



HOKKAIDO UNIVERSITY

Title	Estimating Video Authenticity via the Analysis of Visual Quality and Video Structure
Author(s)	PENKOV, MICHAEL
Degree Grantor	北海道大学
Degree Name	博士(情報科学)
Dissertation Number	甲第11975号
Issue Date	2015-09-25
DOI	https://doi.org/10.14943/doctoral.k11975
Doc URL	https://hdl.handle.net/2115/59968
Type	doctoral thesis
File Information	Michael_Penkov.pdf



Doctoral Thesis

**Estimating Video Authenticity via the
Analysis of Visual Quality and Video Structure**

**画質と映像構造の解析に基づく
映像信頼性の推定に関する研究**

Michael Penkov

**Laboratory of Media Dynamics,
Graduate School of Information Science and Technology,
Hokkaido University**

July 24, 2015

Contents

1	Introduction	1
1.1	Research Background	1
1.2	Our Contribution	3
1.3	Thesis Structure	5
2	Visual Quality Assessment	6
2.1	Introduction	6
2.2	What is Visual Quality?	6
2.3	Conventional Algorithms	8
2.3.1	Edge Width Algorithm	9
2.3.2	No-reference Block-based Blur Detection	9
2.3.3	Generalized Block Impairment Metric	11
2.3.4	Blocking Estimation in the FFT Domain	12
2.3.5	Blocking Estimation in the DCT Domain	12
2.4	Datasets	14
2.4.1	Artificial Dataset	14
2.4.2	Real-world Dataset (<i>Klaus</i>)	14
2.5	Experiment: Comparison of No-reference Methods	16
2.5.1	Aim	16
2.5.2	Method	18
2.5.3	Results	18
2.6	Experiment: Robustness to Logos and Captions	19
2.6.1	Aim	19
2.6.2	Method	19
2.6.3	Results	20
2.7	Investigation: Comparison of the Causes of Visual Quality Loss	20
2.8	Investigation: Sensitivity to Visual Content	21
2.9	Conclusion	22
3	Shot Identification	23
3.1	Introduction	23
3.2	Preliminaries	23
3.3	Shot Segmentation	25
3.4	Shot Comparison	26
3.4.1	A Spatial Algorithm	27
3.4.2	A Spatiotemporal Algorithm	27

3.4.3	Robust Hash Algorithm	29
3.5	Experiments	31
3.5.1	Datasets	31
3.5.2	Shot Segmentation	33
3.5.3	Shot Comparison	33
3.6	Utilizing Semantic Information for Shot Comparison	36
3.6.1	Investigation: the Effectiveness of Subtitles	37
3.6.2	Proposed Method	39
3.6.3	Preliminary Experiment	39
3.6.4	Experiment	41
3.7	Conclusion	44
4	The Video Authenticity Degree	45
4.1	Introduction	45
4.1.1	Definition	45
4.2	Developing the Full-reference Model	46
4.2.1	Initial Model	46
4.2.2	Detecting Scaling and Cropping	47
4.2.3	Experiments	50
4.2.4	Conclusion	56
4.3	Developing the No-reference Model	56
4.3.1	Derivation from the Full-reference Model	56
4.3.2	Considering Relative Shot Importance	57
4.3.3	Experiments	58
4.3.4	Conclusion	62
4.4	The Proposed No-reference Method	62
4.4.1	Calculation	62
4.4.2	Parent Video Estimation	64
4.4.3	Global Frame Modification Detection	65
4.4.4	Comparison with State of the Art Methods	67
4.5	Experiments	68
4.5.1	Small-scale Experiment on Artificial Data	69
4.5.2	Large-scale Experiment on Artificial Data	72
4.5.3	Experiment on Real Data	76
4.6	Conclusion	80
5	Conclusion	81
	Bibliography	82
	Acknowledgments	89
	Publications by the Author	90

Chapter 1

Introduction

This thesis summarizes the results of our research into video authenticity [1–7]. This chapter serves to introduce the thesis. More specifically, Sec. 1.1 introduces the background of our research; Sec. 1.2 describes the contribution made by our research. Finally, Sec. 1.3 describes the structure of the remainder of this thesis.

1.1 Research Background

With the increasing popularity of smartphones and high-speed networks, videos uploaded have become an important medium for sharing information, with hundreds of millions of videos viewed daily through sharing sites such as YouTube¹. People increasingly view videos to satisfy their information need for news and current affairs. In general, there are two kinds of videos: (a) parent videos, which are created when a camera images a real-world event; and (b) edited copies of a parent video that is already available somewhere. Creating such copies may involve video editing, for example, by shot removal, resampling and recompression, to the extent that they misrepresent the event they are portraying. Such edited copies can be undesirable to people searching for videos to watch [8], since the people are interested in viewing an accurate depiction of the event in order to satisfy their information need. In other words, they are interested in the video that retains the most information from the parent video – the most *authentic* video.

For the purposes of this thesis, we define *authenticity* of an edited video as the proportion of information the edited video contains from its parent video. To the best of our knowledge, we are the

¹www.youtube.com

first to formulate the research problem in this way. However, there are many conventional methods are highly relevant, and they are discussed in detail below.

Since videos uploaded to sharing sites on the Web contain metadata that includes the upload timestamp and view count, a simple method for determining authenticity is to focus on the metadata. More specifically, the focus is on originality [9] or popularity [10]: in general, videos that were uploaded earlier or receive much attention from other users are less likely to be copies of existing videos. Since Web videos contain context information that includes the upload timestamp [11], determining the video that was uploaded first is trivial. However, since the metadata is user-contributed and not directly related to the actual video signal, it is often incomplete or inaccurate. Therefore, using metadata by itself is not sufficient, and it is necessary to examine the actual video signal of the video.

Conventional methods for estimating the authenticity of a video can be divided into two categories: active and passive. Active approaches, also known as *digital watermarking*, embed additional information into the parent video, enabling its editing history to be tracked [12, 13]. However, their use is limited, as not all videos contain embedded information. Passive approaches, also known as *digital forensics*, make an assessment using only the video in question without assuming that it contains explicit information about its editing history [14, 15]. For example, forensic algorithms can detect recompression [16], median filtering [17], and resampling [18]. However, forensic algorithms are typically limited to examining videos individually, and cannot directly and objectively estimate the authenticity of videos. Recently, forensic techniques have been extended to studying not only individual videos, but also the *relationships* between videos that are near-duplicates of each other. For example, *video phylogeny* examines causal relationships within a group of videos, and creates a tree to express such relationships [19]. Other researchers have focused on reconstructing the parent sequence from a group of near-duplicate videos [20]. The reconstructed sequence is useful for understanding the motivation behind creating the near-duplicates: what was edited and, potentially, why.

Another related research area is “video credibility” [21], which focuses on verifying the factual information contained in the video. Since it is entirely manual, the main focus of the research is interfaces that enable efficient collaboration between reviewers.

We approach the problem of authenticity estimation from a radically different direction: *no reference visual quality assessment* [22] (hereafter, “NRQA”). NRQA algorithms evaluate the strength of

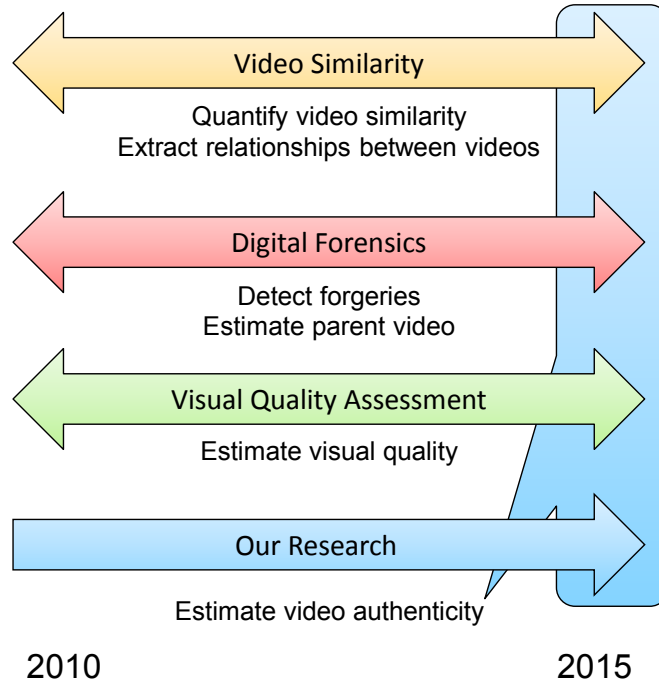


Figure 1.1: A timeline of our research with respect to related research fields.

common quality degradations such as blurring [23]. Our motivation for introducing these algorithms is as follows: editing a video requires recompressing the video signal, which is usually a lossy operation that causes a decrease in visual quality [24]. Since the edited videos were created by editing the parent video, the parent will have a higher visual quality than any of the edited videos created from it. Therefore, visual quality assessment is relevant to authenticity estimation. However, since the video signals of the edited videos can differ significantly due to the editing operations, directly applying the algorithms to compare the visual quality of the edited videos is difficult [25].

Figure 1.1 shows a timeline of our research with respect to the related research fields.

1.2 Our Contribution

This thesis proposes a method that identifies the most authentic video by estimating the proportion of information remaining from the parent video, even if the parent is not available. We refer to this measure as the “authenticity degree”. The novelty of this proposed method consists of four parts: (a) in the absence of the parent video, we estimate its content by comparing edited copies of it; (b) we reduce the difference between the video signals of edited videos before performing a

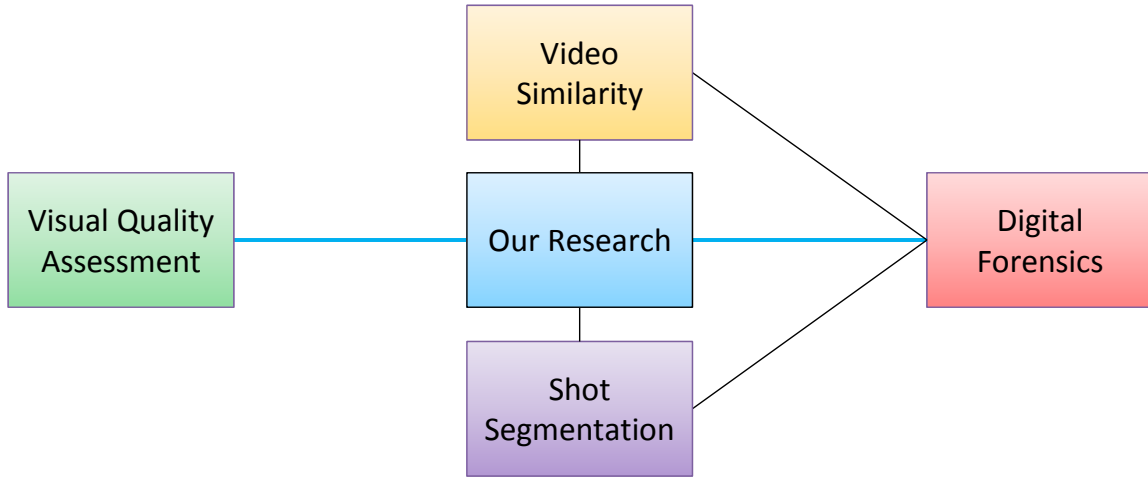


Figure 1.2: A map of research fields that are related to video authenticity.

visual quality comparison, thereby enabling the application of conventional NRQA algorithms; (c) we collaboratively utilize the outputs of the NRQA algorithms to determine the visual quality of the parent video; and (d) we enable the comparison of NRQA algorithm outputs for visually different shots. Finally, since the proposed method is capable of detecting shot removal, scaling, and recompression, it is effective for identifying the most authentic edited video. To the best of our knowledge, we are the first to apply conventional NRQA algorithms to videos that have significantly different signals, and the first to utilize video quality to determine the authenticity of a video. The proposed method has applications in video retrieval: for example, it can be used to sort search results by their authenticity. The effectiveness of our proposed method is demonstrated by experiments on real and artificial data.

In order to clarify our contribution, Figure 1.2 shows a map of research fields related to video authenticity and the relationships between them. Each rectangle corresponds to a research field. The closest related fields to the research proposed in this thesis are quality assessment and digital forensics. To the best of our knowledge, these areas were unrelated until fairly recently, when it was proposed that no-reference image quality assessment metrics be used to detect image forgery [26]. Furthermore, the direct connection between video similarity and digital forensics was also made recently, when it was proposed to use video similarity algorithms to recover the parent video from a set of edited videos [20]. Our contribution is thus further bridging the gap between the visual quality assessment and digital forensics by utilizing techniques from video similarity and shot segmentation. We are looking forward to seeing new and interesting contributions.

1.3 Thesis Structure

This thesis consists of 5 chapters.

Chapter 2 reviews the field of quality assessment in greater detail, with a focus on conventional methods for no-reference visual quality assessment. The chapter reviews several conventional NRQA algorithms and compares their performance empirically. These algorithms enable the method proposed in this thesis to quantify information loss. Finally, it demonstrates the limitations of existing algorithms when applied to videos of differing visual content.

Chapter 3 proposes shot identification, an important pre-processing step for the method proposed in this thesis. Shot identification enables the proposed method to:

1. Estimate the parent video from edited videos,
2. Detect removed shots, and
3. Apply the algorithms from Chapter 2.

The chapter first introduces algorithms for practical and automatic shot identification, and evaluates their effectiveness empirically. Finally, the chapter introduces the shot identification algorithm and evaluates its effectiveness as a whole.

Chapter 4 describes the proposed method for estimating the video authenticity. It offers a formal definition of “video authenticity” and the development of several models for determining the authenticity of edited videos in both full-reference and no-reference scenarios, culminating in a detailed description of the method we recently proposed [27]. The effectiveness of the proposed method is verified through a large volume of experiments on both artificial and real-world data.

Finally, Chapter 5 concludes the thesis and discusses potential future work.

Chapter 2

Visual Quality Assessment

2.1 Introduction

In this chapter, I introduce the field of visual quality assessment. This chapter is organized as follows. First, Sec. 2.2 describes what visual quality is, and what causes quality loss in Web videos. Next, Sec. 2.3 introduces conventional algorithms for automatically assessing the visual quality of videos. The remainder of the chapter describes experiments and investigations. More specifically, Sec. 2.4 introduces the datasets used. Section 2.5 compare the algorithms introduced in Sec. 2.3 to each other. Section 2.6 verifies the robustness of the algorithms to logos and captions. Section 2.7 investigates the causes of visual quality loss in Web video. Section 2.8 investigates the sensitivity of algorithms to differences in visual content. Finally, Sec. 2.9 concludes this chapter.

2.2 What is Visual Quality?

The use of images and video for conveying information has seen tremendous growth recently, owing to the spread of the Internet and high-speed mobile networks and devices. Maintaining the appearance of captured images and videos at a level that is satisfactory to the human viewer is therefore paramount. Unfortunately, the quality of digital image is rarely perfect due to distortions during acquisition, compression, transmission and processing. In general, human viewers can easily identify and quantify the quality of an image. However, the vast volume of images make the subjective assessment impractical in the general case. Therefore, automatic algorithms that predict¹ the visual quality of an image are necessary.

¹Since humans are the main consumers of visual information, the algorithm outputs must correlate well with subjective scores.

There are many applications for visual quality assessment algorithms. First, they can be used to monitor visual quality in an image acquisition system, maximizing the quality of the image that is acquired. Second, they can be used for benchmarking existing image processing algorithms, such as, for example, compression, restoration and denoising.

There are three main categories of objective visual quality assessment algorithms: full-reference, reduced-reference, and no-reference. They are discussed in greater detail below. Figure 2.1 shows a summary of the methods. While the figure describes images, the concepts apply equally well to videos.

The full-reference category assumes that an image of perfect quality (the *reference image*) is available, and assesses the quality of the target image by comparing it to the reference [25, 28–30]. While full-reference algorithms correlate well with subjective scores, there are many cases where the reference image is not available.

Reduced-reference algorithms do not require the reference image to be available [31]. Instead, certain features are extracted from the reference image, and used by the algorithm to assess the quality of the target image. While such algorithms are more practical than full-reference algorithms, they still require auxiliary information, which is not available in many cases.

No-reference algorithms² assess the visual quality of the target image without any additional information [23, 32–37]. This thesis focuses on no-reference algorithms, since they are the most practical. In the absence of the reference image, these algorithms attempt to identify and quantify the strength of common quality degradations³. The remainder of this chapter focuses specifically on no-reference algorithms, since they are most practical.

Finally, in the context of videos shared on the Web, there are several causes for decreases in visual quality:

1. Recompression caused by repetitively downloading a video and uploading the downloaded copy. This happens frequently and is known as *reposting*. We refer to this as the *recompression stage*.
2. Downloading the video at lower than maximum quality and then uploading the downloaded

²Also known as “blind” algorithms

³Also known as “artifacts”

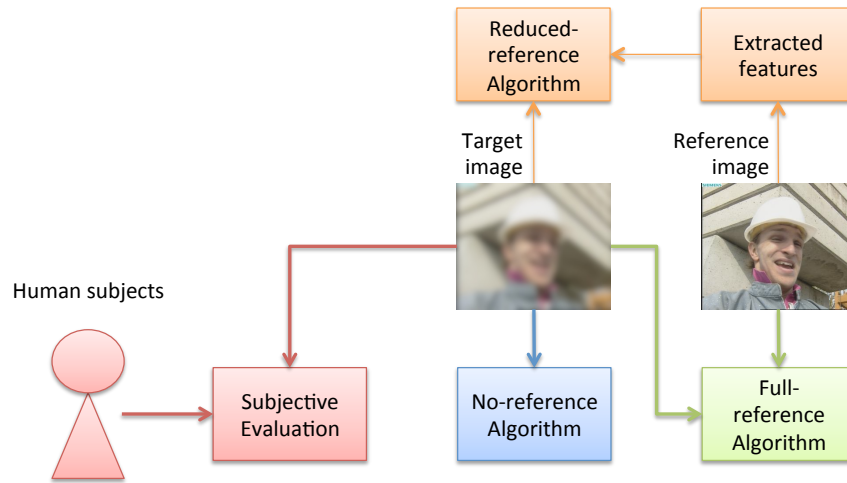


Figure 2.1: A summary of the methods for objective visual quality assessment.

copy. We refer to this as the *download stage*.

3. Significant changes to the video caused by video editing.

2.3 Conventional Algorithms

This section introduces several popular no-reference quality assessment (NRQA) algorithms. Since video quality is very closely related to image quality, many of the methods in this chapter apply to both images and video. In particular, image-based methods can be applied to the individual frames of a video, and their results averaged to yield the quality of the entire video.

Conventional algorithms can be classified by the artifact that they target and the place in the decoding pipeline at which they operate. Figure 2.3 shows the decoding pipeline for a generic image compression method. The input to the pipeline is a compressed image in the form of a bitstream, and the output is the decompressed image in raster format. NRQA algorithms that operate in the spatial domain require the full decoding pipeline to complete. In contrast, some transform-domain NRQA algorithms operate directly after entropy decoding and inverse quantization. Since they do not require an inverse transform, they are computationally less expensive than spatial-domain algorithms.

No-reference methods do not require the original to be available — they work on the degraded video only. These methods work by detecting the strength of artifacts such as blurring or blocking. Blurring and blocking are two examples of *artifacts*, the presence of which indicates reduced visual

quality. Blurring occurs when high-frequency spatial information is lost. This can be caused by strong compression – since high-frequency transform coefficients are already weak in natural images, strong compression can quantize them to zero. The loss of high-frequency spatial information can also be caused by downsampling, as a consequence of the sampling theorem [38].

Blocking is another common artifact in images and video. It occurs as a result of too coarse a quantization during image compression. It is often found in smooth areas, as those areas have weak high frequency components that are likely to be set to zero as a result of the quantization process.

There are other known artifacts, such as ringing and mosquito noise, but they are not as common in Web video. Therefore, the remainder of this chapter will focus on the blurring and blocking artifacts only.

For example, Fig. 2.2(a) shows an original image of relatively high visual quality. Figures 2.2 (b) to (f) show images of degraded visual quality due to compression and other operations.

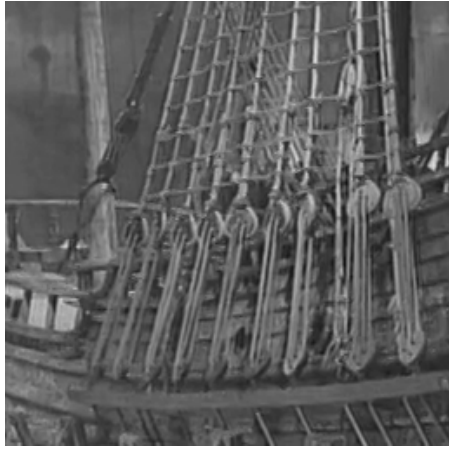
2.3.1 Edge Width Algorithm

The conventional method [34] takes advantage of the fact that blurring causes edges to appear wider. Sharp images thus have a narrower edge width than blurred images. The edge width in an image is simple to measure: first, vertical edges are detected using the Sobel filter. Then, at each edge position, the location of the local row minima and extrema are calculated. The distance between the extrema is the edge width at that location.

One limitation of this algorithm is that it cannot be used to directly compare the visual quality of two images of different resolutions, since the edge width depends on resolution. A potential work-around for this limitation is to scale the images to a common resolution prior to comparison.

2.3.2 No-reference Block-based Blur Detection

While the edge width algorithm described in Section 2.3.1 is effective, it is computationally expensive due to the edge detection step. Avoiding edge detection would allow the algorithm to be applied to video more efficiently. The no-reference block-based blur detection (NR-BBD) algorithm [36] targets video compressed in the H.264/AVC format [39], which includes an in-loop deblocking filter that blurs macroblock boundaries to reduce the blocking artifact. Since the effect of the deblocking filter is proportional to the strength of the compression, it is possible to estimate the amount of blurring



(a) Original



(b) Light JPEG compression



(c) Moderate JPEG compression



(d) Moderate H.264 compression



(e) Strong H.264 compression



(f) Blur

Figure 2.2: An example of images of different visual quality.

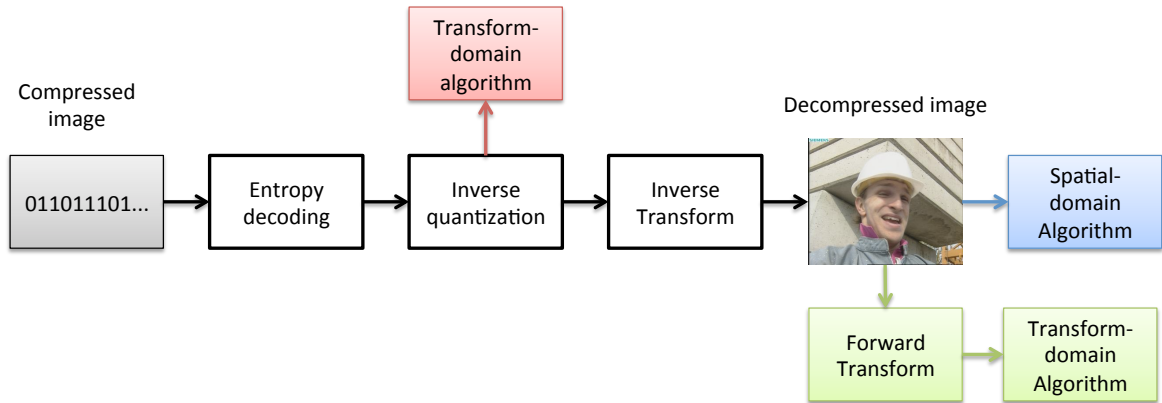


Figure 2.3: Quality assessment algorithms at different stages of the decoding pipeline.

caused by the H.264/AVC compression by measuring the edge width at these macroblock boundaries. Since this algorithm requires macroblock boundaries to be at fixed locations, it is only applicable to video keyframes (also known as I-frames), as for all other types of frames the macroblock boundaries are not fixed due to motion compensation.

2.3.3 Generalized Block Impairment Metric

The Generalized Block Impairment Metric (GBIM) [32] is a simple and popular spatial-domain method for detecting the strength of the blocking artifact. It processes each scanline individually and measures the mean squared difference across block boundaries. GBIM assumes that block boundaries are at fixed locations. The weight at each location is determined by local luminance and standard deviation, in an attempt to include the spatial and contrast masking phenomena into the model. While this method produces satisfactory results, it has a tendency to misinterpret real edges that lie near block boundaries.

As this is a spatial domain method, it is trivial to apply it to any image sequence. In the case of video, full decoding of each frame will need to complete prior to applying the method. This method corresponds to the blue block in Figure 2.3. Since this algorithm requires macroblock boundaries to be at fixed locations, it is only applicable to video keyframes (also known as I-frames), as for all other types of frames the macroblock boundaries are not fixed due to motion compensation.

