FAKULTÄT
FÜR !NFORMATIK

Faculty of Informatics

Features in Visual Media Analysis

DISSERTATION

submitted in partial fulfillment of the requirements for the degree of

Doktorin der Sozial- und Wirtschaftswissenschaften

by

Maia Zaharieva
Registration Number 9707986

to the Faculty of Informatics
at the Vienna University of Technology

Advisor: Prof. Dr. Christian Breiteneder

The dissertation has been reviewed by:

Prof. Dr. Christian Breiteneder

Prof. Dr. Stéphane Marchand-Maillet

Wien, 31.10.2011

Maia Zaharieva

DECLARATION

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

October 31, 2011
Vienna, Austria

Maia Zaharieva

ABSTRACT

Today, film analysis is still a tedious process performed mostly manually by film experts. Existing computer vision approaches aim at improved retrieval and summarization methods rather than at film understanding. While current research is predominantly focused on the question what can we learn and extract from a film as the final product, this thesis aims at the study of the filmmaking process as a source for high-level content information.

The central question of this thesis is what can computer vision methods provide to support film analysis as performed by film expert? We discuss a possible mapping between factors that influence the production, presentation, and perception of movies, their application by means of well-established film techniques, and existing feature extraction methods in computer vision. This novel view on film analysis allows for the exploration and identification of three areas in the domain of automated film analysis and understanding. The first area comprises research tasks that have been subject to active research in the recent past. The second area covers research topics that are not immediately solvable for a fully automated computer vision approach without any prior knowledge. The last area identifies research tasks that are still open in the context of automated film analysis and understanding.

Finally, we introduce three novel research questions and possible solutions: camera take reconstruction, film comparison, and recurring element detection. Performed experiments reveal two significant potentials. First, they can assist film experts by providing support for tasks that are currently performed manually. Second, proposed algorithms blaze the trail for advanced application scenarios such as the analysis of different montage patterns, the identification of missing shots, the reconstruction of the original film cut, or the detection of recurring elements.

ZUSAMMENFASSUNG

Trotz großer Fortschritte in der automatisierten Bild- und Videoverarbeitung werden viele Untersuchungen in der Filmanalyse heute immer noch manuell durchgeführt. Existierende Ansätze und Anwendungen der Computer Vision haben meist das Ziel relevante Informationen zu finden oder große Datenmengen kompakt darzustellen als Filme zu verstehen. Während die aktuelle Forschung im Bereich der automatisierten Filmanalyse sich mit der Frage beschäftigt, was wir aus dem Film als solchem lernen und extrahieren können, untersucht diese Arbeit den Entstehungsprozess eines Filmes als möglichen Ausgangspunkt für die automatisierte Filmanalyse.

Die zentrale Fragestellung dieser Arbeit ist: "*Inwieweit können Methoden der Computer Vision Filmwissenschafter unterstützen?*". Wir diskutieren eine mögliche Verlinkung zwischen Faktoren, welche die Entstehung, Gestaltung und Wahrnehmung von Filmen beeinflussen und existierenden Methoden der Computer Vision. Diese neue Sicht auf die Filmanalyse ermöglicht die Identifikation und die Erforschung von drei Gruppen von Fragestellungen im Kontext der automatisierten Filmanalyse: Die erste Gruppe umfasst Forschungsfragen, welche seit einigen Jahren aktiv untersucht werden. In der zweiten Gruppe sind Forschungsfragen zu finden, welche aus dem heutigen Stand der Wissenschaft nicht unmittelbar gelöst werden können. Die dritte Gruppe repräsentiert Fragestellungen, welche von großem Interesse für Filmwissenschafter sind, jedoch in der Computer Vision bisher nicht untersucht wurden.

Im praktischen Teil dieser Arbeit stellen wir drei neue Forschungsrichtungen und deren mögliche Lösungsansätze ausführlich vor: die Wiederherstellung der originalen Aufnahmesequenz, den Vergleich unterschiedlicher Filmversionen und die Erkennung von wiederkehrenden Elementen in Filmen. Die erzielten Ergebnisse der durchgeführten Experimente weisen zwei wesentliche Charakteristika auf: Erstens, zeitaufwändige Aufgaben in der manuellen Filmanalyse können durch automatisierte Methoden effektiv unterstützt werden. Zweitens, die vorgeschlagen Lösungsansätze öffnen den Raum für weitere Fragestellungen in der Filmanalyse wie zum Beispiel für die Analyse von Montagemustern, die Identifizierung verlorener Bild- und Filmsequenzen und das Erkennen von wiederkehrenden Elemente.

To my father.

ACKNOWLEDGMENTS

I have not failed. I've just found
10,000 ways that won't work.

— Thomas Edison

First and foremost I would like to thank my boss and supervisor, Christian Breiteneder for giving me the opportunity to follow my ideas in this thesis. His support, advice, and valuable feedback have been pivotal for my work. To my second supervisor, Stéphane Marchand-Maillet, for the constructive criticism and discussions that helped me sort out ideas and details on my work.

I would also like to thank my colleagues, Dalibor Mitrović and Matthias Zeppelzauer, for sharing research ideas, sweets, or a beer whenever needed. To Horst Eidenberger for being the first one to really convince me of writing a thesis a long time ago and for sharing his workspace with me although (or maybe because of the fact that) I did not pursue his ideas.

To all my friends for enduring my frustration, sharing the joy, and reminding me of life outside of research.

Finally, I would like to thank my family. There are no words to describe the love, care, and support through the years.

PUBLICATIONS

Some ideas and figures have appeared previously in the following publications:

JOURNAL ARTICLES

1. M. Zeppelzauer, M. Zaharieva, D. Mitrović, and C. Breiteneder: *Retrieval of Motion Composition in Film*. Digital Creativity. accepted. 2011.
2. M. Zaharieva, D. Mitrović, M. Zeppelzauer, and C. Breiteneder: *Film analysis in archive documentaries*. IEEE MultiMedia. 18:38–47. 2011.
3. M. Zaharieva, M. Zeppelzauer, D. Mitrović, and C. Breiteneder: *Archive film comparison*. International Journal of Multimedia Data Engineering and Management. 1(3):41–56. 2010.

PEER-REVIEWED CONFERENCE PUBLICATIONS

1. M. Zaharieva and C. Breiteneder: *Recurring Element Detection in Movies*. In International Multimedia Modeling Conference (MMM'12). accepted. 2012.
2. D. Mitrović, M. Zeppelzauer, M. Zaharieva, and C. Breiteneder: *Retrieval of Visual Composition in Film Analysis*. In International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS'11). 2011.
3. D. Mitrović, S. Hartlieb, M. Zeppelzauer, and M. Zaharieva: *Scene Segmentation in Artistic Archive Documentaries*. In Symposium of the Workgroup Human-Computer Interaction and Usability Engineering (USAB'10). LNCS 6389/2010. pp. 400–410. 2010.
4. M. Zaharieva, M. Zeppelzauer, C. Breiteneder, and D. Mitrović: *Camera take reconstruction*. In International Multimedia Modeling Conference (MMM'10). LNCS 5916/2010. pp. 379–388. 2010.
5. M. Zeppelzauer, M. Zaharieva, D. Mitrović, and C. Breiteneder: *A novel trajectory clustering approach for motion segmentation*. In International Multimedia Modeling Conference (MMM'10). LNCS 5916/2010. pp. 433–443. 2010.
6. M. Zaharieva, M. Zeppelzauer, D. Mitrović, and C. Breiteneder: *Finding the missing piece: Content-based video comparison*. In IEEE International Symposium on Multimedia (ISM'09). pp. 330–335. 2009.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

ASA	American Standards Association
ANMRR	Average Normalized Retrieval Rate
CCV	Color Coherence Vector
DCT	Discrete Cosine Transform
DOG	Difference-of-Gaussian
GBR	Geometry-based region
GLOH	Gradient Location and Orientation Histogram
GoF	Group of Frames
GoP	Group of Pictures
HSV	Hue-Saturation-Value
IBR	Intensity-based region
JPEG	Joint Photographic Experts Group
KLT	Kanade-Lucas-Tomasi
LSD	Line Segment Detector
MPEG	Moving Picture Experts Group
MSER	Maximally Stable Extremal Regions
MSF	Markov Stationary Features
PCA	Principal Components Analysis
RANSAC	RANdom SAmple Consensus
RGB	Red-Green-Blue
SIFT	Scale Invariant Feature Transform
SURF	Speeded Up Robust Features

1 INTRODUCTION

The camera is no innocent eye.

— Jesse J. Prinz

Visual media is a broadly used term usually referring to TV, movies, photography, etc. In general, it can be distinguished between still and moving pictures. Both types involve (predominantly) visual senses, share the same visual features, and often require the same computer vision methods for their automated analysis, understanding, and retrieval. Moving pictures, such as films, videos or TV broadcasts, additionally face new possibilities and challenges introduced by the dimension of time. Following, this work is focused primarily on the study of visual features for film analysis and understanding. It neglects the auditory aspect, although, crucial characteristics and details are explained where necessary.

Film studies is a research discipline that explores the application of various film techniques and conventions, the production and distribution of films, how they are responded by the audience, and what economical, technological, social and aesthetic practices influence their creation and perception [41, 135, 179]. Despite the recent progress in technology and in computer vision methods, film analysis is still a tedious process performed mostly manually by film scholars and film experts. This situation is substantially caused by missing mutual understanding. The two main questions are:

- *What are film experts interested in?* and
- *What can computer vision methods provide for their support?*

Today, the focus of computer vision methods for film and video analysis moves from low-level tasks such as shot boundary detection to high-level analysis such as genre, scene or event recognition. However, such approaches are driven by the requirements of the audience and consumers and aim at improved retrieval and summarization methods rather than film understanding.

This work presents an attempt to build a bridge between the two disciplines: film studies and computer vision. It is based on the theory that films have a manipulative aspect. Next to narrative and involved actors, manifold factors in the process of filmmaking influence the way a film is perceived by the audience. Some film factors have attention models as origin. For example, color, contrast, positioning of an object or its motion can be applied to attract the attention of the audience to a certain area in the scene. Other factors are the result of the artistic nature and creativity of the people involved in the creation process. The study of such film factors and the exploration of the feasibility of corresponding computer vision approaches for their detection and analysis is the subject of this thesis. In this work we investigate three aspects of features in visual media:

1. *What features of visual media do influence their production, presentation, and perception most?* In this context, we explore features (mainly) from a filmmaking point of view and provide a basic understanding for common film techniques and their intent.

2. *What features can be applied to represent given characteristics of visual media?* This topic deals with features as a basic module in computer vision. It provides a brief overview over well-established features in automated film and video analysis.

3. *What features can be linked together in praxis?* In the core of this work, we explore the feasibility and the potential of a mapping between the previously discussed views on features in visual media.

1.1 SUMMARY OF CONTRIBUTIONS

The contributions of this thesis can be summarized as follows:

1. A *novel view on visual media analysis*. The understanding of crucial elements in the process of filmmaking and the technical knowledge about methods of computer vision allows for the definition of a thorough mapping between the two disciplines. Such a mapping does not solely improve the mutual understanding of the two research areas. Moreover, it allows for the identification of feasible research tasks for an automated film analysis and outlines the limitations of existing computer vision methods.

2. The identification and exploration of *novel applications in the context of film analysis and understanding*. Following the study and the analysis discussed above, this work presents novel application scenarios in the context of automated film analysis and understanding. Performed experiments and evaluations show the potential of the proposed methods to assist a number of tedious tasks currently performed manually by film experts such as the comparison of different film versions or the identification of unique shots. Furthermore, such methods can be applied as an intermediate step towards a high-level film analysis such as the analysis of montage patterns of different filmmakers or editors.

1.2 THESIS OUTLINE

Chapter 2 explores the features of visual media that most influence their production, presentation, and perception. In the process of this study we discuss the process of filmmaking and the intentions behind different film techniques in detail. This chapter provides a basic understanding of the elements of media aesthetics in filmmaking and allows for the identification of possible tasks for automated visual media analysis.

Chapter 3 briefly outlines a set of features for visual media representation from a computer vision point of view. In the following, discussed features are applied in experiments and evaluations in the remaining chapters.

Chapter 4 creates the link between visual media features in filmmaking and in computer vision. Various applications of the elements of media aesthetics in filmmaking by means of well-established film techniques are mapped to existing approaches for their automated detection and analysis in computer vision. This study allows for the identification of three areas in the domain of automated film analysis and understanding:

1. tasks that have been subject to active research in the last years;
2. tasks that are not directly solvable for a fully automated computer vision approach; and
3. research tasks that are still open in the context of automated film analysis and understanding.

Chapter 5 presents three novel application scenarios in this context. All experiments are performed in an unconventional data set of archived documentaries. The explored data set bears challenges from both artistic and technological point of view and outlines possible limitations of existing approaches in computer vision. All three case studies show the high potential to improve the process of film analysis, understanding, and retrieval.

Finally, *Chapter* 6 provides a summary of the work presented in this thesis and discusses some ideas and directions for future research.

Part I

BACKGROUND

2 MEDIA AESTHETICS IN FILM

A system of aesthetics can never confine,
within one interpretation,
notions which must include them all.

— Jean Mitry [105]

Aim of the chapter

In this chapter we focus on the first research question of this thesis: *What features of visual media do influence their production, presentation, and perception?* This question is important for several aspects: to provide a basic understanding of the elements of media aesthetics in filmmaking, to identify possible tasks for automated visual media analysis, and to stress the significance of the mapping proposed in this thesis. Next to the narrative story of a film and featuring actors, its overall presentation strongly influences the perception (aesthetic experience) of a movie. Since the whole process of filmmaking is a chain of aesthetic decisions, the understanding of such background factors provides high-level structural and semantic information that can be missed by conventional automated retrieval methods. After defining the term of media aesthetics, we address the basic media elements as proposed by Zettl: light and color, space, time and motion, and sound [178, 179]. Furthermore, we present several advanced media concepts that are the result of the combination of and interaction among various fundamental media elements. Following, we provide a brief background on the basic media elements from a human perception point of view. The idea is to understand what human perception is most sensitive to in order to identify those computational features that can best describe visual media. Finally, we discuss existing research on automated processing of media elements as defined by the concept of computational media aesthetics.

What is media aesthetics?

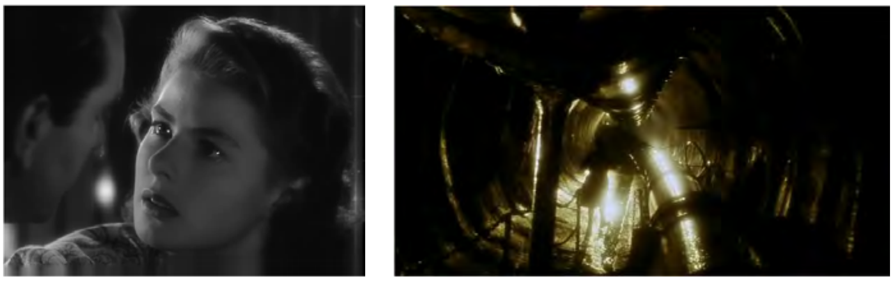
(Applied) *media aesthetics* is the study of elements that influence the production, perception, and presentation of media [179]. Manifold factors influence the creation of visual media and our perceptual reaction to them. Firstly, the creation of any subject of art is accompanied by a series of decisions by its creator: which format to use (e.g. sculpture, painting, image, film, etc.), what do depict, and how to arrange it within the chosen format. Secondly, media elements are influenced by the time and the context in which they were created. For example, the introduction of color to the film production process essentially influenced the perception of a movie and encouraged the artistry of the filmmakers. Finally, our own experience affects the way we perceive media elements [105]. However, in this chapter we focus on the intended purpose of the presented media elements.

2.1 FUNDAMENTAL MEDIA AESTHETIC ELEMENTS

As *fundamental media aesthetic elements* we refer to the aesthetic elements as proposed by Herbert Zettl: light and color, space, time and motion, and sound [179]. Although, first published in the early 1970s, today, Zettls's book is still considered as one of the best and most comprehensive books on film aesthetics.

2.1.1 *Light and Color*

Both lighting and color imply certain aesthetic messages and are commonly used to guide the observer's attention, to set up a specific mood, or to create motifs. Lighting is mostly manipulated as the interaction of light and shadows. For example, in *Casablanca* (1942) lighting draws the attention of the audience to the tears in the eyes of Ilsa which can be easily missed otherwise (see Figure 2.1a). Another example can be found in *Delicatessen* (1991) where lighting and color settings are used to increase the tension of the scene (see Figure 2.1b).



(a) In *Casablanca* (1942) sidelight draws the attention to the tears in the eyes of Ilsa.

(b) In *Delicatessen* (1991) lighting increases the tension of the scene.

Figure 2.1: Lighting in film.

Light level

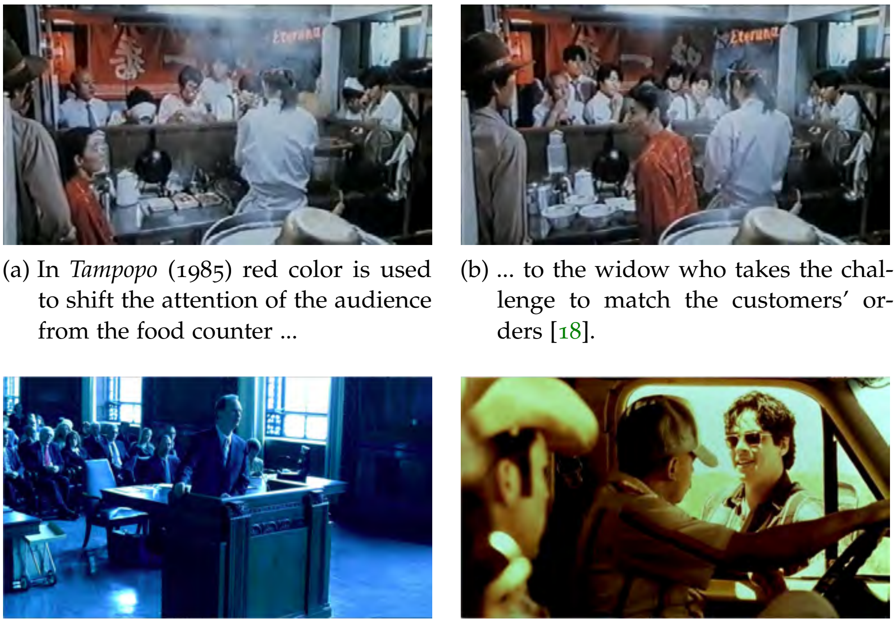
Common techniques for lighting manipulation include e.g. high- and low-key lighting referring to the overall light level in a scene. *High-key* lighting usually conveys ambience ranging from normalcy to energy and enthusiasm (see Figures 2.2a-2.2b) while *low-key* lighting is mostly associated with crime and mystery (see Figures 2.1a-2.1b).

Light direction

Depending on the light direction a difference can be made between *above-eye-level* (associated with normalcy) and *below-eye-level* lighting (well-known from horror movies). The combination thereof creates the three main types of lighting techniques: Chiaroscuro, flat, and silhouette. *Chiaroscuro lighting* is characterized by selective illumination, overall low light level, and distinct and dense shadows. These elements contribute essentially to the direction of the audience's attention and to the creation of expressive visual compositions. On the contrary, the light direction in *flat lighting* is not exactly definable which leads to light or transparent shadows. Flat lighting suggests normalcy, efficiency, and upbeat mood. Finally, in *silhouette lighting*, as the name implies, background is brightly lighted and the actors or objects in the foreground are unlighted. Thus, silhouette lighting is characterized by its extreme on light and dark contrast and its accent on contour rather than texture and volume.

Color functions

Similar to lighting, color can cause a specific feeling, guide the attention of the audience, and stress the quality of an event or object. For example, the color settings of the kitchen scene in *Tampopo* (1985) make the widows' red dress to stand out and draw the attention of the audience to her (see Figures 2.2a-2.2b). A change in color settings can also support the narrative development of the film, such as change in scenes or leap in time. In *Traffic* (2000) different colors are used to influence the mood of the audience and stress the difference in the location settings (see Figures 2.2c-2.2d).



(a) In *Tampopo* (1985) red color is used to shift the attention of the audience from the food counter ...

(b) ... to the widow who takes the challenge to match the customers' orders [18].

(c) In *Traffic* (2000) color post processing is used to influence mood. All scenes that are shot in Washington DC have a blue tone ...

(d) ... while all the scenes shot in Mexico have a hazy yellow look.

Figure 2.2: Color in film.

What color are your feelings?

Despite the fact, that from a perception point of view, color is not as strongly perceived as form, color is one of the main resources of art expressivity [105]. The use of color by means of harmony, dynamics, transformation, and contrast strongly influences the way a film or a picture is experienced by the audience. However, the relativity of color should not be disregarded. Color depends on the context: it depends on the surface, on the surrounding colors, on the lighting conditions, and on the subjectivity of the color perception of different people. Despite the fact that certain colors set off the expression of a particular emotion better than others (for example a violent scene is usually associated with cool colors), the interpretation of colors varies according to the person perceiving it and according to the subjectivity of the creator, i.e. a filmmaker can use color according to color theory (harmonically), contrapuntally (to provoke shock or to create tension), or with little or no intentional meaning. As Zettl states: "In general, the 'psychological' interpretation of colors is a very slippery business" [179].

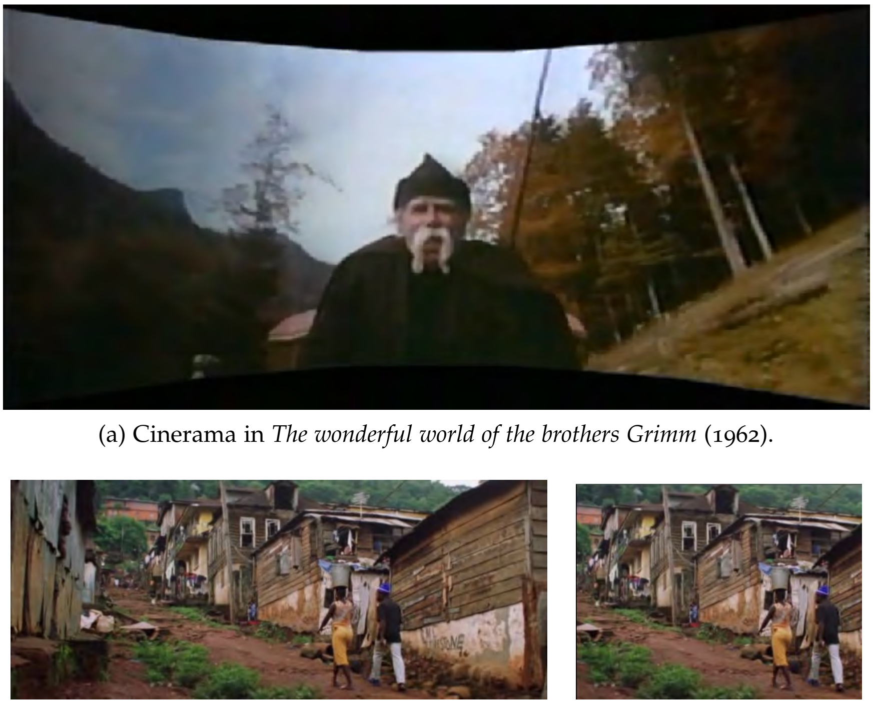
2.1.2 *Space*

Aspect ratio

Various factors contributes to the presentation and arrangement of space within a frame. First, the dimension and shape of a frame control the available space and its overall composition. Early films, such as *Casablanca* (1942), were shot in an Academy ratio (1.33:1 or 1.37:1) and had an almost quadratic shape (see for an example Figure 2.1a). In the early 1950's, Cinerama employed three cameras and three projectors allowing the effect of peripheral vision for the audience and had an aspect ratio of approximately 1:2.85 (see Figure 2.3a) [14]. Today, wide screens emphasize horizontal compositions and allow for the arrangement of multiple areas of interest. Figures 2.3b-2.3c show the difference in the visible areas between different frame dimensions.

Field of view

The specific dimensions of a frame together with the camera position (i.e. angle, level, hight, and distance of shooting) define the field of view of a film space. The decision for a specific camera position causes a drastic difference in the framing of the image and in the perception of the filmed event [18]. Apart from the narrative significance, camera techniques can guide the attention of the audience and increase visual interest. For example, a close up shot can bring details, that can be easily missed otherwise (see for an example Figure 2.4a). Furthermore, similar to the interpretation of color meanings (see Section 2.1.1), certain camera techniques are often assigned absolute meanings. An example for such association is the low angle shot with a powerful character. However, in fact, Bordwell claims that there is no such absolute or general meaning [18]. In some cases, the camera techniques carry such meanings. However, there are no hard and fast rules which preserves the uniqueness and richness of many individual films (see for an example Figure 2.4b).

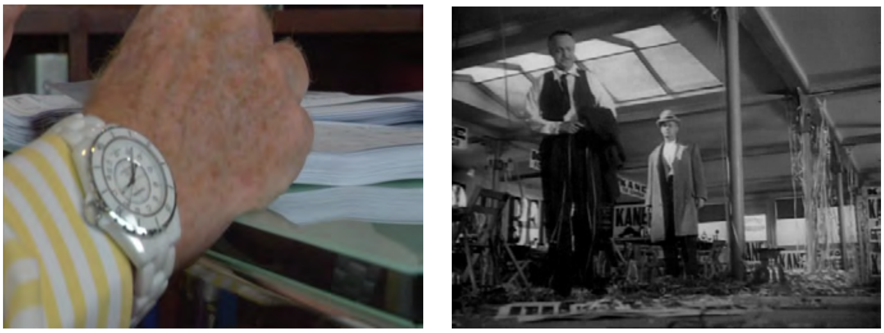


(a) Cinerama in *The wonderful world of the brothers Grimm* (1962).

(b) A scene from *Blood Diamond* (2006) with the original aspect ratio from 2:35.1...

(c) ... and cropped at aspect ratio of 4:3 (TV standard).

Figure 2.3: Space in film: aspect ratio.



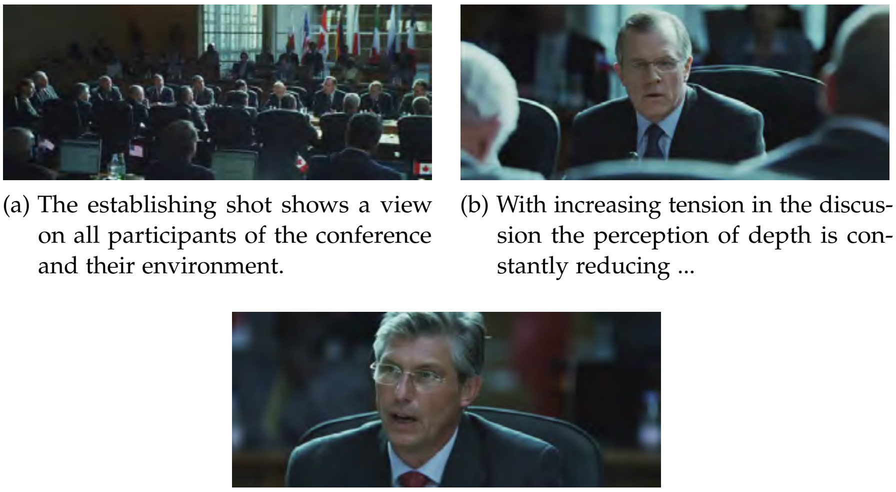
(a) An extreme close up from *Let's Make Money* (2008) revealing details that can be easily missed in any other shot type.

(b) In *Citizen Kane* (1941) low-angle shot are often used to convey Kane's power. However, the lowest angle occur at the point of his most humiliating defeat – the lost gubernatorial campaign [18].

Figure 2.4: Space in film: camera techniques.

Depth of field

Next to field of view, another essential characteristic of space is the perception of depth of field[1]. Various elements, such as lighting, setting, objects and camera positioning, contributes to the impression of space depth. Such depth cues originate mostly in the visual perception and can additionally stress or even damp the perception of depth[2] (see Figure 2.5). Finally, characters and camera movements additionally influence the perception of space. Changes in the camera angle, level, height or distance during the shot support the information about the space of the image (visible and not visible) and make its objects and characters become sharper and more vivid.



(a) The establishing shot shows a view on all participants of the conference and their environment.

(b) With increasing tension in the discussion the perception of depth is constantly reducing ...

