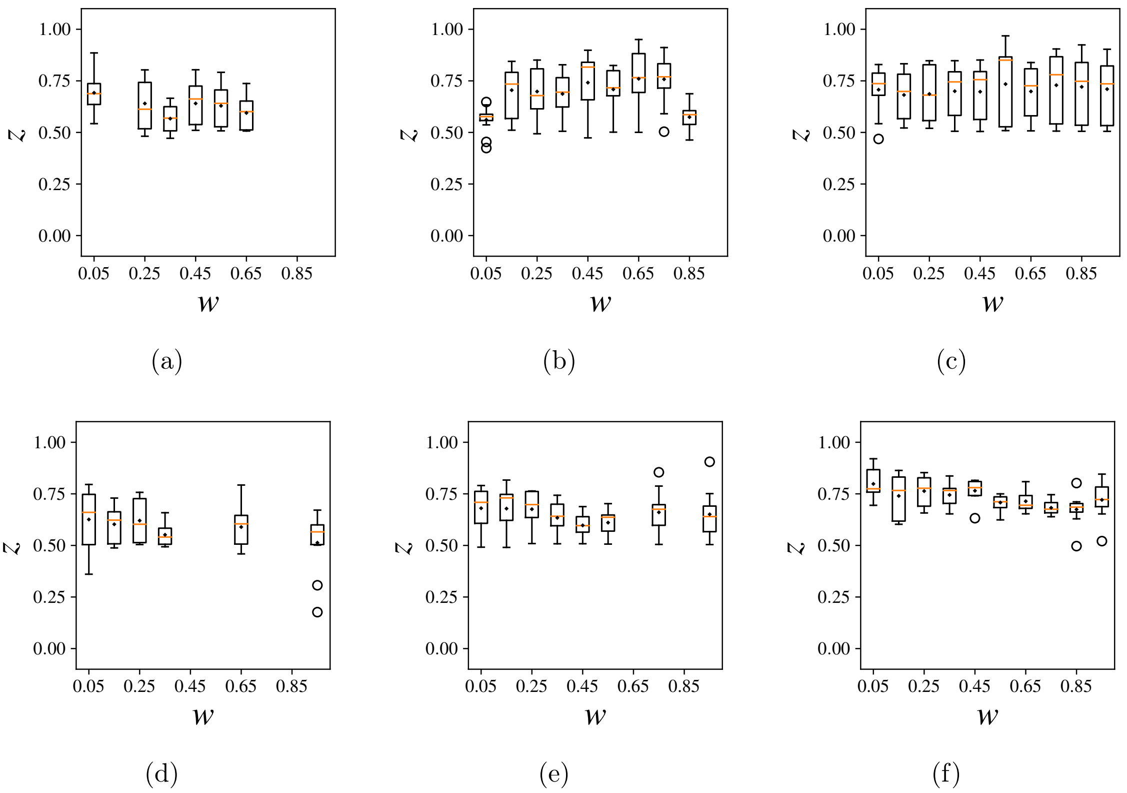
Contents

Here we discuss the advances in video summarization achieved from our study and the limitations faced: The consequences of users' disagreement on the relevance of video segments, the appropriate assessment scale for collecting annotations, and the evaluation of video summarizers using multi-valued scale annotations. Finally, we point out what can be improved in future work.

5.1 WHEN USERS DISAGREE ABOUT THE RELEVANCE OF INFORMATION

The studies conducted by Gygli et al. (2014) and Song et al. (2015) use the Cronbach's alpha coefficient to ensure the annotations' quality in the SumMe and TVSum50 data sets, respectively. Both data sets reached a Cronbach's alpha higher than 0.7 on average (which is the minimum acceptable value in psychometric studies); however, 9 of the 20 videos on SumMe showed unacceptable quality. Since Cronbach's alpha coefficient is a direct measure of user disagreement, values lower than 0.7 reveal that users consider different video segments relevant in SumMe. This issue illustrates how challenging the evaluation of video summarizers is in certain situations. For example, users might judge the video segments' relevance in a way unrelated to the video information if it goes against their interests. Due to the size of the sample of users, we cannot say whether there are distinct patterns of behavior or whether the notes are inconsistent.