2.3.4 Blocking Estimation in the FFT Domain

In [33], an FFT frequency domain approach is proposed to solve the false-positive problem of the method introduced in Sec. 2.3.3. Because the blocking signal is periodic, it is more easily distinguished in the frequency domain. This approach models the degraded image as a non-blocky image interfered with a pure blocky signal, and estimates the power of the blocky signal by examining the coefficients for the frequencies at which blocking is expected to occur.

The method calculates a residual image and then applies a 1D FFT to each scanline of the residual to obtain a collection of residual power spectra. This collection is then averaged to yield the combined power spectrum. Figure 2.4 shows the combined power spectrum of the standard Lenna image compressed at approximately 0.5311 bits per pixel. N represents the width of the FFT window, and the x-axis origin corresponds to DC. The red peak at $0.125N$ is caused by blocking artifacts, as the blocking signal frequency is $0.125 = \frac{1}{8}$ cycles per pixel⁴. The peaks appearing at $0.25N$, $0.375N$, $0.5N$ are all multiples of the blocking signal frequency and thus are harmonics.

This approach corresponds to the green blocks in Figure 2.3. It can be seen that it requires an inverse DCT *and* FFT to be performed. This increases the complexity of the method, and is its main disadvantage. The application of this method to video would be performed in similar fashion as suggested for [32].

2.3.5 Blocking Estimation in the DCT Domain

In an attempt to provide a blocking estimation method with reduced computational complexity, [40] proposes an approach that works directly in the DCT frequency domain. That is, the approach does not require the full decoding of JPEG images, short-cutting the image decoding pipeline. This approach is represented by the red block in Figure 2.3.

The approach works by hypothetically combining adjacent spatial blocks to form new shifted blocks, as illustrated in Figure 2.5. The resulting block $\hat{b}$ contains the block boundary between the two original blocks b_1 and b_2 .

Mathematically, this could be represented by the following spatial domain equation:

$$\hat{b} = b_1q_1 + b_2q_2 \tag{2.1}$$

⁴Block width for JPEG is 8 pixels

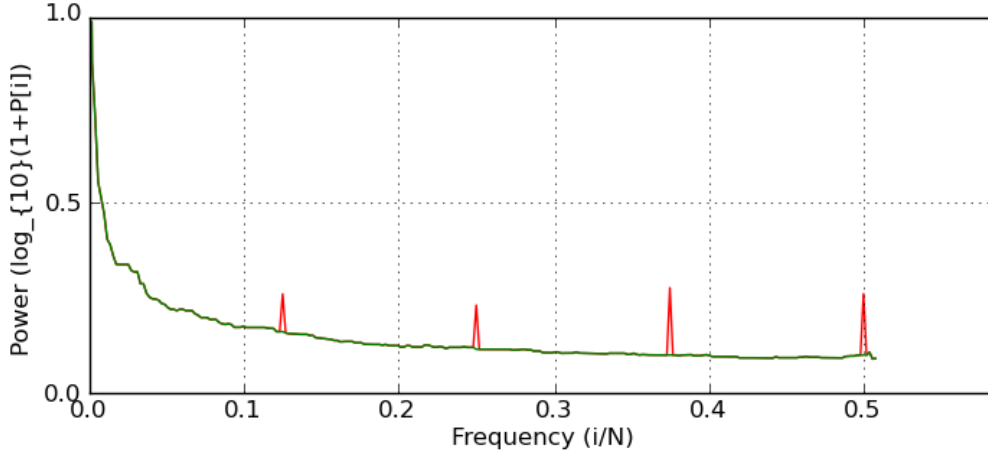


Figure 2.4: Blocking estimation in the FFT domain

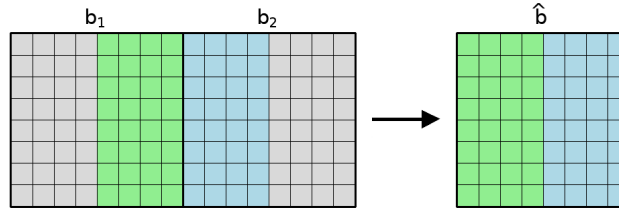


Figure 2.5: Forming new shifted blocks.

$$q_1 = \begin{bmatrix} O & O_{4 \times 4} \\ I_{4 \times 4} & O \end{bmatrix}, q_2 = \begin{bmatrix} O & I_{4 \times 4} \\ O_{4 \times 4} & O \end{bmatrix}$$

where I is the identity matrix and O is the zero matrix. The DCT frequency domain representation of the new block can be calculated as:

$$\hat{B} = B_1 Q_1 + B_2 Q_2 \quad (2.2)$$

where $\hat{B}$, B_1 , B_2 , Q_1 , Q_2 are DCT domain representations of $\hat{b}$, b_1 , b_2 , q_1 , q_2 , respectively. Now, each $\hat{B}$ can be calculated by using only matrix multiplication and addition, as B_1 and B_2 are taken directly from the entropy decoder output in Figure 2.3, and Q_1 , Q_2 are constant. Thus $\hat{B}$ can be calculated without leaving the DCT domain. Since the new blocks will include the block boundaries of the original image, the strength of the blocking signal can be estimated by examining coefficients of the new blocks at expected frequencies.

The benefit of this approach is its efficiency, making it particularly well-suited to quality assessment of encoded video. The disadvantage is that it's somewhat codec-dependent in assuming the DCT coefficients will be available. This won't always be the case, e.g. when dealing with uncompressed images, or images compressed using a different transform. In such a case, the approach can still be used, but it will be less efficient as it will require forward DCTs to be computed. Another benefit of working directly in the DCT domain is that the power of higher frequency components can be more easily determined. This makes it possible to combine this approach with blur estimation approaches that work in the DCT domain to produce an image quality assessment method that is sensitive to both blur and blocking degradation.

2.4 Datasets

This section introduces the datasets that were used for the experiments.

2.4.1 Artificial Dataset

The artificial dataset was created by uploading a 1080p HD video to YouTube. Copies of the uploaded video were then downloaded in several different subsampled resolutions (more specifically, 240p, 360p, 480p and 720p). Furthermore, 5 generations of copy were obtained by uploading the previous generation and then downloading it, where the zeroth generation is the original video. Table 2.1 shows a summary of the artificial dataset. This dataset is available online⁵. Figure 2.6 shows zoomed crops from each video.

2.4.2 Real-world Dataset (*Klaus*)

The dataset was created by crawling Youtube⁶ for all footage related to the Chilean pen incident⁷. It includes a total of 76 videos, corresponding to 1 hour 16 minutes of footage. After manually examining the downloaded content, a number of distinct versions of the footage emerged. Table 2.2 shows a subset of these versions with their resolutions, logo and caption properties. The *CT1* and *168H* columns list the presence of CT1 (channel) and 168H (program) logos, while the *Comments* column lists the presence of any other logos and/or notable features. Figure 2.7 shows frames from

⁵<http://tinyurl.com/lup9sch>

⁶<http://www.youtube.com>

⁷http://en.wikipedia.org/wiki/2011_Chilean_Pen_Incident

Table 2.1: A summary of the artificial dataset.

Video	Comment
V_0	Original video (prior to upload)
V_1	Uploaded copy of V_0 , downloaded as 1080p
V_2	Uploaded copy of V_0 , downloaded as 720p
V_3	Uploaded copy of V_0 , downloaded as 480p
V_4	Uploaded copy of V_0 , downloaded as 360p
V_5	Uploaded copy of V_0 , downloaded as 240p
V_6	2nd generation copy
V_7	3rd generation copy
V_8	4th generation copy
V_9	5th generation copy

Table 2.2: Different versions of the *Klaus* video.

Version	Resolution	Comments
CT1	854×480	None
RT	656×480	Cropped, RT logo
Euronews	640×360	Euronews logo
MS NBC	1280×720	None
AP	854×478	AP logo, captions
Horosho	854×470	Russian captions
CT24	640×360	CT24 logo, captions
GMA	640×360	GMA captions
Flipped	320×240	Horizontally flipped
TYT	720×480	Letterboxed

each of the 10 versions described in Table 2.2. While the videos show the same scene, the subtle differences in logos and captions are clearly visible.

Based on view count (not shown), CT1 was the most popular version. Most other versions appear to be scaled near-duplicate copies of this version, with notable exceptions: the CT24 version replaces the CT1 logo with the CT24 logo; the Euronews version does not have any of the logos from the CT1 version; and the AP version does not have the CT1 logo, but has the 168H logo. The absence of CT1 and/or 168H logo in some of the videos is an interesting property of the test set because it hints at the possibility of there being a number of intermediate sources for the footage.

The ground truth values for this dataset were obtained through subjective evaluation. A total of 20 subjects participated in the evaluation.

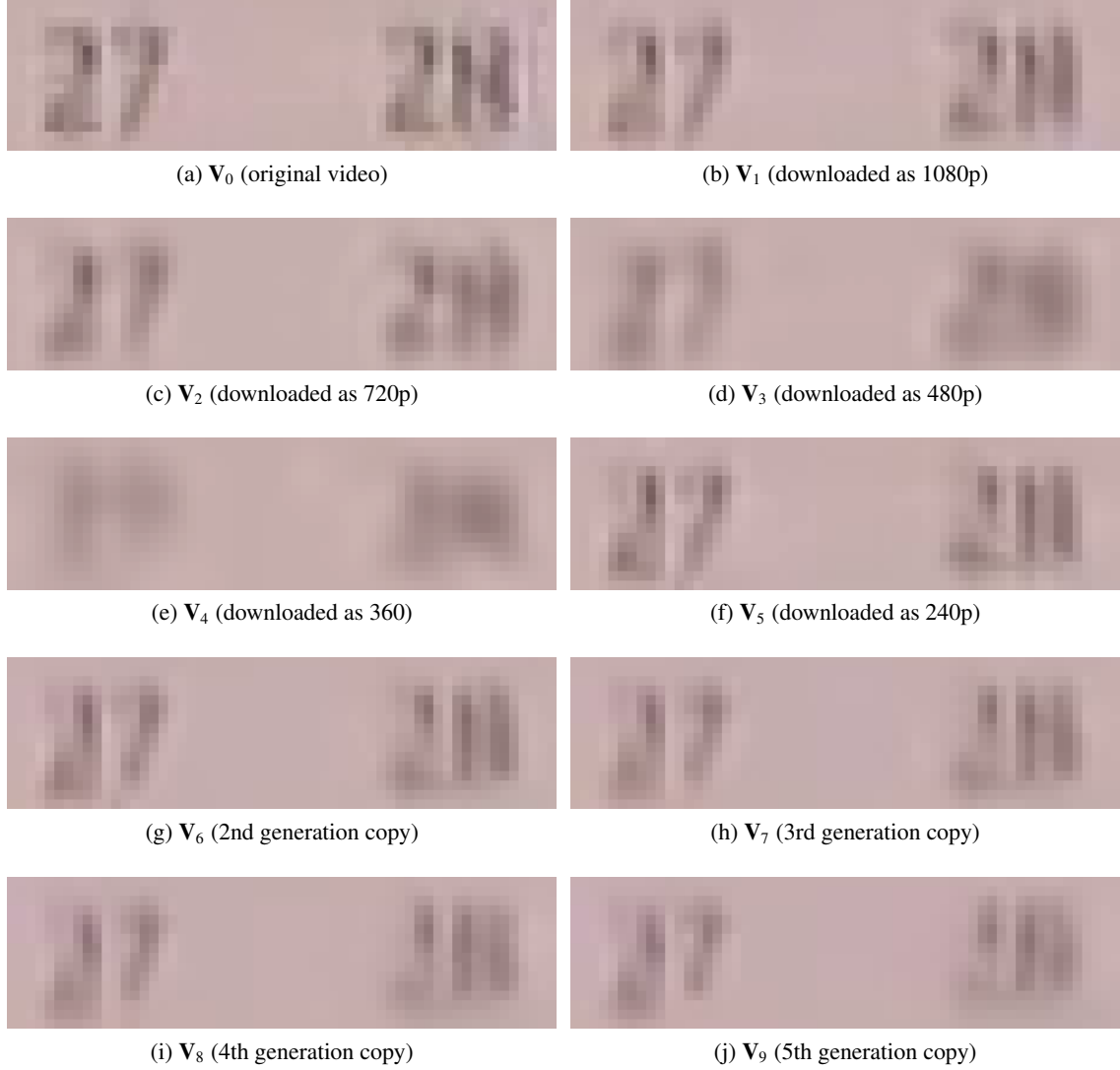


Figure 2.6: Zoomed crops from each video of Table 2.1.

2.5 Experiment: Comparison of No-reference Methods

2.5.1 Aim

This experiment aims to determine the most effective NRQA algorithm out of those presented in Sec. 2.3⁸.

⁸excluding the algorithm described in Sec. 2.3.5, since it can only be applied directly to JPEG images, not video



(a) AP



(b) CT1



(c) CT24



(d) Euronews



(e) Flipped



(f) GMA



(g) Horosho



(h) MS NBC



(i) RT



(j) TYT

Figure 2.7: Frames from 10 different versions of the video.

Table 2.3: Results of the comparison using the artificial dataset.

Algorithm	$ r $	$ \rho $
<i>edge-width</i>	0.26	0.14
<i>edge-width-scaled</i>	0.92	0.83
<i>gbim</i>	0.11	0.13
<i>nr-bbd</i>	0.77	0.83
<i>be-fft</i>	0.10	0.54

2.5.2 Method

To compare the effectiveness of the NRQA algorithms, I measured the sample correlation coefficient⁹ of the outputs from each algorithm with the ground truth values for each dataset. The sample correlation coefficient of two sequences of equal length $x = \{x_1, \dots, x_N\}$ and $y = \{y_1, \dots, y_N\}$, where N is the length of the sequences, is calculated as:

$$r(x, y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}. \quad (2.3)$$

Additionally, I calculated the rank order correlation coefficient¹⁰ as:

$$\rho(x, y) = r(X, Y), \quad (2.4)$$

where X and Y are the respective rank orders for each item of x and y , respectively. Coefficients of ± 1 indicate perfect correlation, while coefficients close to zero indicates poor correlation.

2.5.3 Results

Table 2.3 shows the results for the experiment using the artificial dataset. More specifically, it shows the correlation of the results obtained using each conventional algorithm to the ground truth. As expected, the edge width algorithm performs much better with scaling than without, since the dataset includes videos of different resolutions.

Similarly, Table 2.3 shows the results for the experiment using the artificial dataset. In this case, the edge width algorithm with scaling is the only algorithm that correlates strongly with the ground truth. Therefore, the best-performing algorithm for both the artificial and real datasets was the edge width algorithm.

⁹Also known as Pearson's r .

¹⁰Also known as Spearman's ρ .

Table 2.4: Results of the comparison using the real data set.

Algorithm	$ r $	$ \rho $
<i>edge-width</i>	0.11	0.23
<i>edge-width-scaled</i>	0.63	0.56
<i>gbim</i>	0.01	0.02
<i>nr-bbd</i>	0.08	0.38
<i>be-fft</i>	0.01	0.23



Figure 2.8: The mask used for the robustness experiment (black border for display purposes only)

2.6 Experiment: Robustness to Logos and Captions

2.6.1 Aim

The presence of logos and captions (see Fig. 2.7) was expected to interfere with the edge width algorithm, since logos and captions also contain edges. The aim of the experiment was to quantify the effect of this interference and investigate methods of mitigating it.

2.6.2 Method

In order to reduce the effect of logos and captions on the edge width algorithm, a mask was created as follows:

$$D(n, m) = |I_n - I_m| \star G(\sigma) \quad (2.5)$$

$$M = \sum_{n=1}^N \sum_{m=n+1}^N \text{thresh}(D(n, m), \epsilon) \quad (2.6)$$

where I_m refers to the spatio-temporally aligned frames; $G(\sigma)$ is a zero-mean Gaussian; $\star$ is the spatial convolution operator; thresh is the binary threshold function; σ and ϵ are empirically determined constants with values of 4.0 and 64, respectively. Zero areas of the mask indicate areas of

Table 2.5: Edge width outputs with and without masking.

Version	Comparative	Masked	Rank
CT1	4.35	4.57	1
RT	3.79	4.03	2
Euronews	4.76	4.97	3
MS NBC	4.76	4.91	4
AP	4.96	5.22	5
Horosho	4.91	5.39	6
CT24	5.21	5.65	7
GMA	5.20	7.35	8
Flipped	7.25	7.38	9
TYT	7.84	8.13	10
r	0.86	0.93	1.00
ρ	0.97	0.98	1.00

the frame that should be included in the edge width algorithm calculation. Such areas are shown in white in Figure 2.8. Colored areas are excluded from the calculation.

The edge width algorithm described in Sec. 2.3.1 was applied to each image in the dataset, both with and without masking. The results were compared with the ground truth, which was obtained through subjective evaluation.

2.6.3 Results

Table 2.5 shows the edge width algorithm outputs with and without masking. The rank column shows the rank of each version, ordered by the outcome of the subjective evaluation. The r and ρ rows show the sample correlation coefficient and rank order correlation coefficient, respectively, between the outputs of the algorithm and the rank. Since the correlation coefficients are higher when masking is used, they support the hypothesis that the logos and captions do interfere with the edge width algorithm. However, the small benefit from applying the mask is offset by the effort required to create the mask. More specifically, creating the mask requires comparing the images to each other exhaustively, which is not practical when dealing with a large number of images.

2.7 Investigation: Comparison of the Causes of Visual Quality Loss

As mentioned in Sec. 2.2, there are several causes of visual quality loss in Web video. This experiment utilizes the artificial data set and focuses on the download and re-compression stages. Table 2.6 shows

Table 2.6: Download stage visual quality degradation.

Resolution	SSIM
1080p	1.00
780p	0.96
480p	0.92
360p	0.87
240p	0.81

Table 2.7: Recompression stage visual quality degradation.

Generation	SSIM
0	1.00
1	0.95
2	0.94
3	0.93
4	0.92
5	0.91

the SSIM output for each downloaded resolution of V_0 . Furthermore, Table 2.7 shows the SSIM output for each copy generation. These tables allow the effect of each stage on the visual quality to be compared. For example, downloading a 480p copy of a 1080p video accounts for as much quality loss as downloading a 4th generation copy of the same 1080p video.

2.8 Investigation: Sensitivity to Visual Content

Visual content is known to affect NRQA algorithms in general and the edge width algorithm in particular [23]. To illustrate this effect, Fig. 2.9 shows images of similar visual quality, yet significantly different edge widths: 2.92, 2.90, 3.92 and 4.68 for Figs. 2.9 (a), (b), (c) and (d), respectively. Without normalization, the images of Figs. 2.9(c) and (d) would be penalized more heavily than the images of Figs. 2.9(a) and (b). However, since these images are all uncompressed, they are free of compression degradations, and there is no information loss with respect to the parent video. Therefore, the images should be penalized equally, that is, not at all. Normalization thus reduces the effect of visual content on the NRQA algorithm and achieves a fairer penalty function.



(a) *Mobcal* (2.92)



(b) *Parkrun* (2.90)



(c) *Shields* (3.92)



(d) *Stockholm* (4.68)

Figure 2.9: Images with the same visual quality, but different edge widths (shown in parentheses).

2.9 Conclusion

This chapter introduced visual quality assessment in general, and several no-reference visual quality assessment algorithms in particular. The algorithms were compared empirically using both artificial and real-world data. Out of the examined algorithms, the edge width algorithm described in Sec. 2.3.1 proved to be the most effective.

Chapter 3

Shot Identification

3.1 Introduction

This chapter introduces shot identification, an important pre-processing step that enables the method proposed in this thesis. This chapter is organized as follows. Section 3.2 introduces the notation and some initial definitions. Sections 3.3 and 3.4 describe algorithms for automatic shot segmentation and comparison, respectively. Section 3.5 presents experiments for evaluating the effectiveness of the algorithms presented in this chapter, and discusses the results. Section 3.6 investigates the application of subtitles to shot identification, as described in one of our papers [6]. Finally, Sec. 3.7 concludes the chapter.

3.2 Preliminaries

This section provides initial definitions and introduces the notation used in the remainder of the thesis. First, Fig. 3.1 shows that a single video can be composed of several shots, where a shot is a continuous sequence of frames captured by a single camera. The figure shows one video, six shots, and several hundred frames (only the first and last frame of each shot is shown).

Next, borrowing some notation and terminology from existing literature [20], we define the *parent video* $\mathbf{V}_0$ as a video that contains some new information that is unavailable in other videos. The parent video consists of several shots $\mathbf{V}_0^1, \mathbf{V}_0^2, \dots, \mathbf{V}_0^{N_{V_0}}$, where a shot is defined as a continuous sequence of frames that were captured by the same camera over a particular period of time. The top part of Fig. 3.2 shows a parent video that consists of $N_{V_0} = 4$ shots. The dashed vertical lines indicate shot boundaries. The horizontal axis shows time.

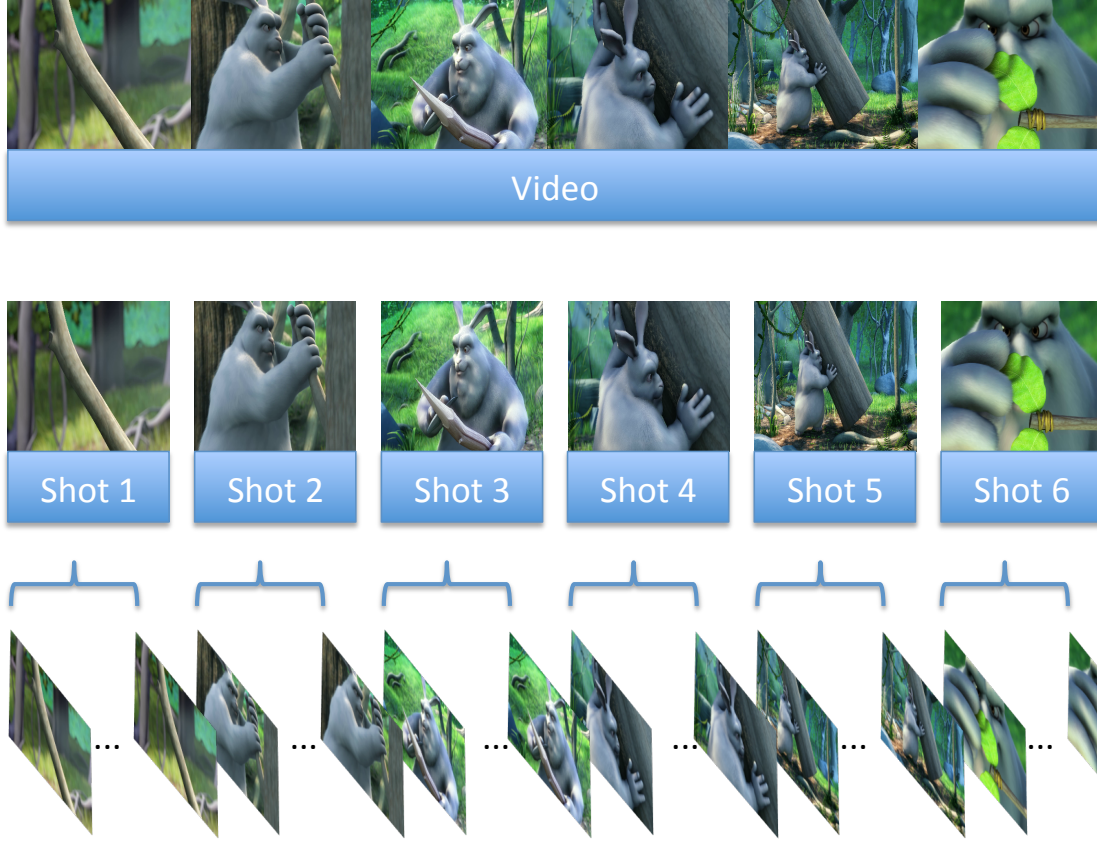


Figure 3.1: Examples of a video, shots, and frames.

The parent video can be divided into its constituent shots as shown in the bottom part of Fig. 3.2. Section 3.3 describes an algorithm for automatic shot segmentation. After segmentation, the shots can be edited and then rearranged to create *edited videos* $\mathbf{V}_1, \mathbf{V}_2, \dots, \mathbf{V}_M$ as shown in Fig. 3.3. Each shot of the edited videos will be visually similar to some shot of the parent video. We use the notation $\mathbf{V}_{j_1}^{k_1} \simeq \mathbf{V}_{j_2}^{k_2}$ to indicate that $\mathbf{V}_{j_1}^{k_1}$ is visually similar to $\mathbf{V}_{j_2}^{k_2}$. Section 3.4 describes an algorithm for determining whether two shots are visually similar.