(c) ... by means of camera techniques (motion and field of view) and lighting settings.

Figure 2.5: Space in film: from deep to shallow space. The sequence shows the G8 Conference on Diamonds in *Blood Diamond* (2006).

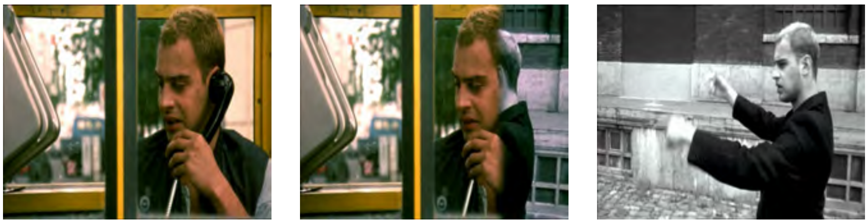
2.1.3 *Time and Motion*

Motion distinguishes film from other visual media – it has duration and is, thus, dependent on time. Film time is not the time of the action or the production; it is the perceived time. Hence, the control of time is essential in making the audience perceive the desired pace and rhythm of the story [39]. Time manipulation usually involves shortened plot time in comparison with the story time. Such time jumps may omit time spans from seconds to centuries. The audience must recognize that time has passed. Often, such time manipulation relies on human experience in omitting scenes that are of no importance to the narrative

[1] Please, note that by depth of field we do not mean the film term referring to "the distance within which objects appear in sharp focus" [14] but rather the term depth from visual perception point of view.

[2] For more details on space perception see Section 2.3.2

development. For example, the time from waking up to breakfast in the morning is not necessarily shown to the audience since the filmmaker can rely on the experience and imagination of the audience. In other cases, time manipulation can be achieved by means of various cinematic techniques, such as accelerated montage (associated with fast time passing), diverse shot transitions, color manipulation, or slow motion (see for an example Figure 2.6).



(a) As Manni explains Lola what's happened earlier ...

(b) ... both a wipe shot transition and color manipulation ...

(c) ... are used to visualize the time jump in the story line.

Figure 2.6: Time manipulation in film. Run Lola Run (1998).

2.1.4 *Sound*

In the early days of filmmaking, salient films were projected to the accompaniment of music. The absence of sound prevented the audience to experience a real feeling of duration or time passing [105]. Today, the presence, or even the absence, of sound can create powerful effects. For example, to hear someone's sobbing may provoke much stronger feelings than to see him/her crying; a quiet sequence in a film can increase the tension, etc. Sound can actively shape the perception and interpretation of an image, guide the attention of the audience and even form expectations. Sound can "clarify image events, contradict them, or render them ambiguous" [18].

Rhythm

Sound strongly influences the perception of rhythm in a movie. Rhythm involves (patterns of) beat and tempo and is most recognizable in film music. However, sound effects can also exhibit rhythmic characteristics (e.g. church bells, horse gallop, etc.). Finally, speech also has rhythm since people exhibit characteristic frequencies and amplitudes and distinct patters of pacing and syllabic stress [18]. Usually, the rhythm of sound is in accordance with the rhythm of editing and motion within a scene. Sound can also support or impeach the credibility of a scene by supporting or contradicting the visual impressions. Usually, sound that is unfaithful to its visual source is used for a dramatic transition or comic effect. Finally, sound possesses temporal and spatial dimensions. The spatial dimensions relate to the source of the sound. The temporal dimensions originate in the relation between

Credibility

Space / Time

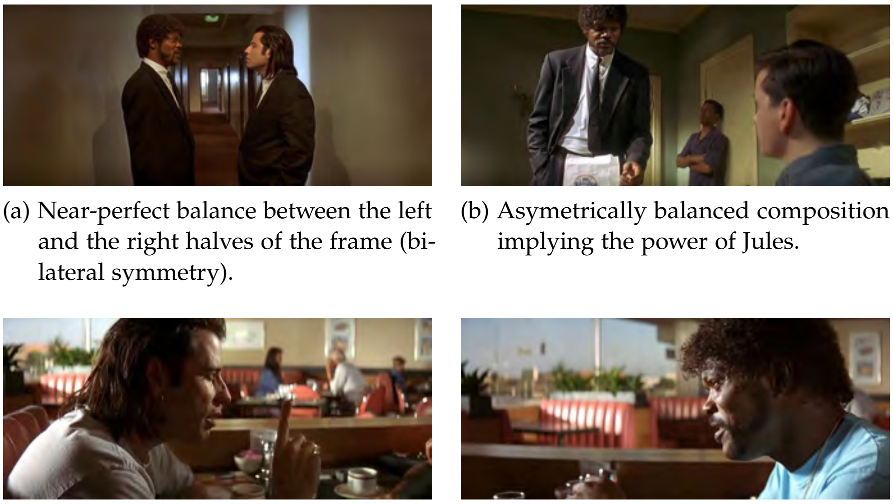
the sound and the event that take place in the same time (simultaneous / nonsimultaneous sounds).

2.2 ADVANCED MEDIA CONCEPTS

The fundamental media elements, as presented in the previous section, are rarely applied individually. Rather, the combination of and interaction among various basic media elements create powerful and expressive tools that shape the perception of visual media. We call such interplay of fundamental media elements an *advanced media concept* and focus on those concepts that are extensively presented in film theory and film study: composition, continuity, motif, and rhythm, tempo and pace.

2.2.1 *Composition*

Composition refers to the use of light and color and the arrangement of objects, characters, and the camera position for photographic and dramatic expression [14]. Filmmakers usually try to balance the image. The extreme type of image balance is called bilateral symmetry (see Figure 2.7a) [18]. More common, however, is a more loose image balance that can also create strong effects, for example, to stress the significance of a character or object, to imply the power of a character (see Figure 2.7b). Finally, unbalanced shots are often applied to increase the tension or dynamics of a scene (see Figures 2.7c-2.7d).



(a) Near-perfect balance between the left and the right halves of the frame (bilateral symmetry).

(b) Asymetrically balanced composition implying the power of Jules.

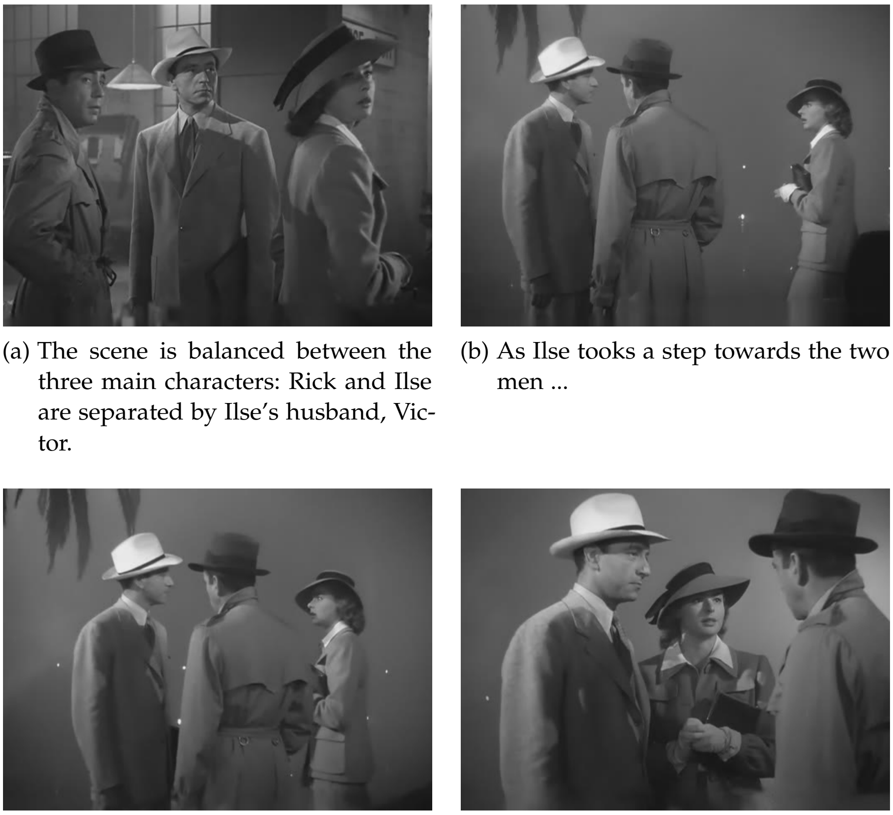
(c) Unbalanced dynamic composition in the restaurant that makes the audience look back ...

(d) ... and forth between Vincent and Jules.

Figure 2.7: Film compositions in *Pulp Fiction* (1994).

Motion

While compositions in still images, such as photographs and paintings, can be thorough planned, the creation of a composition in a film scene is additionally complicated by the dimension of motion. Through character and camera motion, compositions in film scenes become more dynamic and ever changing [96]. For example, camera motion is often required to follow or to lead characters and to make adjustments for motion in the frame (see for an example Figure 2.8).



(a) The scene is balanced between the three main characters: Rick and Ilse are separated by Ilse's husband, Victor.

(b) As Ilse tooks a step towards the two men ...

(c) ... the camera moves around to keep the balance of the scene and to anticipate the character movement ...

(d) ... and to face Ilse and her husband leaving Casablanca together.

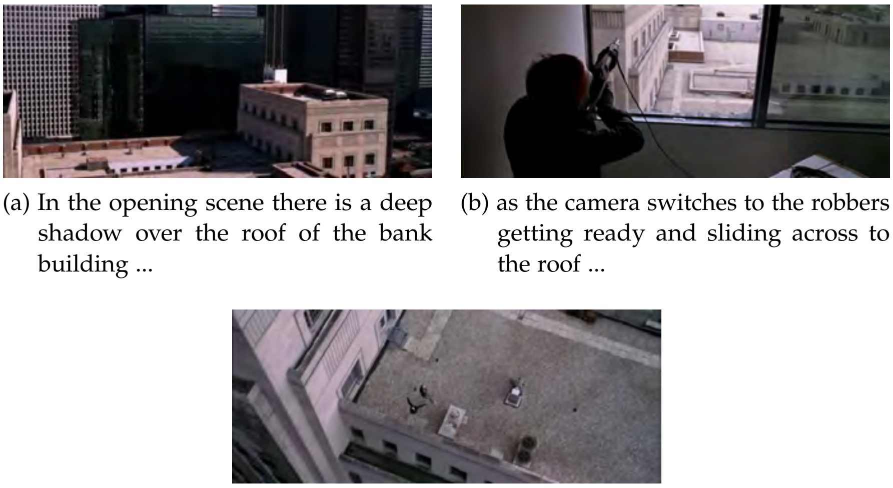
Figure 2.8: Balancing the scene using character and camera motion. The final scene at the airport in *Casablanca* (1942).

Contrast, color

Contrast and color are further compositional elements. Human eyes are sensitive to (even small) differences and changes in contrast and color. On a dark background brightly lit objects will stand out while the darker regions tend to fade and vice versa. The same principle applies for color: a bright object on a subdued background draws the attention of the audience (see for an example Figures 2.2a-2.2b).

2.2.2 *Continuity*

The overall structure of a film is defined by its physical (e.g. shots) and logical (e.g. scenes) units. Even if the audience is aware of it, the film structure should not be presented too obviously. The art of filmmaking consists of creating a unity and a feeling of continuity [105]. In general, continuity refers to the coherence and clearness of the story [14]. It is achieved by maintaining the unity and coherence of manifold factors such as narration, dramatic, space, time, movement, action, and point of view. Since sometimes scenes set in the same time might be shot several days apart, discontinuities can appear in the final version, such as misplaced items or actors, new details can appear / disappear, etc. (see Figure 2.9 for an example).



(a) In the opening scene there is a deep shadow over the roof of the bank building ...

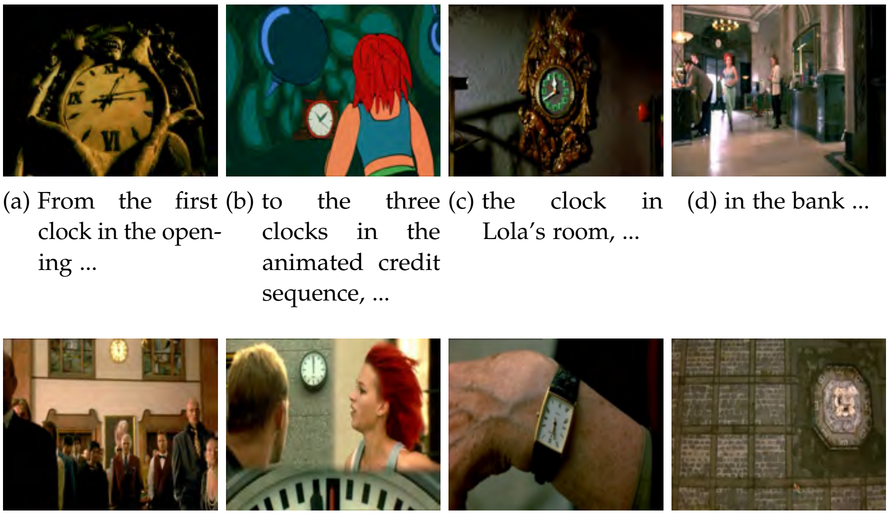
(b) as the camera switches to the robbers getting ready and sliding across to the roof ...

(c) the shadow is completely gone.

Figure 2.9: Continuity error in *The Dark Knight* (2008): scene shot at different day times.

2.2.3 *Motif*

A motif is defined as any "recurring element in a motion picture that gains in dramatic importance through its repetition" [14]. It can be an object, a color, a place, a person, a sound, a movement, pattern of lighting, camera position, or even a story line. A motif can be easily recognizable such as the ringing sound in *Children of Men* (2006) or the use of color red in *Run Lola Run* (1998). However, in most cases, motifs require semantic understanding and rely on the experience and attentiveness of the audience. For instance, note the variety in the visual appearance of the clock-motif in *Run Lola Run* (1998) in Figure 2.10.



(a) From the first clock in the opening ...

(b) to the three clocks in the animated credit sequence, ...

(c) the clock in Lola's room, ...

(d) in the bank ...

(e) and the casino, ...

(f) to the clock Manni watches as he waits for Lola, ...

(g) the watch of the old lady in the front of the bank, ...

(h) and an areal shot of a fountain looking like a clock.

Figure 2.10: Clock-motif in *Run Lola Run* (1998).

2.2.4 *Rhythm, Tempo and Pace*

There are no generally accepted definitions for the terms rhythm, tempo, and pace. Moreover, they are often confused and used interchangeably in the literature. Many sources refer to the terms with the assumption that the reader knows what is meant.

Tempo/Pace

Zettl defines pace as the perceived speed of an event and tempo as the perceived duration [179]. Adams defines tempo as the filmic counterpart to the musical term, namely as "rate of delivery" [5]. Both definitions relate tempo/pace to the perceived time and speed in a film. Zettl argues that perceived time is dependent on event density which can be manipulated by camera and object motion and by "motion" induced by editing [179]. Encyclopedia Britannica adds two more factors that can influence the tempo/pace that an audience senses in a film: "the actual speed and rhythm of movement and cuts within the film, the accompanying music, and the content of the story" [20].

Rhythm

The definition of rhythm in film analysis is very blurry. Beaver confuses rhythm and pace and defines pace as the rhythm of the film [14]. Bordwell and Thompson state that "the issue of rhythm in cinema is enormously complex and still not well understood" [18]. Mitry gives one of the most exhaustive discussions on rhythm in a film [105]. According to Mitry, the term rhythm is not to be confused with speed or to be related to any metric pattern. Moreover, rhythm is not made up of simple relationships of durations but rather of "relationships of intensity contained within relationships of duration". The intensity of a shot is influenced by the amount of movement in the shot and on the length of time it lasts. Thus, two shots of the same length may be perceived as longer or shorter depending on the dynamics of their content and the presentation of the content itself. The essential factor in rhythm is for Mitry "not actual duration itself but the impression of duration, it is this quality and it alone, not a predetermined metric length, which serves as a referent". The more static the content and the narrower the presentation, the less degree of attention and the shorter perception time it demands, the longer the shot appears to be.

The Odessa Steps

The most cited example for a rhythmic montage is the sequence *Odessa Steps* from the salient film *The Battleship Potemkin* (1925) by the russian filmmaker Sergei Eisenstein (e.g. [14, 105]). The sequence shows the slaughter between the town people from Odessa and the White Guards. A look into the shot length montage shows no visible pattern for the strongly perceived rhythm (see Figure 2.11). Eisenstein builds the sequence using a parallel montage – shots of the White Guards are alternating with shots of the folk and victims. The analysis of the parallel sequences on their own – despite consideration of the shot content intensity – does not reveal any explicit rhythm (see Figures 2.12a-2.12b). The dramatic art in the sequence is additionally intensified by the employment of drastic long close-up and detail shots of victims in contrast to the short close-up shots from boots or rifles of the White Guards (see Figure 2.12c). The shots of the marching soldiers stand out against the rhythmic background of the town people. This example illustrates the challenge in the automated detection, measuring and visualization of rhythm. The perception of rhythm is the product of content intensity and the narration itself. In this example, the rhythmic marching of the soldiers enforces the rhythm feeling to the crowd. Although both Figure 2.12b and Figure 2.12c may lead one to assume rhythm in the sequence due to the alternating ups and downs in the depicted distributions, their actual irregularity does not allow for the definition of a reliable detection algorithm.

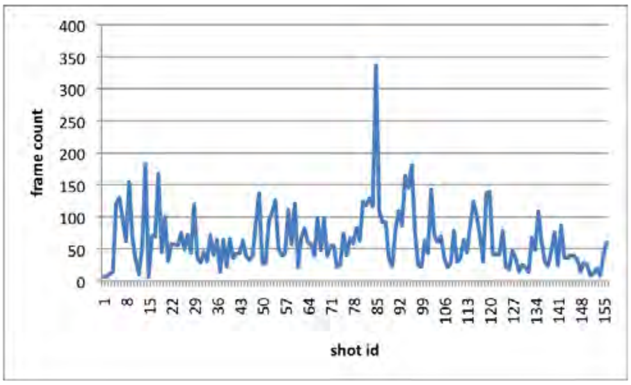


Figure 2.11: Shot length distribution in the *Odessa Steps* from *The Battleship Potemkin* (1926). There is no visible pattern originating in the shot length alone.

2.3 perception of visual media

What is perception?

The term *perception* refers to "the ability to see, hear, or become aware of something through the senses" [99]. It is the joint product of sensory stimuli and the recognition and interpretation of those stimuli. While the physical aspect of the process is well-defined and understood (e.g. how the human eye sees an object), the psychological aspect (e.g. how an object is interpreted by the human brain) still poses manifold challenges for computational media understanding. While we do not present an exhaustive introduction into the topics of visual and sound perception, we briefly address only the perception of the fundamental media aesthetic elements: light and color, space, time and motion, and sound.

2.3.1 *Light and Color*

The primary component of visual perception is light. Physical objects absorb and reflect light and thereby become visible for the human eye. The brightness of an object depends on (according to [10]):

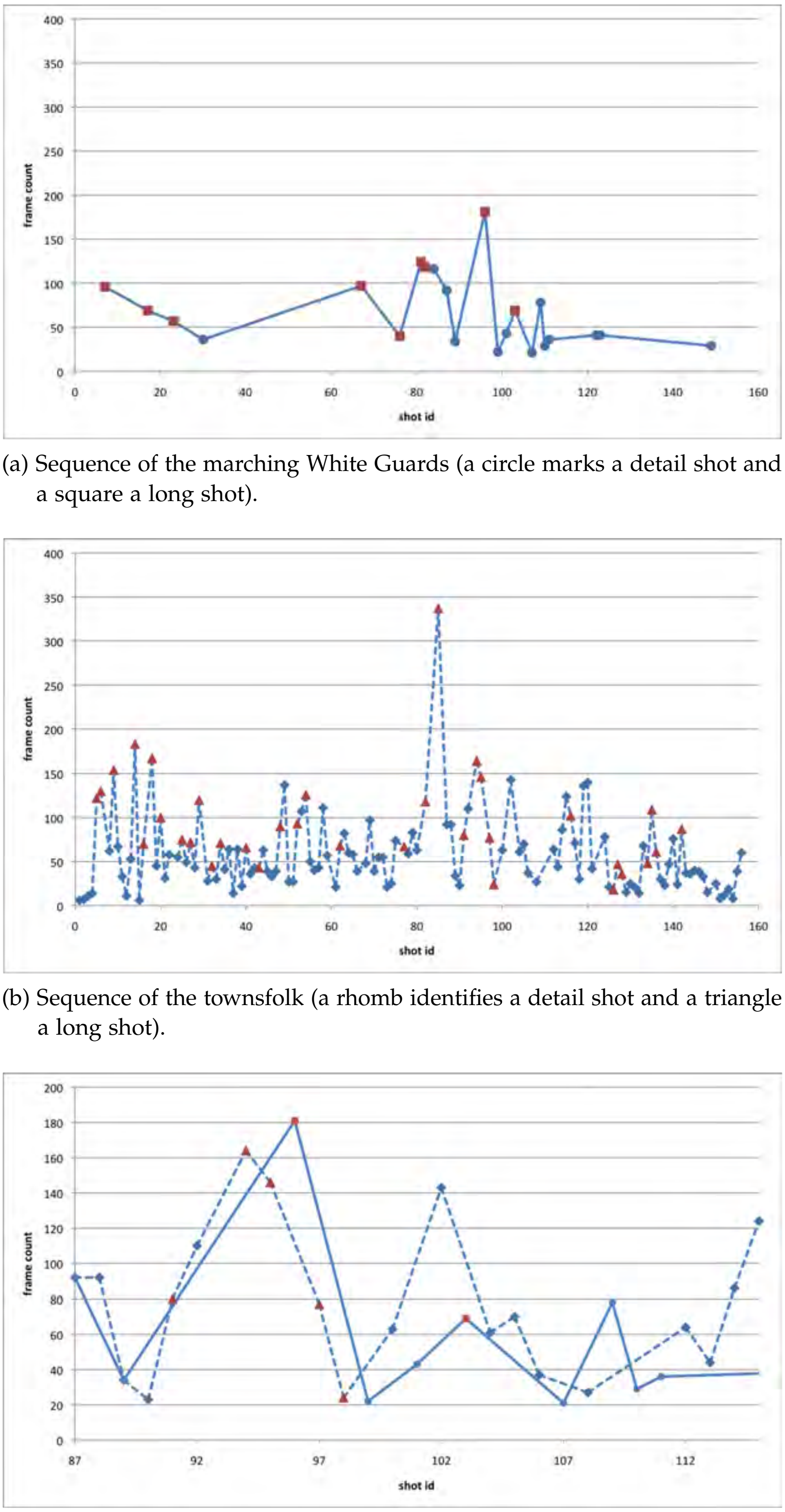
1. the distribution of light,
2. the various optical and physiological processes in the human eye and nervous system, and
3. the capacity of an object to emit and reflect light.

Illumination vs. luminance

The last factor is called *luminance* and is often mistaken with the term of *illumination*. Illumination is a varying property of light. It refers to the amount of light (luminous flux per unit area) falling on a surface and is measured in lux or footcandle. On the contrary, luminance is a constant property of any surface and refers to the light coming off the surface in the direction of the observers. It is measured in candela per square meter (cd/m^2). In general, the percentage of light an object throws back, remains the same. However, different objects can have different illumination – from nearly perfectly specular surfaces such as some metal objects to perfectly absorbing surfaces such as any black objects. Noteworthy is the fact, that even in the presence of varying lighting, the human eye sees objects in relation to the lighting sources of the whole setting and can quite well judge its brightness. *Brightness* is a subjective attribute of light as perceived by the human eye, i.e. it is perceived and not measured. Humans, usually, assign labels ranging from very bright to very dark. Since luminance is the measurable term that most closely corresponds to brightness, brightness is often referred to as perceived luminance.

Color perception

Pure light contains all colors of the visible spectrum although it is perceived as colorless. When it hits an objects, some colors are reflected and some are blocked. The reflected colors contribute to the viewer's



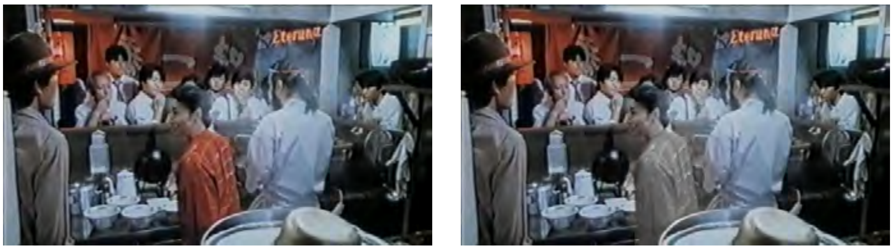
(a) Sequence of the marching White Guards (a circle marks a detail shot and a square a long shot).

(b) Sequence of the townsfolk (a rhomb identifies a detail shot and a triangle a long shot).

(c) Rhythmic parallel montage (solid line depicts the shots of the White Guards, dashed line those of the town people; blue circles on solid lines and blue rhombs on dashed lines mark detail shots of the White Guards and town people respectively; red squares and red triangles the corresponding long shots).

Figure 2.12: Rhythm analysis of the sequence *Odessa Steps* from *The Battleship Potemkin* (1926).

perception of color. Color appearance is strongly affected by the context in which it is perceived: general lighting settings, surrounding objects and their colors, size of the object itself, etc. The mechanisms of the human eye allow for the distinction of millions of different colors. Amazingly, however, the perceptual categories by which humans grasp the world develop from the simple to the complex [10]. There is no guarantee that two persons perceive or name a particular color exactly the same way. More reliable distinction (and the most simplest one) is between brightness and darkness. The number of colors that can be distinguished reliably is usually limited to very few colors – the primaries plus, sometimes, the secondaries connecting them. Human perception can easily identify subtly different shapes. However, it is very limited in the identification of a particular color by memory or at some spatial distance. Finally, human perception is more sensitive to color relations and contrast than to absolute luminance (see Figure 2.13 for an example for color contrast applied to draw the attention of the audience).



(a) The original scene from *Tampopo* (1985) applies color contrast to draw the attention of the audience to the widow at the food counter.

(b) A change in the color settings of the widow's dress, i.e. less contrast to the surrounding objects, may lead to loss of the audience' attention in this crowded scene.

Figure 2.13: Color perception: contrast.

2.3.2 *Space*

Any object can be described within a three-dimensional space by means of its characteristics such as size, shape, position, motion, and direction. The space itself has attributes as well, for example, depth, distance, location direction, etc. In the course of time some of these characteristics may change. Hence, space perception refers to the study of the four-dimensional perceptual world (three-dimensional space plus time) experienced by an observer [10, 62].

Figure 2.14 visualizes some of the most essential factors for the perception of spatial depth in a two-dimensional frame. *Blocking* objects or persons are probably the strongest depth cue by dividing the frame space in multiple depth levels (from blocking object in the foreground to occluded objects in the farther background). Furthermore, the farther away the objects are in the distance, the smaller they appear and the denser their textures become (note the *size* and *texture* gradients of the persons sitting at the table in Figure 2.14). Additionally, *shadows* and *mirroring* provide information about the positioning of the object in the frame. Also the *linear perspective*, that is based on the observation that parallel lines converge to a common vanishing point, supports the perception of depth (note the ceiling lighting and its mirroring on the table). Finally, *motion* is another indicator for space and distance. Noteworthy in this context are the kinetic and the stereokinetic effects. The *kinetic effect* refers to the ability of the human visual system to reconstruct complex three-dimensional shapes or rigid objects from motion information rather than the perception of multiple single elements (e.g. lines, dots, etc.) in space. In contrast, the *stereokinetic effect* is a visual illusion: rotating two-dimensional shapes that are assembled in a specific way can create the illusion of a solid, three-dimensional object.