When F_β, RCC, and CLUSA match annotations from divergent users, video summarizers' performance decreases. Notably, we can identify this issue in the RCC and CLUSA. In RCC, the video segments that users diverged are those whose average correlation is



(a) (b) (c)

(d) (e) (f)

Figure 5.1 Relating z_i CLUSA's scores for each w_i compression range on collected annotations: (a) and (d) Binary assessment scale, (b) and (e) Three-point Likert scale, and (c) and (f) Five-point Likert scale, with leave-one-out approach.

close to or below zero. In CLUSA, the matching score's variance on each compression range indicates how much the users diverged. Figure 5.1 shows box plots for each compression range (the x-axis) in CLUSA. The higher the interquartile range of each box (the box's height), the higher the divergence between users. Comparing each plot from top to bottom, respectively, users diverge differently in each video. This suggests that there is no single compression rate to summarize all videos.

5.2 A SUITABLE ASSESSMENT SCALE FOR COLLECTING USER ANNOTATIONS

The box plots in Fig. 5.1(c) and 5.1(f) were generated from annotations collected using five-point Likert scale. We found different variances for both videos. In the top video, the variance in each compression range is higher compared to the bottom video. Contrarily, the variance is similar when the same video is annotated on different scales, as illustrated in Figs. 5.1(e) and 5.1(f). The assessment scale does not seem to affect the users' divergence, only user annotations' internal consistency.

On average, Cronbach's alphas on the binary scale are lower than on the three-point Likert scale, which is lower than on the five-point Likert scale. Despite the collecting guidelines of SumMe and TVSum50 can not be directly compared, SumMe's binary scale annotations also had a lower Cronbach's alpha compared to TVSum50's multi-valued annotations. Thus, collecting annotations using multi-valued scales is recommended, the five-point Likert scale being the most convenient scale for video summarization. In this case, metrics should definitely be able to evaluate video summarizers using multi-valued scale annotations.

5.3 ASSESSING PERFORMANCE BY MATCHING RELEVANCE SCORES DIRECTLY

The F_β metric is limited for video summarization as it only evaluates binary scales. This is restrictive mainly for data sets like TVSum50, which tackles this issue by converting a multi-valued scale into a binary scale with a preset compression rate. On the other hand, two other metrics - RCC and CLUSA - were devised to deal with any types of annotated data set, be the scales multi-valued or binary. Accordingly, both of them meet the video summarization assumption of preserving relevant video information when evaluating video summarizers applied to data sets annotated with binary scales. In this case, CLUSA sets a different weight for each compression range predefined to evaluate a video summarizer. Hence different data sets could present different baselines according to the way the set of compression ranges spans the annotations. Contrarily, RCC have an advantage over CLUSA in data sets annotated with binary scales because the baseline is constant (zero) for any set of user annotations. This means that summarization studies using RCC would not need to calculate a baseline for each data set, though it is noteworthy that both metrics play the same role for data sets annotated in a binary scale.

In the case of multi-valued annotations, the RCC do not distinguish the importance of video segments marked by the users when assessing the relevance scores predicted by video summarizers. As a consequence, RCC do not meet the criterion of conciseness of user annotation. In contrast, CLUSA weighs video summaries according to their compression rates. In other words, the CLUSA's adaptive nature allows to evaluate the conciseness criterion, unlike any other metric.

On TVSum50 data set, some video summarizers do not generate summaries with compression rates other than those used to train the method, as happens to dppLSTM. In this case, both metrics have opposite behaviors. While RCC overestimate dppLSTM for approaching annotations without distinguishing the importance of each video segment, CLUSA penalizes dppLSTM for having generated video summaries only on compression rates which span user annotations. In summary, although CLUSA is suitable for video summarization, its weighting approach leads to issues when evaluating some video summarizers. For future work, an alternative solution would be to set weights only for compression ranges available in the video summarization data sets.

5.4 FUTURE WORK

Our study and the previous studies introduced herein assumed that all video segments are annotated from a single relevance perception, measuring all objects' relevance 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 the environment (i), whereas others would consider relationships (ii) as the most important characteristic of the video. Video summarization studies might incorporate video captioning techniques, which already approach this on several collections of videos. In that case, metrics for video summarization must be custom also to perform text matching, similar to matching metrics in the field of natural language processing.

While previous studies summarize all videos in a single compression rate, our study showed that this does not satisfy all users. Considering this, an alternative might be to devise video summarizers aiming at the compression ranges whose variances are low. In this way, the video summary generated by summarizers would satisfy most users.