Next, we use the notation I_j^k to refer to the *shot identifier* of $\mathbf{V}_j^k$, and I_j to denote the set of all the shot identifiers contained in $\mathbf{V}_j$. These shot identifiers are calculated as follows. First, a graph is constructed, where each node corresponds to a shot in the edited videos, and edges connect two shots $\mathbf{V}_{j_1}^{k_1}$ and $\mathbf{V}_{j_2}^{k_2}$ where $\mathbf{V}_{j_1}^{k_1} \simeq \mathbf{V}_{j_2}^{k_2}$. Finally, connected components are detected, and an arbitrary integer assigned to each connected component — these integers are the shot identifiers. Figure 3.4 illustrates the process of calculating the shot identifiers for the edited videos shown in Fig. 3.3. In this

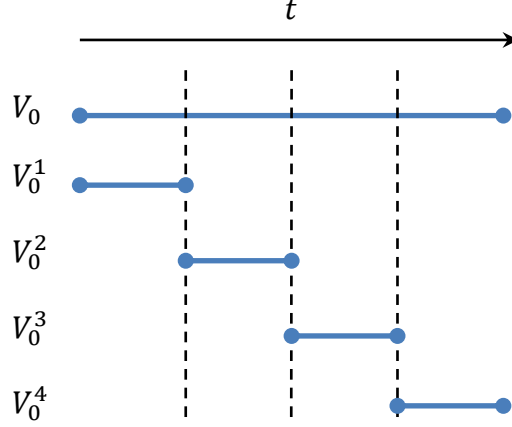


Figure 3.2: An example of a parent video and its constituent shots.

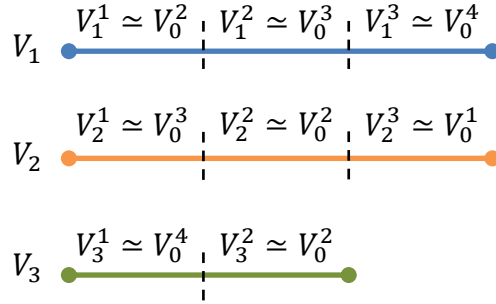


Figure 3.3: An example of videos created by editing the parent video of Fig. 3.2.

example, $I_1^1 = I_3^2 = I_2^2 = i_1$, $I_1^2 = I_2^1 = i_2$, $I_1^3 = I_3^1 = i_3$, $I_2^3 = i_4$ and i_1 through to i_4 are different from each other. The color of each node indicates its connected component. The actual value of the shot identifier assigned to each connected component and the order of assigning identifiers to the connected components are irrelevant.

3.3 Shot Segmentation

Shot boundaries can be detected automatically by comparing the color histograms of adjacent frames, and thresholding the difference [41]. More specifically, each frame is divided into 4×4 blocks, and a 64-bin color histogram is computed for each block. The difference between corresponding blocks is then computed as:

$$F(\mathbf{f}_1, \mathbf{f}_2, r) = \sum_{x=0}^{63} \frac{\{H(\mathbf{f}_1, r, x) - H(\mathbf{f}_2, r, x)\}^2}{H(\mathbf{f}_1, r, x)}, \quad (3.1)$$

provide a good survey of existing algorithms [48].

3.4.1 A Spatial Algorithm

This section introduces a conventional algorithm for comparing two images based on their color histograms [49]. First, each image is converted into the HSV color space, divided into four quadrants, and a color histogram is calculated for each quadrant. The radial dimension (saturation) is quantized uniformly into 3.5 bins, with the half bin at the origin. The angular dimension (hue) is quantized into 18 uniform sectors. The quantization for the value dimension depends on the saturation value: for colors with saturation is near zero, the value is finely quantized into uniform 16 bins to better differentiate between grayscale colors; for colors with higher saturation, the value is coarsely quantized into 3 uniform bins. Thus the color histogram for each quadrant consists of 178 bins, and the feature vector for a single image consists of $178 \times 4 = 712$ features, where a feature corresponds to a single bin.

Next, the l_1 distance between two images $\mathbf{f}_1$ and $\mathbf{f}_2$ is defined as follows:

$$l_1(\mathbf{f}_1, \mathbf{f}_2) = \sum_{r=1}^4 \sum_{x=1}^{178} H(\mathbf{f}_1, r, x) - H(\mathbf{f}_2, r, x), \quad (3.3)$$

where r refers to one of the 4 quadrants, and $H(\cdot, r, x)$ corresponds to the x th bin of the color histogram for the r th quadrant. In order to apply Eq. (3.3) to shots, the simplest method is to apply it to the first frames of the shots being compared, giving the following definition of visual similarity:

$$\mathbf{V}_{j_1}^{k_1} \simeq \mathbf{V}_{j_2}^{k_2} \iff l_1(\mathbf{f}_{j_1}^{k_1}, \mathbf{f}_{j_2}^{k_2}) < \phi, \quad (3.4)$$

where τ is an empirically determined threshold; $\mathbf{f}_{j_1}^{k_1}$ and $\mathbf{f}_{j_2}^{k_2}$ are the first frames of $\mathbf{V}_{j_1}^{k_1}$ and $\mathbf{V}_{j_2}^{k_2}$, respectively.

3.4.2 A Spatiotemporal Algorithm

This section introduces a simple conventional algorithm for comparing two videos [50]. The algorithm calculates the distance based on two components: spatial and temporal. The distances between their corresponding spatial and temporal components of the two videos are first compared separately, and combined using a weighted sum. Each component of the algorithm is described in detail below.

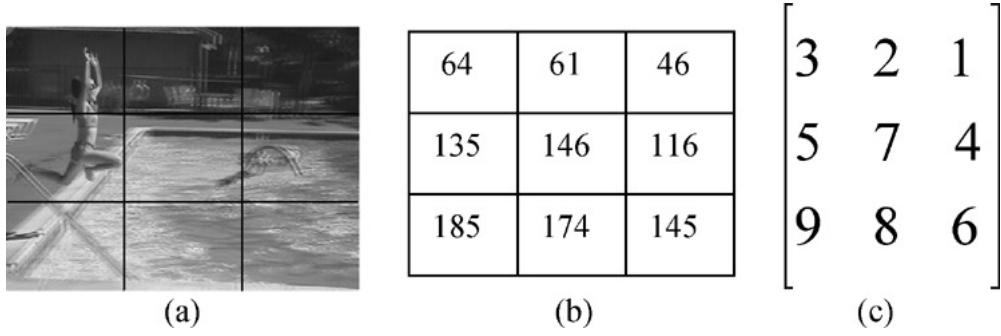


Figure 3.5: Visualizing the spatial component (from [50]).

The spatial component is based on image similarity methods and calculated for each frame of the video individually. The algorithm first discards the color information, divides the frame into $p \times q$ equally-sized blocks and calculates the average intensity for each block, where p and q are pre-defined constants. Each block is numbered in raster order, from 1 to m , where the constant $m = p \times q$ represents the total number of blocks per frame. The algorithm then sorts the blocks by the average intensity. The spatial component for a single frame is then given by the sequence of the m block numbers, ordered by the average intensity. Figure 3.5 (a), (b) and (c) visualizes a grayscale frame divided into 3×3 blocks, calculated average intensity, and the order of each block, respectively. The entire spatial component for the entire video is thus an $m \times n$ matrix, where n is the number of frames in the video.

The distance between two spatial components is then calculated as:

$$d_S(S_1, S_2) = \frac{1}{Cn} \sum_{j=1}^m \sum_{k=1}^n |S_1[j, k] - S_2[j, k]|, \quad (3.5)$$

where S_1 and S_2 are the spatial components being compared, $S_1[j, k]$ corresponds to the j th block of the k th frame of S_1 , and C is a normalization constant that represents the maximum theoretical distance between any two spatial components for a single frame.

The temporal component utilizes the differences in corresponding blocks of sequential frames. It is a matrix of $m \times n$ elements, where each element $T[j, k]$ corresponds to:

$$\delta_j^k = \begin{cases} 1 & \text{if } V_j[k] > V_j[k-1] \\ 0 & \text{if } V_j[k] = V_j[k-1], \\ -1 & \text{if otherwise} \end{cases} \quad (3.6)$$

where $V_j[k]$ represents the average intensity of the j th block of the k th frame of the video V . Figure 3.6 illustrates the calculation of the temporal component. Each curve corresponds to a different j . The

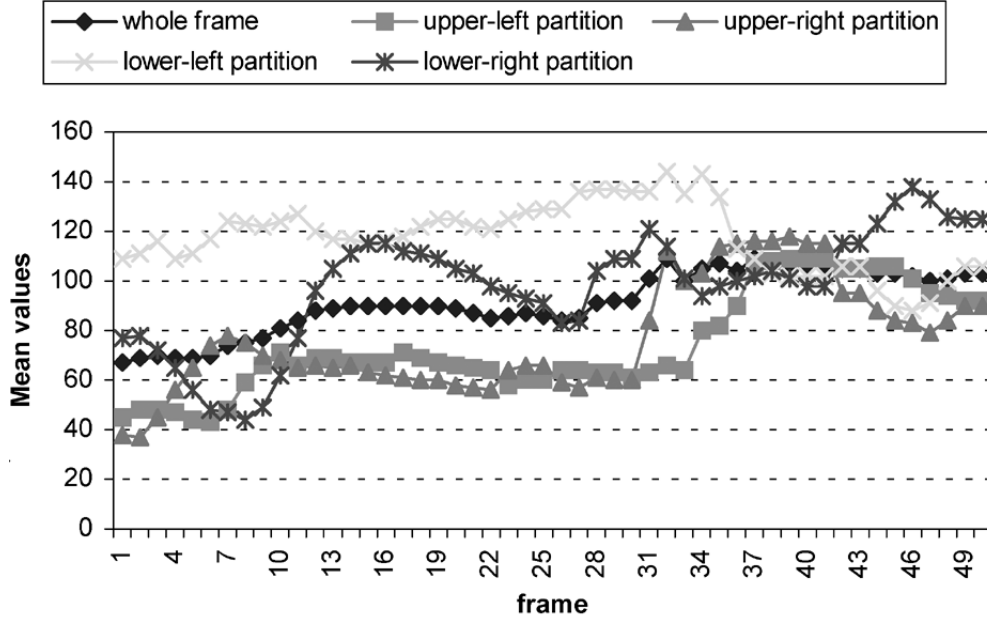


Figure 3.6: Visualizing the temporal component (from [50]).

horizontal and vertical axes correspond to k and $V_j[k]$, respectively.

Next, the distance between two temporal components is calculated as:

$$d_T(T_1, T_2) = \frac{1}{C(n-1)} \sum_{j=1}^m \sum_{k=2}^n |T_1[j, k] - T_2[j, k]|. \quad (3.7)$$

The final distance between two shots is calculated as:

$$d_{ST}(\mathbf{V}_{j_1}^{k_1}, \mathbf{V}_{j_2}^{k_2}) = \alpha D_S(S_{j_1}^{k_1}, S_{j_2}^{k_2}) + (1 - \alpha) D_T(T_{j_1}^{k_1}, T_{j_2}^{k_2}), \quad (3.8)$$

where $S_{j_1}^{k_1}$ and $T_{j_1}^{k_1}$ are the spatial and temporal components of $\mathbf{V}_{j_1}^{k_1}$, respectively, and α is an empirically-determined weighting parameter.

Finally, the algorithm enables the following definition of visual similarity:

$$\mathbf{V}_{j_1}^{k_1} \simeq \mathbf{V}_{j_2}^{k_2} \iff D_{ST}(\mathbf{V}_{j_1}^{k_1}, \mathbf{V}_{j_2}^{k_2}) < \tau, \quad (3.9)$$

where τ is an empirically determined threshold.

3.4.3 Robust Hash Algorithm

The robust hashing algorithm yields a 64-bit hash for each shot [51]. The benefits of this method are low computational and space complexity. Furthermore, once the hashes have been calculated, access

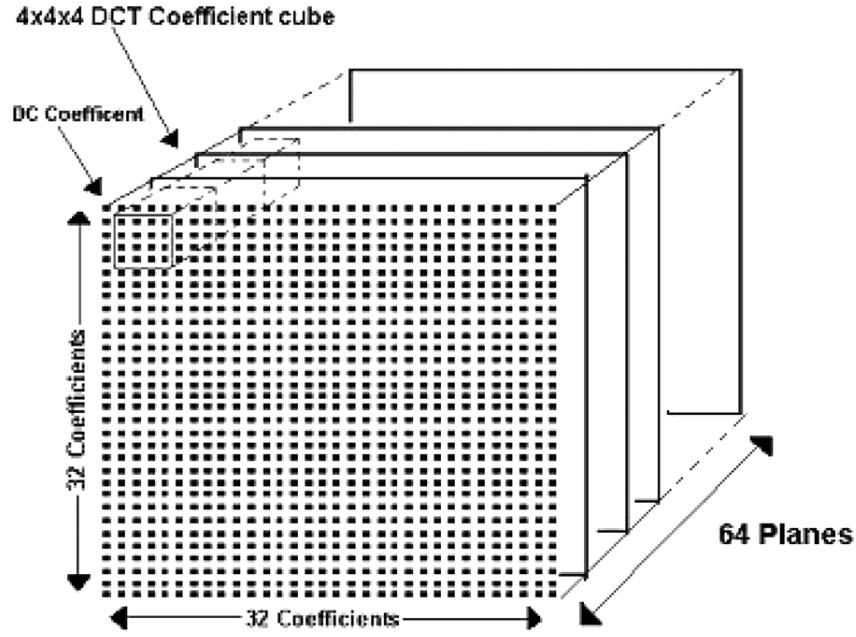


Figure 3.7: Extracting the DCT coefficients (from [51]).

to the original video is not required, enabling the application of the method to Web video sharing portals (e.g. as part of the processing performed when a video is uploaded to the portal). After the videos have been segmented into shots, as shown in the bottom of Fig. 3.2, the hashing algorithm can be directly applied to each shot. The algorithm consists of several stages: (a) preprocessing, (b) spatiotemporal transform and (d) hash computation. Each stage is described in more detail below.

The processing step focuses on the luma component of the video. First, the luma is downsampled temporally to 64 frames. Then, the luma is downsampled spatially to 32×32 pixels per frame. The motivation for the preprocessing step is to reduce the effect of differences in format and post-processing.

The spatiotemporal transform step first applies a Discrete Cosine Transform to the preprocessed video, yielding $32 \times 32 \times 64$ DCT coefficients. Next, this step extracts $4 \times 4 \times 4$ lower frequency coefficients, as shown in Fig. 3.7. The DC terms in each dimension are ignored.

The hash computation step converts the $4 \times 4 \times 4$ extracted DCT coefficients to a binary hash as follows. First, the step calculates the median of the extracted coefficients. Next, the step replaces each coefficient by a 0 if it is less than the median, and by a 1 otherwise. The result is an sequence of 64 bits. This is the hash.

Table 3.1: Summary of the datasets used in the experiments.

Name	Videos	Total dur.	Shots	IDs
<i>Bolt</i>	68	4 h 42 min	1933	275
<i>Kerry</i>	5	0 h 47 min	103	24
<i>Klaus</i>	76	1 h 16 min	253	61
<i>Lagos</i>	8	0 h 6 min	79	17
<i>Russell</i>	18	2 h 50 min	1748	103
Total	175	9 h 41 min	4116	480

The calculated hashes enable shots to be compared using the Hamming distance:

$$D(\mathbf{V}_{j_1}^{k_1}, \mathbf{V}_{j_2}^{k_2}) = \sum_{b=1}^{64} h_{j_1}^{k_1}[b] \oplus h_{j_2}^{k_2}[b], \quad (3.10)$$

where $\mathbf{V}_{j_1}^{k_1}$ and $\mathbf{V}_{j_2}^{k_2}$ are the two shots being compared; $h_{j_1}^{k_1}$ and $h_{j_2}^{k_2}$ are their respective hashes, represented as sequences of 64 bits each; and $\oplus$ is the exclusive OR operator. Finally, thresholding the Hamming distance enables the determination of whether two shots are visually similar:

$$\mathbf{V}_{j_1}^{k_1} \simeq \mathbf{V}_{j_2}^{k_2} \iff D(\mathbf{V}_{j_1}^{k_1}, \mathbf{V}_{j_2}^{k_2}) < \tau, \quad (3.11)$$

where τ is an empirically determined threshold between 1 and 63.

3.5 Experiments

This section presents the results of experiments that compare the effectiveness of the algorithms introduced in this chapter. Section 3.5.1 introduces the datasets used in the experiments. Then, Sections 3.5.2 and 3.5.3 evaluate the algorithms implemented by Equations (3.1) and (3.11), respectively.

3.5.1 Datasets

Table 3.1 shows the datasets used in the experiment, where each dataset consists of several edited videos collected from YouTube. The “Shots” and “IDs” columns show the total number of shots and unique shot identifiers, respectively. The *Klaus* dataset was first introduced in detail in Sec. 2.4.2. The remaining datasets were all collected in similar fashion. Figures 3.8(a)–(d), (e)–(h), (i)–(l), (m)–(p), and (q)–(t) show a small subset of screenshots of videos from the *Bolt*, *Kerry*, *Klaus*, *Lagos* and *Russell* datasets, respectively. The ground truth for each dataset was obtained by manual processing.

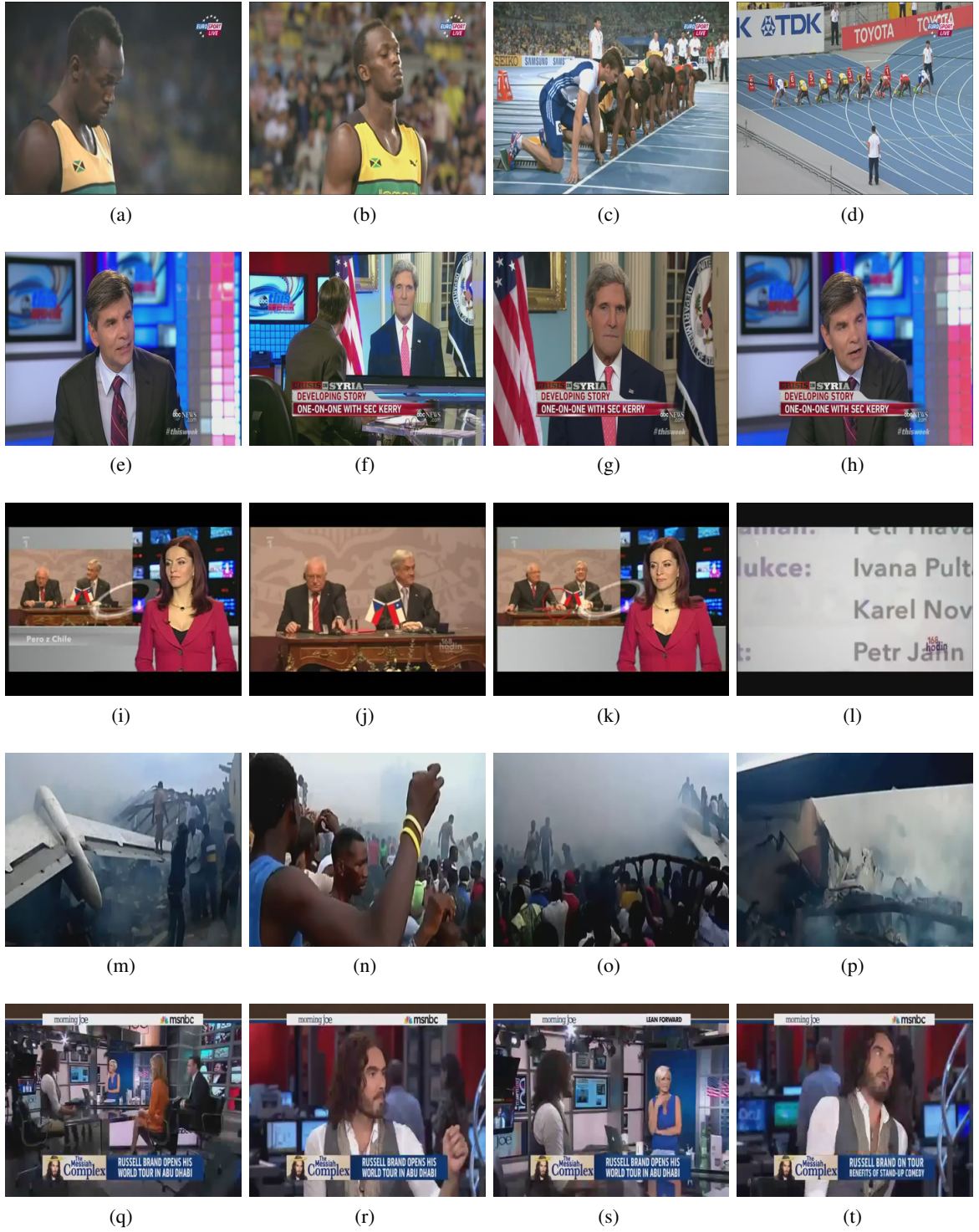


Figure 3.8: Frames from the real datasets. Top to bottom, the five image rows show screenshots from the *Bolt*, *Kerry*, *Klaus*, *Lagos* and *Russell* datasets, respectively, of Table 3.1.

Table 3.2: Accuracy of the automatic shot boundary detector ($\nu = 5.0$).

Dataset	Precision	Recall	F_1 score
<i>Bolt</i>	0.81	0.65	0.72
<i>Klaus</i>	0.88	0.77	0.81
<i>Kerry</i>	0.91	0.81	0.86
<i>Lagos</i>	0.92	0.69	0.79
<i>Russell</i>	0.98	0.89	0.93

3.5.2 Shot Segmentation

The parameter ν determines the shot boundary detection threshold. It is applied to the result of Eq. (3.2). For this experiment, its value was empirically set to 5.0. Since the actual shot boundaries are known from ground truth, it is possible to evaluate the shot boundary detector in terms of precision, recall and F_1 :

$$P = \frac{TP}{TP + FP}, \quad (3.12)$$

$$R = \frac{TP}{TP + FN}, \quad (3.13)$$

and

$$F_1 = 2 \times \frac{P \times R}{P + R}, \quad (3.14)$$

respectively, where TP is the number of true positives: actual boundaries that were detected by the automatic algorithm; FP is the number of false positives: boundaries that were detected by the automatic algorithm but that are not actual shot boundaries; and FN is the number of false negatives: actual boundaries that were not detected by the automatic algorithm. The mean precision, recall and F_1 score for each dataset is shown in Table 3.2. These results show that while the precision of the algorithm is relatively high for all datasets, the recall depends on the data. Through manual examination of the results, the main cause of false negatives (low recall) was identified as “fade” or “dissolve” shot transitions. The *Bolt* and *Lagos* datasets contain many such transitions, leading to the relatively low recall for those datasets.