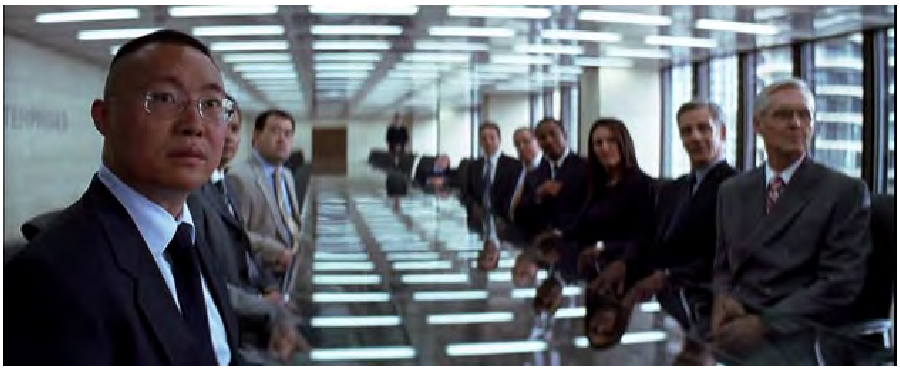
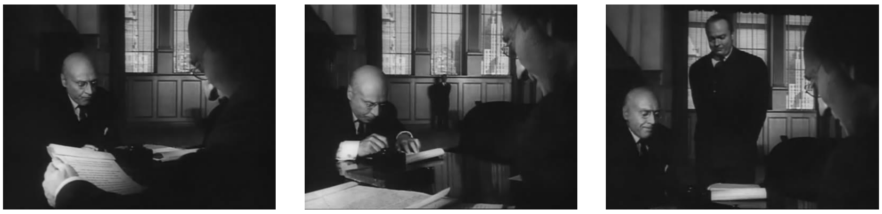


Figure 2.14: Strong space perception in *The Dark Knight* (2008).

Perceptual constancy

The human ability to perceive people and objects as having normal size regardless of changes in the view point or distance is called *perceptual constancy*. Known objects are judged by experience while unknown objects are judged by putting them into context of a known reference or by judging the area the object occupies (see for an example Figure 2.15).

Noteworthy in the context of space perception are also some peculiar characteristics of human perception. In general, pictorial elements are read from left to right (regardless from the way of writing). As consequence, any objects at the right side of the frame looks heavier or more important. A diagonal from the bottom left to the top right suggests an uphill while the opposite diagonal suggests a downhill. Finally, movement to the right is perceived as being easier than movement to the left. This phenomenon is known as *frame asymmetry* and has been subject to considerable academic controversy [10, 179].



(a) At the beginning of the scene the windows in the background appear to be of ordinary size (judged by experience).

(b) As Kane appears in the background, the windows become giant (judged by reference).

(c) Finally, as Kane moves in the foreground, the windows appear to be of ordinary size again.

Figure 2.15: Space perception: size constancy. *Citizen Kane* (1941).

2.3.3 *Time and Motion*

Time perception

The term time perception refers to the awareness of time passing [21]. In contrast to e.g. color or motion perception, time perception is not a tangible attribute and does not have a direct trigger. Time is actually not perceived as such but rather changes or events in time. Ricard A. Block identifies four interacting factors that influence the perception of time [17]:

1. personality characteristics such as sex, interests, psychological and physiological state, and experience,
2. characteristics of the time period itself, i.e. number and complexity of the events happening, their modality, duration, and constellation,
3. activities during the time period such as attentional demands or performance of competing tasks, and
4. changes in time-related behaviors that occur when temporal judgment or estimation is required (simultaneity, rhythm, order, etc.).

Motion perception

Closely related to the concept of time is motion perception since motion is only possible within a given time span. Motion appeals to a basic human instinct and, thus, quickly attracts attention. In general, motion implies change in the surrounding environment and, thus, may require for an action. Such changes can have a positive connotation such as the appearance of a friend or a negative one such as an approaching danger. The lower limit for the detection of (isolated) motion is set to about 35ms temporally and about 1 min of arc spatially [148]. However, these values vary depending on the illumination, on the motion duration and velocity, and on the region of the retina stimulated. Noteworthy is the fact that humans are more reliable at the detection of relative motion, i.e. the detection of a moving object (or group of objects) relative to another object(s) [130]. Two essential characteristics of motion are direction and speed. The perceived direction of a motion is depending on the context in which the movement occurs and is subject to the law of simplicity (i.e. grouping similar things together) [10]. Following, the motion of e.g. a flock of birds is perceived as a global motion despite the individual moves of the wings. As stated above, the perception of motion is possible only within a limited range of speed (e.g. the sun seems to stand still although the earth is in permanent motion, or the quick move with a flash light appears to create a still lighted line). However, also the size of the moving object influences the perceived speed, i.e., in general, large objects seem to move slower than small ones. Finally, motion pictures allow for the experience of motion that is otherwise not perceivable by either accelerating or reducing the speed of recording [10]. This makes it possible to see a plant growing and dying within a minute or the cracking of a glass in thousands of bits for several seconds or even minutes.

Motion perception in motion pictures

The perception of motion in a film is actually the result of an optical illusion. Moving pictures are in fact sequences from still images but perceived as continuous motion (see for an example Figure 2.16). Various theories exist about how this illusion comes about. An early theory that tries to explain the phenomenon of apparent motion is the *persistence of vision*. It basically refers to the after-images, i.e. images that still appear on the retina of the eye for a fraction of a second after the original image has ceased. This optical illusion goes back to Peter Mark Roget in 1824 and was later adopted by Terry Ramsaye in 1926 for the explanation of perceived motion in film. More recent works explore the question if different mechanisms react to apparent and to real motion. Apparent motion can be differentiated in *short-range* (i.e. motion over short distances and of brief duration, producing motion aftereffects) and *long-range* apparent motion (i.e. motion over long distances and of long duration, no motion aftereffects) [8]. Recent research suggests common perception of short-range apparent motion and those of real motion while there is a notable difference to the perception of long-range motion [8]. Motion pictures fall within the limits of the short-range category since the changes between consecutive frames are small enough. Hence, according to this theory, motion in film is transformed by the same physical mechanisms as the real continuous movement does. Cavanagh and Mather argue that the differences between short- and long-range motion processing are a direct result of different stimuli and are not the evidence for two different motion processes [26, 27]. Instead, the authors propose a classification based on first- and second-order motion stimuli. A *first-order* motion process is the result of displacements of first-order differences in luminance and color (first-order statistics). Similarly, a *second-order* motion process responds to differences in second-order characteristics such as texture, contrast, motion or binocular properties.

Figure 2.16: Motion perception in motion pictures: frame sequence from *A man with a movie camera* (1929).

2.3.4 *Sound*

We are usually unaware of the amount of sound information that we are processing in our everyday experience: street noise, people talking, church bells, etc. Several aspects of sound, as we perceive it, are relevant to the film's use of sound: loudness, pitch, and timbre [18]. They interact with each other to create the audio picture of a film, enable us to recognize characters, and shape our experience of a film as a whole.

2.3.4.1 *Loudness*

Abrupt shifts in volume are often used to startle the audience of a movie which makes the estimation of a sound loudness an useful tool for the film analysis. The loudness of the sound describes its perceived intensity (from quiet to loud), i.e. loudness is a subjective quantity and cannot be measured directly [109]. In general, loud sounds have higher amplitude and softer sounds have lower amplitude. Hence, *magnitude estimation* can be applied to determine the relationship between physical intensity and perceived loudness. However, sounds of the same intensity can appear to be differently loud due to the sensitivity of the human ear to different sound frequencies. This is indicated in the *equal-loudness contours* (see Figure 2.17). The figure illustrates two essential characteristics of the dependance of loudness upon its frequency in conjunction with the sound intensity. Firstly, the sensitivity to loudness drops significantly below approx. 350Hz and above approx. 15kHz. Secondly, the dependance of loudness upon frequency is different at different intensity levels (note the change of the curves shape as the intensity increases). Another two factors that can influence the loudness of a sound are its bandwidth and its duration. Two sounds of the same intensity but different spectra can appear differently loud to the audience. Above a *critical bandwidth* of 160Hz at a centre frequency of 1kHz, loudness increases with increasing bandwidth [45]. Below the critical bandwidth, changes in the bandwidth (at fixed overall intensity level) does not result in a change in loudness. Finally, the loudness of a sound is time dependent: the loudness decreases for durations smaller than 100ms while for longer durations the loudness is almost independent of duration [45].

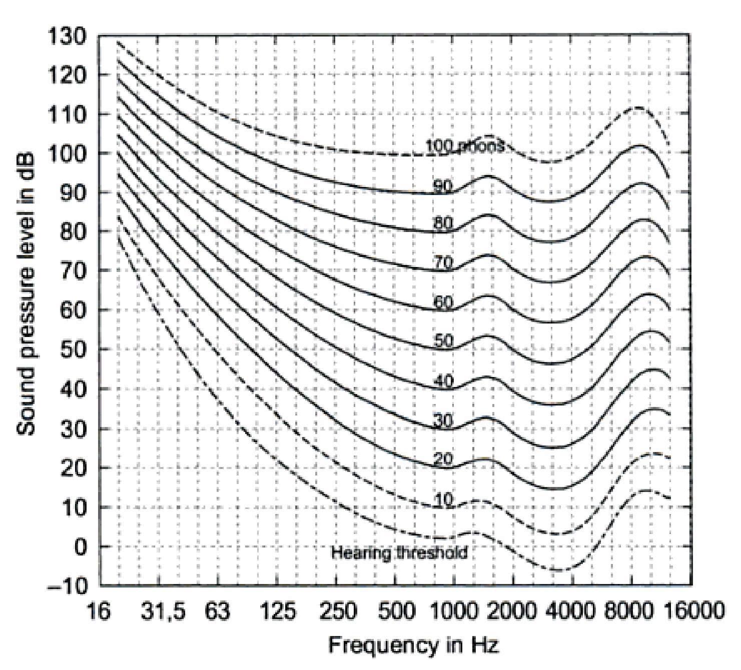


Figure 2.17: Equal-loudness contours (from [109]).

2.3.4.2 *Pitch*

Similar to loudness, pitch is a perceptual sound attribute that allows for the ordering of sounds on a scale from low to high. In a film, pitch supports the differentiation among different sounds and objects. It can also be used to create a joke (for example, a young boy trying to speak in a deep voice). Changes in the pitch are often related to changes in the shots (e.g. from a medium shot to a close-up) [18]. In general, pitch is closely related to *frequency* – large frequencies result in a higher pitch than low frequencies. Pitch also depends on *sound pressure level*. In general, increasing the intensity of a sound decreases its pitch for low frequencies (below approx. 2kHz) and increases its pitch for high frequencies (above approx. 4kHz) [109]. The presence of additional sounds, *partial masking*, can also influence the perception of pitch shifts. In average, additional sounds with a lower frequency than the original (test) sound results in a positive pitch shift, while partial masking produced by sounds at a higher frequency than that of the test sound yields a negative pitch shift [45].

2.3.4.3 *Timbre*

In the psychoacoustics timbre is also known as sound color or sound quality. It depends mainly on the harmonic components of the sound or the relative number of overtones. The American Standards Association (ASA) defines timbre as "that attribute of sensation in terms of which a listener can judge that two sounds having the same loudness and pitch are dissimilar" [1]. The timbre of a sound allows for the recognition of familiar sounds and, moreover, for the differentiation of musical instruments (for example, if a specific note is played on a piano or a guitar). Timbre is a multidimensional perceptual attribute of sound that is closely related to the excitation pattern of that sound [109].

2.4 COMPUTATIONAL MEDIA AESTHETICS

Definition

The term computational media aesthetics was first defined by Chitra Dorai and Svetha Venkatesh in 2001 [111]. It comprises the algorithmic study of media elements and the analysis of their application for "clarifying, intensifying, and interpreting an event for the audience". Research in the field of computational media aesthetics is performed mostly by the group of the original authors, see e.g. [3, 4, 5, 6, 39, 106, 107, 108, 152, 154]. However, earlier works (e.g. [7, 122]) as well as further recent works (e.g. [125]) also explore film editing techniques as a basis for film annotation and retrieval although they do no explicitly refer to the term computational media aesthetics. For example, Radev et al. propose a general film model that compromises structural, semantic, and syntactic elements of film [122]. The model represents a conceptual schema graph whose nodes are the basic film structural elements and features derived from an analysis of the film theory. A more practical approach is proposed by Aigrain et al. for the detection of macro video segments or scenes [7]. The authors employ a set of "medium knowledge-based" rules that account for shot transitions, shot repetition, contiguous shot settings, shot length, soundtrack (music detector), and camera motion. Although the authors present a thorough exploitation of explicit film grammar, they report no evaluation results. Recently, Rasheed et al. exploited a set of low-level features including average shot length, color variance, motion content and lighting key for the genre recognition (comedy, action, drama, horror) of film previews [125]. Presented experiments with over hundred film previews prove that the combination of visual cues with cinematic principles is a powerful tool for genre classification.

Film grammar

Depending on the retrieval task, existing research exploring film editing techniques usually combines the analysis of several media elements. Examples for research work focused on a single media element are the exploitation of color and sound. Based on the assumption that color is closely related to mood, Truong et al. cluster scenes together that share the same color information [154]. Furthermore, they detect "color events", i.e. shots in which the filmmaker intentionally uses color to draw the attention of the audience. The authors hypothesize

Color

that the rarer a color composition is in film, the stronger it attracts the attention of the audience. Moncrieff et al. propose a method for the automated detection of affective sound events in a film based on the dynamics of the sound energy of audio [106, 107, 108]. Despite the subjective nature of the affects associated with sound energy events, experiments on four horror movies prove their correlation. Pfeiffer et al. explore audio editing practices for scene determination [120]. The authors argue that it is possible to identify scenes which are created through sound overlaps and explore the change patters of audio features to automatically determine relations between consecutive shots and group them into scenes.

Sound

Feature combinations

Various feature combinations that exploit editing techniques have been proposed recently. The majority of the explored applications are focused on affective film analysis and automated genre recognition. Dominant features are average shot length, color and motion information, see e.g. [34, 125, 141, 149, 152, 161, 162, 166]. For the determination of scene types (e.g. dialog, action) several authors additionally explore the montage patterns of the corresponding scenes [31, 32, 169]. Finally, film tempo has been explored by means of shot length and motion information for the localization of dramatic events and section boundaries, e.g. [4, 5, 6].

3 MEDIA FEATURES

If I could have expressed the same in a song,
I would have written a song instead.

— Bob Dylan

Aim of the chapter

In computer vision, media representation involves the selection of those features that best describe relevant aspects of media content. In general, there are two broad approaches for media representation: global or local features. Global descriptors represent media by a single feature computed from the whole image. On the opposite, local features represent media by a set of descriptors computed at some points of interest in the corresponding media. In this chapter, we introduce global, Section 3.1, and local features, Section 3.2, that have been successfully applied in recent research for the representation and retrieval of visual media. Since motion detection and description can be based on either global or local features, motion features are discussed separately in Section 3.3. In this chapter, we do not aim at a comprehensive coverage of existing features in computer vision but focus on those set of features that are addressed through performed experiments and evaluations discussed in the remaining chapters.

3.1 GLOBAL FEATURES

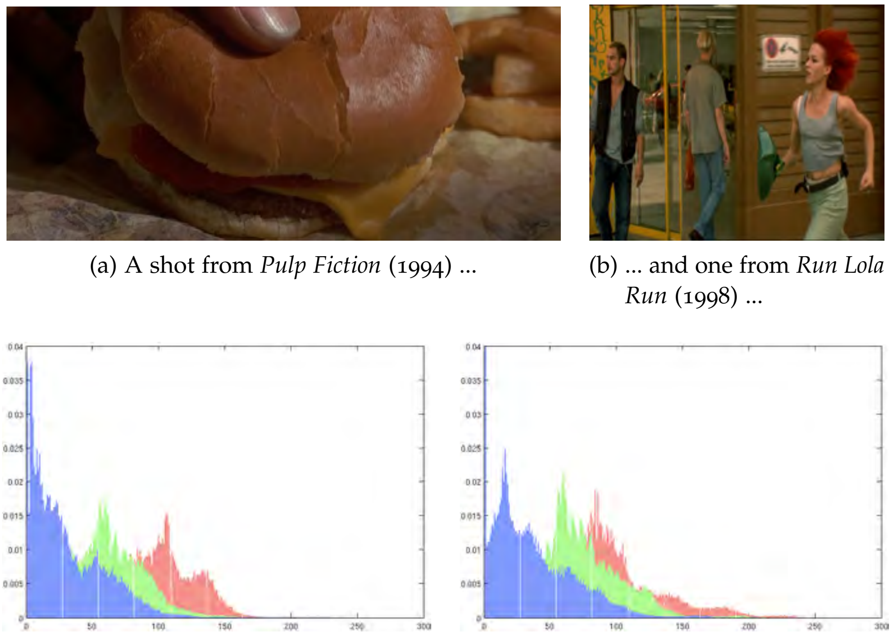
Global feature based approaches characterize media by their entire image. Despite the low computational costs and compact representation, such approaches (usually) possess limited application due their sensitivity to local changes, such as occlusion, illumination changes, etc. In general, such approaches are only applicable for the recognition of rigid objects and often require preliminary segmentation.

3.1.1 *Color features*

Color analysis can be performed for any kind of color spaces, such as Red-Green-Blue (RGB) and Hue-Saturation-Value (HSV). For monochromatic images the intensity (grey-level) information is usually explored.

Histograms

Color histograms represent the distribution of colors in an image. In general, histograms are robust to rotation but not to illumination changes and occlusion. Furthermore, global histograms contain no spatial information and, thus, different images can share the same color distribution even if they are semantically not related (see for an example Figure 3.1[1]). Applications of color histograms comprise e.g. image retrieval [139], object tracking [68] and image fingerprinting [52]. To introduce spatial information images are often divided into M × N parts. Following, each part can be described separately.



(a) A shot from *Pulp Fiction* (1994) ...

(b) ... and one from *Run Lola Run* (1998) ...

(c) ... and the corresponding color (RGB) histograms for the shot from *Pulp Fiction* ...

(d) ... and those from *Run Lola Run*.

Figure 3.1: An example for similar color histograms despite different shot content.

CCV

A different way of incorporating spatial information into the color histogram is the Color Coherence Vector (CCV) proposed by Pass et al. [119]. CCVs split a color histogram into two parts: a coherent vector and a non-coherent vector. A pixel is regarded as coherent if it belongs to a large region of the same color. A CCV of an image is the combination of the histograms over all coherent and all incoherent pixels of the image.

Correlograms

Color patches

MSF

Further color descriptors that consider spatial information are color correlograms and color patches. The correlogram expresses the spatial correlation of pairs of colors [65]. The autocorrelogram is a variation of the color correlogram which considers the correlation between identical colors only [65]. Color patches measure the coarse color distribution of an image. The image is divided into M × N regions and for each region the mean intensity is computed [59]. Recently, Li et al. proposed a semi-global feature, Markov Stationary Features (MSF), that has been shown to be effective for the task of TRECVID[2] concept

[1] The similarity between the presented color histograms is measured using histogram intersection (see Section 3.3, equation 3.3).

[2] http://trecvid.nist.gov/.

detection [86]. MSF extends the histogram-based features by characterizing the spatial co-occurrence of histogram patterns using Markov chain models. It therefore encodes spatial structure information both within and between histogram bins.

Moments

Color moments are based on the assumption that the color distribution of an image can be interpreted as a probability distribution and, thus, can be characterized by its moments [145]. Since the low order moments capture most of the color distribution, usually, only the first three moments are used (mean, standard deviation, and skewness). Similar to the color histograms, color moments are sensitive to illumination changes and partial occlusions. Color moments have been successfully applied e.g. for image and video indexing and retrieval [133, 145], and cut detection [51].

Dominant color regions

Li et al. detect and track dominant color objects in the HSV color space [89]. The authors represent each shot by a dominant color histogram based on detected objects depending on their temporal duration. A more robust approach to illumination changes are methods based on color ratio gradients [55, 64]. Color ratio gradients are derived from the color constant ratios of corresponding (key-) frames. Due to its insensitiveness to object position and geometry, shadows, and illumination changes, the approach has been successfully applied for shot and object representation.

Color ratio gradients

MPEG-7

Finally, MPEG defines five color feature descriptors: Dominant Color, Scalable Color, Color Structure, Color Layout, Group of Frames (GoF)/ Group of Pictures (GoP) Color [66]. The first three represent the color distribution in an image. The later two descriptors describe the color relation between sequences (or groups) of images. The *Dominant Color Descriptor* describes the representative colors in an image or image region. It allows for the specification of a small number of dominant color values and their statistical properties such as distribution and variance. The *Scalable Color Descriptor* is a color histogram in the HSV color space encoded by a Haar transform. Similarly, the *Color Structure Descriptor* is also based on color histograms. However, it uses a small structuring window to identify localized color distribution. The extension of the descriptor to a group of frames or a group of images is the *GoF/GoP Color Descriptor*. Finally, the *Color Layout* captures the spatial layout of the representative colors in an image. The representation is based on the coefficients of the Discrete Cosine Transform (DCT). MPEG-7 color descriptors have been successfully applied (mostly in combination with other features) for video clip matching [16].

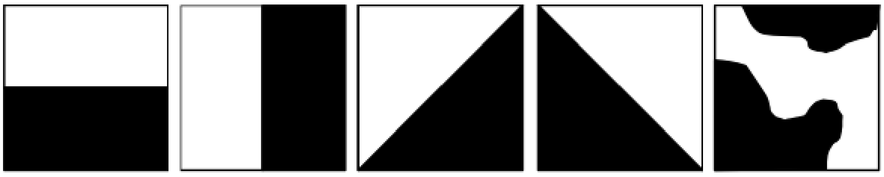
Evaluations

In an evaluation of the performance of the MPEG-7 color descriptors in a visual surveillance scenario the Color Structure outperformed the remaining descriptors in terms of Average Normalized Retrieval Rate (ANMRR) [9]. Similarly, an empirical evaluation of the MPEG-7 color descriptors and color correlograms [65] shows the superior performance of the Color Structure Descriptor in the retrieval of semantic image categories in terms of recall and precision [116]. Recently, Van de Sande et al. investigated the invariance (to intensity, color changes and shifts) and the distinctiveness of different color descriptors [158]. The authors compared the performance of histogram-based descriptors, color moments and color moment invariants, and color descriptors based on Scale Invariant Feature Transform (SIFT). The performed experiments reaffirm that global color descriptors alone are often not sufficient for content-based retrieval. Despite their partial invariance to illumination changes histogram- and moment-based descriptors are clearly outperformed by the color-enhanced local descriptors on both image and video category retrieval.

3.1.2 *Edge features*

Edge features can represent both shape and texture characteristics of visual content and are commonly used for the detection of region boundaries [43, 144], text areas [134, 171], shot boundaries [60, 138], etc. In general, an edge describes intensity variation in close surroundings of a pixel. The computation of edges is based on gradients: edge magnitude is the magnitude of the gradient, and the edge direction is perpendicular to the gradient direction.

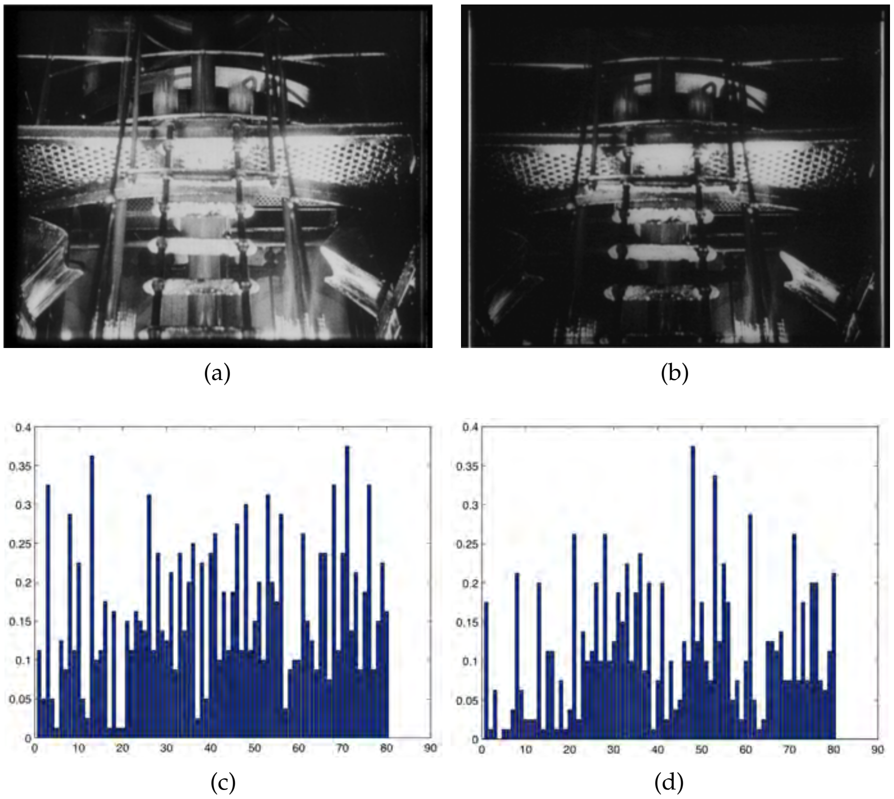
MPEG-7 The MPEG-7 edge histogram describes the local distribution of orientations of edges and distinguishes between five types of edges: horizontal, vertical, 45-degree diagonal, 135-degree diagonal, and non-directional edges (see Figure 3.2) [66]. Each image/frame is divided into 4 × 4 non-overlapping sub-images and for each sub-image a local edge histogram with 5 bins for the corresponding edge types is computed. MPEG-7 edge histogram has been proved to be effective for image similarity retrieval [97] and – in combination with further visual descriptors – for video shot boundary [138] and copy detection [16]. A general drawback of edge features is their sensitivity to illumination changes. Figure 3.3 illustrates the influence of changes in the illumination on the corresponding MPEG-7 edge histograms.



(a) horizontal (b) vertical (c) 45-degree (d) 135-degree (e) non-directional

Figure 3.2: MPEG-7 edge types [66].

Line detection Line detection is an allied area to edge detection. It is usually applied for the recognition of specific shapes (e.g. sport fields, scratches, wires, etc.), camera orientation, vanishing points, etc. Conventional methods


(a)

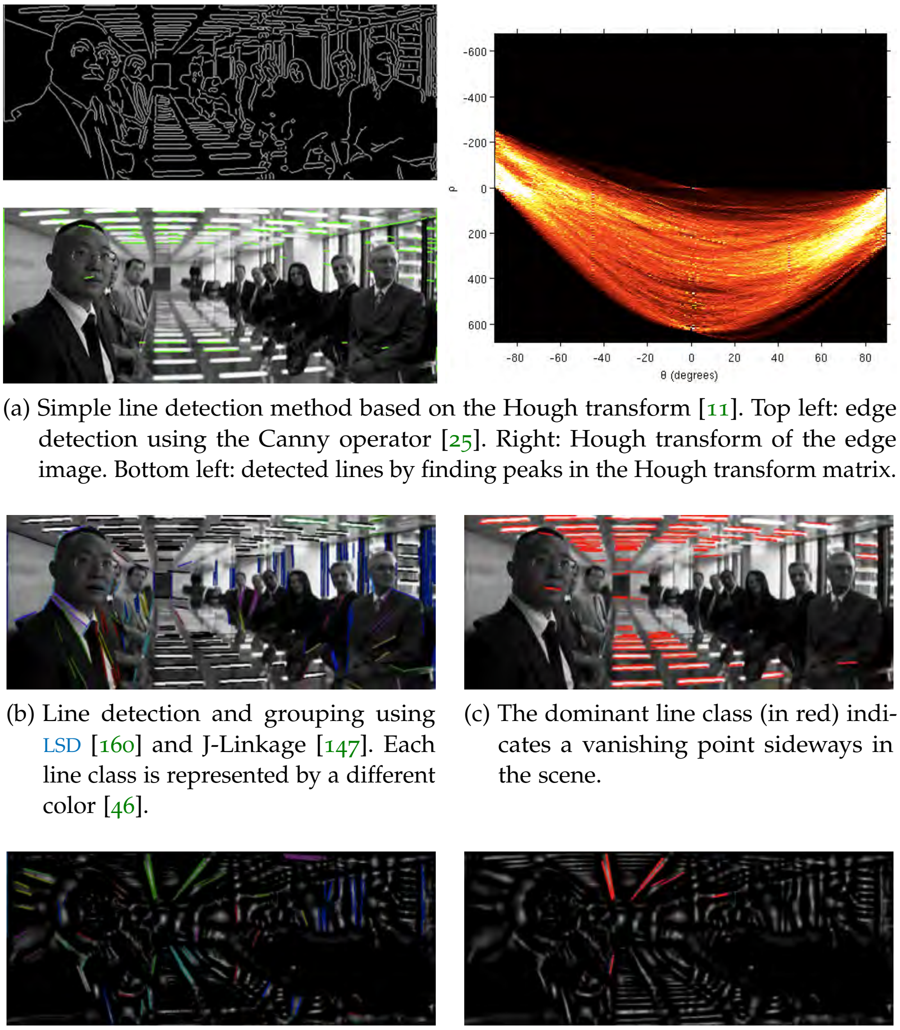
(b)

(c)

(d)

Figure 3.3: An example for differences in MPEG-7 edge histograms due to illumination changes: two frames from two different film compilations and the corresponding MPEG-7 edge histograms.

for line detection are usually based on a Hough transform [11]. Such methods extract lines exceeding predefined length and gap thresholds (see for an example Figure 3.4a). Depending on the selected thresholds and present textures in the image, such methods can lead to a significant number of falsely detected or missed lines. In general, line detection is not necessarily based on preceding edge detection. Burns et al., for example, present an algorithm based only on the gradient orientations at each pixel [24]. Although the algorithm is able to extract low contrast lines, it still depends on fixed thresholds. The method is improved by von Gioi et al. [160] who propose a linear-time Line Segment Detector (LSD) which is a combination of Burn's method and the meaningful alignments by Desolneux et al. [38]. This method is very fast and accurate with a minimal amount of falsely detected or missed lines (see Figures 3.4b-3.4c). Line detection approaches aim at the detection of "real" lines. However, in some situations humans perceive lines where there are none. In the discussed example in Figure 3.4, lines are defined by the lightings on the ceilings and the corresponding mirroring on the table. Such lines are essential for the perception of depth in the scene space (or a vanishing point in the back of the scene). A possible solution includes the consideration of present symmetry, e.g. by measuring the phase symmetry of points in an image [76] (see Figures 3.4d-3.4e).