BIBLIOGRAPHY

ABDALLA, K.; MENEZES, I.; OLIVEIRA, L. Modelling perceptions on the evaluation of video summarization. *Expert Systems with Applications*, v. 131, p. 254–265, 10 2019. ISSN 09574174. Available from Internet: <https://linkinghub.elsevier.com/retrieve/pii/S095741741930301X>.

AGNIHOTRI, L.; DIMITROVA, N.; KENDER, J. R. Design and evaluation of a music video summarization system. In: *Proceedings of International Conference Multimedia and Expo*. [S.l.: s.n.], 2004. v. 3, p. 1943–1946.

ANASTASI, A. *Testagem psicológica*. ArtMed, 2000. ISBN 9788573076158. Available from Internet: <https://books.google.com.br/books?id=p5oWtwAACAAJ>.

ARMAN, F. et al. Content-based browsing of video sequences. In: *Proceedings of the second ACM international conference on Multimedia*. New York, New York, USA: ACM Press, 1994. p. 97–103. ISBN 0897916867. Available from Internet: <http://portal.acm.org/citation.cfm?doid=192593.192630>.

AWAD, G. et al. Instance search retrospective with focus on TRECVID. *International Journal of Multimedia Information Retrieval*, Springer, v. 6, n. 1, p. 1–29, 3 2017. ISSN 2192-6611. Available from Internet: <http://link.springer.com/10.1007/s13735-017-0121-3>.

BORSBOOM, D. *Measuring the Mind: Conceptual Issues in Contemporary Psychometrics*. Cambridge University Press, 2005. ISBN 9781139444637. Available from Internet: <https://books.google.com.br/books?id=Fv3bAWvIDo4C>.

CHANG, P.; HAN, M.; GONG, Y. Extract highlights from baseball game video with hidden Markov models. *IEEE International Conference on Image Processing*, v. 1, p. I–609–I–612, 2002. ISSN 1522-4880. Available from Internet: <http://ieeexplore.ieee.org/document/1038097/>.

Chong-Wah Ngo; Yu-Fei Ma; Hong-Jiang Zhang. Automatic video summarization by graph modeling. In: *Proceedings Ninth IEEE International Conference on Computer Vision*. IEEE, 2003. p. 104–109. ISBN 0-7695-1950-4. Available from Internet: <http://ieeexplore.ieee.org/document/1238320/>.

CHU, W.-S.; Yale Song; JAIMES, A. Video co-summarization: Video summarization by visual co-occurrence. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2015. p. 3584–3592. ISBN 978-1-4673-6964-0. ISSN 10636919. Available from Internet: <http://ieeexplore.ieee.org/document/7298981/>.

CO-INVESTIGATOR, N. *Automatic Performance Evaluation for Video Summarization.* [S.l.], 2013. v. 53, 1689–1699 p.

CRONBACH, L. J. Coefficient alpha and the internal structure of tests. *Psychometrika*, Wiley, v. 16, n. 3, p. 297–334, 9 1951. ISSN 0033-3123. Available from Internet: <https://www.jstor.org/stable/1419921?origin=crossrefhttp://content.apa.org/journals/ccp/14/1/73dhttp://link.springer.com/10.1007/BF02310555>.

DAVIS, J.; GOADRICH, M. The relationship between Precision-Recall and ROC curves. In: *Proceedings of the 23rd international conference on Machine learning - ICML '06.* New York, New York, USA: ACM Press, 2006. p. 233–240. ISBN 1595933832. Available from Internet: <http://arxiv.org/abs/1609.07195http://portal.acm.org/citation.cfm?doid=1143844.1143874>.

FAJTL, J. et al. Summarizing Videos with Attention. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics).* [s.n.], 2019. v. 11367 LNCS, p. 39–54. ISBN 9783030210731. Available from Internet: <http://link.springer.com/10.1007/978-3-030-21074-8_4>.

FAWCETT, T. An introduction to ROC analysis. *Pattern Recognition Letters*, v. 27, n. 8, p. 861–874, 2006. ISSN 01678655.

GEORGE, D.; MALLERY, P. *SPSS for Windows Step by Step: A Simple Guide and Reference 18.0 Update.* 11th. ed. Upper Saddle River, NJ, USA: Prentice Hall Press, 2010. ISBN 0205011241, 9780205011247.

GYGLI, M. et al. Creating Summaries from User Videos. In: *Proceedings of the European Conference on Computer Vision.* [s.n.], 2014. p. 505–520. ISBN 978-3-319-10584-0. Available from Internet: <http://link.springer.com/10.1007/978-3-319-10584-0_33>.