3.5.3 Shot Comparison

The parameter τ of Eq. (3.11) determines the Hamming distance threshold for visual similarity. In order to investigate the significance of τ , we exhaustively compared the Hamming distances between

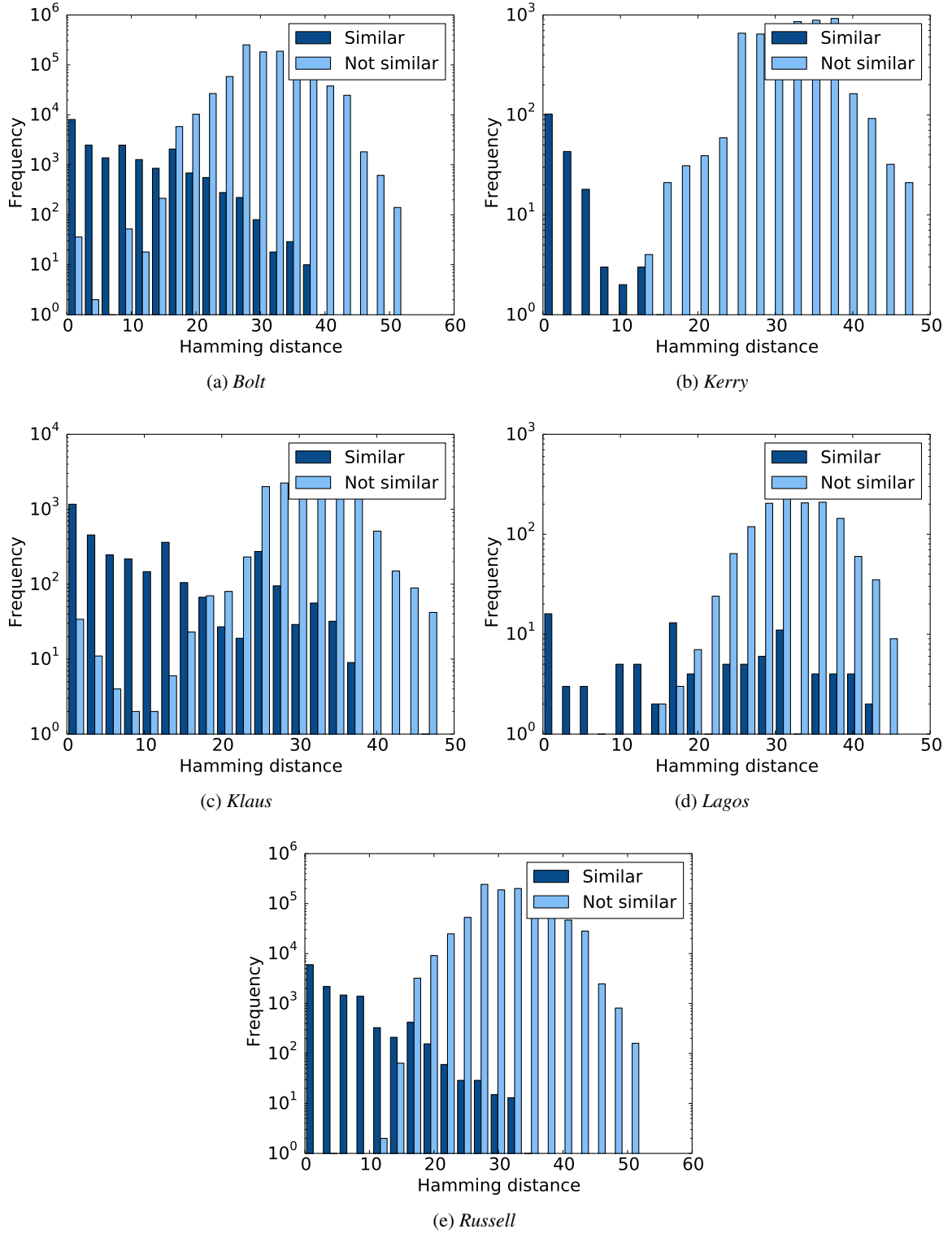


Figure 3.9: A histogram of Hamming distances for the datasets in Table 3.1.

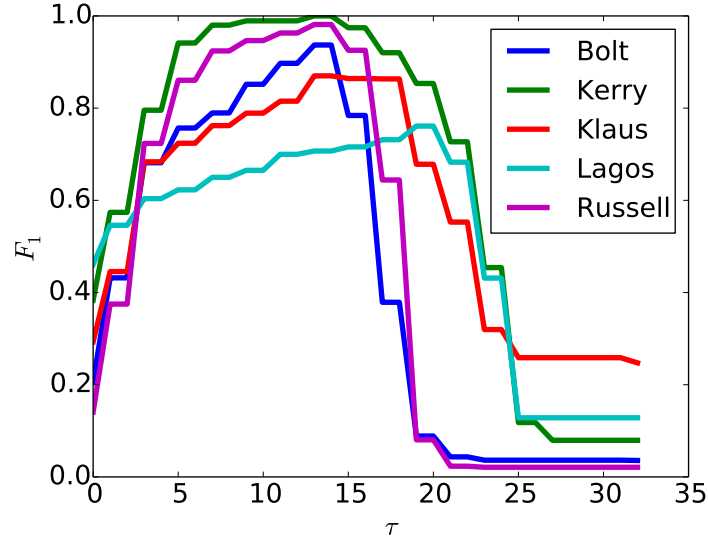


Figure 3.10: F_1 as a function of τ for the datasets in Table 3.1.

all shots in each dataset and plotted their histograms. The histograms are shown in Fig. 3.9, with frequency on a logarithmic scale. Two histograms are shown for each dataset: (i) the “similar” histogram corresponds to Hamming distances between shots that are visually similar; and (ii) the “not similar” histogram corresponds to Hamming distances between shots that are not visually similar. Figure 3.9 shows that as the Hamming distance increases, the proportion of comparisons between visually similar shots decreases, and the proportion of comparisons between visually dissimilar shots increases. However, it is obvious that picking a value for τ that allows all shots to be identified correctly is impossible, since the two distributions always overlap.

In order to quantitatively judge the accuracy of the automatic method for calculating shot identifiers, we conducted an experiment using the query-response paradigm, where the query is a shot, and the response is the set of all shots that have the same shot identifier, as calculated by the automatic method. Using the manually-assigned shot identifiers as ground truth, we then measured the precision, recall and F_1 score as Eqs. (3.12), (3.13), and (3.14), respectively. In this case, true positives are shots in the response that have the same manually-assigned shot identifier as the query shot; false positives are shots in the response that have different manually-assigned shot identifiers to the query shot; and false negatives are shots that were not part of the response, but that have the same manually-assigned shot identifiers as the query shot. The average values across all shots for each dataset were

Table 3.3: Accuracy of the automatic shot identifier function.

Dataset	Precision	Recall	F_1 score
<i>Bolt</i>	0.99	0.92	0.94
<i>Kerry</i>	1.00	1.00	1.00
<i>Klaus</i>	1.00	0.84	0.87
<i>Lagos</i>	1.00	0.62	0.71
<i>Russell</i>	1.00	0.98	0.98

then calculated.

Finally, Fig. 3.10 shows the F_1 score for each dataset for different values of τ . From this figure, we set the value of τ to 14 in the remainder of the experiment, as that value achieves the highest average F_1 score across all the datasets. Table 3.3 shows the obtained results for that value of τ . From this table, it is obvious that some datasets are easier to calculate shot identifiers for than others: for example, *Kerry*; and some are more difficult: for example, *Lagos*. More specifically, in the *Lagos* case, the recall is particularly low. This is because the dataset consists of videos that have been edited significantly, for example, by inserting logos, captions and subtitles. In such cases, the hashing algorithm used is not robust enough, and yields significantly different hashes for shots that a human would judge to be visually similar. Nevertheless, Table 3.3 shows that automatic shot identifier calculation is possible.

3.6 Utilizing Semantic Information for Shot Comparison

Web video often contains shots with little temporal activity, and shots that depict the same object at various points in time. An example of such shots includes anchor shots and shots of people speaking (see Figure 3.11). Since the difference in spatial and temporal content is low, the algorithms introduced in Sections 3.4.1 and 3.4.2 cannot compare such shots effectively. However, if such shots contain different speech, then they are semantically different. This section investigates the use of semantic information for shot comparison. More specifically, the investigation will examine using subtitles to distinguish visually similar shots.

There are several formats for subtitles [52]. In general, subtitles consist of phrases. Each phrase consists of the actual text as well as the time period during which the text should be displayed. The subtitles are thus synchronized with both the moving picture and the audio signal. Figure 3.12 shows an example of a single phrase. The first row contains the phrase number. The second row contains



Figure 3.11: Unique shots from the largest cluster (first frames only).

2
00:00:03,000 -- > 00:00:06,000
Secretary of State John Kerry.
Thank you for joining us, Mr. Secretary.

Figure 3.12: An example of a phrase.

the time period. The remaining rows contain the actual text. In practice, the subtitles can be created manually. Since this is a tedious process, Automatic Speech Recognition (ASR) algorithms that extract the subtitles from the audio signal have been researched for over two decades [53]. Furthermore, research to extract even higher-level semantic information from the subtitles, such as speaker identification [54], is also ongoing.

Subtitles are also useful for other applications. For example, [55] implements a video retrieval system that utilizes the moving picture, audio signal and subtitles to identify certain types of scenes, such as gun shots or screaming.

3.6.1 Investigation: the Effectiveness of Subtitles

This section introduces a preliminary investigation that confirms the limitations of the algorithm introduced in Sec. 3.4.1 and demonstrates that subtitle text can help overcome these limitations. This section first describes the media used in the investigation, before illustrating the limitations and a potential solution.

Table 3.4: Subtitle text for each shot in the largest cluster.

Shot	Duration	Text
3	4.57	And I'm just wondering, did you...
6	30.76	that the President was about...
9	51.55	Well, George, in a sense, we're...
13	59.33	the President of the United...
15	15.75	and, George, we are not going to...
17	84.38	I think the more he stands up...
19	52.39	We are obviously looking hard...
21	83.78	Well, I would, we've offered our...
25	66.50	Well, I've talked to John McCain...
27	66.23	This is not Iraq, this is not...

The video used in the investigation is an interview with a political figure, obtained from YouTube. The interview lasts 11:30 and consists of 28 semantically unique shots. Shot segmentation was performed manually. The average shot length was 24.61s. Since the video did not originally have subtitles, the subtitles were created manually.

Next, the similarity between the shots was calculated exhaustively using a conventional method [44], and a clustering algorithm applied to the results. Since the shots are semantically unique, ideally, there should be 28 singleton clusters – one cluster per shot. However, there were a total of 5 clusters as a result of the clustering. Figure 3.13 shows the first frames from representative shots in each cluster. The clusters are numbered arbitrarily. The representative shots were also selected arbitrarily. The numbers in parentheses show the number of shots contained in each cluster. This demonstrates the limitation of the conventional method, which focuses on the moving picture only.

Finally, Table 3.4 shows the subtitle text for each of the shots contained in the largest cluster, which are shown in Fig. 3.11. The shot number corresponds to the ordinal number of the shot within the original video. The table clearly shows that the text is significantly different for each shot. Including the text in the visual comparison will thus overcome the limitations of methods that focus on the moving picture only. In this investigation, since the subtitles were created manually, they are an accurate representation of the speech in the video, and strict string comparisons are sufficient for shot identification. If the subtitles are inaccurate (out of sync or poorly recognized words), then a more tolerant comparison method will need to be used.



Figure 3.13: Representative shots from each cluster. The parentheses show the number of shots in each cluster.

3.6.2 Proposed Method

The proposed method compares two video shots, S_i and S_j . The shots may be from the same video or different videos. Each shot consists of a set of frames $F_{(\cdot)}$ and a set of words $W_{(\cdot)}$, corresponding to the frames and subtitles contained by the shot, respectively. The method calculates the difference between the two shots using a weighted average:

$$D(S_i, S_j) = \alpha D_f(F_i, F_j) + (1 - \alpha) D_w(W_i, W_j), \quad (3.15)$$

where α is a weight parameter, D_f calculates the visual difference, and D_w calculates the textual difference. The visual difference can be calculated using any of the visual similarity methods mentioned in Sec. 3.4. The textual difference can be calculated as the bag-of-words distance [56] commonly used in information retrieval:

$$D_w(W_i, W_j) = \frac{|W_i \cap W_j|}{|W_i \cup W_j|}, \quad (3.16)$$

where $|\cdot|$ calculates the cardinality of a set.

Equation 3.15 thus compares two shots based on their visual and semantic similarity.

3.6.3 Preliminary Experiment

This section describes an experiment that was performed to verify the effectiveness of the proposed method. This section first reviews the purpose of the experiment, introduces the media used, describes the empirical method and evaluation criteria and, finally, discusses the results.

The authenticity degree method introduced in Chapter 4 requires shots to be identified with high precision and recall. The shot identification component must be robust to changes in visual quality,

Table 3.5: Video parameters

Video	Fmt.	Resolution	Duration	Shots
V_1	MP4	1280×720	01:27	6
V_2	MP4	1280×720	11:29	23
V_3	FLV	640×360	11:29	23
V_4	MP4	1280×720	11:29	23
V_5	FLV	854×480	11:29	23

Table 3.6: ASR output example for one shot

Well, I, I would, uh, we've offered our friends, we've offered the Russians previously...
well highland and we've austerity our friendly the hovering the russians have previously...
highland and we've austerity our friendly the hovering the russians have previously...
well highland and we poverty our friendly pondered the russians have previously...
well highland and we've austerity our friendly the hovered russians have previously...

yet sensitive to changes in visual and semantic content. The aim of this experiment is thus to compare shots that differ in the above parameters using the method proposed in Section 3.6.2 and determine the precision and recall of the method.

The media used for the method includes five videos of a 11min 29s interview with a political figure. All videos were downloaded from YouTube¹. The videos differ in duration, resolution, video and audio quality. None of the videos contained subtitles. Table 3.5 shows the various parameters of each video, including the container format, resolution, duration and number of shots. During pre-processing, the videos were segmented into shots using a conventional method [41]. Since this method is not effective for certain shot transitions such as fade and dissolve, the results were corrected manually. The subtitles for each shot were then created automatically using an open-source automatic speech recognition library². Table 3.6 shows an example of extracted subtitles, with a comparison to manual subtitles (top line). As can be seen from the table, the automatically extracted subtitles differ from the manually extracted subtitles significantly. However, these differences appear to be systematic. For example, “offered” is mis-transcribed as “hovering” in all of the four automatic cases. Therefore, while the automatically extracted subtitles may not be useful for understanding the semantic content of the video, they may be useful for identification.

The experiment first compared shots to each other exhaustively and manually, using a binary scale:

¹www.youtube.com

²<http://cmusphinx.sourceforge.net/>

Table 3.7: A summary of experiment results.

Method	P	R	F_1
Visual only	0.37	1.00	0.44
Proposed	0.98	0.96	0.96

similar or not similar. Shots that were similar to each other were grouped into clusters. There was a total of 23 clusters, with each cluster corresponding to a visually and semantically unique shot. Each node in the graph corresponds to a shot, and two similar shots are connected by an edge. These clusters formed the ground truth data which was used for evaluating the precision and recall of the proposed method. Next, the experiment repeated the comparison using method proposed in Section 3.6.2, using an α of 0.5 to weigh visual and word differences equally. To calculate visual similarity between two shots, the color histograms of the leading frames of the two shots were compared using a conventional method [49].

To evaluate the effectiveness of the proposed method, the experiment measured the precision and recall within a query-response paradigm. The precision, recall and F1 score were calculated as Eqs. 3.12, 3.13 and 3.14. Within the query-response paradigm, the query was a single shot, selected arbitrarily, and the response was all shots that were judged similar by the proposed method. True positives were response shots that were similar to the query shot, according to the ground truth. False positives were response shots that were not similar to the query shot, according to the ground truth. False negatives were shots not in the response that were similar to the query shot, according to the ground truth. The above set-up was repeated, using each of the shots as a query.

The average precision, recall and F-score were 0.98, 0.96, 0.96, respectively. In contrast, using only visual features ($\alpha = 1.0$) yields precision, recall and F-score of 0.37, 1.00, 0.44, respectively. The results are summarized in Table 3.7. These results illustrate that the proposed method is effective for distinguishing between shots that are visually similar yet semantically different.

3.6.4 Experiment

This section describes the new experiment. The purpose and method of performing the experiment are identical to those of Sec. 3.6.3; differences with that experiment include a new dataset and a new method of reporting the results. The following paragraphs describe the media used, review the



Figure 3.14: An example of a frame with open subtitles.

experiment & evaluation methods, and discuss the results.

The media used for the experiment consists of two datasets. Dataset 1 is identical to the dataset used in Sec. 3.6.3 (see Table 3.5). Dataset 2 includes 21 videos of an interview with a popular culture figure. All videos were downloaded from YouTube³. The videos differ in duration, resolution, video and audio quality. While none of the videos contained English subtitles, several videos contained *open subtitles* in Spanish, as shown in Fig. 3.14. Such subtitles are part of the moving picture — it is impossible to utilize them in Eq. 3.15 directly. Table 3.8 shows the various parameters of each video, including the container format, resolution, duration and number of shots. The method of pre-processing each video is described in Sec. 3.6.3.

The experiment first compared shots to each other exhaustively and manually, using a binary scale: similar or not similar. Shots that were similar to each other were grouped into clusters. There was a total of 98 clusters, with each cluster corresponding to a visually and semantically unique shot. Next, a graph was created, where each node in the graph corresponds to a shot, and two similar shots are connected by an edge. These clusters formed the ground truth data which was used for evaluating the precision and recall of the proposed method. Next, the experiment repeated the comparison using method proposed in Section 3.6.2, using several different values of α . To calculate visual similarity between two shots, the color histograms of the leading frames of the two shots were compared using

³www.youtube.com

Table 3.8: Video parameters for dataset 2.

Video	Fmt.	Resolution	Duration	Shots
V_1	MP4	636×358	01:00.52	15
V_2	MP4	596×336	08:33.68	96
V_3	MP4	640×360	08:23.84	93
V_4	MP4	640×358	15:09.17	101
V_5	MP4	1280×720	15:46.83	103
V_6	MP4	640×360	08:42.04	97
V_7	MP4	1280×720	06:26.99	8
V_8	MP4	1280×720	09:19.39	97
V_9	MP4	1280×720	08:59.71	99
V_{10}	MP4	596×336	08:33.68	96
V_{11}	MP4	596×336	08:33.68	96
V_{12}	MP4	596×336	08:33.68	96
V_{13}	MP4	596×336	08:33.68	96
V_{14}	MP4	596×336	08:33.68	96
V_{15}	MP4	596×336	08:33.61	97
V_{16}	MP4	596×336	08:33.68	96
V_{17}	MP4	1280×720	09:06.45	99
V_{18}	MP4	634×350	00:10.00	1
V_{19}	MP4	480×320	08:33.76	96
V_{20}	MP4	596×336	08:33.68	96
V_{21}	MP4	596×336	08:33.72	97

a conventional method [49].

To evaluate the effectiveness of the proposed method, the experiment measured the precision and recall within a query-response paradigm. The precision, recall and F1 score were calculated as Eqs. 3.12, 3.13 and 3.14. Within the query-response paradigm, the query was a single shot, selected arbitrarily, and the response was all shots that were judged similar by the proposed method. True positives were response shots that were similar to the query shot, according to the ground truth. False positives were response shots that were not similar to the query shot, according to the ground truth. False negatives were shots not in the response that were similar to the query shot, according to the ground truth. The above set-up was repeated, using each of the shots as a query.

Tables 3.9 and 3.10 show the results for datasets 1 and 2, respectively. These results, in particular those of Table 3.10 demonstrate the trade-off in utilizing both text and visual features for shot identification. If only visual features are used, then recall is high, but precision is low, since it is impossible to distinguish between shots that are visually similar but semantically different — this was the focus

Table 3.9: A summary of experiment results for dataset 1.

Method	$\bar{P}$	$\bar{R}$	$\bar{F}_1$
Text only ($\alpha = 0.0$)	1.00	0.85	0.89
Visual only ($\alpha = 1.0$)	0.39	1.00	0.45
Proposed ($\alpha = 0.5$)	0.98	0.96	0.96

Table 3.10: A summary of experiment results for dataset 2.

Method	$\bar{P}$	$\bar{R}$	$\bar{F}_1$
Text only ($\alpha = 0.0$)	0.86	0.28	0.30
Visual only ($\alpha = 1.0$)	0.20	0.80	0.18
Proposed ($\alpha = 0.5$)	0.88	0.31	0.36

of Sec. 3.6.3 and the motivation for utilizing ASR results in shot identification. If only text features are used, then the precision is high, but recall is low, since not all shots contain a sufficient amount of text to be correctly identified. For example, many shots in dataset 2 contained no subtitles at all. Furthermore, the accuracy of ASR varies significantly with the media. For example, dataset 1 contains a one-on-one interview with a US political figure, who speaks for several seconds at a time and is rarely interrupted while speaking. Dataset 2 contains a three-on-one interview with a British comedian, who speaks with an accent and is often interrupted by the hosts. By utilizing both visual and text features, the proposed method brings the benefit of a higher average F_1 score. However, the recall is still too low to be useful in practical shot identification.

3.7 Conclusion

This section described the process of shot identification and introduced relevant algorithms. The effectiveness of the individual algorithms and the shot identification process as a whole was confirmed via experiments on a real-world dataset. The integration of subtitles into shot comparison improved the precision and recall when compared to the spatial only approach (see Tables 3.9 and 3.10, it was still lower than the results obtained using the robust hash algorithm (see Table 3.3).

Chapter 4

The Video Authenticity Degree

4.1 Introduction

This chapter proposes the *video authenticity degree* and a method for its estimation. We begin the proposal with a formal definition in Sec. 4.1.1 and the development of a full-reference model for the authenticity degree in Sec. 4.2. Through experiments, we show that it is possible to determine the authenticity of a video in a full-reference scenario, where the parent video is available. Next, Section 4.3 describes the extension of the full-reference model to a no-reference scenario, as described in another of our papers [3]. Through experiments, we verify that the model applies to a no-reference scenario, where the parent video is unavailable. Finally, Section 4.4 proposes our final method for estimating the authenticity degree of a video in a no-reference scenario. Section 4.5 empirically verifies the effectiveness of the method proposed in Sec. 4.4. The experiments demonstrate the effectiveness of the method on real and artificial data for a wide range of videos. Finally, Section 4.6 concludes the chapter.

4.1.1 Definition

The authenticity degree of an edited video $\mathbf{V}_j$ is defined as the proportion of remaining information from its parent video $\mathbf{V}_0$. It is a positive real number between 0 and 1, with higher values corresponding to high authenticity¹.

¹Our earlier models did not normalize the authenticity degree to the range [0, 1].

4.2 Developing the Full-reference Model

4.2.1 Initial Model

In the proposed full-reference model, the parent video V_0 is represented as a set of shot identifiers S_i ($i = 1, \dots, N$, where N is the number of shots in V_0). For the sake of brevity, the shot with the identifier equal to S_i is referred to as “the shot S_i ” in the remainder of this section. Let the authenticity degree of V_0 be equal to N . V_0 can be edited by several operations, including: (1) shot removal and (2) video recompression, to yield a near-duplicate video V_j ($j = 1, \dots, M$, where M is the number of all near-duplicates of V_0). Since editing a video decreases its similarity with the parent, the authenticity degree of V_j must be lower than that of V_0 . To model this, each editing operation is associated with a positive penalty, as discussed below.

Each shot in the video conveys an amount of information. If a shot is removed to produce a near-duplicate video V_j , then some information is lost, and the similarity between V_j and V_0 decreases. According to the definition of Section 4.1.1, the penalty for removing a shot should thus be proportional to the amount of information lost. Accurately quantifying this amount requires semantic understanding of the entire video. Since this is difficult, an alternative approach is to assume that all the N shots convey identical amounts of information. Based on this assumption, the shot removal penalty p_1 is given by:

$$p_1(V_0, V_j, S_i) = \begin{cases} 1 & \text{if } S_i \in V_0 \wedge S_i \notin V_j \\ 0 & \text{otherwise} \end{cases}. \quad (4.1)$$

Recompressing a video can cause a loss of detail due to quantization, which is another form of information loss. The amount of information loss is proportional to the loss of detail. An objective way to approximate the loss of detail is to examine the decrease in video quality. In a full-reference scenario, one of the simplest methods to estimate the decrease in video quality is by calculating the mean PSNR (peak signal-to-noise ratio) [24] between the compressed video and its parent. A high PSNR indicates high quality; a low PSNR indicates low quality. While PSNR is known to have limitations when comparing videos with different visual content [25], it is sufficient for our purposes

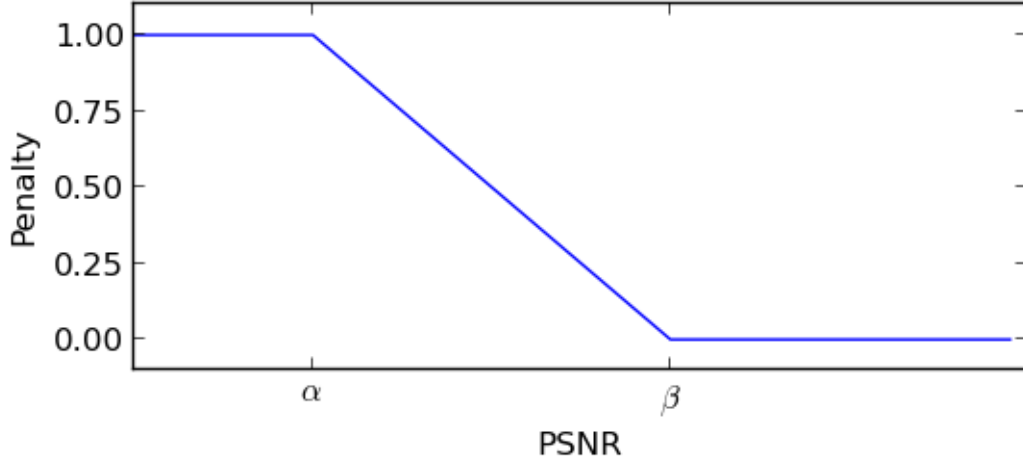


Figure 4.1: Penalty p_2 as a function of PSNR.

since the videos are nearly identical. Thus, the recompression penalty p_2 is given by:

$$p_2(V_0, V_j, S_i) = \begin{cases} 0 & \text{if } S_i \notin V_j \\ 1 & \text{if } Q_{ij} < \alpha \\ 0 & \text{if } Q_{ij} > \beta \\ (\beta - Q_{ij})/(\beta - \alpha) & \text{otherwise} \end{cases}, \quad (4.2)$$

where α is a low PSNR at which the shot S_i can be considered as missing; β is a high PSNR at which no subjective quality loss is perceived; and $V_0[i]$ and $V_j[i]$ refer to the image samples of shot S_i in video V_0 and V_j , respectively, and

$$Q_{ij} = \text{PSNR}(V_0[S_i], V_j[S_i]). \quad (4.3)$$

A significant loss in quality thus corresponds to a higher penalty, with the maximum penalty being 1. This penalty function is visualized in Fig. 4.1.