(a) Simple line detection method based on the Hough transform [11]. Top left: edge detection using the Canny operator [25]. Right: Hough transform of the edge image. Bottom left: detected lines by finding peaks in the Hough transform matrix.

(b) Line detection and grouping using LSD [160] and J-Linkage [147]. Each line class is represented by a different color [46].

(c) The dominant line class (in red) indicates a vanishing point sideways in the scene.

(d) Preceding phase symmetry detection [77] ...

(e) ... allows for the recognition of the vanishing point in the depth of the scene.

Figure 3.4: Line detection in a scene from *The Dark Knight* (2008). The Original frame is depicted in Figure 2.14.

3.2 LOCAL FEATURES

Opposite to global features, local features do not describe entire media but only a local neighborhood around a set of salient (interest) points. Following, local features prove higher reliability in the recognition of media parts (e.g. objects) despite significant changes such as clutter, occlusion, variations in the illumination, etc. In general, the common process workflow for a local feature-based application scenario passes three well-defined stages. First, salient points in both model and test image are identified. Second, their local characteristics are captured by (invariant) feature descriptors. Finally, a matching strategy ascertains if two feature vectors belong to the same keypoint in both images. In the following, we briefly present related work for all three stages.

3.2.1 *Interest Point Detectors*

A wide variety of interest point detectors exist in the literature. In the following we briefly present the most recent ones. For a thorough survey and evaluation of the performance of interest point detectors in the context of repeatability please refer to [102, 129].

Harris-based detectors

The *Harris* detector is based on the use of an auto-correlation matrix [61]. Interest points are detected if the matrix has two significant eigenvalues. Harris points are geometrically stable and reliable in the presence of image rotation, illumination and viewpoint changes [129]. However, the performance of the detector fails when the image resolution changes significantly. To adapt the Harris detector to the scale factor, Mikolajczyk et al. propose the *Harris-Laplace* detector [100, 103]. It selects corners at location where a Laplacian attains extrema in scale space. The *Harris-Affine* detector additionally extends Harris-Laplace by using a second moment matrix to achieve affine invariance [101, 103].

Hessian-based detectors

The *Hessian-Laplace* detector searches for local maxima of the Hessian determinant and selects characteristic scale via a Laplacian [103]. Hessian-Laplace obtains a higher localization accuracy in scale space as Harris-Laplace. Laplacian scale selection acts as a matched filter and works better on blob-like structures than on corners since the shape of a Laplacian kernel fits to the blobs. Similar to Harris-Affine, the *Hessian-Affine* detector extends Hessian-Laplace to achieve invariance to affine transformations [103]. The affine neighborhood is determined by an affine adaptation process based on a second moment matrix. Bay et al. introduced recently a further detector based on the Hessian matrix – the *Fast-Hessian* detector [13]. It approximates a Gaussian second order derivative with box filter. Image convolution with the box filter is computed rapidly using integral images.

MSER

Maximally Stable Extremal Regions (MSER) is a watershed-based algorithm based on intensity values. The algorithm detects connected intensity regions below and above a certain threshold and selects those which remain stable over a set of thresholds [98].

GBR

Tuytelaars et al. proposed two region detectors [156]. The first detector, the *Geometry-based region (GBR)* detector, is an edge-based method. It starts from points detected using the Harris approach and uses the nearby edges. Two nearby edges – which are required for each point, limit the number of potential features. A parallelogram region is bound by these two edges. The parallelogram is determined by several intensity based function. The second detector, the *Intensity-based region (IBR)* detector, is an intensity extrema based algorithm. It

IBR

investigates the intensity profiles along rays going out of the local extremum. An ellipse is fitted to the region determined by significant changes in the intensity profiles.

DOG The *Difference-of-Gaussian (DOG)* detector was introduced by Lowe as keypoint localization method for the SIFT approach [91, 92]. Interest points are identified at peaks (local maxima and minima) of Gaussian function applied in scale space. All keypoints with low contrast or keypoints that are localized at edges are eliminated using a Laplacian function.

Limitations A common criticism to edge-based methods is that it is more sensitive to noise and changes in neighboring texture. Interest point detectors which are less sensitive to changes in texture perform well in a classification scenario since they recognize and capture those features that are common for all instances in a given class. On the opposite, identification relies on those features that are unique for a given object. Thus, the question arises: *How can we measure the distinctiveness of an interest point?* Schmid et al. defines information content as measure of the distinctiveness of an interest point [129]. It is based on the characteristics of the local shape of the image at the interest point. If all descriptors lie close together, they do not convey any information, that is the information content is low. Thus, matching would fail since any point can be matched to many others. However, we cannot simply ignore features with low information content. For example, in a shoeprint or coin classification scenario, there is a large number with similar (or even equal) local appearance. In spite of the low (local) information content, single descriptors can play an essential role in the global context. The main limitation of local features turns out to be their locality. Moreover, methods which detect most interest points do not necessarily perform the best. First, we are faced with the problem of overfitting (i.e. a single interest point can be matched to many others). Second, since misleading features may appear (e.g. due to background changes), the information captured per interest point plays an essential role. Thus, in the next section we give a short overview of state-of-the-art local feature descriptors.

3.2.2 *Local Descriptors*

Given a set of interest points, the next step is to choose the most appropriate descriptor to capture the characteristics of a provided region. Different descriptors emphasize different image properties such as intensity, edges or texture. In the following, we focus on four descriptors which show outstanding performance with respect to changes in illumination, scale, rotation and blur. For a thorough survey on the performance of local feature descriptors please refer to [102].

SIFT Lowe introduced the *SIFT* descriptor which is based on gradient distribution in salient regions – at each feature location, an orientation is selected by determining the peak of the histogram of local image gradient orientations [92]. Subpixel image location, scale and orientation are associated with each SIFT feature vector (4 × 4 location grid × 8 gradient orientations). SIFT features show outstanding performance in existing evaluations. However the main drawback and critical point is their computation time. Two SIFT modifications – Fast Approximated SIFT and Principal Components Analysis (PCA)-SIFT – claim on approximating or even outperforming accuracy and faster matching. The *Fast Approximated SIFT* descriptor is accelerated by the use of an integral orientation histogram [57]. The authors evaluate their approach on a data set of fifteen images and report a speed up by a factor of eight while the matching and repeatability performance is slightly decreased. PCA-SIFT aims at the reduction of the descriptor dimensionality by applying a PCA to the scale-normalized gradient patches [72]. Presented evaluation results on the INRIA Graffiti data set show that the PCA-SIFT descriptor performs slightly worse than SIFT. Finally, since the SIFT descriptor is not invariant to light color changes, several color-based SIFT descriptors have been introduced recently [2, 19, 23, 54, 157, 159]. Van de Sande et al. present a complementary evaluation on the performance of several color-based features in respect to light intensity changes and light intensity shifts [158]. The authors show that SIFT-based descriptors outperform histogram- and moment-based features on both image and video recognition. In the presented evaluations, the most promising descriptor for the task of category recognition without any a priori knowledge is *OpponentSIFT*. OpponentSIFT describes all of the channels in the opponent color space using SIFT descriptors.

GLOH

Mikolajczyk and Schmid propose another extension of the SIFT descriptor – *Gradient Location and Orientation Histogram (GLOH)* – designed to increase the robustness and distinctiveness of the SIFT descriptor [102]. Instead of dividing the path around the interest points into a 4 × 4 grid, the authors divide it into a radial and angular grid. A log-polar location grid with 3 bins in radial and 8 bins in angular directions is used. The gradient orientations are quantized into 16 bins which gives a 272 bin histogram further reduced in size using PCA to 128 feature vector dimension.

Shape context

Belongie et al. introduce *Shape Context* as feature descriptor for shape matching and object recognition [15]. The authors represent the shape of an object by a discrete set of points sampled from its internal or external boundaries. Starting points for the presented approach are edge pixels as found by the Canny detector [25]. Following, for each point the relative location of the remaining points is accumulated in a coarse log-polar histogram. Mikolajczyk et al. extend the original shape context to capture orientations [102]. Hence, the obtained feature vector is of dimension 36 (location is quantized into 9 bins and orientation into 4 bins).

SURF

Speeded Up Robust Features (SURF) are fast scale and rotation invariant features [13]. The descriptor captures distributions of Haar-wavelet responses within the neighborhood of an interest point. Each feature descriptor has only 64 dimensions which results in fast computation and comparison (4 × 4 location grid × 4 wavelet responses in horizontal and vertical direction).

Evalluations

Schmid et al. performed a complementary evaluation on the performance of local descriptors with respect to rotation, scale, illumination, and viewpoint change, image blur and Joint Photographic Experts Group (JPEG) compression [102]. In most of the tests SIFT and GLOH clearly outperformed the remaining descriptors: shape context, steerable filters, PCA-SIFT, differential invariants, spin images, complex filters, and moment invariants. Furthermore, Stark and Schiele report that the combination of Hessian-Laplace detector with SIFT and GLOH descriptor outperform local features such as Geometric Blur, k-Adjacent Segments and shape context in an object categorization scenario [143] . For their evaluation the authors used three different data sets containing quite distinguishable objects such as cup, fork, hammer, knife, etc. On the contrary, we performed an evaluation of the classification and identification power of various combinations of interest point detectors and local feature descriptors on a dataset of ancient coins bearing large intra- and small interclass variations [71]. The achieved results show the overall outstanding performance of DOG and SIFT in both identification and classification tests closely followed by Hessian-based detectors in a combination with the Shape Context descriptor. The main critical point of the SIFT features is their computation time. However, since feasibility and not time is the focus of the experiments performed within the scope of this work, we apply SIFT as thecurrently most reliable local feature.

Spatio-temporal features

More recently, spatial local features have been extended to the temporal dimension. Spatio-temporal features find application in event and action recognition [80, 115, 163], video summarization [79], video signatures and video copy detection [82, 85, 164]. Laptev extends the scale-invariant Harris-Laplace interest point detector by a 3 × 3 spatio-temporal second-moment matrix [80]. The author applies an iterative method to find points that are both spatial maxima of the Harris corner detector [61] and extrema of a scale-normalized Laplacian function in space. Recently, Leon et al. proposed a method for video signatures based on video tomography [85]. Video tomography captures spatio-temporal changes and is, thus, a measure for local and global motion in videos. A tomography image is composed by taking a fixed line from each frame, e.g. at the center of the frame, and arranging them from top to bottom to create an image (shot signature). Law-To et al. report outstanding performance of differential-based descriptors on video copy detection in comparison to ordinal intensity signatures and Laptev's spatio-temporal features [83]. The differential descriptors are computed in three steps [82]. First, keypoints are identified using the Harris corner detector [61] in each frame of a shot. Following, local features are computed at four spatial positions around the keypoints as Gaussian differential decomposition of the grey-level signal until the second order:

$$f = \left(\frac{\delta I}{\delta x}, \frac{\delta I}{\delta y}, \frac{\delta I}{\delta xy}, \frac{\delta^2 I}{\delta x^2}, \frac{\delta^2 I}{\delta y} \right). \tag{3.1}$$

The resulting feature vector is a 20-dimensional descriptor. Finally, keypoints along frames are associated to build trajectories that are represented by the average descriptors. Additionally, the authors assign a behavior label to each trajectory and distinguish between moving and motionless points.

3.2.3 *Matching Strategies*

Distance measures

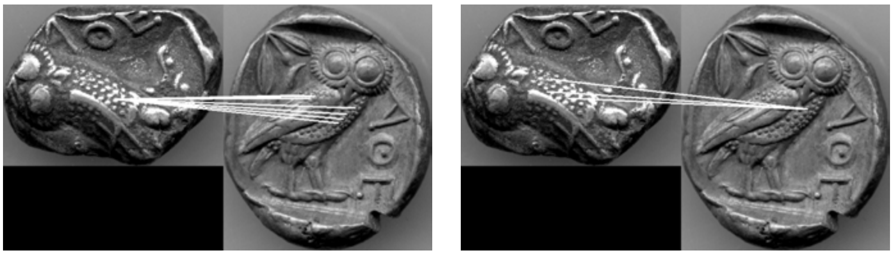
The *Euclidean distance* is a widely used measure to determine whether two feature vectors belong to the same keypoint in different images [13, 72, 92, 102, 103]. Other techniques, such as the *Mahalanobis distance*, can also be applied [12, 47, 100, 101, 102, 150]). However, the Mahalanobis distance bears mainly two disadvantages. First, it uses a single covariance matrix to all interest points. Second, it is based on the assumption that the errors of the descriptors should follow a normal distribution; an assumption that is verified neither theoretically nor experimentally. Terasawa et al. show in experimental tests that the distribution of rotation invariant descriptors is not a normal distribution [150].

Threshold-based strategy

Given a set of distances between corresponding keypoints, a very simple matching strategy is to accept all matches above a pre-defined threshold: *threshold-based matching* (e.g. [102]). Adjusting the threshold allows for the selection of appropriate trade-off between false positives and false negatives matches. Essential characteristic of this approach is that a descriptor can have several matches. Figure 3.5 illustrates the challenge. Given the two ancient greek coins, the owl pictured bears similar characteristics on both its breast and back. The provided descriptor results in a total of 9 matches, all of which are locally correct. However, there is only one correct match in the global context. Furthermore, multiple descriptors in the training image can be matched against the same descriptor from the test image. The misleading high number of matching descriptors penalizes recognition algorithms which rely on the total number of matches. Thus, a further processing step, such as cross correlation [101, 136], is required to increase the reliability of detected matches.

Nearest neighbor-based strategies

Nearest neighbor matching is a strategy to limit the number of possible matches. A descriptor is matched against its nearest neighbor if the



(a) single descriptor from the test image is matched against multiple descriptors from the training image, and ...

(b) ... visa versa [90].

Figure 3.5: Multiple matching of local features.

distance between them is below a given threshold [72, 102, 103]. *Nearest neighbor ratio matching* considers additionally the distance to the second nearest neighbor [12, 13, 92, 102]. Again, there can be multiple matches when different descriptors from the test image are matched against the same descriptor from the training image (see for example Figure 3.5b). To overcome this problem, one can either ignore all ambiguous matches (e.g. [136]) or keep the one with lowest distance. Both approaches lead to a loss of potentially stable matches. More recent matching strategies use methods of geometric filtering based on the local spatial arrangements of the matched descriptors [101, 128, 136], or on multiple view geometric relations [47]. Limiting assumption of this strategy is that model and test image follow homography or epipolar transformation.

3.3 MOTION FEATURES

Motion analysis comprises a set of research tasks of diverse complexity. While the detection of present motion in a temporal sequence of images or a video sequence is the simplest task, the distinction between camera and object motion is a challenging and yet not a reliably solvable task.

Motion content feature

Recently, a feature describing the motion content of a shot based on histogram intersection has been applied for shot similarity measurement [124, 175]. The motion content feature MF for a shot s is defined as:

$$\mathrm{MF}_s = \frac{1}{N}\sum_{f=1}^{N}(1 - \mathrm{HistInter}(f, f+1)), \tag{3.2}$$

where N is the number of frames in the shot and the histogram intersection $\mathrm{HistInter}$ for any two frames (or still images) x and y is defined as [146]:

$$\mathrm{HistInter}(x, y) = \sum_{b \in bins} \min(H_x(b), H_y(b)). \tag{3.3}$$

In general, the most basic motion detection method is based on frame subtraction (given a stationary camera position and constant illumination). The difference frame f_d for two frames at time point s and t can be defined as binary image with non-zeros elements representing areas with substantial motion:

Differential method

$$f_d(i,j) = \begin{cases} 0 & \text{if } |f_s(i,j) - f_t(i,j)| \leqslant \epsilon \\ 1 & \text{otherwise,} \end{cases} \tag{3.4}$$

where ϵ is a predefined threshold (small positive number) [142]. To gather information about the direction of motion a cumulative difference frame can be constructed:

$$f_d^{cum}(i,j) = \sum_{k=1}^{n} a_k |f_1(i,j) - f_k(i,j)|, \tag{3.5}$$

where a_k is the significance weight for the corresponding frame k. While a difference frame carries information about motion presence, extractable motion characteristics are not very reliable [142].

Optical flow

Another approach for motion analysis is the optical flow computation [22, 63, 93]. Optical flow aims at the establishment of the motion field[3] for a given video sequence and results in the determination of motion direction and motion velocity at (possibly) all image points. In general, optical flow computation is based on the assumptions of constant illumination and relatively small motion between consecutive frames. Furthermore, an accurate optical flow estimation is computationally expensive [22]. A possible solution for these limitations is the determination of a sparse motion field by means of tracking of interest points instead of all image points.

[3] A *motion field* is the (ideal) two-dimensional representation of three-dimensional motion where each image point is assigned a velocity vector [142].

Part II

FILM ANALYSIS

4

VISUAL-BASED COMPUTATIONAL MEDIA AESTHETICS

It is frequently at the edges of things
that we learn most about the middle ...

— Walter Murch [110]

Aim of the chapter

The first part of this work provided a basic understanding of the elements of media aesthetics in filmmaking that most influence the production and perception of a movie. Additionally, we discussed these elements from a visual perception point of view and presented existing approaches of computer vision for their analysis and representation. In this chapter, we present a direct linkage/mapping between the previously addressed media aesthetic elements, their application in filmmaking by means of well-established film techniques, and existing approaches in computer vision. This view of media analysis allows for the exploration and identification of three areas in the domain of automated film analysis and understanding. The first area comprises research tasks that have been subject to thorough research in the last years such as scene recognition or that have even been declared as solved by the research community such as the detection of shot cuts. The second area covers research tasks that are not directly solvable for a fully automated computer vision approach. Such tasks require additional knowledge, e.g. about the shape and structure of an object for the detection of the camera angle or about the normal movements of an object for the distinction between fast and slow motion. Finally, the third area identifies research tasks that are still open in the context of automated film analysis and understanding. The exploration of the last area and the discussion of initial approaches for the presented research tasks is the main focus of this chapter.

While film analysis and media understanding is still in the early stages of development, the focus of existing approaches slowly shifts from basic research tasks such as the identification of meaningful video clips for digital preservation to advanced film analysis such as scene and event detection, genre and emotion recognition, etc. Currently, semantics is explored mostly by means of incorporated metadata (e.g. title, synopsis, script, manually generated metadata, etc.) [56, 168] or available domain knowledge (e.g. events in sport videos or the fact that a scream belongs to a horror or an action scene rather than to a romantic one) [151, 165]. Such information originates in the final product, i.e. the study of the characteristics of a particular domain (e.g. news or sport videos) or the comparison amongst different film types (e.g. comedy and horror movies). Following, while existing approaches predominantly explore the question what can we learn from the film as finished product, we aim at the study of the origins of a film and the filmmaking process as a source for high-level information and semantics.

4.1 FILM FACTOR SPACE

As already discussed in Chapter 2, it is the combination of and interaction among manifold media aesthetic elements that exert influence on audience and film critics. The (way of) story telling is significant for how well a movie is perceived by the audience. In general, media aesthetic elements can be grouped into three categories according to their origins in the filmmaking process:

1. *(Pre-)production.* The selection of actors as well as locations for filming is part of the pre-production phase. Actors influence strongly the perception and attendance of a movie. Both presence (e.g. favorite actor) and play (e.g. great/weak performance) are important aspects most audience is intuitively aware of. During the production (or shooting) stage each scene is usually shot multiple times in slightly different versions which results in many thousands of feet of exposed film. During shooting, many people contribute to the shaping of the film such as the script supervisor (continuity from shot to shot), photography director (lighting and camera techniques), sound mixer, visual-effects supervisor, etc. [18].

2. *Post-production (Editing).* Post-production describes the process of editing of previously recorded material, including the use of special effects and audio dubbing [35]. Usually, this stage in the filmmaking does not happen independently but rather parallel to the production itself. In this way, adjustments or additional shots can be undertaken during the production phase. The goals of the editor is to find the rhythm of the movie, to create a narrative continuity and arrange it in a way that will create the dramatic emphasis so that the film will be effective. Editing decisions range from straightforward presentation of material to the alteration of the meaning of that material.

3. *Technology.* Three aspects of the technology employed in filmmaking can be distinguished. *First*, technology is present in every aspect of filmmaking, e.g. camera, lights, editing and audio recording equipment, etc. *Second*, the target device (e.g. projector or TV) influences e.g. the visual composition by setting limits on the frame size. *Third*, the medium (film vs. video) may influence the perception of the visual quality of a movie.

Since no film effect or technique is the product of a single media aesthetic element but rather the combination of and interaction between various factors, the presented categories construct a triangular space where the position of an element indicates the contributing factors. In Figure 4.1 we show the distribution of some film techniques[1]. *Continuity*, for example, refers to the coherence and clearness of the story [14, 18] (for details see Section 2.2.2). The narrative continuity of the story line as well as the general setup (e.g. lighting, camera position, actors' appearance, etc.) are subjects of the production stage. However, the editing phase ensures the smooth flow of space, time and action over a series of shots using e.g. an establishing shot, eyeline match, shot/reverse-shot pattern, etc. A *motif* describes a recurring element that gains in dramatic importance through its repetition [14]. A motif can be some visual component (e.g. color, character or object), action (motion pattern), or sound (for details see Section 2.2.3). Again, it is the combination of two factors that raises a film element to a motif: the production has to envision for the motif, and the post-production ensures the rhythm of recurrence.

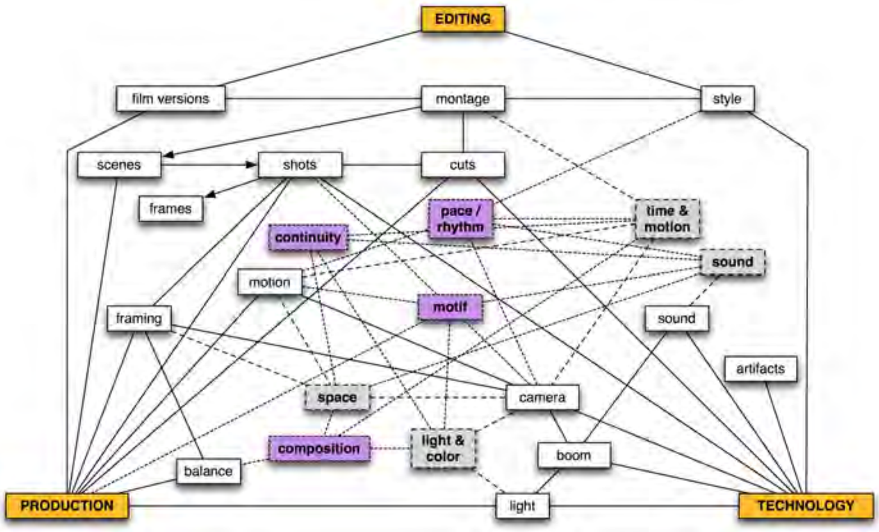


Figure 4.1: Film factors space: gray boxes indicate fundamental media elements, purple boxes advanced media concepts.

In the following sections, we will discuss a possible mapping between the presented media aesthetic elements in filmmaking, their application by means of well-established film and editing techniques, and feasible approaches in computer vision for each element in detail. Figure 4.2 summarizes the color conventions used for the mapping graphs in the following sections[2].

[1] Please note, that we do not aim at a comprehensive coverage of existing film techniques but rather at a simple and clear visualization of the film factor space.

[2] Please note, that the distinction between *visual*- and *motion*-based approaches is based on the time component of the corresponding features, e.g. static vs. motion features.

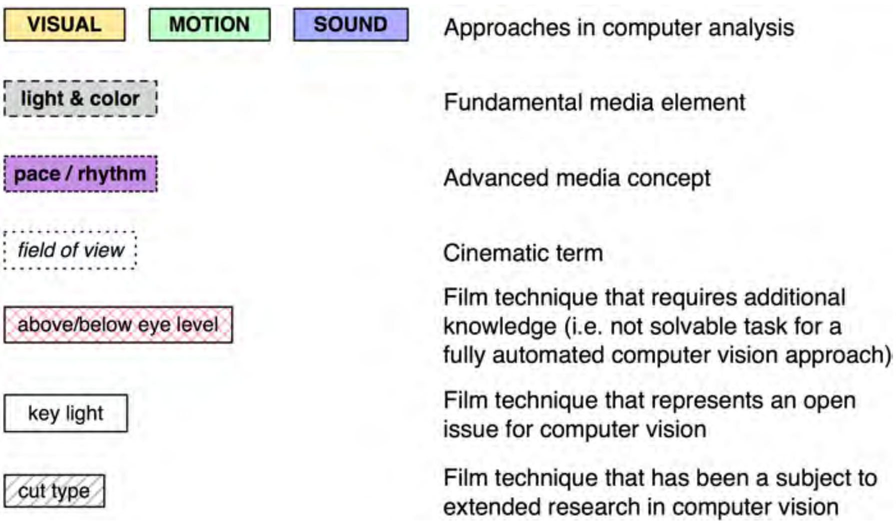

VISUAL
MOTION
SOUND
Approaches in computer analysis
light & color
Fundamental media element
pace / rhythm
Advanced media concept
field of view
Cinematic term
above/below eye level
Film technique that requires additional knowledge (i.e. not solvable task for a fully automated computer vision approach)
key light
Film technique that represents an open issue for computer vision
cut type
Film technique that has been a subject to extended research in computer vision

Figure 4.2: Mapping legend.

4.2 fundamental media elements

4.2.1 *Light and Color*

The following Figure 4.3 depicts the mapping between color and light as basic media elements and computer vision approaches (highlighted in yellow) that can be applied for the automated detection of the corresponding film editing technique.