HANJALIC, A. Shot-boundary detection: unraveled and resolved? *IEEE Transactions on Circuits and Systems for Video Technology*, v. 12, n. 2, p. 90–105, 2002. ISSN 10518215. Available from Internet: <http://ieeexplore.ieee.org/document/988656/>.

Haojin Yang; MEINEL, C. Content Based Lecture Video Retrieval Using Speech and Video Text Information. *IEEE Transactions on Learning Technologies*, v. 7, n. 2, p. 142–154, 4 2014. ISSN 1939-1382. Available from Internet: <doi.ieeecomputersociety.org/10.1109/TLT.2014.2307305http://ieeexplore.ieee.org/document/6750040/>.

HE, L. et al. Auto-summarization of audio-video presentations. In: *Proceedings of the ACM international conference on Multimedia.* [s.n.], 1999. p. 489–498. ISBN 1581131518. Available from Internet: <http://portal.acm.org/citation.cfm?doid=319463.319691>.

HUTZ, C. S.; BANDEIRA, D. R.; TRENTINI, C. M. *Psicometria.* Artmed Editora, 2015. (Avaliação Psicológica). ISBN 9788582712368. Available from Internet: <https://books.google.com.br/books?id=cVlICgAAQBAJ>.

JABRAYILOV, R.; EMONS, W. H. M.; SIJTSMA, K. Comparison of Classical Test Theory and Item Response Theory in Individual Change Assessment. *Applied Psychological Measurement*, v. 40, n. 8, p. 559–572, 11 2016. ISSN 0146-6216. Available from Internet: <http://journals.sagepub.com/doi/10.1177/0146621616664046>.

KELLEHER, J. D.; NAMEE, B. M.; D'ARCY, A. *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies*. MIT Press, 2015. (The MIT Press). ISBN 9780262331746. Available from Internet: <https://books.google.com.br/books?id=3EtQCgAAQBAJ>.

KENDALL, M. G. A New Measure of Rank Correlation. *Biometrika*, v. 30, n. 1/2, p. 81, 6 1938. ISSN 00063444. Available from Internet: <https://www.jstor.org/stable/2332226?origin=crossref>.

KENDALL, M. G. The treatment of ties in ranking problems. *Biometrika*, v. 33, n. 3, p. 239–251, 1945. ISSN 0006-3444. Available from Internet: <https://academic.oup.com/biomet/article-lookup/doi/10.1093/biomet/33.3.239>.

KIM, G.; SIGAL, L.; XING, E. P. Joint Summarization of Large-Scale Collections of Web Images and Videos for Storyline Reconstruction. In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2014. p. 4225–4232. ISBN 978-1-4799-5118-5. ISSN 10636919. Available from Internet: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6909934>.

KLINE, P. *Handbook of psychological testing, second edition*. 2. ed. London: Routledge, 2013. 1–744 p. ISBN 9781317798057. Available from Internet: <https://www.taylorfrancis.com/books/9781315812274>.

LIU, T.; ZHANG, H. J.; QI, F. A Novel Video Key-Frame-Extraction Algorithm Based on Perceived Motion Energy Model. *IEEE Transactions on Circuits and Systems for Video Technology*, v. 13, n. 10, p. 1006–1013, 2003. ISSN 10518215.

LIU, W. et al. Multi-task deep visual-semantic embedding for video thumbnail selection. In: *Proceedings of the Conference on Computer Vision and Pattern Recognition*. [s.n.], 2015. p. 3707–3715. ISBN 978-1-4673-6964-0. Available from Internet: <http://ieeexplore.ieee.org/document/7298994/>.

MAHASSENI, B.; LAM, M.; TODOROVIC, S. Unsupervised video summarization with adversarial LSTM networks. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, v. 2017-Janua, p. 2982–2991, 2017.

MOSLEY, L. *A balanced approach to the multi-class imbalance problem*. Tese (Doutorado) — Iowa State University, Digital Repository, Ames, 2013. Available from Internet: <https://lib.dr.iastate.edu/etd/13537/>.

NGUYEN, C.; NIU, Y.; LIU, F. 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. [S.l.: s.n.], 2012. p. 647–650.