Finally, $A(V_j)$, the authenticity degree of V_j , is calculated as follows:

$$A(V_j) = N - \sum_{i=1}^N p_1(V_0, V_j, S_i) + p_2(V_0, V_j, S_i). \quad (4.4)$$

Note that if V_0 and V_j are identical, then $A(V_j) = N$, which is a maximum and is consistent with the definition of the authenticity degree from Sec. 4.1.1.

4.2.2 Detecting Scaling and Cropping

Scaling and cropping are simple and common video editing operations. Scaling is commonly performed to reduce the bitrate for viewing on a device with a smaller display, and to increase the appar-

ent resolution (supersampling) in order to obtain a better search result ranking. Cropping is performed to purposefully remove unwanted areas of the frame. Since both operations are spatial, they can be modeled using images, and the expansion to image sequences (i.e. video) is trivial. For the sake of simplicity, this section deals with images only.

Scaling can be modeled by an affine transformation [57]:

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} S_x & 0 & -C_x \\ 0 & S_y & -C_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}, \quad (4.5)$$

where (x, y) and (x', y') are pixel co-ordinates in the original and scaled images, respectively; S_x and S_y are the scaling factors in the horizontal and vertical directions, respectively; C_x and C_y are both 0. Since the resulting (x', y') may be non-integer and sparse, interpolation needs to be performed after the above affine transform is applied to every pixel of the original image. This interpolation causes a loss of information. The amount of information lost depends on the scaling factors and the type of interpolation used (e.g. nearest-neighbor, bilinear, bicubic).

Cropping an image from the origin can be modeled by a submatrix operation:

$$I' = I[\{1, 2, \dots, Y\}, \{1, 2, \dots, X\}], \quad (4.6)$$

where I' and I are the cropped and original image, respectively; Y and X are the number of rows and columns, respectively, in I' . Cropping from an arbitrary location (C_x, C_y) can be modeled through a shift by applying the affine transform in Eq. 4.5 (with both S_x and S_y set to 1) prior to applying Eq. (4.6). Since this operation involves removing rows and columns from I , it causes a loss of information. The amount of information lost is proportional to the number of rows and columns removed.

Since both scaling and cropping cause a loss of information, their presence must incur a penalty. If I is an original image and I' is the scaled/cropped image, then the loss of information due to scaling can be objectively measured by scaling I' back to the original resolution, cropping I if required, and measuring the PSNR [24]:

$$Q = \text{PSNR}(J, J'), \quad (4.7)$$

where J' is I' restored to the original resolution and J is a cropped version of I . Cropping I is necessary since PSNR is very sensitive to image alignment. The penalty for scaling is thus included

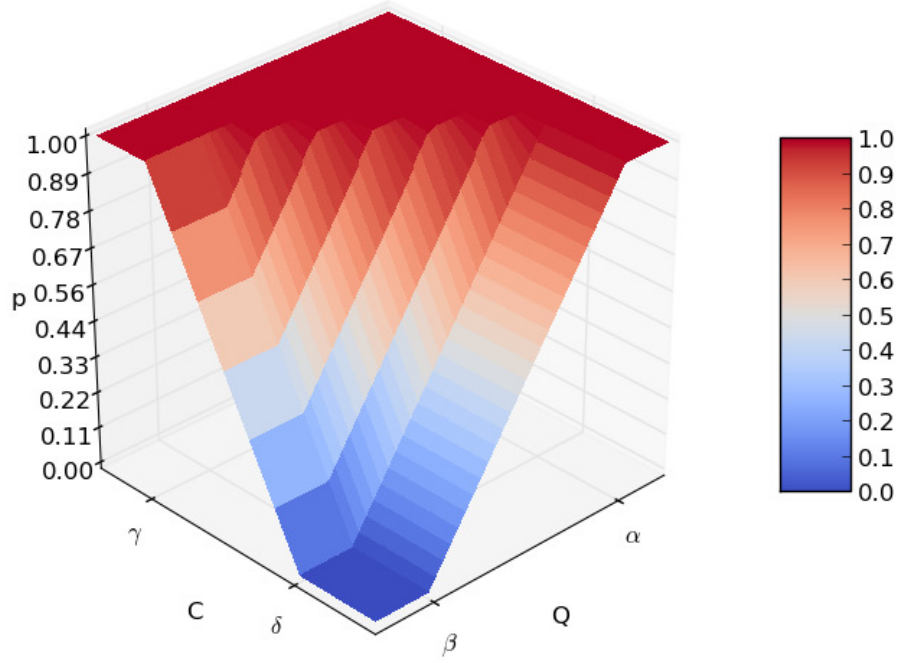


Figure 4.2: Penalty p as a function of Q and C .

in the calculation of Eq. (4.7), together with the penalty for recompression. The loss of information due to cropping is proportional to the frame area that was removed (or inversely proportional to the area remaining).

To integrate the two penalties discussed above with the model introduced in Sec. 4.2.1, I propose to expand the previously-proposed one-dimensional linear shot-wise penalty function, shown in Figure 4.1, to two dimensions:

$$p(I, I') = \max \left\{ 0, \min \left(1, \frac{Q - \beta}{\alpha - \beta} + \frac{C - \delta}{\gamma - \delta} \right) \right\}, \quad (4.8)$$

where C is the proportion of the remaining frame area of I' ; (α, β) and (γ, δ) are the limits for the PSNR and proportion of the remaining frame, respectively. Eq. (4.8) outputs a real number between 0 and 1, with higher values returned for low PSNR or remaining frame proportion. The proposed penalty function is visualized in Fig. 4.2.

Implementing Eqs. (4.7) and (4.8) is trivial if the coefficients used for the original affine transform in Eq. (4.5) are known. If they are unknown, they can be estimated directly from I and I' in a full-reference scenario. Specifically, the proposed method detects keypoints in each frame and calculates their descriptors [58]. It then attempts to match the descriptors in one frame to the descriptors in the

other frame using the RANSAC algorithm [59]. The relative positions of the descriptors in each frame allow an affine transform to be calculated. The remaining crop area is calculated directly from the coefficients of the affine transform. Furthermore, the coefficients of this affine transform allow the remaining crop area to be calculated as:

$$C_i^j = \frac{w'_j \times h'_j}{w_0 \times h_0}, \quad (4.9)$$

where (w'_j, h'_j) correspond to the dimensions of the keyframe from the near duplicate, after applying the affine transform; and (w_0, h_0) correspond to the dimensions of the keyframe from the parent video.

4.2.3 Experiments

4.2.3.1 Testing the Initial Model

This section introduces an experiment that demonstrates the validity of model proposed in Sec. 4.2.1. It consists of three parts: (1) selecting the model parameters; (2) creation of artificial test media; and (3) discussion of the results.

The values for the model parameters α and β were determined empirically. First, the *shields* test video (848×480 pixels, 252 frames) was compressed at various constant bitrates to produce near-duplicate copies. A single observer then evaluated each of the near-duplicates on a scale of 1–5, using the Absolute Category Rating (ACR) system introduced in ITU-T Recommendation P.910. The results of the evaluation, shown in Fig. 4.4, demonstrate that the observer rated all videos with a PSNR lower than approximately 23dB as 1 (poor), and all videos with a PSNR higher than approximately 36dB as 5 (excellent). The parameters α and β are thus set to 23 and 36, respectively.

To create the test set, first the single-shot test videos *mobcal*, *parkrun*, *shields* and *stockholm* (848×480 pixels, 252 frames each) were joined, forming V_0 . These shots are manually assigned unique signatures S_1, S_2, S_3 and S_4 , respectively. Next, 9 near-duplicate videos V_j ($j = 1, \dots, 9$) were created by repetitively performing the operations described in Sec. 4.2.1. For video compression, constant-bitrate H.264/AVC compression was used, with two bitrate settings: high (512Kb/s) and low (256Kb/s).

The results of the experiment are shown in Table 4.1. Each row of the table corresponds to a video in the test set, with $j = 0$ corresponding to the parent video. The columns labeled S_i ($i = 1, 2, 3, 4$)

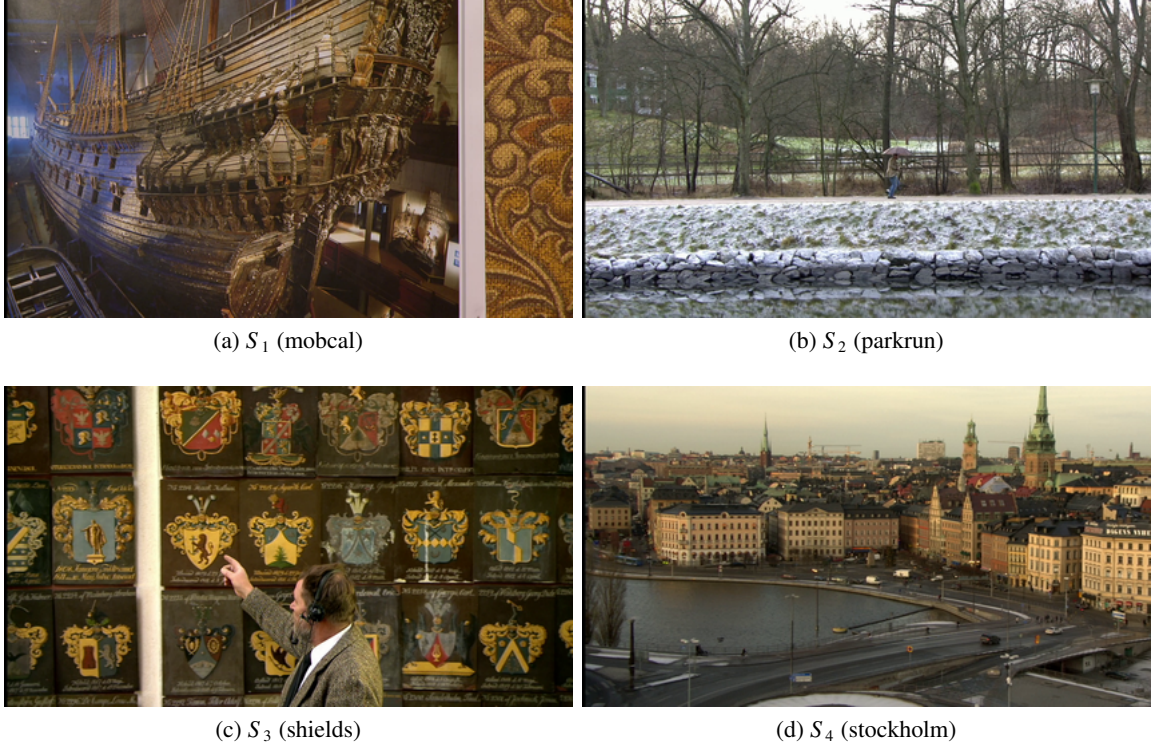


Figure 4.3: Single-shot videos used in the experiment.

show the PSNR of the shot S_i in V_j with respect to V_0 . A “–” indicates removed shots. The last column shows the authenticity degree for V_j , calculated using Eq. (4.4). These results show that the proposed model is consistent with the definition of the authenticity degree. Specifically, the authenticity degree is at a maximum for the parent video, and editing operations such as shot removal and recompression reduce the authenticity degree of the edited video. Aside from V_0 , V_3 has the highest authenticity degree, which is consistent with the fact that it retains all the shots of V_0 with only moderate video quality loss.

4.2.3.2 Testing the Extended Model on Images

This section describes an experiment performed to verify the effectiveness of the model proposed in Sec.4.2.2. A test data set was created by editing a frame (hereafter, original image) of the *mobcal* test video using different scaling and crop parameters to Eq. (4.5), creating 42 edited images. Then, the penalty of Eq. 4.8 was calculated for each edited image using two different methods. Method 1 used the coefficients of Eq. (4.5) directly. Method 2 estimated the coefficients by comparing feature points

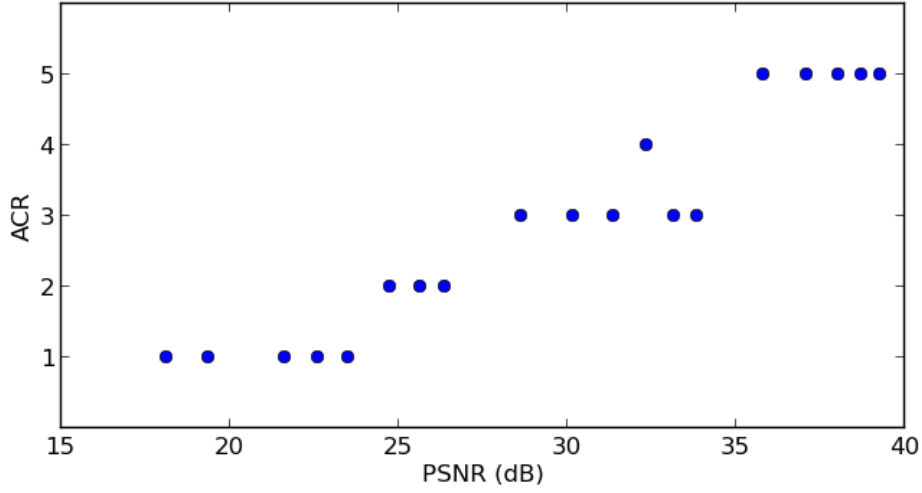


Figure 4.4: Absolute Category Rating (ACR) results.

Table 4.1: Shot-wise PSNR and the authenticity degree.

j	S_1	S_2	S_3	S_4	$A(V_j)$
0	∞	∞	∞	∞	4.00
1	–	∞	–	∞	2.00
2	∞	–	∞	∞	3.00
3	35.57	27.19	32.85	36.33	3.05
4	–	27.19	–	36.33	1.32
5	–	26.40	–	36.87	1.26
6	35.57	–	32.85	36.33	2.72
7	35.57	–	34.48	36.81	2.85
8	–	23.97	–	33.26	0.86
9	–	26.40	–	–	0.26

in the original image and edited image. The empirical constants $\alpha, \beta, \gamma, \delta$ were set to 23dB, 36dB, 0.5 and 0.8, respectively. The aspect ratio was maintained for all combinations (i.e. $S_x = S_y$).

The results for a selected subset of the edited images are shown in Table 4.2. The columns p_1 and p_2 show the penalties calculated using Methods 1 and 2, respectively. Higher penalties are assigned to images with a low magnification ratio (zoomed out images) and cropped images. Furthermore, the penalties in the p_2 column appear to be significantly higher than in the p_1 column. The cause becomes obvious through comparing the PSNRs obtained using Methods 1 and 2, shown in columns Q_1 and Q_2 respectively. The higher penalties are caused by the fact that the PSNRs for Method 2 are significantly lower in some cases, marked with bold. This can be explained by a lack of precision in

Table 4.2: Results of the experiment.

S	A_1	A_2	Q_1	Q_2	p_1	p_2
0.7	0.7	0.70	29.80	28.40	0.81	0.92
0.7	0.9	0.90	31.57	25.67	0.01	0.47
0.8	0.6	0.60	32.72	28.10	0.92	1.00
0.8	0.7	0.70	33.20	29.62	0.55	0.83
0.8	0.8	0.80	31.27	30.00	0.36	0.46
0.8	0.9	0.90	33.02	30.21	0.00	0.11
0.9	0.6	0.60	31.84	31.21	0.99	1.00
0.9	0.7	0.70	31.32	30.67	0.69	0.75
0.9	0.8	0.80	35.23	31.63	0.06	0.33
1.1	0.5	0.50	37.99	31.75	0.85	1.00
1.1	0.6	0.60	36.94	29.40	0.59	1.00
1.1	0.7	0.70	38.91	32.15	0.11	0.62
1.1	0.8	0.80	35.28	28.29	0.06	0.59
1.1	1.0	1.00	40.64	26.70	0.00	0.05
1.2	0.6	0.60	41.70	31.42	0.23	1.00
1.2	0.7	0.70	34.46	29.65	0.45	0.82
1.2	0.8	0.80	39.17	29.57	0.00	0.49
1.2	0.9	0.90	42.05	29.92	0.00	0.13
1.3	0.5	0.50	40.42	23.07	0.66	1.00
1.3	0.8	0.78	34.12	17.56	0.14	1.00
1.3	0.9	0.90	42.40	30.02	0.00	0.13
1.3	1.0	1.00	45.14	27.25	0.00	0.01

estimating the affine transform coefficients. PSNR is well-known to be very sensitive to differences in image alignment. Using an alternative full-reference algorithm such as SSIM [29] may address this discrepancy.

Finally, columns A_1 and A_2 show the proportion of the frame area remaining after cropping, calculated using Methods 1 and 2, respectively. It is clear that the cropping area is being estimated correctly. This will prove useful when performing scaling and cropping detection in a no-reference scenario.

4.2.3.3 Testing the Extended Model on Videos

This section describes an experiment that was performed to evaluate the effectiveness of the proposed method. In this experiment, human viewers were shown an parent video and near duplicates of the parent. The viewers were asked to evaluate the similarity between the parent and each near duplicate. The goal of the experiment was to compare these subjective evaluations to the output of the proposed

Table 4.3: Test set and test results for the Experiment of Sec. 4.2.3.3.

Video	Editing	Raw	Norm.	Prop.
V_0	Parent video	–	–	–
V_1	3 shots removed	3.93	0.41	0.82
V_2	7 shots removed	3.14	−0.47	0.50
V_3	3/4 downsampling	4.93	1.41	0.90
V_4	1/2 downsampling	3.93	0.48	0.85
V_5	90% crop	3.93	0.47	0.32
V_6	80% crop	2.57	−1.05	0.00
V_7	Medium compression	3.14	−0.29	0.86
V_8	Strong compression	2.42	−0.97	0.71

Table 4.4: Mean and standard deviations for each viewer.

Viewer	μ	σ
1	4.13	0.52
2	3.50	0.93
3	3.75	0.71
4	3.00	1.07
5	3.50	1.07
6	3.25	1.58
7	3.37	1.19

method. The details of the experiment, specifically, the data set, subjective experiment method, and results, are described below.

The data set was created from a single YouTube news video (2min 50s, 1280×720 pixels). First, shot boundaries were detected manually. There were a total of 61 shots, including transitions such as cut, fade and dissolve [42]. Next, parts of the video corresponding to fade and dissolve transitions were discarded, since they complicate shot removal. Furthermore, in order to reduce the burden on the experiment subjects, the length of the video was reduced to 1min 20s by removing a small number of shots at random. In the context of the experiment, this resulting video was the “parent video”. This parent video had a total of 16 shots. Finally, near duplicates of the parent video were created by editing it, specifically: removing shots at random, spatial downsampling (scaling), cropping, and compression. Table 4.3 shows the editing operations that were performed to produce each near duplicate in the “Editing” column.

During the experiment, human viewers were asked to evaluate the near duplicates: specifically, the number of removed shots, visual quality, amount of cropping, and overall similarity with the parent,

Table 4.5: Rank-order of each near duplicate.

	V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8
Subj.	4th	6th	1st	2nd	3rd	8th	5th	7th
Prop.	4th	6th	1st	3rd	7th	8th	2nd	5th

Table 4.6: Correlation with subjective results.

Method	r	ρ
Subjective	1.00	1.00
Objective	0.52	0.66

using a scale of 1 to 5. The viewers had access to the parent video. Furthermore, to simplify the task of detecting shot removal, the viewers were also provided with a list of shots for each video.

Table 4.4 shows the mean and standard deviation for each viewer. From this table, we can see that each viewer interpreted the concept of “similarity with the parent” slightly differently, which is expected in a subjective experiment. To allow the scores from each viewer to be compared, they need to be normalized in order to remove this subjective bias. The results are shown in Table 4.3, where the “Raw” column shows the average score for each video, and the “Norm.” column shows the normalized score, which is calculated by accounting for the mean and standard deviation for each individual viewer (from Table 4.4). The “Prop.” column shows the authenticity degree that was calculated by the proposed method. Table 4.5 shows the near duplicates in order of decreasing normalized subjective scores and the proposed method outputs, respectively. From these results, we can see that for the subjective and objective scores correspond for the top 4 videos. For the remaining 4 videos, there appears to be no connection between the subjective and objective scores. To quantify this relationship, Table 4.6 shows the correlation and rank-order correlation between the normalized subjective scores and the proposed method outputs, as was done in the previous seminars. The correlation is lower than what was reported in previous experiments [4], before scaling and cropping detection were introduced.

The decrease in correlation could be caused by an excessive penalty on small amounts of cropping in Eq. (4.21). This hypothesis is supported by Table 4.5, where the rank-order of V_5 , a slightly cropped video, is significantly lower in the subjective case. On the other hand, the rank-orders for V_6 , a moderately cropped video, correspond to each other. This result indicates that viewers tolerated a slight amount of cropping, but noticed when cropping became moderate. On the other hand, the

different rank-orders of V_7 indicate that the penalty on compression was insufficient. Capturing this behavior may require subjective tuning of the parameters $\alpha, \beta, \gamma, \delta$ and possibly moving away from the linear combination model, for example, to a sigmoid model.

4.2.4 Conclusion

This section introduced several models for determining the authenticity degree of an edited video in a full-reference scenario. While such models cannot be applied to most real-world scenarios, they enable greater understanding of the authenticity degree and the underlying algorithms. Section 4.3 further expands these models to no-reference scenario, thus enabling their practical application.

4.3 Developing the No-reference Model

4.3.1 Derivation from the Full-reference Model

Since V_0 is not available in practice, $A(V_j)$ of Eq. (4.4) cannot be directly calculated, and must be estimated instead.

First, in order to realize Eq. (4.4), the no-reference model utilizes the available Web video context, which includes the number of views for each video, to estimate the structure of the parent video by focusing on the most-viewed shots. The set of most-viewed shots $\widehat{V}_0$ can be calculated as follows:

$$\widehat{V}_0 = \left\{ S_i \mid \left\lceil \frac{\sum_{V_j, S_i \in V_j} W(V_j)}{W_{total}} \right\rceil \geq \beta \right\}, \quad (4.10)$$

where β is a threshold for determining the minimum number of views for a shot, $W(V_j)$ is the number of times video V_j was viewed, and W_{total} is the total number of views for all the videos in the near-duplicate collection.

Next, the method of estimating the shot-wise penalty p is described. First, N is estimated as $\widehat{N} = |\widehat{V}_0|$, where $|\cdot|$ denotes the number of shots in a video. Next, an estimate of the penalty function p that is based on the above assumptions is described. The relative degradation strength Q_{ij} ($i = 1, \dots, M$, $j = 1, \dots, N$) is defined as:

$$Q_{ij} = \frac{Q_{NR}(V_j[S_i]) - \mu_i}{\sigma_i}, \quad (4.11)$$

where μ_i and σ_i are normalization constants for the shot S_i , representing the mean and standard deviation, respectively; $V_j[S_i]$ refers to the image samples of shot S_i in video V_j ; and Q_{NR} is a

no-reference quality assessment algorithm. Such algorithms estimate the quality by measuring the strength of degradations such as wide edges [34] and blocking [32]. The result is a positive real number, which is lower for higher quality videos. The penalty for recompression can then be defined as:

$$\hat{q}(V_j, S_i) = \begin{cases} 0 & \text{if } Q_{ij} < -\gamma \\ 1 & \text{if } Q_{ij} > \gamma \\ \frac{Q_{ij} + \gamma}{2\gamma} & \text{otherwise} \end{cases}, \quad (4.12)$$

where γ is a constant that determines the scope of acceptable quality loss. A significant loss in quality of a most-viewed shot thus corresponds to a higher penalty $\hat{q}$, with the maximum penalty being 1. Next, $\hat{p}$, the estimate of the shot-wise penalty function p , can thus be calculated as:

$$\hat{p}(V_j, S_i) = \begin{cases} 1 & \text{if } S_i \notin V_j \\ \hat{q}(V_j, S_i) & \text{otherwise} \end{cases}. \quad (4.13)$$

Equation (4.13) thus penalizes missing most-viewed shots with the maximum penalty equal to 1. If the shot is present, the penalty is a function of the relative degradation strength Q_{ij} . The penalty function $\hat{p}$ does not require access to the parent video V_0 , since the result of Eq. (4.10) and no-reference quality assessment algorithms are utilized instead. Finally, $\widehat{A}$, the estimate of the authenticity degree A , can be calculated as:

$$\widehat{A}(V_j) = \widehat{N} - \sum_{i=1}^{\widehat{N}} \hat{p}(V_j, S_i). \quad (4.14)$$

4.3.2 Considering Relative Shot Importance

The authenticity degree model proposed in Sec. 4.3.1 suffers from several weaknesses. For example, it relies on several assumptions that are difficult to justify. One of the main assumptions is that all shots contain the same amount of information. In general, this is not always true, since some shots obviously contain more information than others.