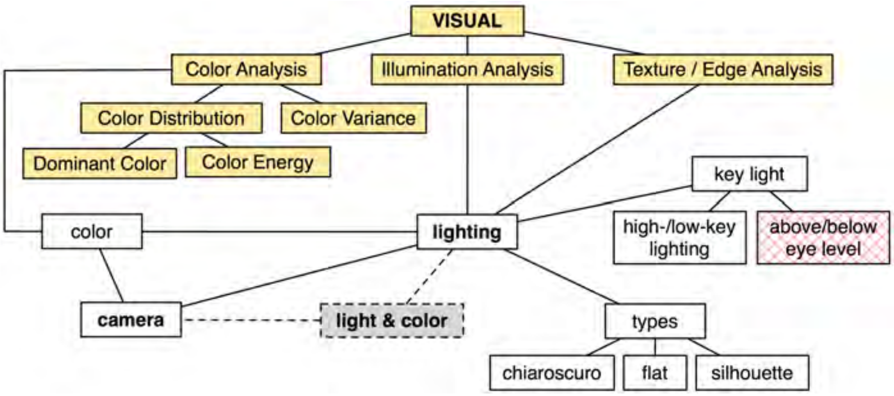
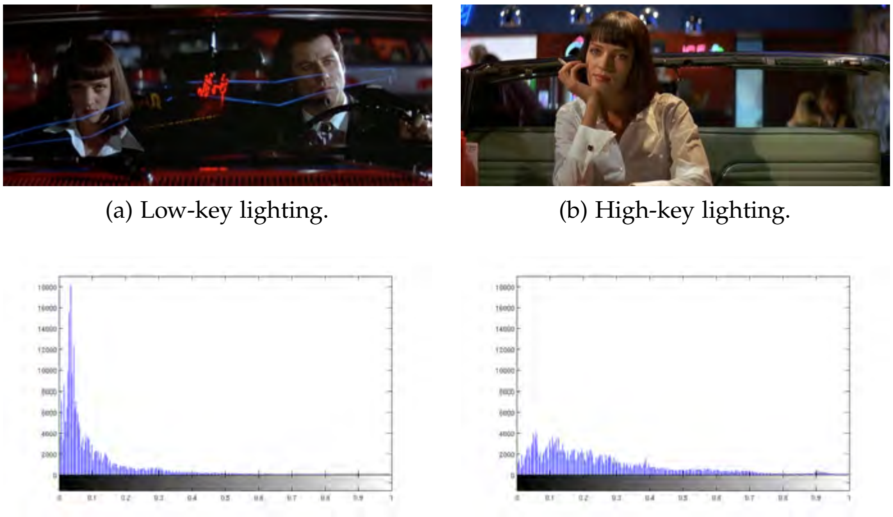


Figure 4.3: Media element vs. computer vision: Light and Color.

Light level

As already discussed in Section 2.1.1 light is mostly manipulated as the interaction of light and shadows for the purpose of orientation and to establish a specific mood. The overall light level of a scene or a frame can be easily measured by the corresponding intensity histogram (see for an example Figure 4.4). Rasheed et al. apply the product of mean and variance of an intensity histogram to distinguish between *low-* and *high-key lighting* [125]. Similarly, mean and standard deviation are applied in [141, 166]. The authors essentially assume that low-key lighting results in lower mean and variance values in comparison to high-key lighting. In fact, the variance in a low-key frame is lower due to a more balanced distribution of the illumination values. However, the mean is sensitive to outliers and may be an unreliable indicator for the light level (note the identical mean values in Figures 4.4c-4.4d). To overcome this limitation Wang et al. propose the use of the median (as an indicator for the general level of brightness) and the proportion of shadow area (dark pixels with values below a predefined threshold of 0.18) within a frame [161].



(a) Low-key lighting.

(b) High-key lighting.

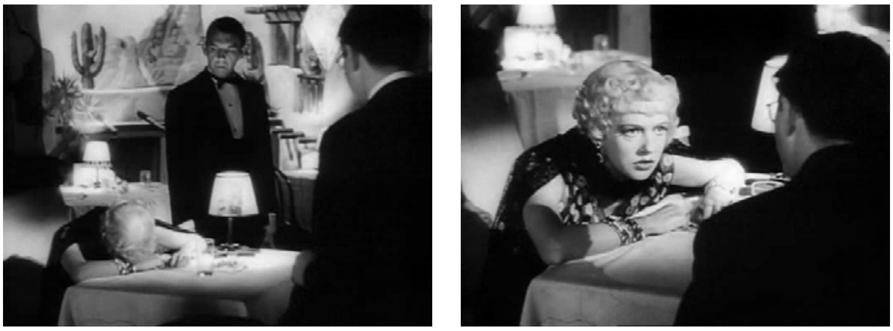
(c) 1×1 intensity histogram showing predominantly dark regions (mean:855; variance: 4.9053e+06; median: 107.5; shadow area: 0.85).

(d) 1×1 intensity histogram showing more balanced intensity distribution (mean: 855; variance: 8.1768e+05; median: 454.5; shadow area: 0.47).

Figure 4.4: Examples for high- and low-key lighting. First row: frames from *Pulp Fiction* (1994). Second row: corresponding intensity histograms.

Light direction

The detection of the position of the main light source in a scene, i.e. *above-/below-eye-level lighting* (see Figure 4.5), requires for a priori knowledge of the structure of the lighted objects in the scene. Thus, a fully automated detection of the light direction is only possible if predefined models of the objects exist in the scene. However, the three main lighting techniques – flat, Chiaroscuro, and silhouette – are characterized by their illumination, light level, and the resultant shadow types and, thus, bear characteristics that can be generalized and automatically detected. *Flat lighting* has no exactly definable light direction which results in light shadows and an overall high light level (see Figure 4.6a). The corresponding intensity histogram shows a broad distribution of intensity values (see Figure 4.6d). Figure 4.6b shows an example for *Chiaroscuro lighting* (selective illumination, overall low light level, and distinct and dense shadows). A global intensity histogram (even if a very rough feature descriptor) indicates narrow intensity range, and thus, low-contrast (see Figure 4.6e). Furthermore, the low intensity values point at an overall low light level. Finally, *silhouette lighting* is characterized by its extreme of light and dark contrast and its accent on contour (see Figure 4.6c). The intensity histogram shows two peaks in the low and high regions of the intensity range and low distribution of the values in between (see Figure 4.6f). Since distinct contours and less textures are typical for the silhouette lighting, edge and texture analysis can be additionally applied to distinguish among the various lighting techniques (see Figures 4.6g-4.6i).



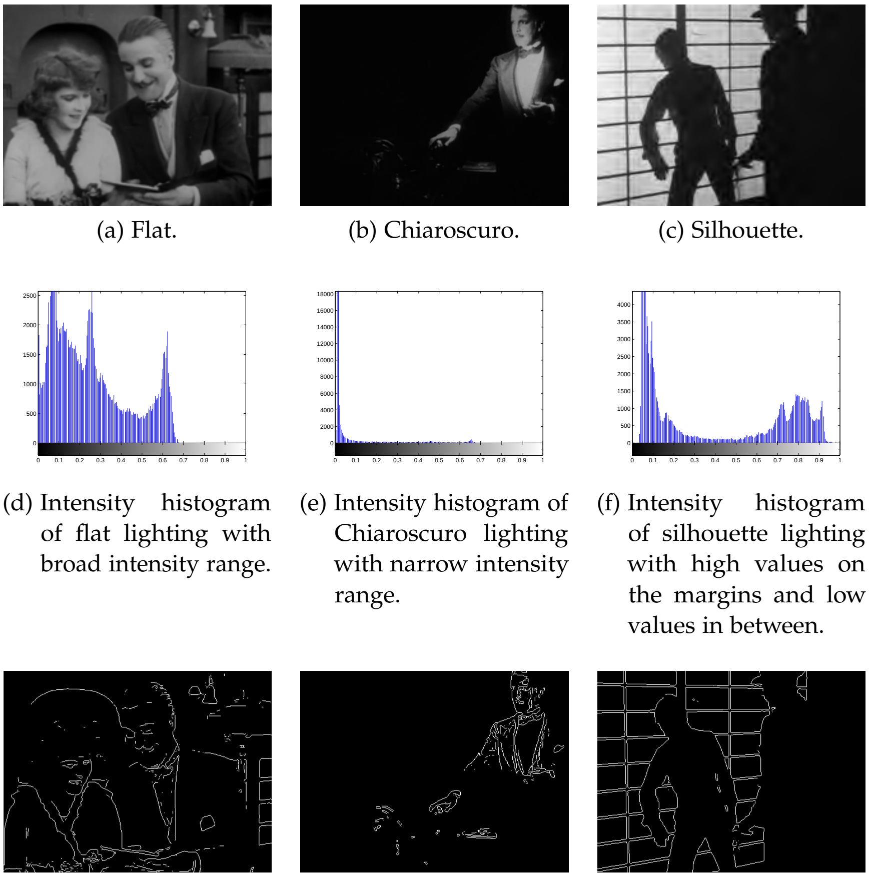
(a) Below-eye-level lighting.

(b) Above-eye-level lighting.

Figure 4.5: Examples for above-/below-eye-level lighting from *Citizen Kane* (1941). The main light source of the nightclub scene is the table lamp. The beginning with a long shot sets the leading characters (the waiter and the reporter Thompson) into below-eye-level lighting. Moving into the story and close up sets the camera focus on Kane's ex-wife Susan whose head is below the level of the table lamp.

Color analysis

Due to its strong dependency on context, color analysis is only feasible to a limited extent. Color-based retrieval (e.g. color-based mood analysis) presumes the color application by filmmakers in compliance with the subjective experience of the audience. This assumption fails for two reasons. Firstly, any filmmaker expresses his own creativity and may follow established conventions or contradict them to establish a certain feeling with the audience. Secondly, the subjectivity of color perception cannot be unified for the entire audience. Figure 4.7 shows four example scenes from *Run Lola Run* (1998) depicting Lola and her boyfriend Manni. All scenes from Lola and these from Manni show similar color distributions for the character itself despite different shooting settings and distinct color setting in comparison to the other character. Hence, such information can be used to retrieve related scenes. However, as already discussed in Section 3.1.1, different scenes can share the same color distributions even if they are not semantically related (see Figure 3.1). Thus, color can only be one of the features for a specific retrieval task (e.g. analysis within a single movie and possibly within the works of the same filmmaker) but not the sole solution.



(a) Flat.

(b) Chiaroscuro.

(c) Silhouette.

(d) Intensity histogram of flat lighting with broad intensity range.

(e) Intensity histogram of Chiaroscuro lighting with narrow intensity range.

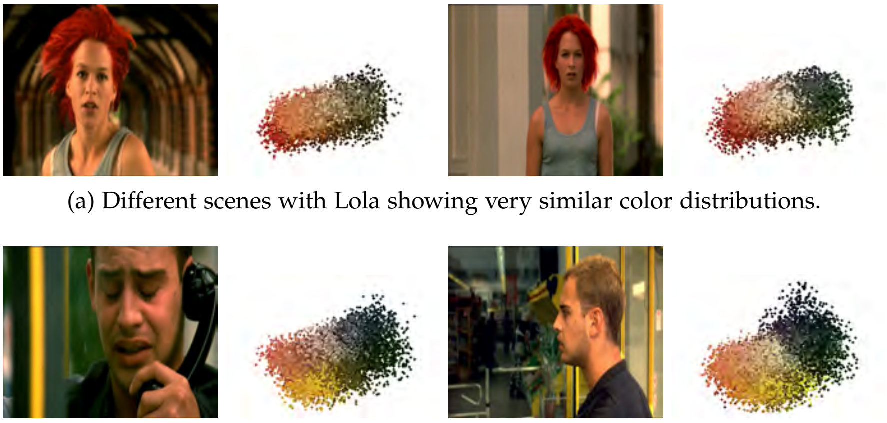
(f) Intensity histogram of silhouette lighting with high values on the margins and low values in between.

(g) Despite its texture, the frame has less distinct contours than a silhouette lighting.

(h) Due to the soft lighting, a Chiaroscuro frame shows less prominent contours.

(i) Distinct contours are typical for silhouette lighting.

Figure 4.6: Examples for lighting techniques. First row: frames from *The Cheat* (1915). Second row: corresponding intensity histograms. Third row: Sobel-based edge detection.



(a) Different scenes with Lola showing very similar color distributions.

(b) Similarly, scenes with her boyfriend, Manni, are (mostly) shot in a different color cloud space.

Figure 4.7: Examples for context-dependent color distributions in *Run Lola Run* (1998) [49].

Recently, three further color features are often applied in the context of content-based video retrieval, e.g. for the tasks of genre classification and affective film analysis: dominant color [149, 162], color variance [125, 141], and color energy [34, 161, 166]. Color variance represents the variety of color used in a video. For an example, a comedy is usually more colorful than a horror movie. Rasheed et al. employ color variance as the generalized variance of the *CIE Luv* color space [125]. Zettl defines color energy as "the relative aesthetic impact a color has on us" [179]. As contributing factors, Zettl identifies the hue, saturation and brightness attributes as well as the size of the color area and the relative contrast between background and foreground colors.

4.2.2 *Space*

Figure 4.8 shows both visual- (yellow highlighted) and motion-based (green highlighted) approaches for the detection and analysis of editing techniques focused on the arrangement of space within a scene.

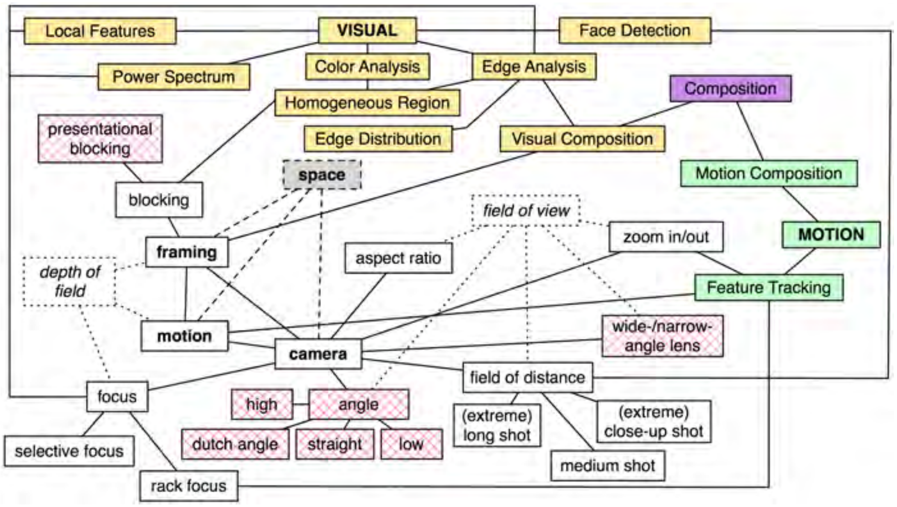


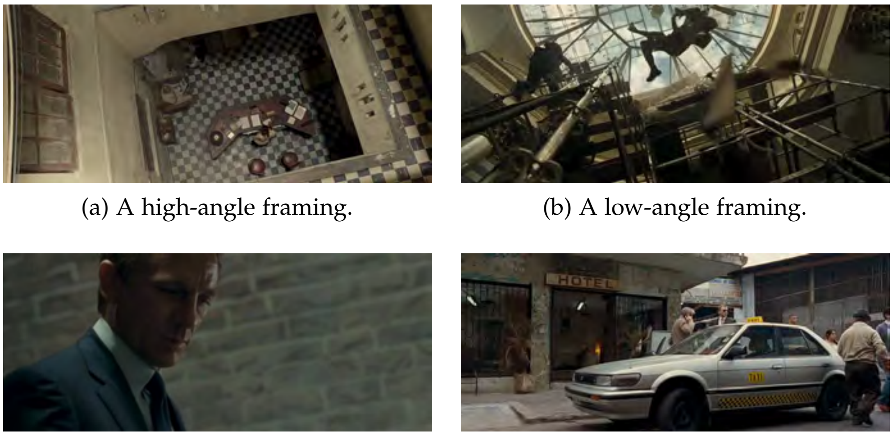
Figure 4.8: Media element vs. computer vision: Space.

Aspect ratio

As discussed in Section 2.1.2 the available space in a frame is first and foremost controlled by the dimension and shape of a frame whose determination is quite a trivial task in computer vision. Within the specific dimensions of a frame the camera angle and distance of shooting define the *field of view* of a scene.

Camera angle

The detection of camera angle (see Figure 4.9) poses an unsolvable task for computer vision. Humans make use of their experience in everyday life to recognize known and to estimate unknown objects, their size and position and, in result, the camera view of a particular scene. Despite recent advances in object modeling, camera angle detection cannot be performed reliably at this state of scientific and technical knowledge.



(a) A high-angle framing.

(b) A low-angle framing.

(c) A Dutch angle (canted framing).

(d) A straight-angle framing.

Figure 4.9: Examples for camera angles from *Quantum of Solace* (2008).

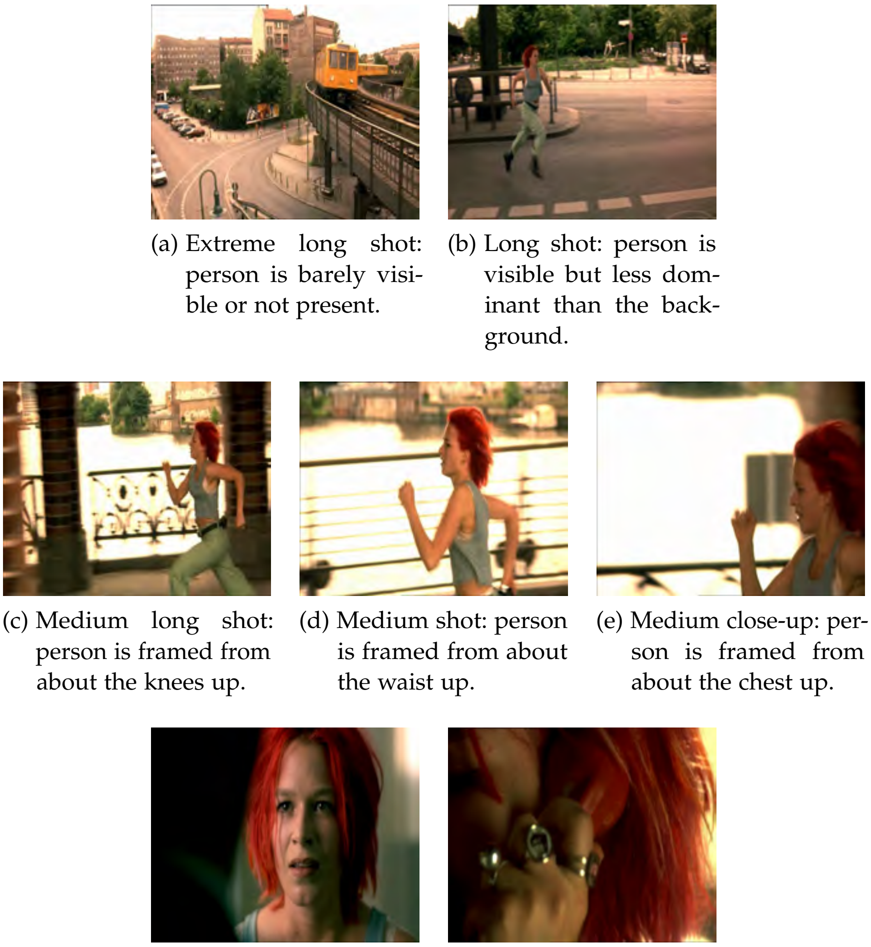
Field of distance

The distance from a camera to an object or a person is referred to as field of distance and results in different types of shots: close up shot, medium shot, long shot, etc. There are no unified and reliable definitions for the different shot types [179]. In this analysis we follow the definitions from Bordwell and Thompson [18] who use the human body as standard measure to distinguish among the different shot types (see for examples Figure 4.10). Recently, Cherif et. al proposed the use of the human body's proportions for the classification of shot types according to the camera distance [33]. Thus, the use of a reliable face detection algorithm may allow for an approximation of the size of a head and an estimation of the size of the body and its relation to the height of the frame. However, the application of the approach is restricted by the reliability of the underlying face detection algorithm. Common limitations of existing face detection methods are the sensitivity to lighting conditions, face orientation, size range, and/or partial occlusion.

The second key characteristic of space is the perception of *depth of field* affected by the arrangement and movements of objects, characters, and camera within a scene but also by the camera focusing mechanism.

Blocking

Blocking refers to the arrangement of characters and objects within a frame [14]. Despite the endless possibilities of characters and objects positioning, two blocking types bear an outstanding expressiveness. A strong sense of depth is created by the placement of a character (or an object) at the edge of a frame in the immediate foreground and the arrangement of further characters and objects in the mid- and background (see for an example Figure 4.11a). Figure 4.11b shows another illustration on how to effectively convey a message through a frame composition. The presentational blocking of a character, i.e. a character looking directly into the camera lens, creates a dynamic relationship with the audience. Furthermore, when a character actually addresses



(a) Extreme long shot: person is barely visible or not present.

(b) Long shot: person is visible but less dominant than the background.

(c) Medium long shot: person is framed from about the knees up.

(d) Medium shot: person is framed from about the waist up.

(e) Medium close-up: person is framed from about the chest up.

(f) Close-up shot: emphasis on head, hands, feet, etc..

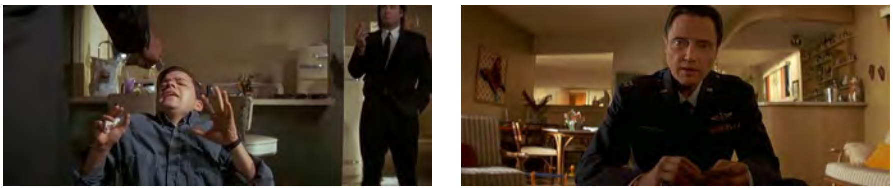
(g) Extreme close-up shot: emphasis on a part of a face or an object.

Figure 4.10: Examples for different shot types from *Run Lola Run* (1998).

the audience (by words or gestures), the audience is not a passive observer but an active participant of the story. While the blocking of a character can be detected using a combination of texture and color clustering approach (see Figure 4.12), the detection of presentational blocking is an unsolvable task due to the diversity in the appearance and the complex motion possibilities of head, eyes, and camera and the combination thereof.

Camera focus

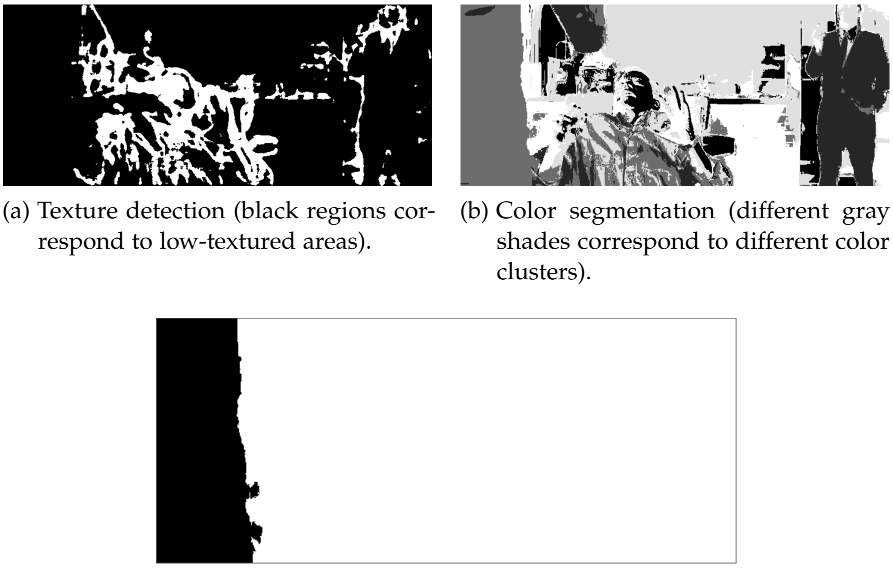
As discussed earlier, camera focusing mechanism is another essential property of a video camera that can influence the perception of space depth within a scene. Scenes that are shot using deep focus cinematography, i.e. all depth planes appear in sharp focus, may require some effort to distinguish specific depth planes. The use of selective focus allows for the choice of a single plane and lets the other planes blur [18]. Recent applications for focus/blur detection include primarily image restoration and quality measurements. Existing approaches



(a) Blocking of a character at the edge of a frame: Jules (on the left side of the frame, shown in back) points his gun at Brett (sitting on a table in the middle of the frame).

(b) Presentational blocking of a character: Captain Kooks (facing the camera) talking to young Butch (the camera is the kid's point of view).

Figure 4.11: Examples for blocking from *Pulp Fiction* (1994).



(a) Texture detection (black regions correspond to low-textured areas).

(b) Color segmentation (different gray shades correspond to different color clusters).

(c) Detected blocking region by the combination of texture and color information. The frame border is introduced for a better visualization and is not a part of the original frame.

Figure 4.12: An approach for the detection of a blocking character/object at an edge of a frame using the example of Figure 4.11a. For a region to be a blocking region, it has to be positioned at the edge of the frame, of similar color, low texture, and certain size.

can be distinguished in edge-based algorithms and methods based on power spectrum analysis. Edge-based approaches rely on the fact that in-focus regions bear higher contrast and more distinct edges than out-of-focus regions. Methods based on power spectrum show that in the Fourier domain camera defocusing results in the eduction of power in higher Fourier frequencies. Quality measurement, i.e. the classification of scenes in out-of-focus and in-focus, is less relevant to a high-level film analysis. However, the detection of selective and racking focus gives a clue about the applied camera techniques and the intended audience attention model. *Selective focus* refers to the camera focus on a single detail of the scene while *racking focus* describes the focus shifts within a scene (see Figures 4.13a-4.13b). A simple approach for the detection of racking focus is the combination of edge detection and local features. While firstly detected edges disappear and new edges appear with a focus shift (see Figures 4.13c-4.13d), blur-independent local features assure the consistency of the scene (see Figure 4.13e).

Finally, as discussed in Section 2.1.2, the perception of space is also influenced by the movements of characters and camera within a scene. Since the following Section 4.2.3 has a closer look into the subject of motion detection and analysis, we refrain from a further discussion at this point to avoid redundancy.

4.2.3 *Time and Motion*

The concepts of time and motion are notably correlating. Since motion is happening in time, the perception of time is strongly influenced by the presence of camera and/or object motion. Figure 4.14 shows both visual- (yellow highlighted) and motion-based (green highlighted) approaches for the detection and analysis of editing techniques focused on motion analysis and time manipulation within a scene.

Time manipulations

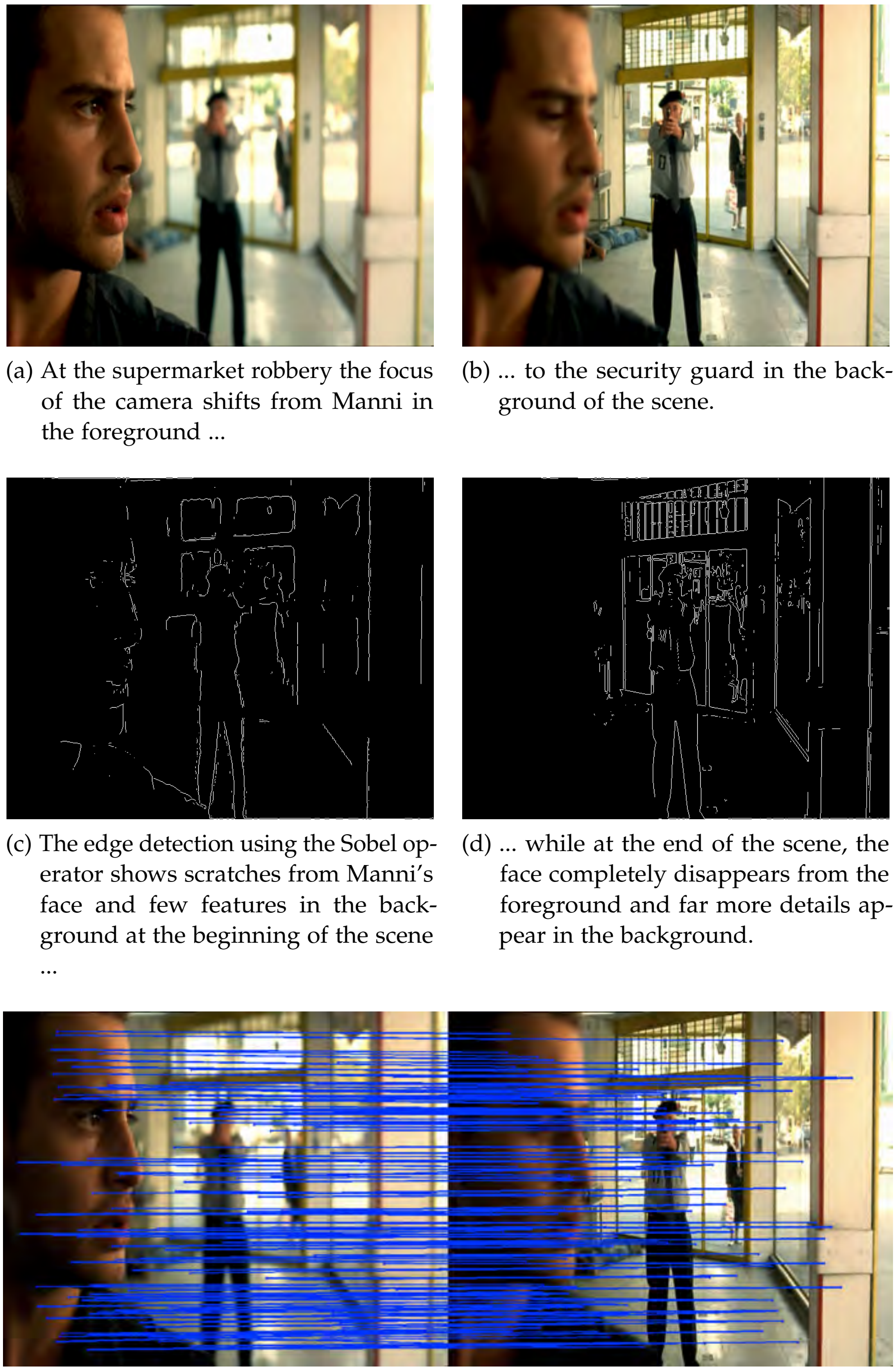
Depending on the pace of time, manipulations of time can be distinguished in accelerated and decelerated time perception (with respect to normal tempo of time passing), simultaneous time passing of two or more events, and time stops.