OTANI, M. et al. Video Summarization Using Deep Semantic Features. In: LAI, S.-H. et al. (Ed.). *Computer Vision – ACCV 2016*. Cham: Springer International Publishing, 2017. p. 361–377. ISBN 978-3-319-54193-8. Available from Internet: <http://link.springer.com/10.1007/978-3-319-54193-8_23>.

OTANI, M. et al. Rethinking the Evaluation of Video Summaries. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2019. p. 7588–7596. ISBN 978-1-7281-3293-8. Available from Internet: <http://arxiv.org/abs/1903.11328https://ieeexplore.ieee.org/document/8954229/>.

PAL, G. et al. Video Shot Boundary Detection: A Review. In: SATAPATHY, S. C. et al. (Ed.). *Emerging ICT for Bridging the Future - Proceedings of the Annual Convention of the Computer Society of India*. Cham: Springer International Publishing, 2015. p. 119–127. ISBN 978-3-319-13731-5. Available from Internet: <http://link.springer.com/10.1007/978-3-319-13731-5_14>.

PASQUALI, L. *Psicometria: Teoria dos testes na psicologia e na educação*. Editora Vozes, 2017. ISBN 9788532656124. Available from Internet: <https://books.google.com.br/books?id=D_Y4DwAAQBAJ>.

PEARSON, K. *On the theory of contingency and its relation to association and normal correlation; On the general theory of skew correlation and non-linear regression*. [S.l.]: Cambridge University Press, 1904.

PLUMMER, B. A.; BROWN, M.; LAZEBNIK, S. Enhancing video summarization via vision-language embedding. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, v. 2017-Janua, p. 1052–1060, 2017.

ROCHAN, M.; YE, L.; WANG, Y. Video Summarization Using Fully Convolutional Sequence Networks. In: FERRARI, V. et al. (Ed.). *Computer Vision – ECCV 2018*. Cham: Springer International Publishing, 2018. p. 358–374. ISBN 978-3-030-01258-8. Available from Internet: <http://arxiv.org/abs/1805.10538http://link.springer.com/10.1007/978-3-030-01258-8_22>.

SHARGHI, A.; LAUREL, J. S.; GONG, B. Query-Focused Video Summarization: Dataset, Evaluation, and a Memory Network Based Approach. In: *Proceedings of the Conference on Computer Vision and Pattern Recognition*. IEEE, 2017. p. 2127–2136. ISBN 978-1-5386-0457-1. Available from Internet: <http://arxiv.org/abs/1707.04960http://ieeexplore.ieee.org/document/8099712/>.

SIMMS, L. J. et al. Does the number of response options matter? Psychometric perspectives using personality questionnaire data. *Psychological Assessment*, v. 31, n. 4, p. 557–566, 4 2019. ISSN 1939-134X. Available from Internet: <http://doi.apa.org/getdoi.cfm?doi=10.1037/pas0000648>.

SONG, Y. et al. TVSum: Summarizing web videos using titles. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, v. 07-12-June, p. 5179–5187, 2015. ISSN 10636919.

SOOMRO, K.; ZAMIR, A. R.; SHAH, M. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. *CoRR*, abs/1212.0, 2012. Available from Internet: <http://dblp.uni-trier.de/db/journals/corr/corr1212.html#abs-1212-0402>.

SPEARMAN, C. The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*, v. 15, n. 1, p. 72, 1 1904. ISSN 00029556. Available from Internet: <https://www.jstor.org/stable/1412159?origin=crossref>.

STUART, A. The Estimation and Comparison of Strengths of Association in Contingency Tables. *Biometrika*, v. 40, n. 1/2, p. 105, 6 1953. ISSN 00063444. Available from Internet: <https://www.jstor.org/stable/2333101?origin=crossref>.

SUN, M.; FARHADI, A.; SEITZ, S. 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)*, v. 8689 LNCS, n. PART 1, p. 787–802, 2014. ISSN 16113349.

SUNDARAM, H.; CHANG, S. F. Condensing computable scenes using visual complexity and film syntax analysis. In: *Proceedings - IEEE International Conference on Multimedia and Expo*. [S.l.: s.n.], 2001. p. 273–276. ISBN 0769511988. ISSN 1945788X.

TASKIRAN, C. M. Evaluation of automatic video summarization systems. In: INTERNATIONAL SOCIETY FOR OPTICS AND PHOTONICS. *Multimedia Content Analysis, Management, and Retrieval 2006*. [S.l.], 2006. v. 6073, p. 60730K.