This section proposes an improved authenticity degree model that does not require the above assumption. The proposed method inherits the same strategy, but adds a weighting coefficient to each shot identifier:

$$A(\mathbf{V}_j) = 1 - \frac{\sum_{i \in \hat{I}_0} w_i p_i(\mathbf{V}_j)}{\sum_{i \in \hat{I}_0} w_i}, \quad (4.15)$$

where w_i is the weighting coefficient of the shot with shot identifier equal to i . The value of w_i should be proportional to the “importance” of the shot: since some shots are more important than others, the loss of their information should have a greater impact on the authenticity degree. The proposed method calculates w_i as the duration of the shot, in seconds. In other words, the proposed method assumes that longer shots contain more information. This assumption is consistent with existing work in video summarization [60].

4.3.3 Experiments

4.3.3.1 Evaluating the No-Reference Model

This section describes an experiment that illustrates the effectiveness of the proposed method. The aim of the experiment is to look for correlations between the full-reference model, no-reference model and subjective ratings from human observers. First, the data used in the experiments are introduced. Next, the method of performing the experiment is described. Finally, the results are presented and discussed.

Two data sets were used in this experiment. First, the *artificial* set consists of the same 10 videos described in Sec. 4.2.3.1. Second, the *lagos* test set consists of 9 near-duplicate videos that were retrieved from YouTube and includes news coverage related to a recent world news event. The videos are compressed with H.264/AVC, but have different resolutions and bitrates. The audio signal in each of the video was muted.

Two methods were used to detect the strength of correlation between the results and ground truth. The first method is the correlation of the raw scores with the ground truth scores², which is calculated as:

$$r(\vec{X}, \vec{Y}) = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}}, \quad (4.16)$$

where n is the number of videos in a test set; $\vec{X} = [X_1, \dots, X_n]$ represents the ground truth; $\vec{Y} = [Y_1, \dots, Y_n]$ represents the results of the algorithm being evaluated; and $\bar{X}$ and $\bar{Y}$ are the means of $\vec{X}$ and $\vec{Y}$, respectively.

²Also known as Pearson’s r .

Table 4.7: Results for each video in the *artificial* test set.

j	A	$\widehat{A}$	EW	GBIM
0	4.00	3.28	2.31	1.18
1	2.00	1.67	2.28	1.18
2	3.00	2.43	2.33	1.18
3	3.05	1.43	2.52	1.20
4	1.32	0.87	2.50	1.20
5	1.26	0.76	2.54	1.20
6	2.72	0.95	2.53	1.19
7	2.85	1.08	2.51	1.19
8	0.86	0.16	2.70	1.21
9	0.26	0.36	2.58	1.21

Table 4.8: Quantitative experiment results.

(a) <i>Artificial</i>			(b) <i>Lagos</i>		
Method	$ r $	$ \rho $	Method	$ r $	$ \rho $
FR	1.00	1.00	Subjective	1.00	1.00
Subjective	0.82	0.76	Proposed	0.55	0.30
Proposed	0.83	0.89	EW	0.40	0.25
EW	0.61	0.60	GBIM	0.50	0.07
GBIM	0.80	0.74	Views	0.03	0.12
—	—	—	Delay	0.40	0.15

The second method is the the correlation of the ranks of the raw scores with the ranks of the ground truth scores³. First, the ranks of each element in $\vec{X}$ and $\vec{Y}$ are calculated to yield $\vec{x}$ and $\vec{y}$. The final output of the second method is:

$$\rho(\vec{X}, \vec{Y}) = r(\vec{x}, \vec{y}). \quad (4.17)$$

For both r and ρ , a value of ± 1 indicates strong correlation with the ground truth, and is therefore a desirable result. A value of zero indicates poor correlation with the ground truth.

To perform the experiment, the proposed no-reference model was first applied to the data to obtain an estimate of the authenticity degree for each video of the *artificial* set. The results are shown in Table 4.7. The columns A and $\widehat{A}$ show the actual and estimated authenticity degrees, respectively. The columns EW and GBIM show the result of directly applying the no-reference quality assessment algorithms of edge width [34] and blocking strength [32], respectively, to each video. From these

³Also known as Spearman's ρ .

Table 4.9: Summary of real data sets.

Name	Description	Videos
<i>Kerry</i>	Interview with J. Kerry	5
<i>Bolt</i>	U. Bolt false start	68
<i>Russell</i>	Interview with R. Brand	18
<i>Klaus</i>	Chilean pen incident	76
<i>Lagos</i>	Dana Air 992 crash	8
Total		175

results, it can be seen that the proposed model is consistent with the definition of the authenticity degree. Specifically, the authenticity degree is at a maximum for the parent video, and editing operations such as shot removal and recompression reduce the authenticity degree of the edited video. The measured correlation with the ground truth (the actual authenticity degree obtained using the full-reference method) is shown in Table 4.8(a). These results show that even without access to the parent video, the proposed no-reference model produces results that are comparable to the full-reference model.

Next, Table 4.8(b) shows the correlation results for the *lagos* test set, with the subjective ratings being used as the ground truth. Views and delay correspond to sorting the results using the number of views and the upload timestamp of the video, respectively, and are obtained directly from the Web video context. While the correlation between the ground truth and the results obtained using the proposed method is lower than that of the *artificial* test set, the proposed method still performs better than the conventional methods.

The significantly lower result can be explained by several factors. As described in Sec. 4.2.3.1, the videos in the *artificial* test set were created by concatenating unrelated shots. In contrast, the videos in the *lagos* test set consisted of many semantically related shots that fit into a news story. In the latter case, the shots no longer carry the same amounts of information: some shots are more important than others. Since the proposed method treats all most-viewed shots as equally important when determining the shot removal penalty, the penalty assigned can be different to that of a human participant. Considering the shot structure in the proposed model may lead to stronger correlation with subjective results.

Table 4.10: Results (r value) for the experiment.

	Comparative	Proposed
<i>Bolt</i>	0.47	0.50
<i>Kerry</i>	0.75	0.69
<i>Klaus</i>	0.73	0.75
<i>Lagos</i>	0.35	0.51
<i>Russell</i>	0.40	0.72

Table 4.11: Results (ρ value) for the experiment.

	Comparative	Proposed
<i>Bolt</i>	0.54	0.49
<i>Kerry</i>	0.90	0.90
<i>Klaus</i>	0.77	0.81
<i>Lagos</i>	0.29	0.52
<i>Russell</i>	0.43	0.48

4.3.3.2 Evaluating Relative Shot Importance

This experiment compares the performance of comparative and proposed methods using subjective evaluations of several datasets. The datasets are shown in Table 4.9. The subjective evaluations of each video were obtained from a total of 20 experiment subjects. The subjects were asked to rate each video on 3 points: (a) relative visual quality; (b) number of deleted shots; and (c) estimated proportion of information remaining from the parent. The subjects were asked to rate (a) and (b) independently, and to rate (c) based on their answers to (a) and (b). All answers were given using a scale of 1 to 5, since this scale is commonly used for visual quality assessment⁴. For each video, the mean answer to (c) across all subjects was recorded as that particular video's estimated proportion of information remaining from the parent.

Next, in order to evaluate the two methods, each method was used to estimate the authenticity degree of each video. Both methods were configured using the same parameters, manually-obtained shot boundaries and shot identifiers. The sample correlation coefficient and rank-order correlation coefficient between the obtained authenticity degrees and the subjective estimates were calculated as Eqs. (4.16) and (4.17), respectively.

The results are shown in Tables 4.10 and 4.11. The better value for each dataset is shown in

⁴See ITU-T Recommendation BT.500: "Methodology for the subjective assessment of the quality of television pictures"

bold. In the majority of cases, the proposed method outperforms the comparative method. The most significant improvement was in the *Lagos* dataset.

4.3.4 Conclusion

This section extended the full-reference models introduced in Sec. 4.2 to a no-reference scenario. Experimental results demonstrated that it is possible to accurately estimate the authenticity of an edited video in cases where the parent video is unavailable, provided there are other edited videos available. Furthermore, including the shot duration in the model significantly improved the effectiveness of the model.

A significant limitation of the model is that it relies on many assumptions in order to utilize the Web context, more specifically, the view count. Furthermore, relying on the Web context restricts the applications of the model to videos that have been uploaded to the Web, which significantly complicates the collection of data. Section 4.4 solves these problems by removing the dependency on the Web context from the model.

4.4 The Proposed No-reference Method

4.4.1 Calculation

Figure 4.5 illustrates the strategy of calculating the authenticity degree. Each shot in the parent video $\mathbf{V}_0$ conveys an amount of information. Editing a video causes a loss of some of this information: for example, recompressing the video causes a loss of detail; removing a shot causes a loss of all the information contained in that shot. The edited video $\mathbf{V}_j$ thus contains less of the parent video’s information than the parent video itself. The proposed method estimates the amount of information lost and penalizes each edited video accordingly. In order to do this, the proposed method attempts to detect the editing operations that were performed. Each detected editing operation contributes to a penalty, which is calculated individually for each shot. If $\mathbf{V}_j$ and $\mathbf{V}_0$ are identical, then $A(\mathbf{V}_j)$ is at its maximum value of 1, since the penalties become zero. Finally, the penalties are aggregated to calculate the authenticity degree for the entire video.

The input to the proposed method is a set of videos $\mathbb{V} = \{\mathbf{V}_1, \dots, \mathbf{V}_M\}$ that are all edited versions of the same parent $\mathbf{V}_0$, which is unknown. In practice, it is possible to obtain $\mathbb{V}$ by searching for a set

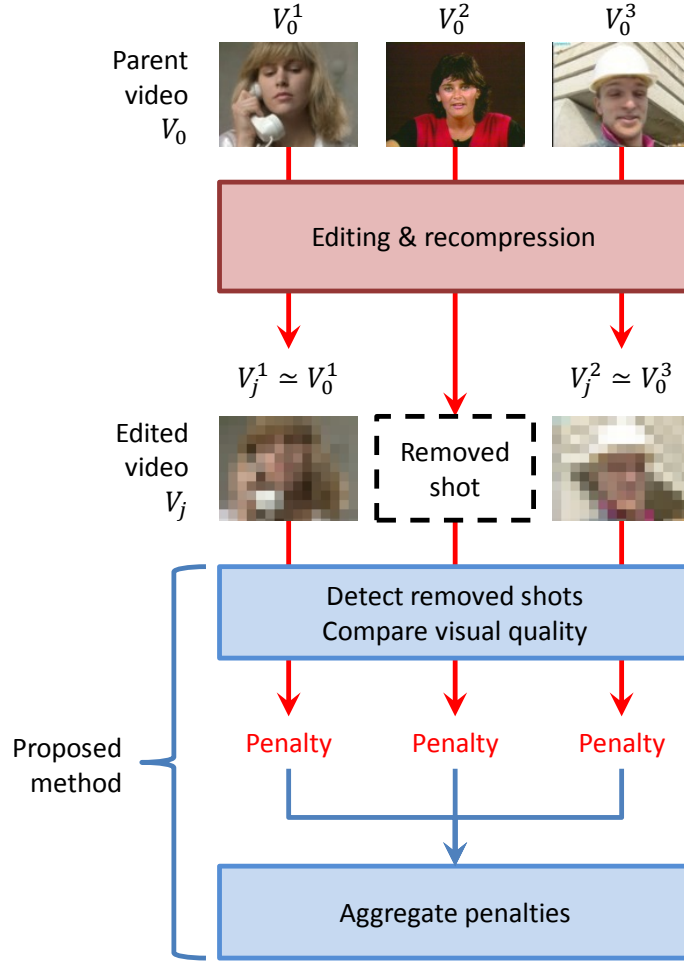


Figure 4.5: The strategy of calculating the authenticity degree.

of keywords describing $\mathbf{V}_0$ on a video sharing portal, for example “Usain Bolt false start” or “Czech president steals pen”. For each edited video $\mathbf{V}_j$ ($j = 1, \dots, M$), the proposed method detects editing operations, determines the corresponding penalties and calculates the authenticity degree as:

$$A(\mathbf{V}_j) = 1 - \frac{\sum_{i \in \hat{I}_0} w_i p_i(\mathbf{V}_j)}{\sum_{i \in \hat{I}_0} w_i}, \quad (4.18)$$

where $\hat{I}_0$ is an estimate of the set of shot identifiers present in the parent video, described in detail in Sec. 4.4.2; $p_i(\cdot)$ is an estimate of the penalty function for the shot whose identifier is equal to i ; and w_i is the weighting coefficient of the shot with shot identifier equal to i . The value of w_i should be proportional to the “importance” of the shot: since some shots are more important than others, the loss of their information should have a greater impact on the authenticity degree. The proposed method calculates w_i as the duration of the shot, in seconds. In other words, the proposed method assumes

that longer shots contain more information. This assumption is consistent with existing work in video summarization [60].

Specifically, each $p_i(\cdot)$ is calculated as:

$$p_i(\mathbf{V}_j) = \begin{cases} 1 & \text{if } i \notin I_j \\ \gamma g_i(\mathbf{V}_j) & \text{otherwise} \end{cases}, \quad (4.19)$$

where the notation $i \notin I_j$ means “ $\mathbf{V}_j$ doesn’t contain a shot with an identifier equal to i ”; $g_i(\mathbf{V}_j)$ is the global frame modification penalty, described in detail in Sec. 4.4.3, for the shot of $\mathbf{V}_j$ whose identifier is equal to i . The parameter γ determines the trade-off between information loss through global frame modifications and information loss through shot removal. Since, in practice, global frame modifications can never remove all the information in a shot entirely, γ should be a positive value less than 1. Equation (4.19) thus assigns the maximum penalty for shot removal, and a fraction of the maximum penalty for global frame modifications.

There may be significant differences between the edited videos: for example some shots may be added/missing (e.g. Fig. 4.5). As demonstrated in Sec. 2.8, this prevents the effective application of conventional NRQA algorithms, since the output of such algorithms will be influenced by the added/missing shots. Before NRQA algorithms can be applied effectively, this influence must be reduced. The proposed method achieves this in two ways. First, when calculating the global frame modification penalty, the proposed method works on a shot ID basis, and only compares the raw visual quality of shots that have the same shot identifier (common content). Second, the proposed method prevents missing/added shots from affecting the outcome of NRQA algorithms. More specifically, if a shot is added (not part of the estimated parent $\hat{I}_0$), then it is ignored completely by Eq. (4.18). If a shot is missing (part of the estimated parent), then a shot removal penalty is assigned by Eq. (4.19).

4.4.2 Parent Video Estimation

In order to detect shot removal, the proposed method first defines B_i as the number of edited videos that contain a shot with an identifier equal to i . Then, $\hat{I}_0$, the estimate of the set of shot identifiers present in $\mathbf{V}_0$, is calculated as:

$$\hat{I}_0 = \bigcup_{j=1}^M \{i | i \in I_j, \frac{B_i}{M} > \delta\}, \quad (4.20)$$

where δ is an empirically determined constant between 0 and 1, and M is the total number of edited videos. Therefore, Eq. (4.20) selects only those shots that appear in a significant proportion of edited videos. The motivation for this equation is to avoid shots that were inserted into only $\mathbf{V}_j$ from another, less relevant, parent video.

This does not prevent the proposed method from effectively handling edited videos with multiple parents — since the proposed method estimates the *proportion* of remaining information, it does not need to distinguish between the different parent videos, and treats them as a single larger parent.

The obtained $\hat{I}_0$ can be thought of as a “reference”, but this reference is not an existing video — it is calculated by the proposed method through comparison of the available videos, without any access to the parent $\mathbf{V}_0$. Therefore, the proposed method can be thought of as “reduced reference” (as opposed to “full reference” or “no reference”).

The proposed method thus estimates $\hat{I}_0$.

4.4.3 Global Frame Modification Detection

This section focuses on the detection and penalization of global frame modifications, specifically, scaling and recompression. Scaling a video requires interpolation and recompression. Since both scaling and recompression cause a loss of information through loss of detail, the penalty for scaling and recompression should be proportional to this loss. The proposed method detects scaling and recompression by comparing the visual quality of shots that have the same shot identifier. Based on the visual quality of each shot, the proposed method calculates the global frame modification penalty as:

$$g_i(\mathbf{V}_j) = \begin{cases} 0 & \text{if } Z_j^{k_i} < -1 \\ 1 & \text{if } Z_j^{k_i} > 1 \\ \frac{Z_j^{k_i} + 1}{2} & \text{otherwise} \end{cases}, \quad (4.21)$$

where k_i is the ordinal number of the shot in $\mathbf{V}_j$ that has the shot identifier equal to i (it satisfies $I_j^{k_i} = i$); and $Z_j^{k_i}$ represents the normalized visual quality degradation strength of the k_i th shot in $\mathbf{V}_j$. Shots with visual quality degradation strength less than one standard deviation from the mean are not penalized, since such weak degradations are typically not noticed by human viewers. A significant loss in quality for a frame thus corresponds to a higher penalty, with the maximum being 1.0. This penalty is visualized in Fig. 4.6.

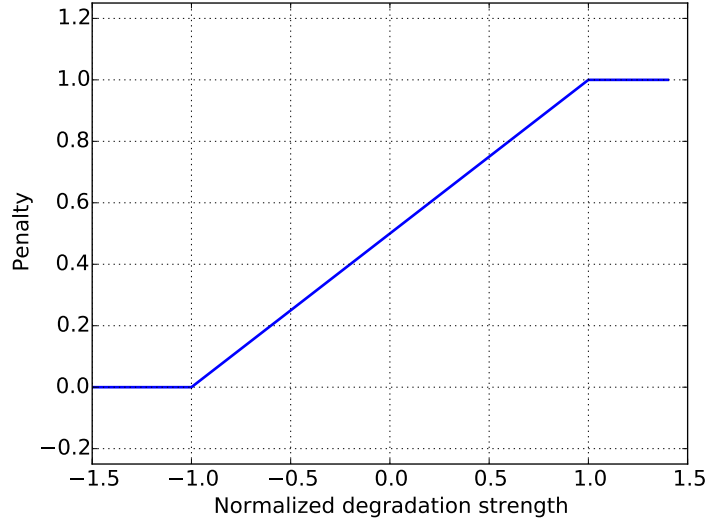


Figure 4.6: Global frame modification penalty $p(\cdot)$.

In order to calculate $g_i(\mathbf{V}_j^{k_i})$, the proposed method first calculates $Q_j^{k_i}$, the raw visual quality of $\mathbf{V}_j^{k_i}$ as:

$$Q_j^{k_i} = \text{mean}(Q_{NR}(\mathbf{V}_j^{k_i}[l_1]), \dots, Q_{NR}(\mathbf{V}_j^{k_i}[l_L])), \quad (4.22)$$

where Q_{NR} is an NRQA algorithm that measures edge width [23]; and $l_1, \dots, l_L$ are the frame numbers of the intra-predicted frames (frames that are encoded independently of other frames [39], hereafter referred to as “keyframes”) of the shot. The proposed method focuses only on the keyframes to reduce computational complexity and speed up processing. The output of Q_{NR} corresponds to the strength of visual quality degradations in the frame: a high value indicates low quality, and a low value indicates high quality.

Next, since the strength of detected visual quality degradations depends on the actual content of the frame, the raw visual quality is normalized as:

$$Z_j^{k_i} = \frac{Q_j^{k_i} - \mu_i}{\sigma_i}, \quad (4.23)$$

where μ_i and σ_i are normalization parameters for the shot whose identifier is equal to i , representing the mean and standard deviation, respectively. The proposed method performs this normalization to fairly penalize shots with different visual content, since visual content is known to affect NRQA algorithms, as shown in Sec. 2.8.

The normalization step above is a novel component of the proposed method. Typically, the output of an NRQA algorithm for a single image is not enough to qualitatively determine its visual quality. In other words, without a reference value, it is impossible to judge whether the visual quality of the image is good or bad. The normalization step provides the proposed method with a reference visual quality level, thus enabling (a) the qualitative assessment of a shot’s visual quality; and (b) the comparison of visual quality between shots with different content (see Fig. 2.9). The effect of normalization is shown empirically in Table 4.14.

Finally, the result of Eq. (4.21) is a real number that is lower for higher quality videos. Therefore, it is an effective penalty for recompressed and scaled videos. The proposed method thus detects and penalizes recompression and scaling. While there are other editing operations, such as color transformations, the addition of logos and subtitles, and spatial cropping, the proposed method does not explicitly detect or penalize them. However, applying these editing operations requires recompressing the video, which causes a decrease in visual quality. Since the proposed method explicitly focuses on visual quality degradation, it will penalize the edited videos in proportion to the extent of the visual quality degradation.

4.4.4 Comparison with State of the Art Methods

In theory, many of the conventional methods introduced in Chapter 1.1 could be used as comparative methods, as they share the same general purpose as the proposed method. In practice, however, comparing the conventional methods to the proposed method directly is difficult. For example, forensic methods are typically used to aid the judgement of an experienced human operator as opposed to being applied automatically. Furthermore, they focus on determining whether individual videos have been tampered or not, which is slightly different from the concrete problem the proposed method is attempting to solve: comparing near-duplicate videos and determining which one is the closest to the parent. For this reason, we do not compare our proposed method with any of the forensic methods.

Similarly, while the method for reconstructing the parent video from its near duplicates [20] is very relevant to our work, it doesn’t achieve the same end result. More specifically, it can be used as an alternative implementation of *part* of our proposed method: the shot identification function described in Chapter 3 and the shot removal detector described in Sec. 4.4.2. Since it does not contain

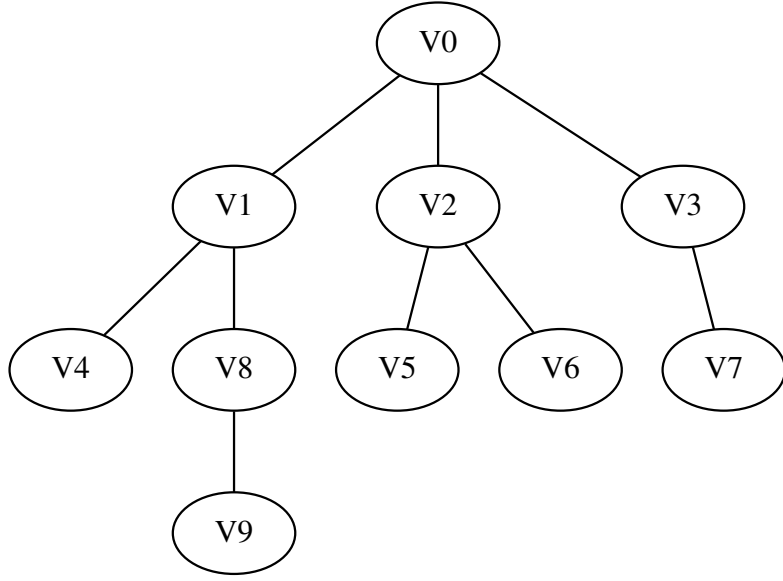


Figure 4.7: The phylogeny tree for the dataset of Table 4.12.

functionality to determine which of the edited videos is most similar to the parent video, it also cannot be directly compared to our proposed method.

Similarly, video phylogeny [19] could be used as a comparative method, with some modifications. After the video phylogeny tree is constructed, the root of the tree will be the parent video, and the edited videos will be its descendants. Authenticity of a video could then be calculated to be inversely proportionate to the depth of the corresponding node in the phylogeny tree. However, there are some limitations to this approach. For example, Fig. 4.7 shows the phylogeny tree for the dataset of Table 4.12. From the graph, it is obvious that V_1 , V_2 and V_3 would have the same authenticity degree, since they are all at a depth of 1 in the tree. However, their expected authenticity degrees are different, as explained by Table 4.12. This problem could be solved by improving the video phylogeny method to consider visual quality. We intend to address this in our future work.

4.5 Experiments

This section describes three experiments that were conducted to verify the effectiveness of the proposed method. Section 4.5.1 evaluates the proposed method in a semi-controlled environment, using a small amount of artificial data that is uploaded to YouTube⁵, with a known parent. Section 4.5.2

⁵<http://www.youtube.com>

evaluates the proposed method in a controlled environment, using a large amount of artificial data, with a known parent. Finally, Section 4.5.3 evaluates the method using real-world data, including a wide range of editing operations, without a known parent.