Time stops

Filmmakers can freeze the time for example to increase tension or to isolate and emphasize a dramatic moment within a sequence [14]. Such *time stops* can be communicated by the use of still (frame) sequences (also associated with the feeling of timelessness) and freeze frames. Both editing techniques imply the absence of motion. However, the use of *freeze frames* is probably more expressive since time stops while the scene is still in motion and full of life. Figure 4.15 shows an example for the application of freeze frames in the final scene from *Run Lola Run* (1999).

Acceleration / deceleration

There are different editing and camera techniques to suggest accelerated or decelerated action or tempo of time passing. One possibility is the manipulation of the amount of frames per second in the record-



(a) At the supermarket robbery the focus of the camera shifts from Manni in the foreground ...

(b) ... to the security guard in the background of the scene.

(c) The edge detection using the Sobel operator shows scratches from Manni's face and few features in the background at the beginning of the scene ...

(d) ... while at the end of the scene, the face completely disappears from the foreground and far more details appear in the background.

(e) Matched local features (SIFT) despite the blur effect.

Figure 4.13: An example for racking focus from *Run Lola Run* (1998).

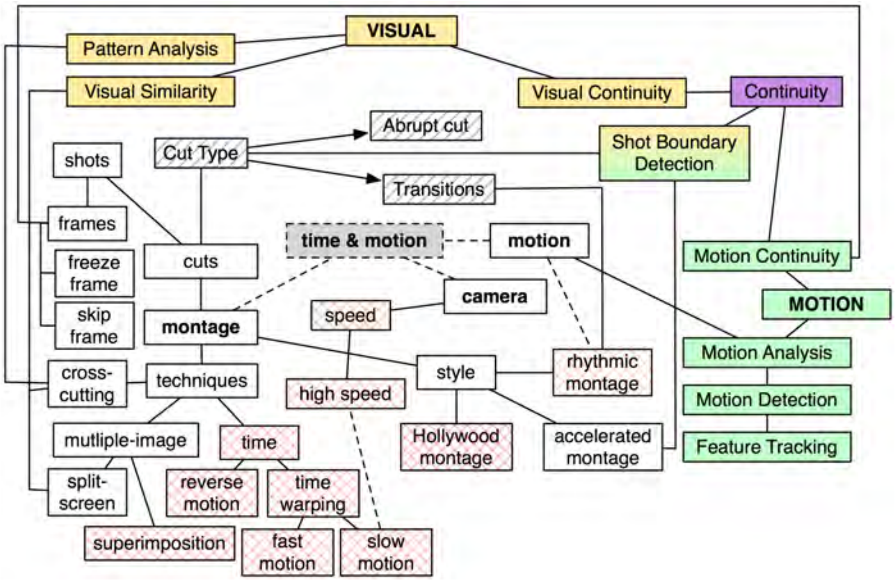


Figure 4.14: Media element vs. computer vision: Time and Motion.

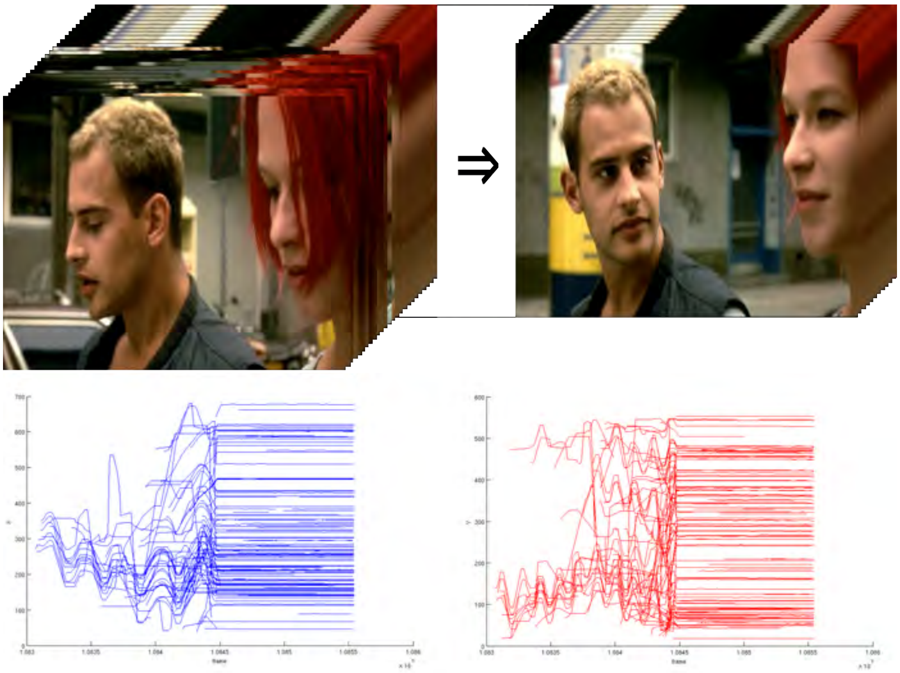


Figure 4.15: An example for the use of freeze frames in the final scene of *Run Lola Run* (1999). Top: Lola and Manni walk side by side. As Manni asks what's in the bag, the sequence freezes on the frame where the audience can see the smile on Lola's face. Bottom: x-rays of the motion trajectories of the keypoints. Both x- and y-direction reveal the point in time where all the motion suddenly stops.

ing and in the playback process. For example, the use of high speed cameras allows for the capturing of scenes which typically cannot be seen with naked eyes, e.g. breaking of glass or a falling drop. In general, at least 18 frames per second are required for the perception of smooth motion. While conventional movies are shot (and played back) at 24 to 25 frames per second, a high speed camera can acquire up to several thousands frames per second. When played at normal speed the scenes create the impression of slowed time and action. The recognition of such scenes requires high-level context understanding and experience and lies beyond the means of automated computer vision techniques.

Another technique to manipulate the perception of time is picture jogging. Picture jog results in a sequence which is not perceived as smooth motion but rather as a jerky sequence. The effect of picture jog can be caused not only by e.g. interrupted transmission but also by intended frame dropping leading to the perception of accelerated action. The detection of such *skip frames* requires for an analysis of the motion continuity and smoothness in consecutive frames or shots (see for an example Figure 4.16). Another editing technique to suggest in a brief period events occurring over a longer time span is the *Hollywood montage* [14]. It usually involves fast cutting and various optical effects conveying the essence of an event or the passage of time (e.g. newsreel) [95]. The detection of such montage sequences requires for semantical knowledge and, thus, cannot be performed fully automatically. In contrast, *accelerated montage* refers to continuously decreasing length of the shots in a given sequence which intensifies the effect of increasing speed of action [14]. While this editing technique can be easily detected by e.g. an analysis of the changes in the average shot length rather than counting the shot frames, manipulations of the presented speed (*fast/slow motion*) or direction of action (*reverse motion*) cannot be recognized automatically by methods of computer vision without a priori knowledge of the "normal" motion model.

Simultaneous events

Split-screen

To suggest the simultaneity of two (or more) events a filmmaker can divide the video frame into several areas showing different visual information. While, nowadays, the *split-screen* technique is mostly known from the TV news, it has its origins in filmmaking. A recent example of a movie that makes an extended use of split-screens is the TV series *24* (2001-2010). *24* demonstrates the conventional application of the split-screen technique: several active areas are divided by sharp boundaries (see Figures 4.17a-4.17b). These characteristics can be used for a visual-based detection of split screens. A possible workflow may include the detection of areas of independent motion and, following, an analysis of the boundaries between such areas (see Figures 4.17c-4.17d). Such an approach allows for the detection of split-screens despite their shape (e.g. rectangle, triangle, oval, etc.) and positioning (e.g. horizontal, vertical, diagonal, etc.). However, the pro-

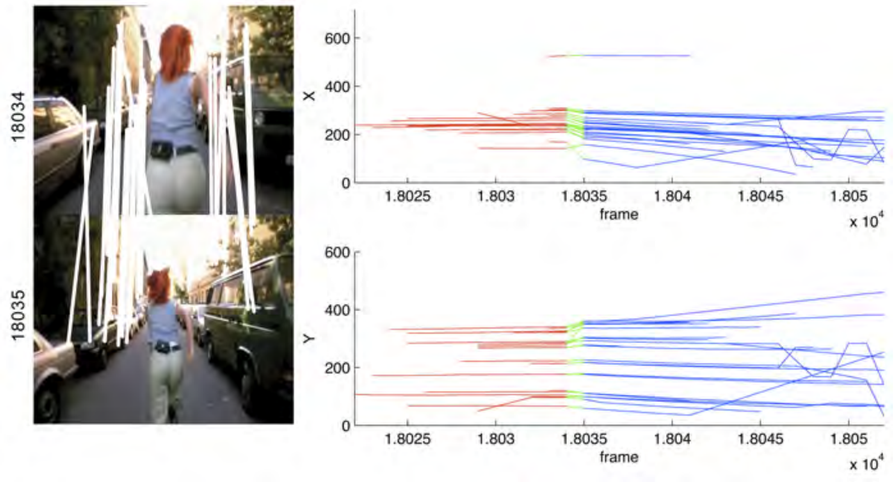


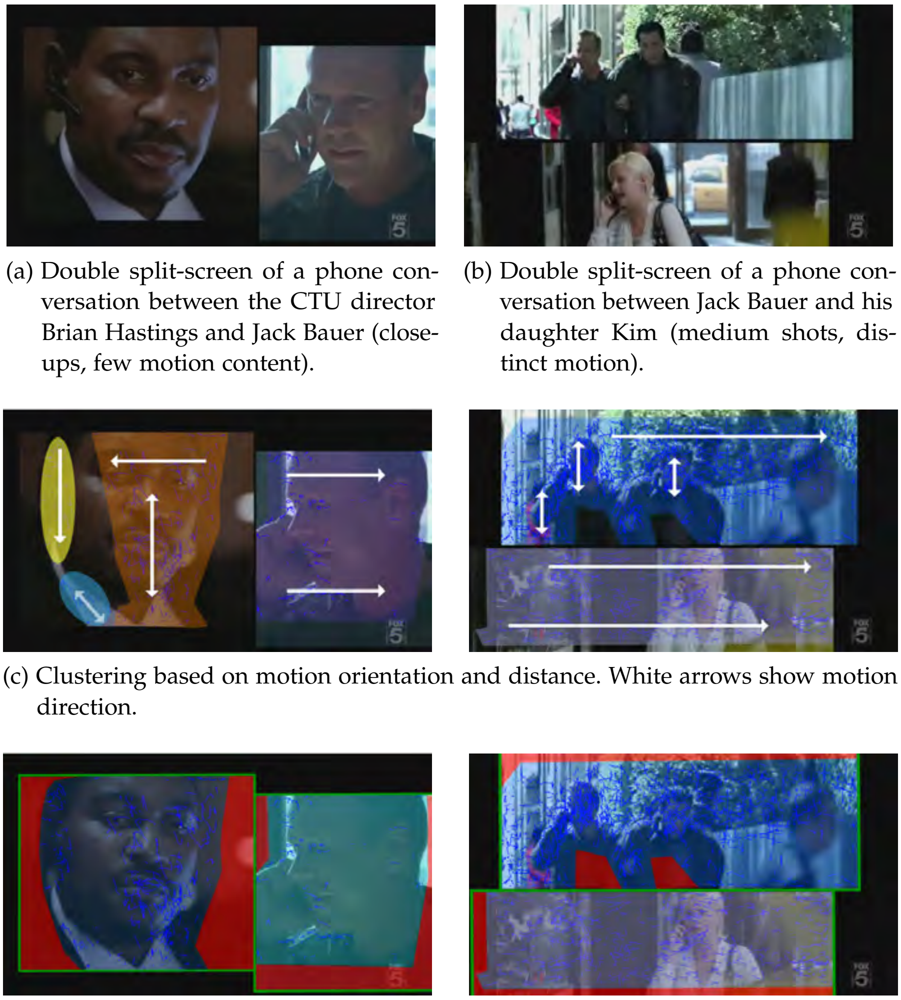
Figure 4.16: An example for skip frames in *Run Lola Run*. Left: matched keypoints in two consecutive frames. Right: motion trajectories of the keypoints (motion jump visualized in green). Before as well as after the presented time moment, the motion trajectories show predominantly smooth motion progress.

posed approach cannot deal with unconventional use of split-screens such as the the one applied in *Run Lola Run* (1998). Figure 4.18 shows two examples for split-screens with superimposed boundaries of the respective visual areas. Such superimposition does not allow for the detection of distinct boundaries between the areas. Since it is not unusual to have areas with independent motion in a conventional video shot, an automated approach that relies solely on the detection of such areas (i.e. without presented boundaries) leads to a high rate of falsely detected split-screens and is, thus, not feasible. Similarly, the detection of *superimposition* of two or more shots appearing in the same frame often requires for semantical knowledge and, thus, cannot be performed fully automatically. Such editing technique is a popular device in trick and experimental films (see for some examples Figure 4.19) [14].

Superimposition

Crosscutting

Another well-established film technique suggesting temporal simultaneity of two or more actions is *crosscutting* [14]. A typical example for such parallel editing is a crosscut between a captured victim in need and a person hurrying to save him/her. Crosscutting reveals high-level semantic information about cause and effect within a specific video sequence [18]. From a technical point of view, a crosscut sequence results in a pattern of the type $AB(A|B)^*$, where A and B are crosscut scenes. Following, crosscut detection can be performed based on shot similarity (resulting in a simple text sequence of the corresponding shot labels) and text-based pattern analysis. Finally, to distinguish crosscutting from further film editing techniques such as shot-reverse-shot as often used in dialog scenes, it is required to assure that the alternating shots, A and B, are visually highly dissimilar



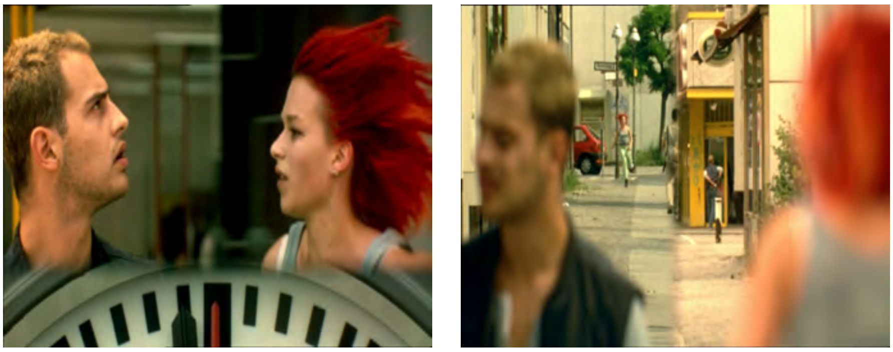
(a) Double split-screen of a phone conversation between the CTU director Brian Hastings and Jack Bauer (close-ups, few motion content).

(b) Double split-screen of a phone conversation between Jack Bauer and his daughter Kim (medium shots, distinct motion).

(c) Clustering based on motion orientation and distance. White arrows show motion direction.

(d) Clusters that are not separated by an uniform area are merged. Red color: non uniform area. Green color: uniform area surrounding a motion cluster.

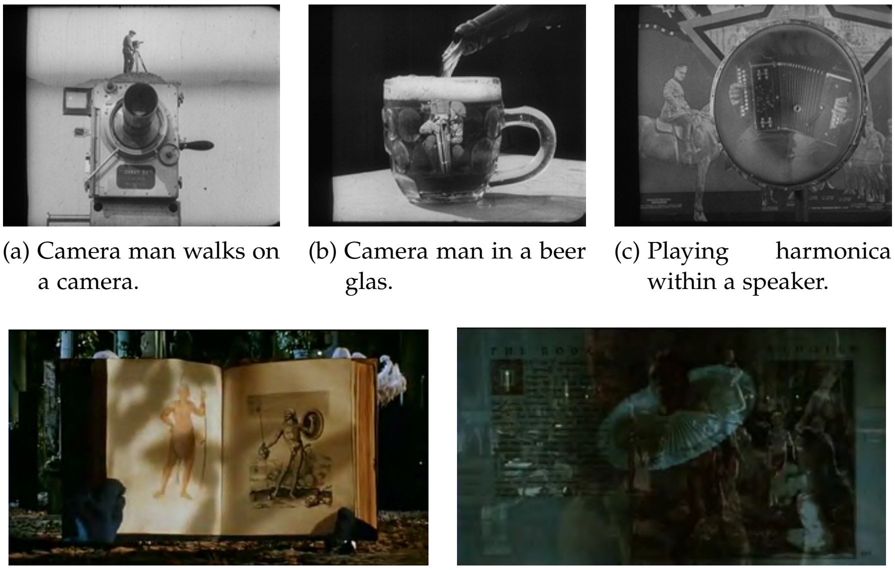
Figure 4.17: Examples for sharply defined split-screens from the TV series *24* (2001-2010).



(a) Triple split-screen. The two top screens show Manni looking at the watch (shown at the bottom line) and Lola running towards Manni.

(b) Double split-screen. Left: Manni in the foreground and Lola in the background. Right: the same scene from 180 degree rotated view.

Figure 4.18: Examples for split-screens in *Run Lola Run* (1998).



(a) Camera man walks on a camera.

(b) Camera man in a beer glas.

(c) Playing harmonica within a speaker.

(d) Double superimposition in the opening scene of the *Book of Utopias*.

(e) Triple superimposition in the opening scenes of the *Book of Mythologies*.

Figure 4.19: Examples for superimposition: 4.19a-4.19c *A Man with a Movie Camera* (1929); 4.19d-4.19e *Prospero's Books* (1991).

and, thus, are not different views of the same scene. Another decisive criterion can be the presence of an establishing shot (typical for dialog scenes).

Camera / object motion

In general, it is not possible to distinguish between camera and object motion reliably (see for an example Figure 4.20). However, from a film aesthetics point of view, the motion content of a sequence as a whole is more relevant than the differentiation between camera and object motion. The motion content of a shot (or a sequence of shots) plays a central role for several film techniques: it can create visual compositions and motifs, facilitates the creation of rhythmic montage and invisible editing (continuity) [14, 18]. All these aspects of motion will be discussed in the following section.

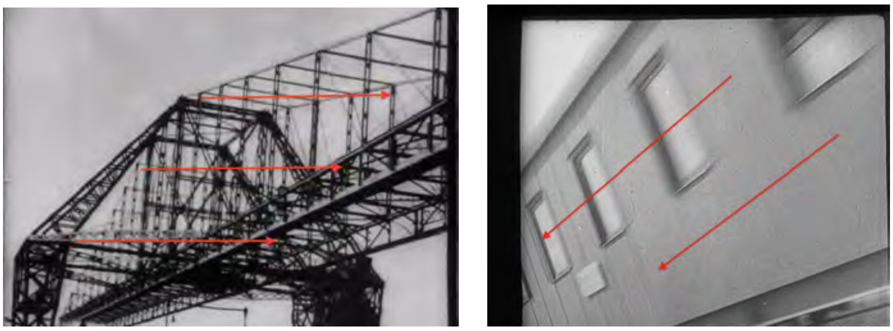
4.3 ADVANCED MEDIA CONCEPTS

4.3.1 *Composition*

The following Figure 4.21 shows the positioning of film composition within the media aesthetic elements as discussed in the previous section and maps the corresponding film editing techniques to basic computer vision approaches.

Framing direction

One of the first characteristics the audience perceives in a particular scene composition is the main direction of the framing. Figures 4.22-4.23 show examples for the four main types of framing direction: *horizontal* (associated with normalcy and rest), *vertical* (dynamic, ex-



(a) Camera passing through construction. *The Eleventh Year* (1928).

(b) Train passing the camera. *A Man with a Movie Camera* (1929).

Figure 4.20: Examples for large motion in a shot (red arrows show motion direction). Despite the similar characteristics, the two examples have different motion origins: 4.20a camera motion and 4.20b object motion.

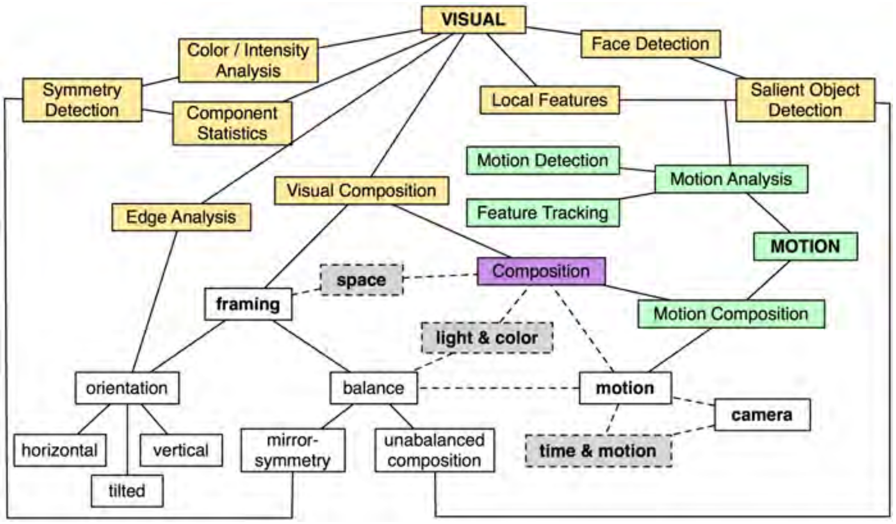
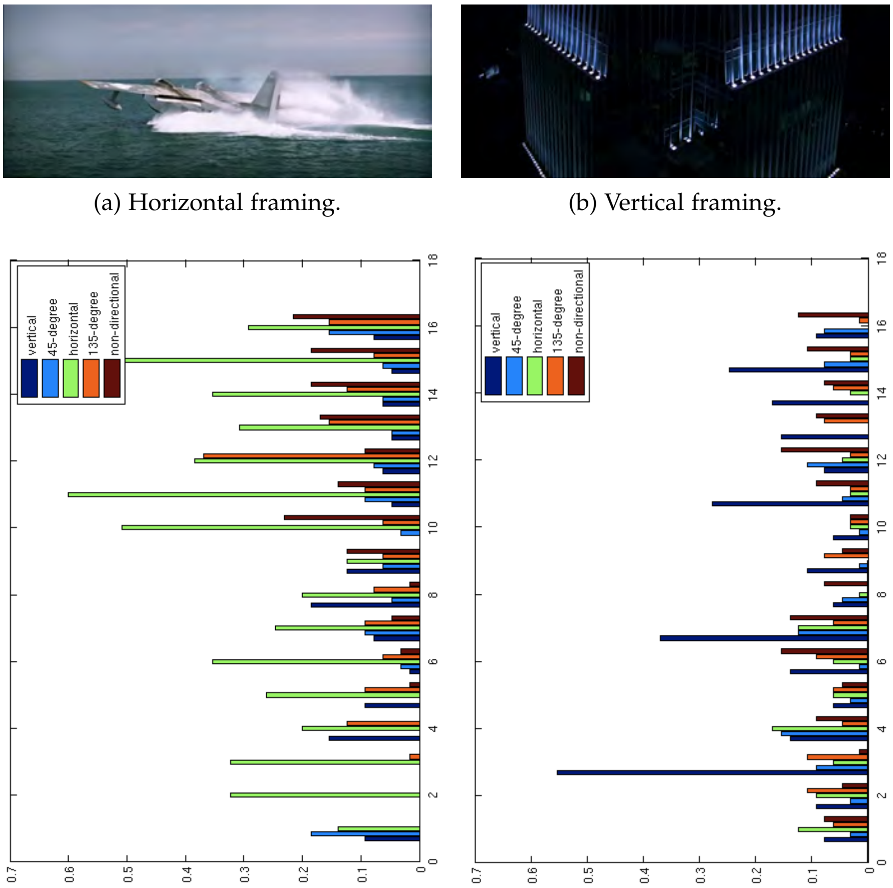


Figure 4.21: Media element vs. computer vision: Composition.

citement), *horizontal/vertical* (normalcy, reflects the everyday world), and *tilted* (disorientation, disturbance) [179]. The characteristics of the framing direction suggest an edge-based detection approach. However, preliminary evaluations show strong dependency on the overall frame composition. Both horizontal and vertical framing are usually clearly defined and easily distinguishable from other framing types. However, the mixed (horizontal/vertical) framing and tilted framing can be easily falsely classified due to misleading edges in the frame composition (see Figures 4.23c-4.23d).

Balance

As already discussed in Section 2.2.1 filmmakers usually try to balance the composition of a scene by distributing objects of interest evenly around the frame [18]. Pleasing compositions are often associated with the rule of thirds where the scene is divided horizontally and vertically into thirds (see for an example Figure 4.24) [179]. This



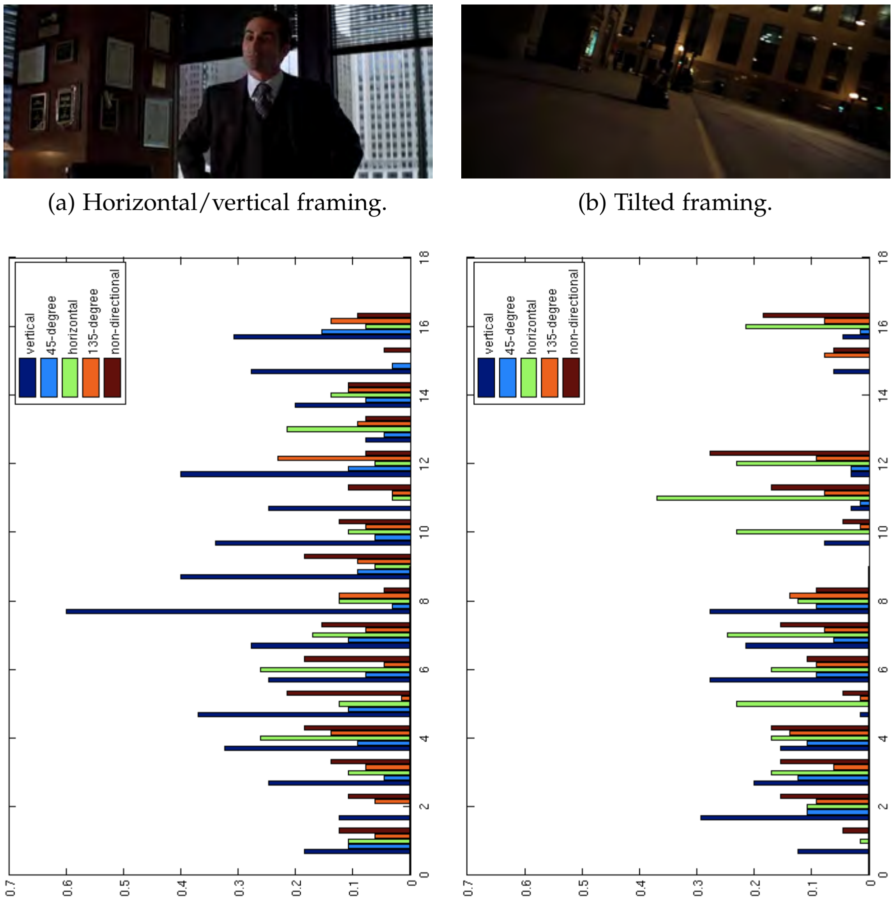
(a) Horizontal framing.