TRUONG, B. T.; VENKATESH, S. Video abstraction. *ACM Transactions on Multimedia Computing, Communications, and Applications*, v. 3, n. 1, p. 3–es, 2 2007. ISSN 15516857. Available from Internet: <http://portal.acm.org/citation.cfm?doid=1198302.1198305%5Cnhttp://dl.acm.org/citation.cfm?id=1198302.1198305http://portal.acm.org/citation.cfm?doid=1198302.1198305>.

URBINA, S. *Essentials of Psychological Testing*. Wiley, 2014. (Essentials of Behavioral Science). ISBN 9781118873090. Available from Internet: <https://books.google.com.br/books?id=c3L0AwAAQBAJ>.

WANG, X.; CHEN, J.; ZHU, C. User-Specific Video Summarization. In: *Proceedings of the International Conference on Multimedia and Signal Processing*. [s.n.], 2011. p. 213–219. ISBN 978-1-61284-314-8. Available from Internet: <http://ieeexplore.ieee.org/document/5957411/>.

WU, T. et al. Hierarchical Union-of-Subspaces Model for Human Activity Summarization. In: *Proceedings of the IEEE International Conference on Computer Vision*. [s.n.], 2016. v. 2016-Febru, p. 1053–1061. ISBN 9781467383905. ISSN 15505499. Available from Internet: <http://ieeexplore.ieee.org/document/7406487/>.

Xiao-Dong Yu et al. Multilevel video representation with application to keyframe extraction. In: *Proceedings of the IEEE International Multimedia Modelling Conference*. [s.n.], 2004. p. 117–123. ISBN 0-7695-2084-7. Available from Internet: <http://ieeexplore.ieee.org/document/1264975/>.

XIONG, Z.; RADHAKRISHNAN, R.; DIVAKARAN, A. Generation of sports highlights using motion activity in combination with a common audio feature extraction framework. In: *Proceedings 2003 International Conference on Image Processing (Cat. No.03CH37429)*. IEEE, 2003. v. 1, p. 5–8. ISBN 0-7803-7750-8. Available from Internet: <http://ieeexplore.ieee.org/document/1246884/>.

YAO, T.; MEI, T.; RUI, Y. Highlight Detection with Pairwise Deep Ranking for First-Person Video Summarization. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. [s.n.], 2016. p. 982–990. ISBN 978-1-4673-8851-1. ISSN 10636919. Available from Internet: <http://ieeexplore.ieee.org/document/7780481/>.

Yong Jae Lee; GHOSH, J.; GRAUMAN, K. Discovering important people and objects for egocentric video summarization. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. [s.n.], 2012. p. 1346–1353. ISBN 978-1-4673-1228-8. Available from Internet: <http://ieeexplore.ieee.org/document/6247820/>.

Youtube. *Youtube for Press*. 2018. Available from Internet: <https://www.youtube.com/intl/en-US/yt/about/press/>.

YUAN, J. et al. A Formal Study of Shot Boundary Detection. *IEEE Transactions on Circuits and Systems for Video Technology*, v. 17, n. 2, p. 168–186, 2 2007. ISSN 1051-8215. Available from Internet: <http://ieeexplore.ieee.org/document/4079667/>.

ZHANG, H. J. et al. An integrated system for content-based video retrieval and browsing. *Pattern Recognition*, v. 30, n. 4, p. 643–658, 4 1997. ISSN 00313203. Available from Internet: <http://linkinghub.elsevier.com/retrieve/pii/S0031320396001094>.

ZHANG, K. et al. Video Summarization with Long Short-Term Memory. In: LEIBE, B. et al. (Ed.). *Computer Vision – ECCV 2016*. Cham: Springer International Publishing, 2016. p. 766–782. ISBN 978-3-319-46478-7. Available from Internet: <http://link.springer.com/10.1007/978-3-319-46478-7_47>.

ZHAO, B.; XING, E. P. Quasi Real-Time Summarization for Consumer Videos. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. [s.n.], 2014. p. 2513–2520. ISBN 978-1-4799-5118-5. ISSN 10636919. Available from Internet: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6909718>.

ZHOU, K.; QIAO, Y.; XIANG, T. Deep Reinforcement Learning for Unsupervised Video Summarization with Diversity-Representativeness Reward. *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, p. 39–54, 12 2017. Available from Internet: <http://link.springer.com/10.1007/978-3-030-21074-8_4http://arxiv.org/abs/1801.00054>.