4.5.1 Small-scale Experiment on Artificial Data

This experiment demonstrates the proposed method using a small artificial dataset, which consists of a single short example parent video and several edited copies of it.

4.5.1.1 Creating the Artificial Data

The single parent video is created by joining the four test videos: *Mobcal*, *Parkrun*, *Shields* and *Stockholm*. The first frames of these four test videos are shown in Figs. 2.9(a), (b), (c) and (d), respectively. The shot identifiers for these four shots are 1, 2, 3, and 4, respectively. Each test video is approximately 10s long, has 1080p resolution, no audio signal and contains a single shot. The created parent video thus consists of four shots and is 40 seconds long. Next, edited videos were created by editing the parent video by adding logos, adding/removing shots entirely and partially, and reversing the number of shots. The editing operations performed on each video are shown in Table 4.12. After editing, each video was reuploaded to YouTube. Thus, with the exception of the parent video, each video in the dataset was recompressed more than once. All the videos are available online⁶. Finally, each video was automatically preprocessed to detect shot boundaries and calculate shot identifiers as described by Sections 3.3 and 3.4.

4.5.1.2 Evaluation Method

To evaluate the proposed method, this section estimates the authenticity degree of the videos in the dataset. Although the parent video V_0 is part of the dataset, it is not used as any sort of reference, and treated the same as any other video. The main purpose of including it in the dataset was to show that the proposed method is capable of identifying it correctly.

After the authenticity degree for each video has been calculated, the videos are ranked in order of highest authenticity degree. These ranks are then compared to the expected ranks, which are determined based on the definition from Sec. 4.1.1. More specifically, the authenticity degree of V_0 is

⁶<http://tinyurl.com/ok92dg2>

Table 4.12: The artificial dataset for the small-scale experiment.

Video	Comments	ER
$\mathbf{V}_0$	Parent video	1
$\mathbf{V}_1$	Reuploaded $\mathbf{V}_0$ to YouTube	2
$\mathbf{V}_2$	Removed 10 frames from each shot of $\mathbf{V}_1$	2
$\mathbf{V}_3$	Reversed order of shots of $\mathbf{V}_1$	4
$\mathbf{V}_4$	Added a shot to $\mathbf{V}_1$	5
$\mathbf{V}_5$	Added a logo to $\mathbf{V}_1$	5
$\mathbf{V}_6$	Downsampled $\mathbf{V}_0$ to 720p	7
$\mathbf{V}_7$	Removed one shot from $\mathbf{V}_0$	8
$\mathbf{V}_8$	Removed two shots from $\mathbf{V}_0$	9
$\mathbf{V}_9$	Removed 60 frames from each shot of $\mathbf{V}_0$	10

Table 4.13: Results for the small-scale experiment on artificial data. The values shown are direct outputs of Eq. (4.18). The rank of each video is shown in parentheses. The r row shows the sample correlation coefficient between the ranks and the expected ranks in Table 4.12.

Video	$\gamma = 0.13$	$\gamma = 0.25$	$\gamma = 0.50$	$\gamma = 0.75$
$\mathbf{V}_0$	0.99 (1)	0.97 (1)	0.94 (1)	0.92 (1)
$\mathbf{V}_1$	0.97 (3)	0.93 (3)	0.87 (3)	0.80 (3)
$\mathbf{V}_2$	0.96 (4)	0.92 (4)	0.84 (4)	0.75 (4)
$\mathbf{V}_3$	0.94 (6)	0.88 (6)	0.77 (6)	0.65 (6)
$\mathbf{V}_4$	0.95 (5)	0.89 (5)	0.78 (5)	0.67 (5)
$\mathbf{V}_5$	0.97 (2)	0.94 (2)	0.88 (2)	0.81 (2)
$\mathbf{V}_6$	0.88 (7)	0.75 (7)	0.50 (8)	0.25 (9)
$\mathbf{V}_7$	0.72 (8)	0.68 (8)	0.61 (7)	0.55 (7)
$\mathbf{V}_8$	0.48 (9)	0.45 (9)	0.40 (9)	0.36 (8)
$\mathbf{V}_9$	0.00 (10)	0.00 (10)	0.00 (10)	0.00 (10)
ρ	0.88	0.88	0.87	0.85

1.0 by definition, since it is the parent video, so its rank must be 1. The remaining expected ranks are determined based on the amount of information removed by the respective editing operation for each video. For example, downsampling removes much more information than simply reuploading, so the rank of $\mathbf{V}_5$ is greater than the rank of $\mathbf{V}_1$. According to the present definition of the authenticity degree, operations such as adding a shot, adding a logo or reversing the order of the shots do not cause a loss of information, but require recompression. Furthermore, the proposed method does not explicitly handle partial shot removal: if a significant part of a shot is removed, then a different shot identifier will be assigned to the shot, and the edited video will be penalized for shot removal, as illustrated by the results for $\mathbf{V}_2$ and $\mathbf{V}_9$. However, since partial shot removal requires recompression of the video,

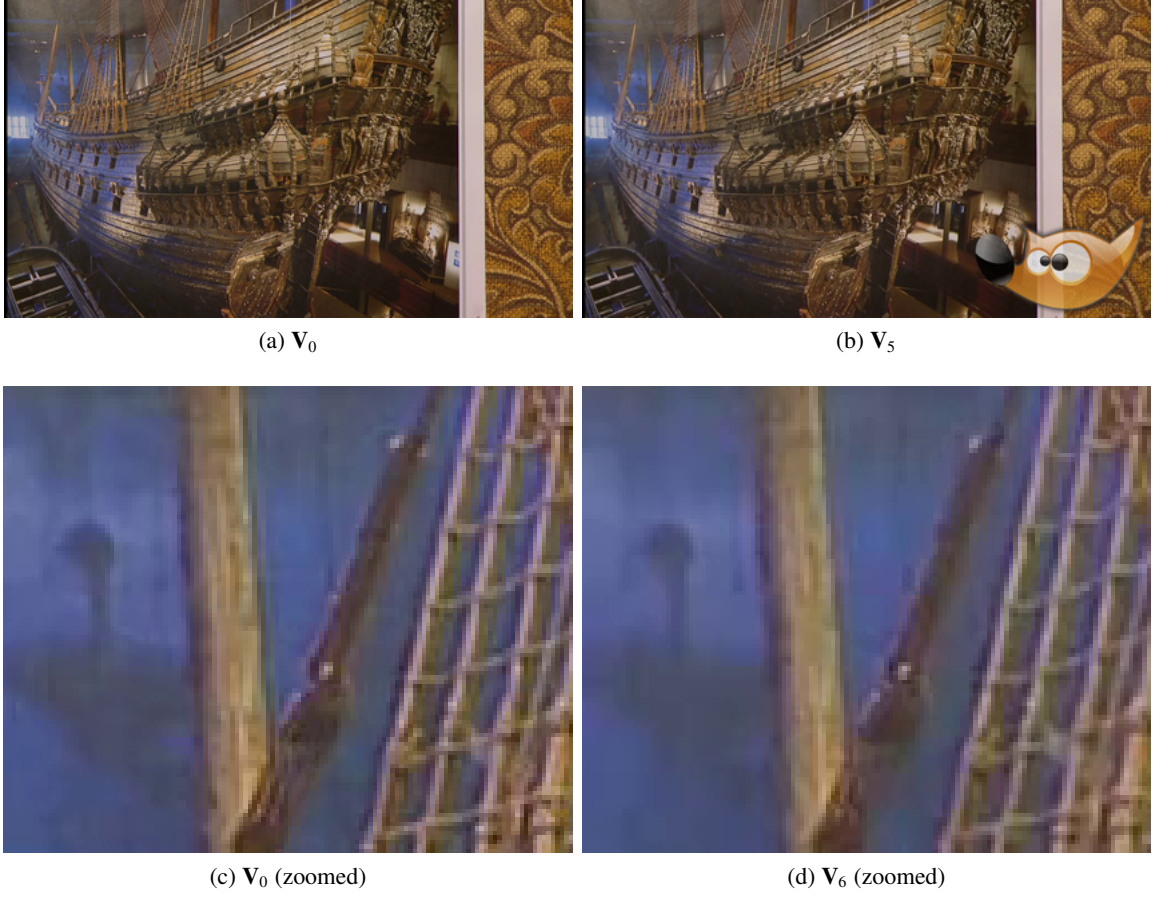


Figure 4.8: Frames from the small-scale artificial dataset.

the proposed method detects this recompression and penalizes the edited video. The expected rank of each video is shown in the “ER” column of Table 4.12. Finally, the value of δ was empirically set to 0.5 for this experiment.

4.5.1.3 Discussion of the Results

Table 4.13 shows the results for various values of γ . The numbers in parentheses indicate the rank of each particular video based on its authenticity degree. The effect of different values of γ can be seen by comparing the authenticity degrees of V_6 and V_8 . The latter has higher video quality, but is missing two shots; the former has lower video quality due to downsampling, but is not missing any shots. For lower values of γ , V_2 has the higher authenticity degree, since the penalty for missing shots outweighs the penalty for recompression. For higher values of γ , the situation is reversed. The ρ row shows the sample correlation coefficient between the expected ranks of Table 4.12 and the ranks based

Table 4.14: The effect of normalization on the small artificial dataset, illustrated by example shots and videos. The “Raw” and “Normalized” columns show the output of Eqs. (4.22) and (4.23), respectively. The “Qualitative” column shows a qualitative assessment of the visual quality of each video.

Video	Raw		Normalized		Qualitative
	<i>Mobcal</i>	<i>Stockholm</i>	<i>Mobcal</i>	<i>Stockholm</i>	
V_0	2.82	3.75	-0.68	-0.89	Excellent
V_1	2.88	3.84	-0.42	-0.45	Good
V_6	3.62	4.54	2.80	2.71	Poor

on the authenticity degrees obtained using each particular value of γ . The correlation coefficient is calculated using Eq. (4.16).

Finally, Table 4.14 demonstrates the effect of normalization on part of the artificial dataset, using two different shots as an example: *Mobcal* and *Stockholm*, shown in Figs. 2.9(a) and (b), respectively. The “Raw” and “Normalized” columns show the output of Eqs. (4.22) and (4.23), respectively. The “Qualitative” column shows a qualitative assessment of the visual quality of each video. The mean and standard deviation for the *Mobcal* shot for the entire artificial dataset were 2.98 and 0.01, respectively. Similarly, the mean and standard deviation for the *Stockholm* shot for the entire artificial dataset were 3.94 and 0.01, respectively. From the raw value alone, it is impossible to determine whether the shot quality is good or poor. Furthermore, a “excellent” raw value for one shot may be a “poor” raw value for a different shot (examine values in **bold**). Therefore, the raw score cannot be used to calculate a fair penalty for all shots. On the other hand, the normalized values better correspond to the qualitative assessments and enable a fairer penalty.

In conclusion, the results of this experiment show that:

1. The proposed method is able to correctly detect and penalize shot removal;
2. The proposed method is sensitive to multiple compression and downsampling; and
3. The proposed method does not explicitly penalize shot insertion, logo insertion, modification of shot order, or partial shot removal, but detects and penalizes the side-effect of recompression.

4.5.2 Large-scale Experiment on Artificial Data

This experiment demonstrates the proposed method using a wider range of data than the previous section. Since collecting a large volume of near-duplicate videos from the Web is difficult, this exper-

Table 4.15: A summary of the parent videos used in the large-scale artificial data experiment.

Dataset	Resolution	Duration	Description
<i>Argo</i>	1920 × 800	2 min 32s	Movie trailer
<i>Bionic</i>	1920 × 1080	2 min 19s	Charity
<i>Budweiser</i>	1920 × 1080	1 min 0s	Product
<i>Daylight</i>	1280 × 720	3 min 16s	Comedy
<i>Drifting</i>	1920 × 1080	3 min 25s	Sport
<i>Ducks</i>	1920 × 1080	1 min 14s	Cartoon
<i>Echo</i>	1280 × 720	2 min 6s	Product
<i>Entourage</i>	1920 × 1080	2 min 27s	Movie trailer
<i>Frozen</i>	1920 × 852	0 min 39s	Movie trailer
<i>Hummingbird</i>	1920 × 1080	3 min 25s	Documentary
<i>Jimmy</i>	1920 × 1080	2 min 21s	Comedy
<i>Lebron</i>	1280 × 720	1 min 21s	Sport
<i>Nfl</i>	1280 × 720	4 min 8s	Sport
<i>Ridge</i>	1920 × 1080	7 min 38s	Nature
<i>Tie</i>	1280 × 720	7 min 27s	Cartoon
<i>Tricks</i>	1920 × 1080	5 min 40s	Sport

iment utilizes artificial data.

4.5.2.1 Creating the Artificial Data

The experiment utilizes 16 datasets. Each dataset consists of a single parent video and edited copies of the parent video. Table 4.15 shows a summary of the parent videos, which were all acquired from YouTube by browsing the *#PopularOnYouTube* video channel. A YouTube playlist with links to each of the parent videos is available online⁷. Figure 4.9 shows the screenshots taken from the artificial videos.

For each dataset, the edited copies were created by editing the parent video with a range of editing operations: downsampling, shot removal, single compression, and double compression. More specifically, the downsampling operation reduced the resolution of the video to the common resolutions used by YouTube. The shot removal operation selected a percentage of shots, at random, and removed them from the video entirely. The compression operation utilized H.264, one of the codecs used by YouTube, to compress the video using a range of Constant Rate Factor (CRF) values. Finally, double compression was also performed with H.264, using a CRF of 18 on both iterations. Each dataset thus consisted of 17 videos (the parent was not included in the dataset). A summary of the editing

⁷<http://tinyurl.com/o7u2mfz>

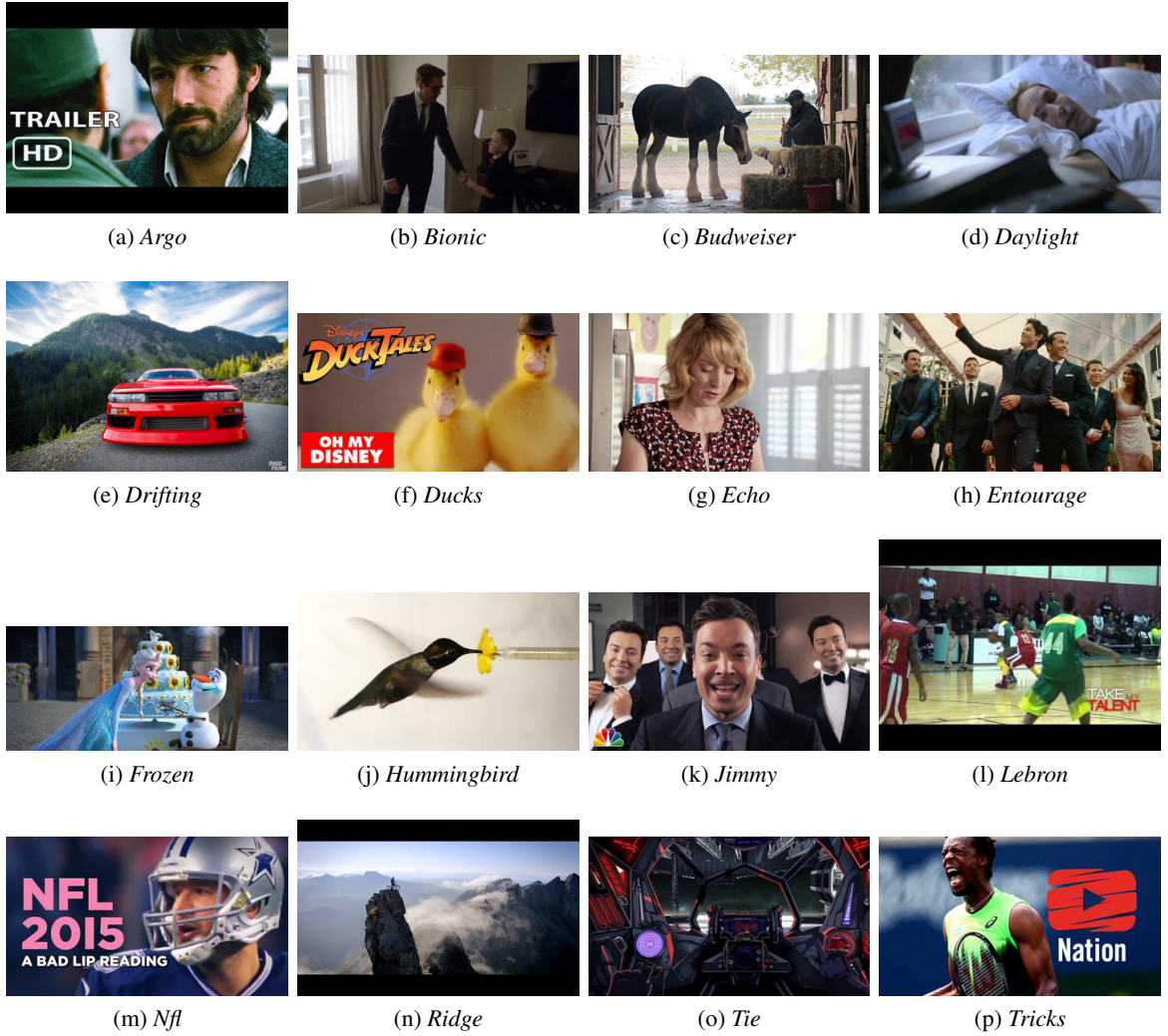


Figure 4.9: Frames from the large-scale artificial datasets.

operations and the parameters used are shown in Table 4.16.

4.5.2.2 Evaluation Method

In contrast to the previous experiment, the amount of data and range of editing operations is much greater. Therefore, it is not possible to objectively rank each of the videos. Instead, we relied on subjective evaluations for each video. The subjects were asked to rate each video on 3 points: (A) relative visual quality; (B) number of deleted shots; and (C) estimated proportion of information remaining from the parent. The subjects were asked to rate (A) and (B) independently, and to rate (C) based on their answers to (A) and (B). All answers were given using a scale of 1 to 5, since this scale

Table 4.16: Editing operations used to generate artificial data.

Operation	Parameter type	Parameter value
Downsampling	Resolution	720p, 480p, 360p
H.264 compression	CRF	18, 26, 34, 40
Shot removal	Percentage	10%, 20%, ..., 90%

Table 4.17: Results for the experiment on large-scale artificial data. r and ρ show the sample correlation coefficient and the rank order correlation coefficient, respectively.

Dataset	Comparative		Proposed	
	r	ρ	r	ρ
<i>Argo</i>	-0.08	0.17	0.82	0.84
<i>Bionic</i>	-0.30	-0.01	0.93	0.90
<i>Budweiser</i>	-0.37	-0.48	0.92	0.86
<i>Daylight</i>	-0.04	0.08	0.84	0.88
<i>Drifting</i>	-0.19	-0.06	0.78	0.69
<i>Ducks</i>	0.12	0.15	0.96	0.93
<i>Echo</i>	0.21	0.38	0.85	0.83
<i>Entourage</i>	-0.41	-0.41	0.90	0.88
<i>Frozen</i>	-0.16	-0.02	0.81	0.77
<i>Hummingbird</i>	-0.36	-0.60	0.86	0.88
<i>Jimmy</i>	-0.57	-0.73	0.76	0.68
<i>Lebron</i>	-0.01	-0.01	0.88	0.82
<i>Nfl</i>	-0.20	0.06	0.97	0.95
<i>Ridge</i>	-0.44	-0.54	0.92	0.91
<i>Tie</i>	0.11	0.42	0.95	0.88
<i>Tricks</i>	-0.38	-0.53	0.91	0.85

is commonly used for visual quality assessment⁸. For each video, the mean answer to (C) across all subjects was recorded as that particular video's estimated proportion of information remaining from the parent. The parent video itself was not shown to the test subjects. A total of 12 subjects participated in the experiment. Next, in order to evaluate the proposed method, the method was used to estimate the authenticity degree of each video. Similar to the previous section, the sample correlation coefficient and rank-order correlation coefficient between the obtained authenticity degrees and the subjective estimates were also calculated using Eqs. (4.16) and (4.17), respectively.

Table 4.18: Summary of real datasets.

Name	Videos	Total dur.	Shots	IDs	ASL	AKI
<i>Bolt</i>	68	4 h 42 min	1933	275	7.99	1.87
<i>Kerry</i>	5	0 h 47 min	103	24	25.23	1.88
<i>Klaus</i>	76	1 h 16 min	253	61	21.47	1.49
<i>Lagos</i>	8	0 h 6 min	79	17	4.58	1.97
<i>Russell</i>	18	2 h 50 min	1748	103	5.80	1.94
Total	175	9 h 41 min	4116	480	13.01	1.83

4.5.2.3 Discussion of the Results

Table 4.17 shows the results of the experiment. For the comparative method, the experiment used the mean edge width for the entire video, with the output scaled such that positive values correspond to high quality. For the proposed method, the γ parameter was empirically set to 0.25. The sample correlation coefficient and the rank-order correlation coefficient are shown in the r and ρ columns, respectively, for each method. The higher value for each dataset is shown in bold. These results show that the proposed method significantly outperforms the comparative method in the majority of cases.

4.5.3 Experiment on Real Data

This experiment utilized real, near-duplicate videos obtained from YouTube. The parent videos for each dataset are unknown. In order to evaluate the proposed method, the same subjective evaluation method as in the previous section was used. A total of 20 subjects participated in the experiment.

4.5.3.1 Description of the Real Data

Table 4.18 shows the datasets used in the experiment, where each dataset consists of several edited videos collected from YouTube. The “Shots” and “IDs” columns show the total number of shots and unique shot identifiers, respectively. The “ASL” and “AKI” columns show the average shot length and average keyframe interval (time between keyframes), respectively, in seconds. Since the average keyframe interval is significantly smaller than the average shot length for all datasets, the majority of shots are represented by at least one keyframe. This validates the optimization strategy employed by Eq. (4.22). Finally, each video was automatically preprocessed to detect shot boundaries and calculate shot identifiers as described by Sections 3.3 and 3.4.

⁸See ITU-T Recommendation BT.500: “Methodology for the subjective assessment of the quality of television pictures”

4.5.3.2 Comparative Methods

For comparative methods, this experiment uses:

1. The number of views of each video,
2. The upload timestamp for each video,
3. The mean visual quality of each video, and
4. The proposed method with $\gamma = 0$ (consider shot removal only).

The number of views and upload timestamp are available from the metadata of each video. The mean visual quality was obtained by calculating the mean edge width [23] for each individual frame, as implemented by Eq. (4.22). Finally, the output of all methods was configured such that positive values correspond to higher authenticity. Therefore, high correlation coefficients correspond to the best methods.

4.5.3.3 Parameter Selection

The parameter γ determines the trade-off between penalizing shot removal and penalizing global frame modifications. Figure 4.10 shows the effect of different values of γ on the results for each of the datasets in Table 4.18. From the figure, it is obvious that the optimal value of γ is different for each dataset. We set $\gamma = 0.7$ in the remainder of the experiments.

The value for the parameter δ , which determines what proportion of edited videos must contain a particular shot identifier in order for that identifier to be part of the parent video, was empirically set to 0.1.

4.5.3.4 Discussion of the Results

Tables 4.19 and 4.20 show the sample correlation coefficient r and rank-order correlation coefficient ρ , respectively. Each line corresponds to one of the datasets in Table 4.18. The number of asterisks indicates the rank of the particular method within a specific dataset: $\star\star\star$, $\star\star$ and $\star$ indicate the best, second best and third best methods, respectively. From the tables, the following observations can be made:

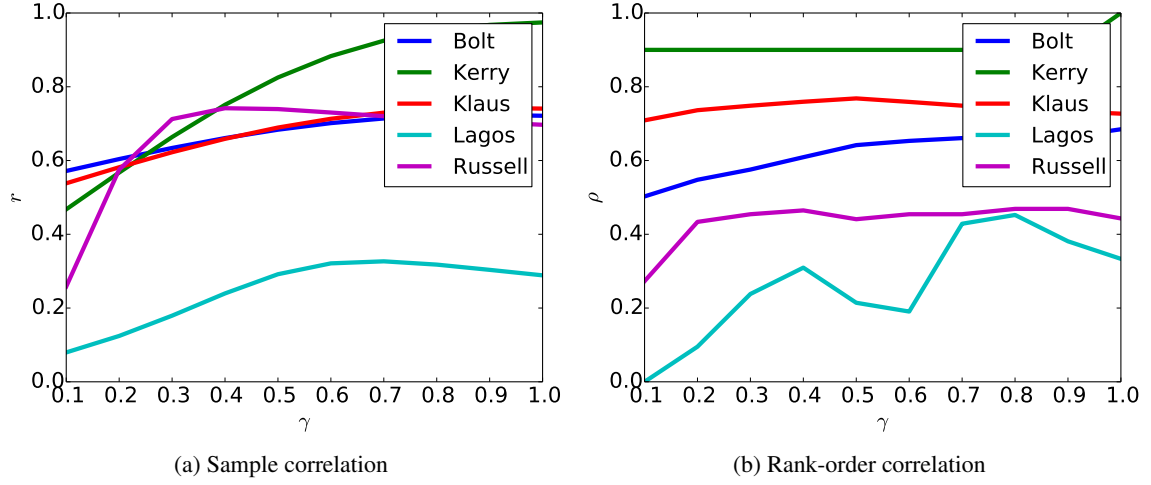


Figure 4.10: Effect of different values of γ on the results for each of the datasets in Table 4.18.