(b) Vertical framing.

(c) The MPEG-7 edge histogram of a horizontal-oriented frame shows predominantly horizontal edges in all areas of the image ...

(d) ... similar to the notedly dominating vertical edges in a vertical framing.

Figure 4.22: Examples for horizontal and vertical framing directions from *The DarkKnight* (2008) with the corresponding MPEG-7 edge histograms.



(a) Horizontal/vertical framing.

(b) Tilted framing.

(c) Despite the presence of all edge types in a horizontal/vertical framing, horizontal and/or vertical edges prevail in all areas of the image.

(d) Although a tilted camera suggests predominantly angled edges in the frame, the overall frame composition strongly influence the detection of edges (e.g. the strong vertical road line, almost horizontal line of the window frames in the back, etc.)

Figure 4.23: Examples for horizontal/vertical and tilted framing directions from *The DarkKnight* (2008) with the corresponding MPEG-7 edge histograms.

knowledge can be used for a coarse estimate of the balance in a scene by, for example, investigating the distribution of textures and colors in the corresponding thirds and in the intersection regions. Recently, Obrador et al. measured layout homogeneity of relevant regions to investigate the role of the overall visual balance in image aesthetics [114]. Since balanced shots are a common aesthetic desire and actually the rule in filmmaking, the detection of both extremes – near-perfect symmetry and unbalanced compositions – is a more appealing task for an automated retrieval system.

Symmetrical compositions

Despite several decades on research in the field, symmetry detection in real-world images remains a challenging task [118]. However, symmetrical compositions in film scenes exhibit characteristics that simplify the task of reliable detection to a certain degree. Recent approaches search for single or multiple axes of symmetry produced by corresponding objects at any arbitrary position in an image. In contrast, a symmetrical composition in a film scene refers to a symmetry at frame (and not at object) level and implies near-perfect mirroring of the right and the left halves of a frame. Hence, the detection of symmetrical composition is reduced to the detection of a mirror-symmetry (or a bilateral reflection symmetry) with the axis of symmetry positioned in the mid frame region. The challenge with the detection of symmetrical compositions in film scenes is that symmetry often occurs at higher level than a perfect symmetry within an object, such as a butterfly, a wheel of a car or a face. Corresponding halves may involve, for example, different objects at similar (but not identical) positions. Thus, symmetrical composition detection requires for the introduction of a looser approach which does not imply perfect shape and object mirroring. Figure 4.24 shows an example for near-perfect symmetrical composition in *Citizen Kane* (1941). Despite the rotation of Kane's face on the background poster, Kane himself and all characters as well as the stage props (e.g. desk, drapes) are carefully positioned. An approach for the detection of such symmetrical compositions is the exploration of present correspondences between regions in the two halves of the scene (see Figure 4.25). The simplest region detection method is the color-based clustering (in case there is no color information: intensity-based clustering). Following, each region can be described using (loose) shape descriptors, area, and location information. A pair-wise comparison with the regions of the second scene half results in the detection of possible correspondences. Eventually, if 1) the majority of the regions has been linked and 2) the correspondences allow for the determination of a steady symmetry line, the scene can be classified as a symmetrical composition.

Unabalanced compositions

The detection of unbalanced compositions requires for the identification and tracking of salient regions (objects or characters). The challenge of salient object detection is that relevant objects usually do not possess homogeneous color or texture characteristics. How-

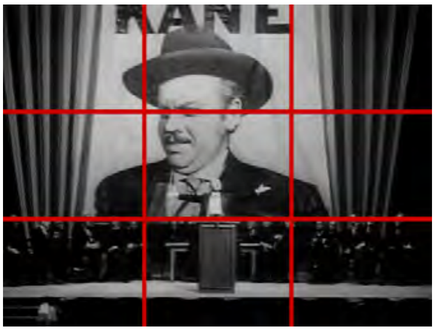
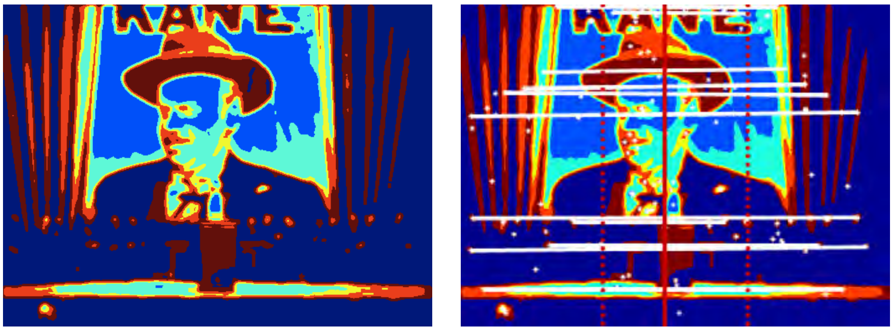


Figure 4.24: Rule of thirds. *Citizen Kane* (1941).



(a) Intensity-based clustering.

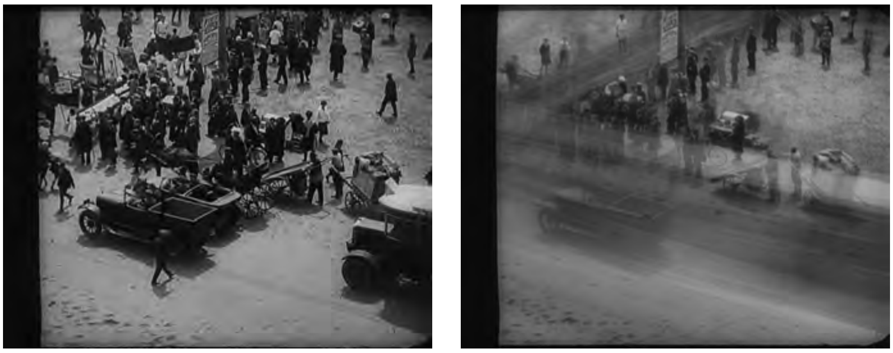
(b) Cluster correspondences.

Figure 4.25: Approach for the detection of symmetrical film compositions. All regions, that are detected using color- or intensity-based clustering, are investigated for correspondences based on shape, area and location. White stars indicate the centroids of the detected regions, white lines detected correspondences. The red dotted lines border the predefined central region of the scene as the only relevant position for the search for a symmetry line. The red solid line shows the detected symmetry line for the scene (slightly off-center). *Citizen Kane* (1941).

ever, in general, two features contribute essentially to the attraction of human attention and are often used to build theattention model of a scene: motion and human faces (e.g. [67, 87, 94]). Following, both features can be used for the detection of unbalanced compositions. For example, motion information and motion clustering methods can be applied to build a database of potentially salient regions. The more often the same region is identified in various shots (moving or not) the more salient it is. This saliency map can be used for the detection of unequally distributed salient regions and, as result, unbalanced scene compositions.

Composition retrieval

Finally, some filmmakers are known for their characteristic compositions. Today, research tasks, such as the locating of all diagonal shots in a specific film by Dziga Vertov, are a tedious process performed manually by film experts. The retrieval of a specific composition can be performed using a predefined template. Mitrović et al. conducted a user study on the performance of several low-level features in the retrieval of visual compositions [104, 173]. The authors show that the KANSEI Shape feature [74] outperforms some well-established MPEG-7 descriptors and, thus, best represents a specific visual template within the given evaluation settings. One drawback of the approach is that it does not account for any motion information but performs solely pairwise comparisons between template and keyframes. However, motion can be essential for the perception of a specific composition (see for an example Figure 4.26). The reflection of motion can be realized by either using a motion-sensitive representation of a video sequence or by extended motion analysis. Zeppelzauer et al. propose an approach for the representation and retrieval of motion compositions based on the detection and tracking of homogeneous motion fields [177]. The description and matching of motion fields is based on spatial and directional information of the corresponding fields.



(a) First frame of the shot (keyframe).

(b) Average frame for the shot.

Figure 4.26: Motion-aided visual composition. While a possible keyframe of the shot shows rather undefined and chaotic arrangement in 4.26a, a motion-sensitive representation clearly identifies a tilted composition of two motion fields on a crossroad (demonstration group vs. cars) in 4.26b. *A Man with a Movie Camera* (1929).

4.3.2 *Continuity*

As discussed in Section 2.2.2 continuity takes different forms in relation to narration, space, time, and motion (see Figure 4.27). Since the understanding of both narrative and temporal continuity requires for high-level semantical comprehension, their analysis is not feasible for computer vision methods. However, temporal and motion continuity bear some characteristics that can be explored fully automatically.

Spatial continuity

Spatial continuity assures the spatial coherence among different shots and, thus, enables the creation and maintaining of the mental map of the audience [179]. In continuity editing, as opposite to the Russian montage, the space of a scene is established along the axis

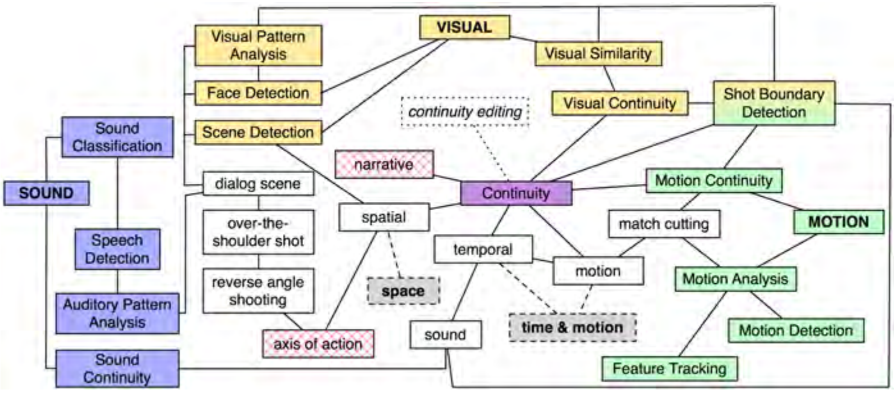


Figure 4.27: Media element vs. computer vision: Continuity.

of action line[3] [18]. This virtual line defines a half circle where the camera(s) can be positioned to shoot the action (see Figure 4.28). It assures the consistency in action direction and character positioning. As result, the visual differences between consecutive shots of the same camera are relatively low. Additionally, shots from different cameras of the same scene bear some visual overlapping. In computer vision, these facts are often used for shot and scene detection. A special type of a scene is the dialog scene. In filmmaking dialogs are often shot using *over-the-shoulder shots* as visualized in Figure 4.28. Alternating over-the-shoulder shots are called *reverse angle shooting* [14]. Recently, various approaches have been proposed for the automated detection of dialog scenes. Existing algorithms exhibit high differences in feature complexity and selected modality (visual, audio, or both). Early approaches explore solely basic visual similarity between shots (e.g. based on color information) and the resulting pattern analysis of the whole video sequence. Furthermore, they are mostly limited to the detection of dialog scenes with two actors. Since a dialog scene and crosscutting may result in a similar pattern, face detection can provide more reliability, e.g. in [36, 78, 181]. Finally, since it is the dialog that characterizes the scene, recently several approaches consider audio analysis and speech detection for dialog scene detection (see e.g. [53, 75, 181]).

Some filmmakers intentionally cross the axis of action to provoke disorientation. In general, the axis line is not fixed for the whole scene – as the characters or the action of the scene moves, the axis line changes accordingly. The large variety of possible motion combinations (camera, objects, characters) and the (current) unreliability in the distinction between camera and object motion and between significant and insignificant motion/object/character makes the automated detection of the axis of action (or the detection of present violation of the rule) not feasible for technical and performance reasons.

[3] Also known as: vector line, principal vector, line of conversation and action, 180 degree line, scene axis, sight line, eye line, and center line [18, 179].

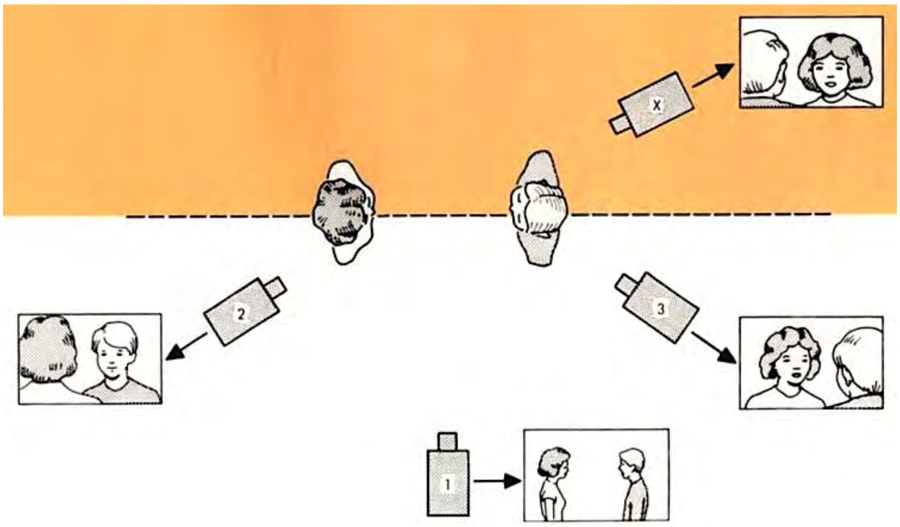


Figure 4.28: Axis of action in a dialog scene (from [18]).

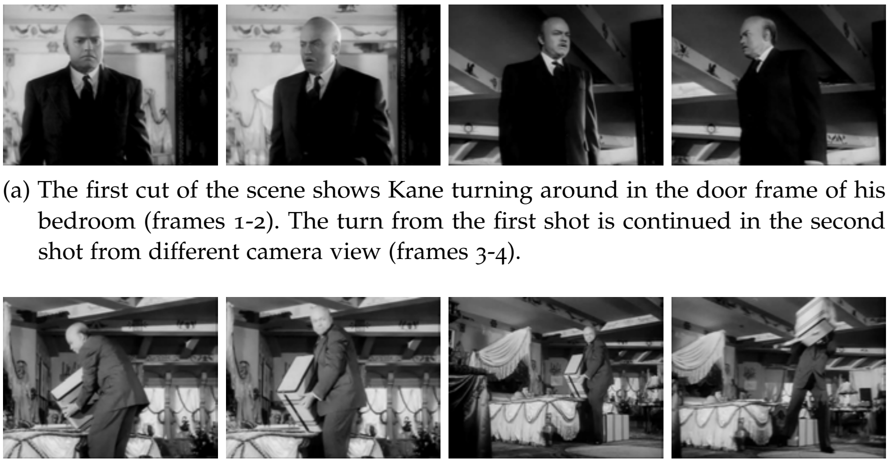
Temporal continuity

Temporal continuity refers to the creation and preservation of time continuity between consecutive shots of the same scene despite any present time manipulations. A standard indicator for temporal continuity is *sound*, e.g. a sound which is associated with a specific scene or a sound over a shot cut [18]. A more advanced tool for the creation and preservation of continuity is the application of *matched cuts*[4]. It describes a cut on identical points of action and creates and supports both spatial and temporal continuity [14, 18]. Figure 4.29 shows two examples for matched cutting (or cutting on motion) where the action from the first shot is smoothly continued in the second shot (despite different camera views or distance of shooting). Zeppelzauer et al. propose an algorithm for the detection of matched cuts based on matching of long-term motion segments [177]. However, the approach has mainly two limitations. First, it requires for user interaction since the user has to sketch the desired continuity. This presumes a priori knowledge about the specific continuity type and limits the retrieval of existing matching cuts. Second, a core descriptor for each motion segment is its median direction which cannot represent a complex motion such as a changing motion direction within the same motion segment. The core element to overcome these problems and to enable a thorough (automated) analysis of present matching cuts is a more precise motion description that accurately describes the motion progress and allows for a scale invariant matching.

Discontinuities

Discontinuities in a video sequence are usually disturbing and disorienting for the viewer (if noticed). Discontinuities can be intentional (e.g. to provoke the audience) or non-intentional (i.e. movie errors in lighting, motion, camera position, props, etc.). For example, Sergei

[4] A matched cut is also referred to as "invisible cut" since it is the motion (rather than the cut itself) that attracts the attention of the audience [14].



(a) The first cut of the scene shows Kane turning around in the door frame of his bedroom (frames 1-2). The turn from the first shot is continued in the second shot from different camera view (frames 3-4).

(b) The second cut shows Kane starting to tear his bedroom apart (frames 1-2). In contrast to the first cut, not the camera view but the distance changes which creates even smoother transition (frames 3-4).

Figure 4.29: Examples for match cutting in *Citizen Kane* (1941). Each sequence shows two consecutive shots depicted by its first and last frame respectively.

Eisenstein is known for his intentional disrupt of the conventions of continuity editing. Eisenstein argued for *montage of collision*, i.e. for contrast and conflict of shots in order to create a new concept[5][14]. Non-intentional discontinuities are typically the result of the fact that shots originating at different times in the production process are merged together to appear to be continuous in time. Following, various continuity errors can appear in the final version such as misplaced items, different lighting and shadows, props can appear or disappear, etc.[6] Recently, Pickup and Zisserman proposed an approach for the detection of continuity errors based on image registration methods [121]. Since moving objects and characters may cause an expected visual difference, the authors apply upper body detectors and trackers to suppress errors caused by motion.

4.3.3 *Motif*

Motifs take very different shapes such as the use of a specific color or sound in a given context, the particular movement of an object or character, camera position, composition type, story line, etc. Since a motif is the product of filmmaker vision, the list of possible motifs is neither exhaustive nor pre-definable. Figure 4.30 shows the mapping between some classical motifs such as color and object and possible computer vision methods for their detection.

[5] A classical example of montage of collision is *October* (1927).

[6] An example for a website that focusses solely on the locating and commenting on errors in movies is http://moviemistakes.com.

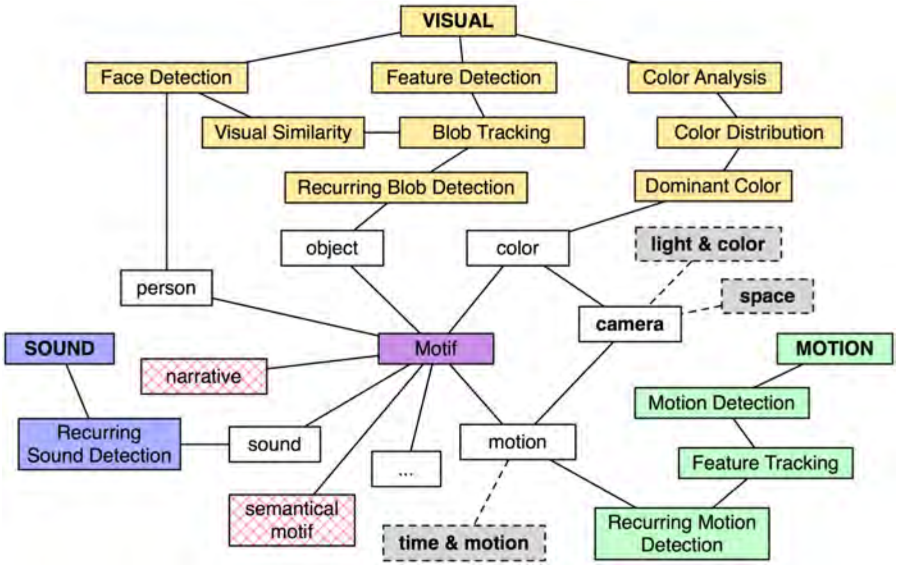


Figure 4.30: Media element vs. computer vision: Motif.

The core characteristic of a motif is its recurrence. However, as already discussed in Section 2.2.3, motifs can have extremely varying appearance and often require for semantical understanding. Examples for such motifs are the time-element in *Run Lola Run* (1998), the words Jules recites before he executes someone in *Pulp Fiction* (1994), or the isolation of Kane stressed by composition and camera positioning in *Citizen Kane* (1941). Such cases cannot be analyzed fully automatically by means of computer vision methods. The lower the level on required semantical understanding and the higher similarity in the appearance (visual or auditory) the more feasible is the automated motif detection (see Figure 4.31).

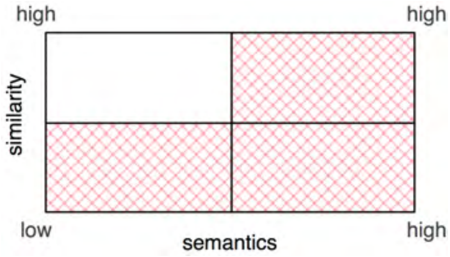
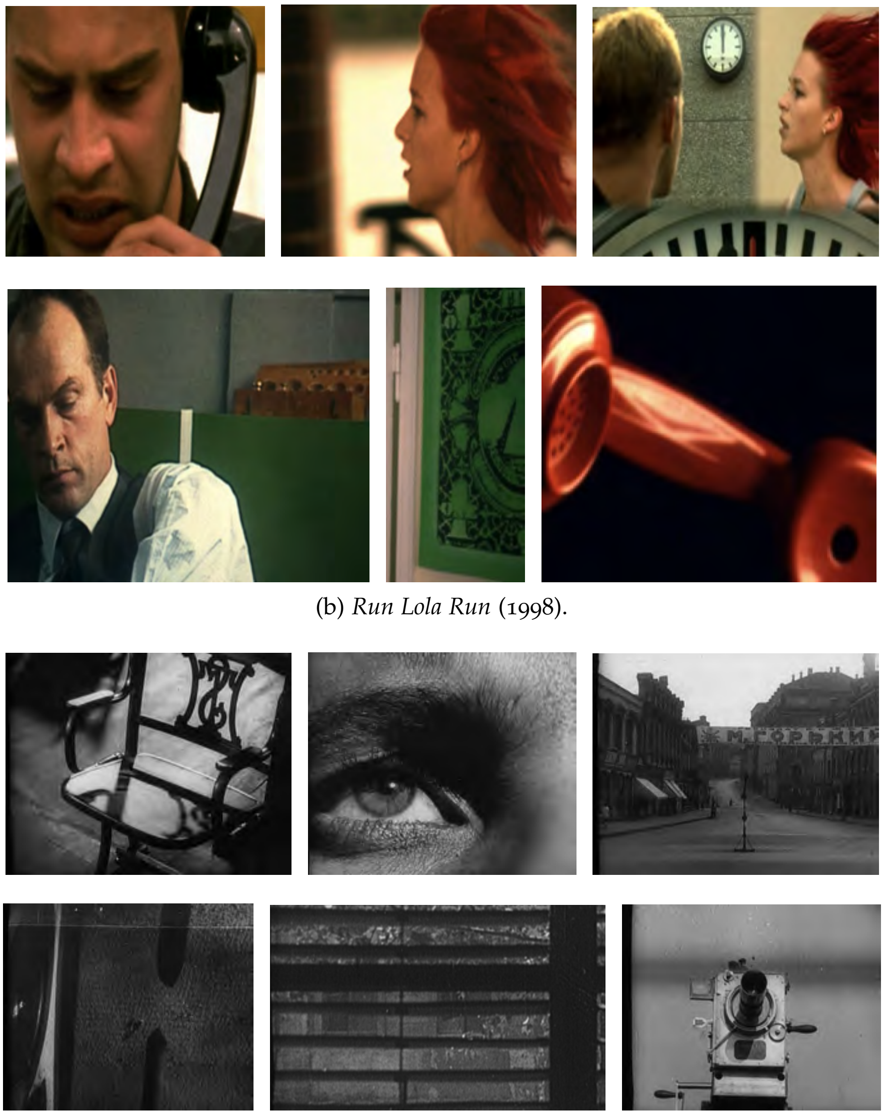


Figure 4.31: Automated motif detection.

From a visual analysis point of view the closest recent research area in computer vision is nearest duplicate detection. This research field aims at the identification of identical or nearly identical images or video sequences. However, in general, a motif does not occupy a whole frame. Thus, a new method is required that builds on the fact that a motif is a recurring element. An automated detection of motifs is only possible within strict constrictions. Similar to the approach of dominant color detection, a dominant blob detection can be performed to identify recurring blobs in a video sequence. Such blobs can be a person, an object, or just a part of them. As result recurring elements can be identified (including leading actors). However, no distinction can be made between significant and insignificant recurrence. Figure 4.32 shows some example results for recurring element detection in two different movies: a contemporary thriller, *Run Lola Run* (1998), and an archive documentary, *A Man with a Movie Camera* (1929). Details on the implementation will be discussed in Section 5.4.



(b) *Run Lola Run* (1998).

(d) *A Man with a Movie Camera* (1929).

Figure 4.32: Examples for detected recurring elements. Please note, that for better visualization all elements have been resized to the same hight, i.e. depicted sizes do not correspond to the actual ones.

4.3.4 *Rhythm, Tempo and Pace*

Figure 4.33 shows the mapping between the film concepts of rhythm tempo and pace and potential computer vision approaches for an automated analysis. As discussed in Section 2.2.4, rhythm depends on the interaction of manifold factors such as the narrative of the story, motion and visual intensity, etc. Although many of the forcing factors can be analyzed on their own, the analysis of their assembly is a slippery issue and, thus, not a well-definable or manageable task for computer vision algorithm. Therefore, in the following, we will focus on the automated analysis of tempo and pace only.

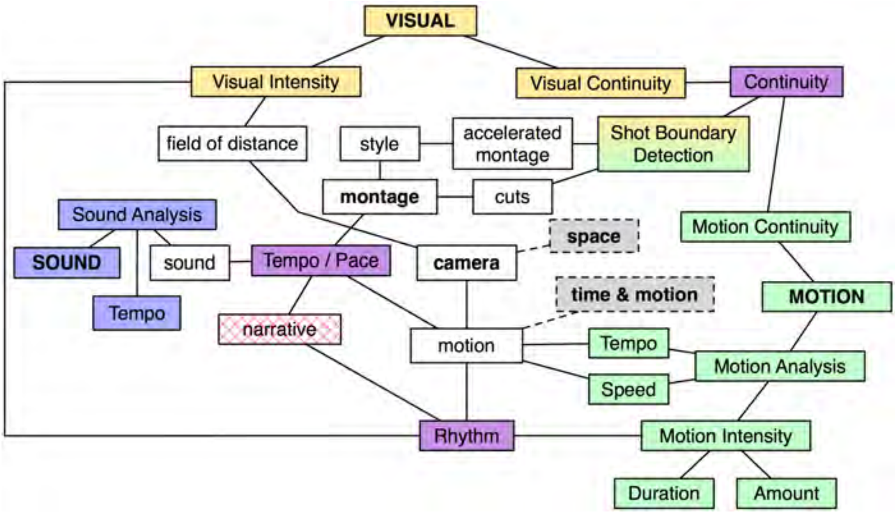


Figure 4.33: Media elements vs. computer vision: Rhythm, Tempo and Pace.

Both tempo and pace are closely related to the perception of time and speed in film. Visually, perceived time is mostly dependent on motion and editing style. Motion content can be measured globally or locally (cp. Section 3.3). Global motion measurement involves usually histogram intersection between consecutive frames. The sum over all shot frames yields information about the motion content of a shot[7]. Another approach to measure motion content is based on local features tracking. This method allows for estimation of the amount of motion in comparison to static features as well as the measurement of the magnitude of motion. Figures 4.34-4.35 show two examples for the difference in the motion content of two consecutive scenes from *Quantum of Solace* (2008)[8]. Independently of the considered motion indicator there are considerable differences between the scenes. The first scene, which is the opening chasing scene of the movie, is characterized by quick shot changes (average shot length: 24 frames/shot), average content change between consecutive frames of a shot of 11%,

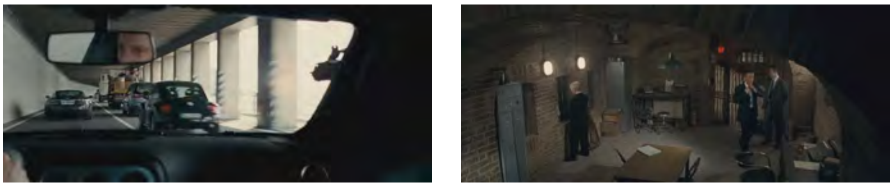
[7] Or more precisely: the change in visual content.

[8] Please note, that for better visualization and easier comparison only the beginning of the first scene is shown.