Table 4.19: Sample correlation coefficients (r) for the experiment on real data.

	Comp. 1	Comp. 2	Comp. 3	Comp. 4	Prop.
<i>Bolt</i>	0.11	0.31	★★ 0.61	★ 0.54	★★★ 0.74
<i>Kerry</i>	★★ 0.79	0.19	★ 0.54	0.37	★★★ 0.97
<i>Klaus</i>	0.27	0.08	★★ 0.47	★ 0.42	★★★ 0.68
<i>Lagos</i>	0.05	-0.51	★★ 0.37	★ 0.35	★★★ 0.43
<i>Russell</i>	-0.13	★ 0.14	★★★ 0.80	-0.04	★★ 0.72

- The correlation coefficients for the number of views and upload timestamps (Comp. 1 and 2) vary significantly between the datasets. This demonstrates that the number of views and upload timestamps are not always reliable for estimating the authenticity of Web video.
- The correlation coefficients for the mean visual quality of each video (Comp. 3) are moderate to high for all datasets. This demonstrates that visual quality is relevant and effective for estimating the authenticity of Web video.
- The correlation coefficients for Comp. 4 are high for the *Lagos* and *Russell* datasets, and moderate for the others. This demonstrates that shot removal is relevant to video authenticity. Furthermore, it shows that the optimal value of the constant γ differs between various datasets.
- The correlation coefficients for the proposed method are high for most of the datasets.

Table 4.20: Rank order correlation coefficients (ρ) for the experiment on real data.

	Comp. 1	Comp. 2	Comp. 3	Comp. 4	Prop.
<i>Bolt</i>	0.27	0.26	★★ 0.67	★ 0.45	★★★ 0.69
<i>Kerry</i>	★★ 0.70	0.20	★ 0.40	0.35	★★★ 1.00
<i>Klaus</i>	0.08	0.10	★ 0.46	★★ 0.55	★★★ 0.67
<i>Lagos</i>	-0.12	-0.38	★ 0.26	★★ 0.27	★★★ 0.43
<i>Russell</i>	0.16	★ 0.33	★★★ 0.55	-0.18	★★ 0.47

In summary, Tables 4.19 and 4.20 show that the results of the proposed method correlate with the subjective estimates better than any of the comparative methods. More specifically, the proposed method significantly outperforms comparative method 3: 28 vs. 18 stars. Therefore, the proposed method realizes more effective authenticity degree estimation than the comparative methods.

Table 4.21: Correlation coefficients of the average evaluation score to the individual evaluation score vs. the proposed method.

	Subjective		Proposed	
Dataset	$\bar{r}$	$\bar{\rho}$	r	ρ
<i>Bolt</i>	0.66	0.64	0.74	0.69
<i>Kerry</i>	0.77	0.77	0.97	1.00
<i>Klaus</i>	0.74	0.74	0.68	0.67
<i>Lagos</i>	0.71	0.72	0.43	0.43
<i>Russell</i>	0.74	0.54	0.72	0.47

Finally, Table 4.21 shows the correlation coefficients of the average evaluation score to the individual evaluation score, and compares them to the proposed method. More specifically, $\bar{r}$ for each dataset was calculated as follows: first, calculate the mean subjective visual quality for each video in the dataset; then, calculate the sample correlation coefficient of each individual subject's response with the mean subjective visual quality, yielding a total of 20 coefficients; and finally calculate the mean of the 20 coefficients. The value for $\bar{\rho}$ was calculated in a similar fashion, replacing the sample correlation coefficient with the rank order correlation coefficient. The coefficients for the proposed method are taken directly from Tables 4.19 and 4.20. This table shows that while the correlation coefficients obtained by the proposed method are not very high, they are comparable to the performance of an average human viewer.

4.6 Conclusion

This chapter proposed a new method for estimating the amount of information remaining from the parent video: the authenticity degree. The novelty of this proposed method consists of four parts: (a) in the absence of the parent video, we estimate its content by comparing edited copies of it; (b) we reduce the difference between the video signals of edited videos before performing a visual quality comparison, thereby enabling the application of conventional NRQA algorithms; (c) we collaboratively utilize the outputs of the NRQA algorithms to determine the visual quality of the parent video; and (d) we enable the comparison of NRQA algorithm outputs for visually different shots. To the best of our knowledge, we are the first to apply such algorithms to videos that have significantly different signals, and the first to utilize video quality to determine the authenticity of a Web video. Experimental results have shown that while conventional NRQA algorithms are closely related to a video's authenticity, they are insufficient where the video signals are significantly different due to editing. Furthermore, the results have shown that the proposed method outperforms conventional NRQA algorithms and other comparative methods. Finally, the results show that the effectiveness of the proposed method is similar to that of an average human viewer. The proposed method has applications in video retrieval: for example, it can be used to sort search results by their authenticity.

The proposed method has several limitations. First, partial shot removal is not penalized: as long as the shot identifiers match, the video containing that shot is not penalized. Next, shot insertion is not penalized at all. Finally, the order of the shots is not considered. Overcoming these limitations is the subject of future work.

Chapter 5

Conclusion

This thesis proposed a method for estimating video authenticity of a video via the analysis visual quality and video structure.

The contents of the thesis are summarized below.

Chapter 1 established the relationship of this thesis to existing relevant research.

Chapter 2 introduced the field of quality assessment in greater detail. The chapter reviews several conventional NRQA algorithms and compares their performance empirically. Finally, it demonstrates the limitations of existing algorithms when applied to videos of differing visual content.

Chapter 3 described shot identification, an important pre-processing step for the method proposed in this thesis. It introduced algorithms for practical and automatic shot identification, and evaluates their effectiveness empirically.

Chapter 4 described the proposed method for estimating the video authenticity. It offers a formal definition of “video authenticity” and a shot-based information model for its estimation. The effectiveness of the proposed method was verified through a large volume of experiments on both artificial and real-world data.

Much remains to be done in the future. More specifically, the proposed method needs to be improved to detect and penalize shot removal, and to consider shot order.

Bibliography

- [1] M. Penkov, T. Ogawa, and M. Haseyama, “Fidelity estimation of online video based on video quality measurement and Web information,” in *Proceedings of the 26th Signal Processing Symposium*, vol. A3-5, 2011, pp. pp. 70–74.
- [2] —, “A note on the application of Web information to near-duplicate online video detection,” vol. 36, no. 9, pp. 201–205, 2012.
- [3] —, “A Study on a Novel Framework for Estimating the Authenticity Degree of Web Videos,” in *Digital Signal Processing Symposium*, 2012.
- [4] —, “A Method for Estimating the Authenticity Degree of Web Videos and Its Evaluation,” in *International Workshop on Advanced Image Technology (IWAIT)*, 2013, pp. 534–538.
- [5] —, “Quantification of Video Authenticity by Considering Video Editing through Visual Quality Assessment,” in *ITC-CSCC*, 2013, pp. 838–841.
- [6] —, “A note on improving authenticity degree estimation through Automatic Speech Recognition,” *ITE Technical Report*, vol. 38, no. 7, pp. 341–346, 2014.
- [7] —, “Estimation of Video Authenticity through Collaborative Use of Available Video Signals,” *ITE Transactions on Media Technology and Applications*, vol. 3, no. 3, pp. 214–225, Jul. 2015. [Online]. Available: https://www.jstage.jst.go.jp/article/mta/3/3/3_214/_article
- [8] R. D. Oliveira, M. Cherubini, and N. Oliver, “Looking at near-duplicate videos from a human-centric perspective,” *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 6, no. 3, 2010. [Online]. Available: <http://portal.acm.org/citation.cfm?id=1823749>

- [9] X. Wu, C.-w. Ngo, and Q. Li, "Threading and autodocumenting news videos: a promising solution to rapidly browse news topics," *IEEE Signal Processing Magazine*, vol. 23, no. 2, pp. 59–68, Mar. 2006. [Online]. Available: <http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=1621449>
- [10] S. Siersdorfer, S. Chelaru, W. Nejdl, and J. S. Pedro, "How useful are your comments?: analyzing and predicting youtube comments and comment ratings," ser. WWW '10. Raleigh, North Carolina, USA: ACM, 2010, Conference proceedings (article), pp. 891–900. [Online]. Available: <http://portal.acm.org/citation.cfm?id=1772690.1772781>
- [11] X. Wu, C.-W. Ngo, A. Hauptmann, and H.-K. Tan, "Real-Time Near-Duplicate Elimination for Web Video Search With Content and Context," *Multimedia, IEEE Transactions on*, vol. 11, no. 2, pp. 196–207, Feb. 2009. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4757425
- [12] D. Kundur and D. Hatzinakos, "Digital watermarking for telltale tamper proofing and authentication," *Proceedings of the IEEE*, vol. 87, no. 7, pp. 1167–1180, Jul. 1999. [Online]. Available: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=771070>
- [13] I. Cox, M. Miller, J. Bloom, J. Fridrich, and T. Kalker, "Digital Watermarking and Steganography," Nov. 2007. [Online]. Available: <http://dl.acm.org/citation.cfm?id=1564551>
- [14] A. Piva, "An Overview on Image Forensics," *ISRN Signal Processing*, vol. 2013, p. 22.
- [15] S. Milani, M. Fontani, P. Bestagini, M. Barni, A. Piva, M. Tagliasacchi, and S. Tubaro, "An overview on video forensics," *APSIPA Transactions on Signal and Information Processing*, vol. 1, p. e2, Aug. 2012. [Online]. Available: http://journals.cambridge.org/abstract_S2048770312000029
- [16] S.-p. Li, Z. Han, Y.-z. Chen, B. Fu, C. Lu, and X. Yao, "Resampling forgery detection in JPEG-compressed images," vol. 3, Coll. of Software, Nankai Univ., Tianjin, China. IEEE, Oct. 2010, Conference proceedings (article), pp. 1166–1170. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5646732

- [17] G. Cao, Y. Zhao, R. Ni, L. Yu, and H. Tian, "Forensic detection of median filtering in digital images," Inst. of Inf. Sci., Beijing Jiaotong Univ., Beijing, China. IEEE, Jul. 2010, Conference proceedings (article), pp. 89–94. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5583869
- [18] A. Popescu and H. Farid, "Exposing digital forgeries by detecting traces of resampling," *Signal Processing, IEEE Transactions on*, vol. 53, no. 2, pp. 758–767, Feb. 2005. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1381775
- [19] Z. Dias, A. Rocha, and S. Goldenstein, "Video Phylogeny: Recovering near-duplicate video relationships," in *2011 IEEE International Workshop on Information Forensics and Security*. IEEE, Nov. 2011, pp. 1–6. [Online]. Available: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6123127>
- [20] S. Lameri, P. Bestagini, A. Melloni, S. Milani, A. Rocha, M. Tagliasacchi, and S. Tubaro, "Who is my parent? Reconstructing video sequences from partially matching shots," in *IEEE International Conference on Image Processing (ICIP)*, 2014.
- [21] N. Diakopoulos and I. Essa, "Modulating video credibility via visualization of quality evaluations," in *Proceedings of the 4th workshop on Information credibility - WICOW '10*. New York, New York, USA: ACM Press, Apr. 2010, p. 75. [Online]. Available: <http://dl.acm.org/citation.cfm?id=1772938.1772953>
- [22] A. K. Moorthy and A. C. Bovik, "Visual quality assessment algorithms: what does the future hold?" *Multimedia Tools Appl.*, vol. 51, no. 2, pp. 675–696, Jan. 2011. [Online]. Available: <http://portal.acm.org/citation.cfm?id=1938166><http://www.springerlink.com/content/wmp6181r08757v39>
- [23] P. Marziliano, F. Dufaux, S. Winkler, and T. Ebrahimi, "A no-reference perceptual blur metric," in *Proceedings of the International Conference on Image Processing. 2002. Proceedings. 2002 International Conference on*, vol. 3, Genimedia SA, Lausanne, Switzerland. Toronto, Ont., Canada: IEEE, 2002, Conference proceedings (article), pp. III–57–III–60 vol.3. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1038902

- [24] Z. Wang and A. C. Bovik, "Modern Image Quality Assessment," *Synthesis Lectures on Image, Video, and Multimedia Processing*, vol. 2, no. 1, pp. 1–156, Jan. 2006. [Online]. Available: <http://www.morganclaypool.com/doi/abs/10.2200/S00010ED1V01Y200508IVM003>
- [25] Q. Huynh-Thu and M. Ghanbari, "Scope of validity of PSNR in image/video quality assessment," *Electronics Letters*, vol. 44, no. 13, p. 800, 2008. [Online]. Available: http://digital-library.theiet.org/content/journals/10.1049/el_20080522
- [26] F. Battisti, M. Carli, and A. Neri, "Image forgery detection by means of no-reference quality metrics," in *SPIE Vol. 8303*, 2012. [Online]. Available: <http://spie.org/Publications/Proceedings/Paper/10.1117/12.910778>
- [27] M. Penkov, T. Ogawa, and M. Haseyama, "Estimation of Video Authenticity through Collaborative Use of Available Video Signals," *ITE Transactions on Media Technology and Applications*, 2015.
- [28] A. Bovik and J. Kouloheris, "Full-reference video quality assessment considering structural distortion and no-reference quality evaluation of MPEG video," in *Proceedings. IEEE International Conference on Multimedia and Expo*, vol. 1. IEEE, 2002, pp. 61–64. [Online]. Available: <http://ieeexplore.ieee.org/articleDetails.jsp?arnumber=1035718>
- [29] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, Apr. 2004. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1284395
- [30] S. Winkler and P. Mohandas, "The Evolution of Video Quality Measurement: From PSNR to Hybrid Metrics," *IEEE Transactions on Broadcasting*, vol. 54, no. 3, pp. 660–668, Sep. 2008. [Online]. Available: <http://ieeexplore.ieee.org/articleDetails.jsp?arnumber=4550731>
- [31] Q. Li and Z. Wang, "Reduced-Reference Image Quality Assessment Using Divisive Normalization-Based Image Representation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 3, no. 2, pp. 202–211, Apr. 2009. [Online]. Available: <http://ieeexplore.ieee.org/articleDetails.jsp?arnumber=4799311>

- [32] H. Wu and M. Yuen, "A generalized block-edge impairment metric for video coding," *Signal Processing Letters, IEEE*, vol. 4, no. 11, pp. 317–320, Nov. 1997. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=641398
- [33] Z. Wang, A. Bovik, and B. Evan, "Blind measurement of blocking artifacts in images," vol. 3, Dept. of Electr. & Comput. Eng., Texas Univ., Austin, TX. IEEE, 2000, Conference proceedings (article), pp. 981–984 vol.3. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=899622
- [34] P. Marziliano, F. Dufaux, S. Winkler, and T. Ebrahimi, "Perceptual blur and ringing metrics: Application to JPEG2000," vol. 19, no. 2. Elsevier, Feb. 2004, Conference proceedings (article), pp. 163–172. [Online]. Available: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.170.4680>
- [35] X. Zhu and P. Milanfar, "A no-reference sharpness metric sensitive to blur and noise," 2009, Conference proceedings (article). [Online]. Available: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.153.5611>
- [36] L. Debing, C. Zhibo, M. Huadong, X. Feng, and G. Xiaodong, "No Reference Block Based Blur Detection," Thomson Corp. Res., Beijing, China. IEEE, Jul. 2009, Conference proceedings (article), pp. 75–80. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5246974
- [37] Z. Wang and A. Bovik, "Reduced- and No-Reference Image Quality Assessment," *IEEE Signal Processing Magazine*, vol. 28, no. 6, pp. 29–40, Nov. 2011. [Online]. Available: [http://ieeexplore.ieee.org/xpl/articleDetails.jsp?tp=&arnumber=6021882&contentType=Journals+&+Magazines&searchWithin=no-reference&ranges=2011_2013_p_Publication_Year&searchField=Search_All&queryText=video+quality+assessment](http://ieeexplore.ieee.org/xpl/articleDetails.jsp?tp=&arnumber=6021882&contentType=Journals+%26+Magazines&searchWithin=no-reference&ranges=2011_2013_p_Publication_Year&searchField=Search_All&queryText=video+quality+assessment)
- [38] J. D. Gibson and A. Bovik, Eds., *Handbook of Image and Video Processing*, 1st ed. Academic Press, Inc., 2000. [Online]. Available: <http://portal.acm.org/citation.cfm?id=556230>
- [39] I. Richardson, *The H.264 Advanced Video Compression Standard*. John Wiley & Sons, 2010. [Online]. Available: <http://books.google.com/books?id=LJoDiPnBzQ8C>

- [40] S. Liu and A. Bovik, "Efficient DCT-domain blind measurement and reduction of blocking artifacts," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 12, no. 12, pp. 1139–1149, Dec. 2002. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1175450
- [41] A. Nagasaka and Y. Tanaka, "Automatic Video Indexing and Full-Video Search for Object Appearances." North-Holland Publishing Co., 1992, Conference proceedings (article), pp. 113–127. [Online]. Available: <http://portal.acm.org/citation.cfm?id=719786>
- [42] R. Lienhart, "Comparison of automatic shot boundary detection algorithms," pp. 290–301, 1999. [Online]. Available: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.87.9460>
- [43] S.-c. Cheung and A. Zakhor, "Efficient video similarity measurement with video signature," in *Proceedings. International Conference on Image Processing*, vol. 1. IEEE, 2002, pp. I–621–I–624. [Online]. Available: http://ieeexplore.ieee.org/xpl/articleDetails.jsp?tp=&arnumber=1038101&contentType=Conference+Publications&searchField=Search_All&queryText=cheung+zakhor
- [44] —, "Fast similarity search and clustering of video sequences on the world-wide-web," *IEEE Transactions on Multimedia*, vol. 7, no. 3, pp. 524–537, Jun. 2005. [Online]. Available: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.67.3547http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1430728>
- [45] X. Wu, A. G. Hauptmann, and C. W. Ngo, "Practical elimination of near-duplicates from web video search," ser. MULTIMEDIA '07. Augsburg, Germany: ACM, 2007, Conference proceedings (article), pp. 218–227. [Online]. Available: <http://portal.acm.org/citation.cfm?id=1291280>
- [46] E. Wold, T. Blum, D. Keislar, and J. Wheaten, "Content-based classification, search, and retrieval of audio," *IEEE Multimedia*, vol. 3, no. 3, pp. 27–36, 1996. [Online]. Available: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=556537>
- [47] X. Wu, A. G. Hauptmann, and C.-W. Ngo, "Novelty detection for cross-lingual news stories with visual duplicates and speech transcripts," in *Proceedings of the 15th international*

- conference on Multimedia - MULTIMEDIA '07*. New York, New York, USA: ACM Press, Sep. 2007, p. 168. [Online]. Available: <http://dl.acm.org/citation.cfm?id=1291233.1291274>
- [48] J. Lu, "Video fingerprinting for copy identification: from research to industry applications," in *IS&T/SPIE Electronic Imaging*, E. J. Delp III, J. Dittmann, N. D. Memon, and P. W. Wong, Eds., Feb. 2009, pp. 725 402–725 402–15. [Online]. Available: <http://proceedings.spiedigitallibrary.org/proceeding.aspx?articleid=1335155>
- [49] S.-c. Cheung and A. Zakhor, "Efficient video similarity measurement with video signature," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 13, no. 1, pp. 59–74, Jan. 2003. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1180382
- [50] C. Kim and B. Vasudev, "Spatiotemporal sequence matching for efficient video copy detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, 2005. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1377368
- [51] B. Coskun, B. Sankur, and N. Memon, "Spatio-Temporal Transform Based Video Hashing," *IEEE Transactions on Multimedia*, vol. 8, no. 6, pp. 1190–1208, Dec. 2006. [Online]. Available: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4014210>
- [52] J. Tiedemann, "Building a Multilingual Parallel Subtitle Corpus," in *Computational Linguistics in the Netherlands*, 2007. [Online]. Available: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.148.6563>
- [53] R. A. Cole, J. Mariani, H. Uszkoreit, A. Zaenen, G. Varile, A. Zampolli, and V. Zue, "Survey of the State of the Art in Human Language Technology." [Online]. Available: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.50.7794>
- [54] M. Everingham, J. Sivic, and A. Zisserman, "Hello! My name is... Buffy – automatic naming of characters in TV video," Sep. 2006. [Online]. Available: <http://eprints.pascal-network.org/archive/00002192/>
- [55] T. Langlois, T. Chambel, E. Oliveira, P. Carvalho, G. Marques, and A. Falcão, "VIRUS," in *Proceedings of the 14th International Academic MindTrek Conference on Envisioning Future*

- Media Environments - MindTrek '10*. New York, New York, USA: ACM Press, Oct. 2010, p. 197. [Online]. Available: <http://dl.acm.org/citation.cfm?id=1930488.1930530>
- [56] C. D. Manning, P. Raghavan, and H. Schutze, *Introduction to Information Retrieval*. Cambridge University Press, 2008. [Online]. Available: <http://portal.acm.org/citation.cfm?id=1394399>
- [57] R. C. Gonzalez and R. E. Woods, *Digital Image Processing (3rd Edition)*, 3rd ed. Prentice Hall, Aug. 2007. [Online]. Available: <http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/013168728X>
- [58] D. Lowe, “Object recognition from local scale-invariant features,” in *Proceedings of the Seventh IEEE International Conference on Computer Vision*. IEEE, 1999, pp. 1150–1157 vol.2. [Online]. Available: <http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=790410>
- [59] M. A. Fischler and R. C. Bolles, “Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography,” *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, Jun. 1981. [Online]. Available: <http://dl.acm.org/citation.cfm?id=358669.358692>
- [60] S. Uchihashi and J. Foote, “Summarizing video using a shot importance measure and a frame-packing algorithm,” in *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No.99CH36258)*, vol. 6. IEEE, 1999, pp. 3041–3044 vol.6. [Online]. Available: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=757482>

Acknowledgments

I would like to sincerely thank my supervisor, Prof. Miki Haseyama, from the Graduate School of Information Science and Technology, Hokkaido University, for the invaluable guidance in writing this thesis.

I would also like to sincerely thank Assistant Prof. Takahiro Ogawa, for the Graduate School of Information Science and Technology, Hokkaido University, for the countless hours of assistance and fruitful discussion over the course of performing the work described in this thesis.

Furthermore, I would like to sincerely thank everyone at the Laboratory of Media Dynamics, Graduate School of Information Science and Technology, Hokkaido University for their invaluable support and assistance.

Finally, I would like to thank the Ministry of Education, Culture, Sports Science and Technology, Japan, for the opportunity to study in Japan on a government scholarship.

Publications by the Author

1. ペンコフ マイケル, 小川 貴弘, 長谷山 美紀 “Fidelity estimation of online video based on video quality measurement and Web information” 第26回信号処理シンポジウム, vol. A3-5, pp. 70–74
2. Michael Penkov, Takahiro Ogawa, Miki Haseyama “A note on the application of Web information to near-duplicate online video detection” 映像情報メディア学会技術報告, vol. 36, no. 9, pp. 201–205
3. Michael Penkov, Takahiro Ogawa, Miki Haseyama “A Study on a Novel Framework for Estimating the Authenticity Degree of Web Videos” 第14回DSPS教育者会議予稿集, pp. 53–54
4. Michael Penkov, Takahiro Ogawa, Miki Haseyama “A Method for Estimating the Authenticity Degree of Web Videos and Its Evaluation,” in International Workshop on Advanced Image Technology (IWAIT), 2013, pp. 534–538
5. Michael Penkov, Takahiro Ogawa, Miki Haseyama “Quantification of Video Authenticity by Considering Video Editing through Visual Quality Assessment”, The 28th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC 2013), pp. 838–841 (2013)
6. Michael Penkov, Takahiro Ogawa, Miki Haseyama “A Note on Improving Video Authenticity Degree Estimation through Automatic Speech Recognition”, 映像情報メディア学会技術報告, vol. 38, no. 7, pp. 341–346
7. Michael Penkov, Takahiro Ogawa, Miki Haseyama, “Estimation of Video Authenticity through Collaborative Use of Available Video Signals”, ITE Transactions on Media Technology and Applications, vol. 3, no. 3, pp. 214–225