77% of all tracked features are classified as motion, and in average a motion vector travels a length of approx. 45 pixels. In contrast, the second scene of the movie is (predominantly) a dialog scene between Bond and "M" and is, in general, characterized by longer shots (in average 49 frames/shot), and less camera and character motion: solely 4% of the visual content changes between consecutive frames, 57% of all tracked features are classified as motion, and the average motion vector magnitude is approx. 4 pixels. B. Adams and Svetha Venkatesh showed that this information can be used not only for scene classification but also for pace/tempo analysis and, in following, for the detection of significant event and dramatic section boundaries [6]. The authors define pace as:

$$P(n) = \frac{\alpha(med_s - s(n)}{\sigma_s} + \frac{\beta(m(n) - \mu_m)}{\sigma_m} \tag{4.1}$$

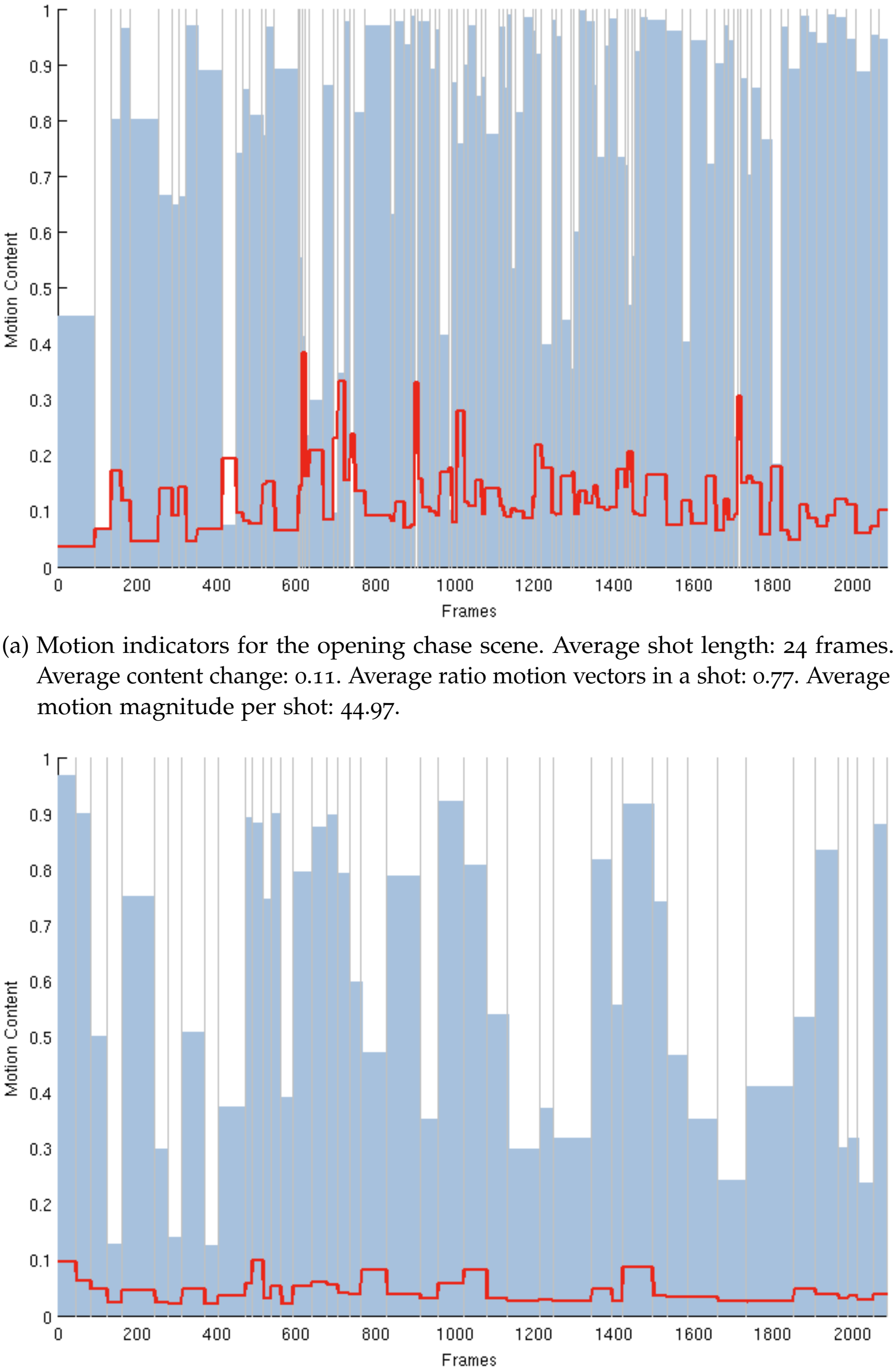
where s is the shot length, m the motion magnitude, n the shot number, σ_s and σ_m the standard deviations of shot length and motion content respectively; and μ_m and med_s are the motion mean and shot length median respectively. α and β are weights indicating the contribution of shot length and motion to the perception of pace. In the performed experiments both weights are given values of 1. The pace function is smoothed with a Gaussian filter to ignore drastic pace changes in a single or small number of shots and to better reflect the human perception of pace. Finally, the authors detect significant event boundaries by edge detection using Deriche's recursive filtering algorithm [37]. Figure 4.36 shows the resulting pace function for the discussed sequence from *Quantum of Solace* (2008) and detected sections using multiscale analysis.



(a) Opening chase sequence.

(b) First scene after the credits.

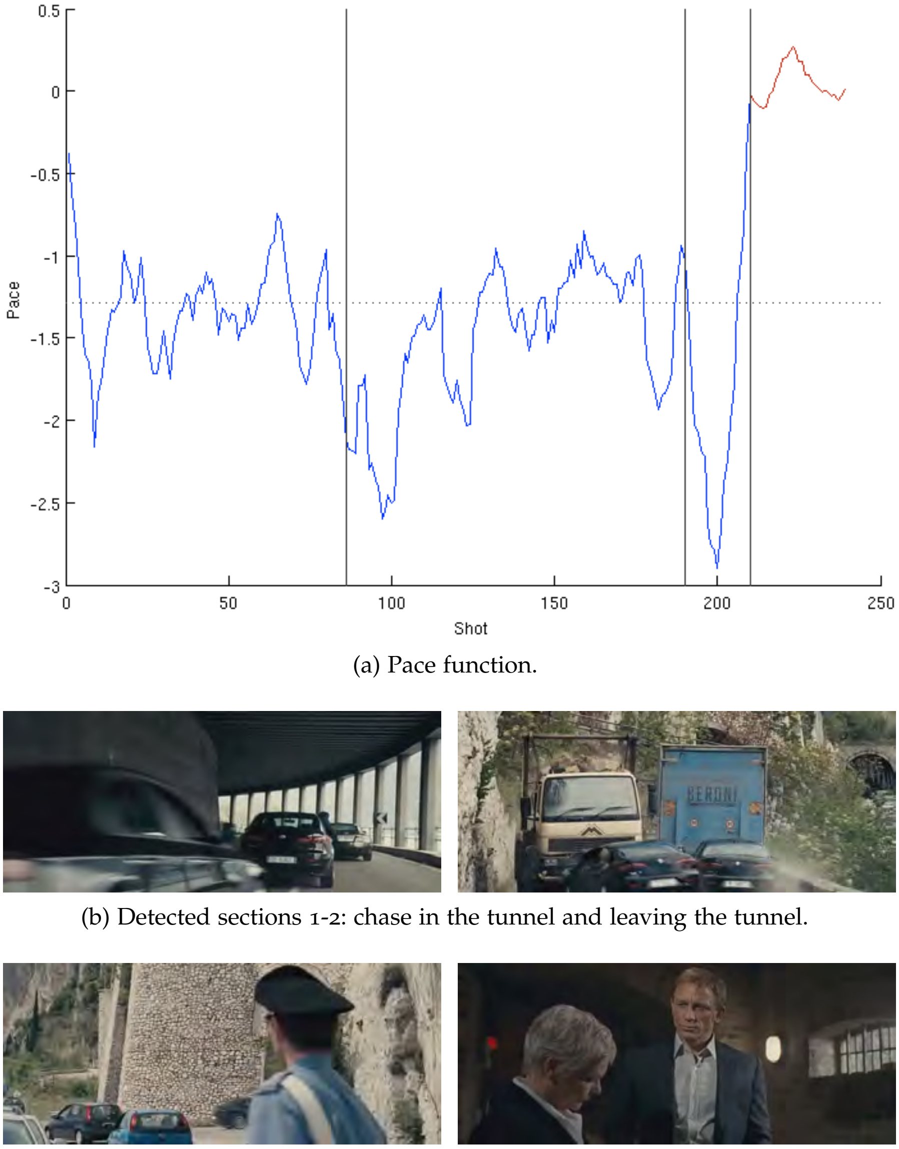
Figure 4.34: Keyframes of the first two scenes of *Quantum of Solace* (2008).



(a) Motion indicators for the opening chase scene. Average shot length: 24 frames. Average content change: 0.11. Average ratio motion vectors in a shot: 0.77. Average motion magnitude per shot: 44.97.

(b) Motion indicators for the first scene after the credits. Average shot length: 49 frames. Average content change: 0.04. Average ratio motion vectors in a shot: 0.57. Average motion magnitude per shot: 3.80

Figure 4.35: Motion content in the first two scenes of *Quantum of Solace* (2008). Gray lines indicate shot borders; red lines the motion content of the corresponding shot by means of histogram intersection; blue bars the ratio of motion vectors to static vectors in a shot by means of feature tracking.



(a) Pace function.

(b) Detected sections 1-2: chase in the tunnel and leaving the tunnel.

(c) Detected sections 3-4: police intervention into the chase scene and scene boundary to the dialog scene.

Figure 4.36: Pace function analysis for the two scenes from *Quantum of Solace* (2008): blue line shows the pace of the first (chasing) scene, red line the dialog scene, gray dotted line the average pace mark for the sequence. Finally, vertical gray lines indicate the results of edge detection on pace function and corresponding story sections from the sequence.

4.4 conclusion

Computer vision approaches for film analysis usually involve both low- (e.g. color and edge detection) and high-level analysis (e.g. feature tracking and representation, object/face detection and recognition). Recent applications originate mostly from the needs and the requirements of a general user. Examples for such applications include the efficient and reliable retrieval of relevant sequences from large media collections, detection of duplicates, summarization and automated classification, etc. Many computer vision approaches perform without any knowledge about film techniques or aesthetic elements. A simplified task, such as the classification of a limited number of sequences in horror and non-horror sequences, can be performed using a basic decision algorithm. However, various applications require for the (intentional or non-intentional) consideration of film techniques. A very simple example is the shot detection which is a fundamental step for any advanced film analysis. Shot detection can be performed fully automatically and (to a certain degree) reliably based on the visual continuity condition within a shot. Another example application for the reflection of film techniques in computer vision is the event detection. Significant pace/tempo changes are often an indicator for the occurrence of a dramatic event (see for examples Figure 4.37). Finally, while *semantics* is still a hype concept in many recent research, its analysis is carried out on a very rudimentary level by e.g. the consideration of available meta data, transcripts, speech recognition, etc.

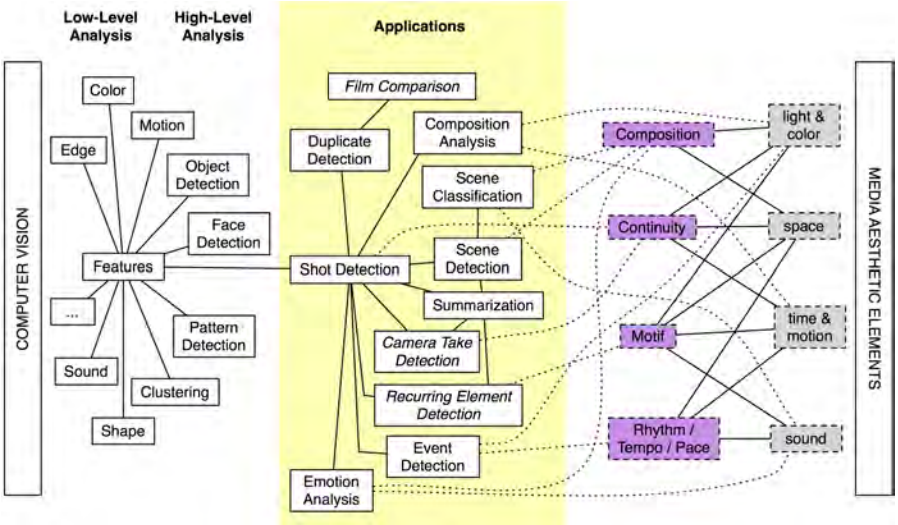


Figure 4.37: Recent computer vision applications. Applications in *italics* will be discussed in detail in the following Chapter 5.

The work presented in this section focused on the question: How far can we go using fully automated tools and where are the frontiers of computer vision. In order to answer this question, we explored a possible linking between fundamental computer vision approaches on the one side and the origins of a film on the other side. Despite the complexity of filmmaking and a certain degree of freedom and creativity that cannot be set into formulas, the analysis of media aesthetic elements and their application in film production by means of film techniques reveals new possibilities for automated film analysis. Many editing techniques require for semantical understanding and are not feasible for an automated computer vision based approach. Some examples include the detection of camera angle, the analysis of editing styles (e.g. Russian style, narrative montage, conceptual montage, montage of attraction), and the detection of time manipulations (see Table 4.1). However, performed analysis and experiments identified a large set of achievable tasks that have not been subject to research so far. Possible queries for automated retrieval of film techniques range from the detection of basic techniques such as the application of Chiaroscuro or silhouette lighting or the detection of selective and racking focus to higher level tasks such as the analysis of visual and motion composition or the detection or recurring elements. Methods implementing such techniques facilitate a fast and systematic retrieval of film material for advanced users, e.g. film archives, film studies, and filmmakers looking for a specific footage.

		Infeasible tasks	Active tasks	Open issues

Fundamental media elements	Light & Color	key light level;	color;
			high-/low-key lighting;
				key lighting types;

	Space	presentational blocking;
		camera angle;
		camera lens;
			zoom;
			field of distance;
				aspect ratio;
				focus type;
				blocking;

	Time	reverse motion;
		time warping;
		rhythmic montage;
		superimposition;
			shot cuts;
			crosscutting;
				freeze frames;
				skip frames;
				split-screen;
				accelerated montage;

Advanced media concepts	Composition		visual composition ;
			motion composition;
				framing orientation;
				balance;
				symmetry;

	Continuity	axis of action;
		narrative continuity;
		montage of collision;
			match cutting;
			dialog scenes;
			visual continuity errors;
				temporal continuity;
				visual continuity;
				reverse-angle shooting;
				over-the-shoulder shots;

	Motif	semantical motif;
		narrative motif;
				recurring objects;

	Rhythm, tempo and pace	rhythm	tempo/pace	

Table 4.1: Tasks in visual-based computational media aesthetics.

5 CASE STUDIES

You've always got to try everything even if you know it's not going to work.

— Anne V. Coates in [117]

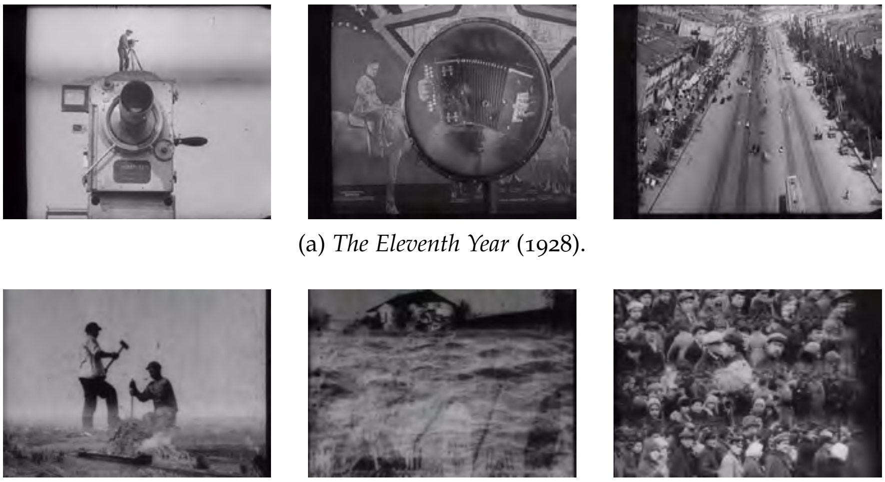
Aim of the chapter

This chapter presents three case studies we conducted in the context of automated film analysis. All experiments are performed on a novel data set of archived documentaries bearing challenges from both artistic and technological point of view. The last experiment is additionally conducted on a contemporary movie. The archive data set is discussed in detail in Section 5.1. The film material at hand allows for the definition of novel research and application topics in the domain of film analysis and understanding. One example for such an application task is the reconstruction of the original camera takes as presented in Section 5.2. The knowledge about the original ordering of the film sequences (as recorded by the camera) facilitates an advanced film analysis such as the detection of montage patterns, the reconstruction of the original montage schema, and, in result, the detection of missing or altered film sequences. The second case study is presented in Section 5.3. It focusses on the comparison of different film versions. This research requirement originates in the fact that many film archives and film museums often possess different versions of the same film material. Differences between film versions can be a result of editorial changes such as different film cuts or a result of the preservation such as missing filmstrips and incomplete copies. Furthermore, the state and the nature of the film material impede the differentiation between very similar shots repeatedly appearing in the film sequence and identical shots having different appearance due to material-specific artifacts such as mold or film tears. In the last experiment, Section 5.4, we investigate recurrences on a more detailed level than a shot or a film sequence. In this case study, we explore the feasibility of detecting dominant characters or objects in a film sequence where dominance is defined by corresponding appearance frequency.

5.1 ARCHIVE VIDEO DATA

Artististry-related challenges

The archive video data set consists of historical artistic documentaries by the Soviet avant-garde filmmaker Dziga Vertov from the 1920s and 1930s. Vertov plays a major role in the history of experimental films and, at the same time, he is considered as the forerunner of the cinema vérité movement in documentaries [35]. Vertov rejects theatrical artificiality such as studios, actors, and staging, and aims at capturing the raw truth with his camera [105]. His films do not contain any narrative structure, which makes them different from material that is usually analyzed in content-based research such as news broadcasts, sports videos, and feature films. Vertov often manipulates his films to demonstrate the artificiality and nonrealism of cinema [35]. He makes use of advanced montage techniques to create complex transitions, multiple exposures, and split screen compositions (see for examples Figure 5.1). Furthermore, his films exhibit a distinctive structure, which is characterized by a high number of short, repeating shots with high visual similarity (see for examples Figure 5.2).



(a) *The Eleventh Year* (1928).

(b) *Man with a Movie Camera* (1929).

Figure 5.1: Vertov's demonstration of cinema artificiality.

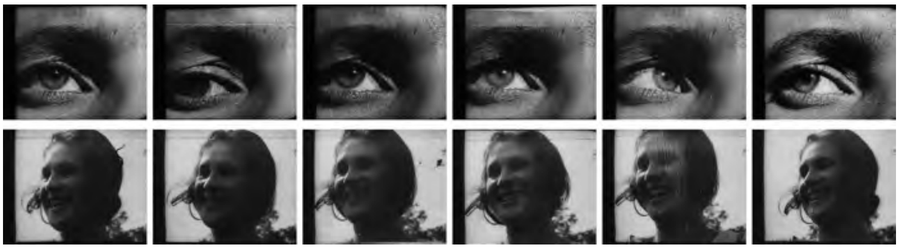
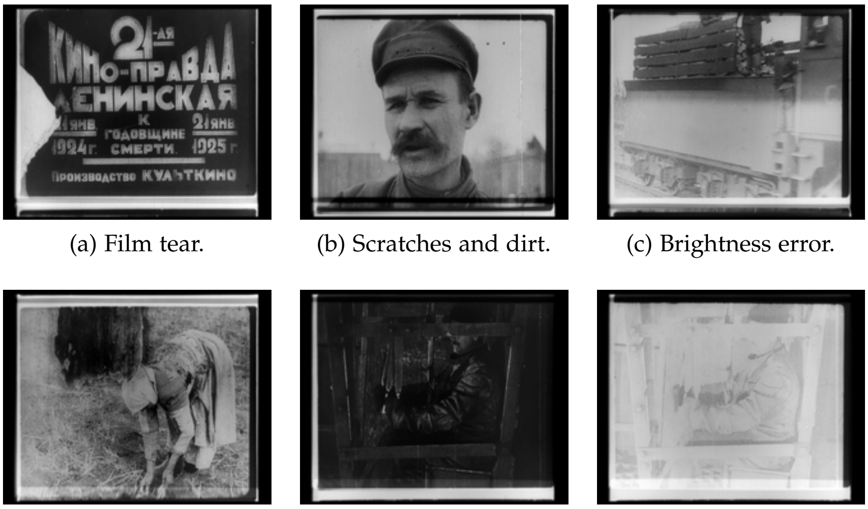


Figure 5.2: Examples for highly similar repeating shots (each shot is represented by its first frame). *Man with a Movie Camera* (1929).

Technology-related challenges

The source film material is 35mm monochrome and mostly silent film which limits the set of available modalities and feasible techniques. The filmstrips were digitized frame-by-frame to make them processable. Existing filmstrips of archived films are usually multiple-generation copies that were never intended for other purposes but backups. Often, the original filmstrips do not exist any more and the available backup copies are the only existing source material left. The state of film material degrades significantly during storage, copying



(a) Film tear. (b) Scratches and dirt. (c) Brightness error.

(d) Scratches. (e) Underexposure. (f) Overexposure.

Figure 5.3: Examples for artifacts in archive film material (all frames exhibit visible framelines). *Film-Truth* (1922-25).

and playback over the decades (see for examples Figure 5.3). Important artifacts in archive film material include:

- *Scratches*, which are usually introduced by dirt in the film projector.
- *Dirt* (dust, liquids, mold), which propagates and increases from one copy to the next.
- *Visible framelines*, which result from copying misaligned filmstrips and the shrinking of the film material. Since the filmstrips are made of organic material they contract over time. Contraction occurs horizontally and vertically and results in shaking and misaligned frames.
- *Low contrast*, which is a result of repeated copying.
- *Flicker*, which results from the fact that film transports in early cameras was performed manually (variable exposure time).
- *Frame displacements*, which result from shrinking of the filmstrips.

5.2 camera take reconstruction

Section outline

In this section we present a new topic in the domain of video retrieval, namely the identification of editing techniques and montage patterns. Contemporary Hollywood-type movies and TV broadcasts usually follow specific editing rules, such as cross-cutting and shot reverse shot [18], resulting in well-established shot editing patterns within a scene. On the contrary, some documentaries, experimental, and art house films often challenge the conventional filmmaking by the use of unusual (non-narrative) camera and editing techniques [35]. Currently, the study of such techniques is a tedious manual process performed by film experts. We propose an automated approach for the analysis of editing techniques and montage sequences that is based on the reconstruction of the original film shooting sequence referred to as camera takes. After defining the terminology and discussing its advantages, we present the two stage algorithm for camera take detection. Following, the algorithm is evaluated on a challenging data set of experimental archive documentaries. Achieved results show the reliability of the algorithm and outline its applicability as supporting tool for the analysis of editing techniques and montage patterns.

What is a camera take?

A *camera take* is defined as a single, continuously-recorded performance with a given camera setup [14]. In the editing process camera takes are often cut into multiple shots and joined together to form a complete movie, i.e. a camera take is a sequence of one or more consecutively recorded video shots. Semantically related and temporally adjacent shots build a video scene. However, shots originating from the same camera take can also be temporally distributed over the entire movie (see Figure 5.4). Examples for such an editing technique are the time jumps and the parallel development of two or more lines of action.

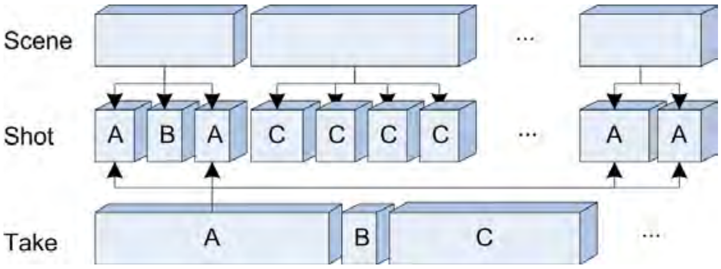


Figure 5.4: Camera takes vs. film scenes.

Why camera takes?

The reconstruction of camera takes yields relationships of shots that proceed at the same place and time. This high-level structural information is beneficial for tasks such as scene segmentation and analysis of montage patterns, editing style, and motion rhythm. Furthermore, reconstructed camera takes allow for compact video representation and nonlinear browsing. The reconstruction is based on the temporal continuity of shots. It does not require the film content to be similar over the entire camera take. For example, several shots cut out from a camera take that contains a long camera pan can have highly dissimilar content. Methods based on keyframes and image features may not find similarities among the shots. The presented approach is able to associate the shots with each other.

Example applications

Various applications for film analysis and retrieval can benefit from the camera takes reconstruction. Examples include:

- *Flashback / -forward detection*: A flashback is defined as a shot that is presented out of chronological order [18]. The detection of camera takes implies the reconstruction of the original chronological order and, thus, allows for a straightforward flashback detection.

- *Montage pattern and rhythm analysis*: The rhythmic relations between two shots indicates highly semantical information. Cinematic rhythm derives from different film techniques such as shot duration, visual and motion content, sound rhythm, and montage patterns. For example, the use of alternating close-ups with shorter shots creates a more intense dialog or conflict sequence.

- *Film analysis and reconstruction*: The reconstruction of the montage schema allows for the identification of incomplete copies and altered versions of the original film material.

- *Video summary*: The association among shots of the same camera take can be further used to create a more compact video summary for non linear browsing.

5.2.1 *Camera Take Detection*

The core element of the algorithm for camera take detection is the motion smoothness analysis between different shots. However, since motion tracking in a long video can become computationally expensive, we introduce an intermediate step to limit the number of candidates for camera takes. To determine possible camera takes we use a fast and yet reliable similarity measure based on edge histograms. Following, we analyze the motion smoothness based on local feature tracking. Figure 5.5 gives an overview over the workflow of the algorithm.

5.2.1.1 *Continuity Analysis*

For the detection of candidate camera takes we first construct the set of all continuity regions for a given shot S_x. The continuity region CR between two shots S_x and S_y is defined as the union of the last n frames of S_x and the first n frames from S_y:

$$CR^{S_x,S_y} = \{f^{S_x}_{a-n+1}, f^{S_x}_{a-n+2}, \ldots, f^{S_x}_{a}, f^{S_y}_{1}, f^{S_y}_{2}, \ldots, f^{S_y}_{n}\} \tag{5.1}$$

where a denotes the number of frames of S_x and f the respective frames in S_x and S_y. S_y represents any other shot from the film. Thus, for a given shot S_x a set of continuity regions (with common S_x last frames) is constructed. In our evaluations, n is set to three which results in a continuity region of the length 6 between any two shots.

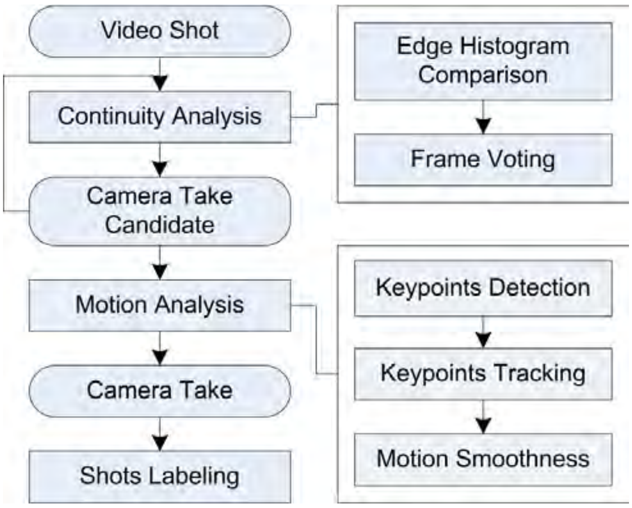


Figure 5.5: Camera take detection workflow.

For every frame from the regions an MPEG-7 edge histogram is computed (see Section 3.1 for details on the feature). Each frame $f^{S_x}_{a-n+1}, f^{S_x}_{a-n+2}, ..., f^{S_x}_{a}$ is compared to every frame from the set of continuity regions for S_x that represents a shot different than S_x. Following, frames vote for the shot with the highest similarity score in terms of Euclidean distance. A shot S_y is accepted to be a following shot of S_x if:

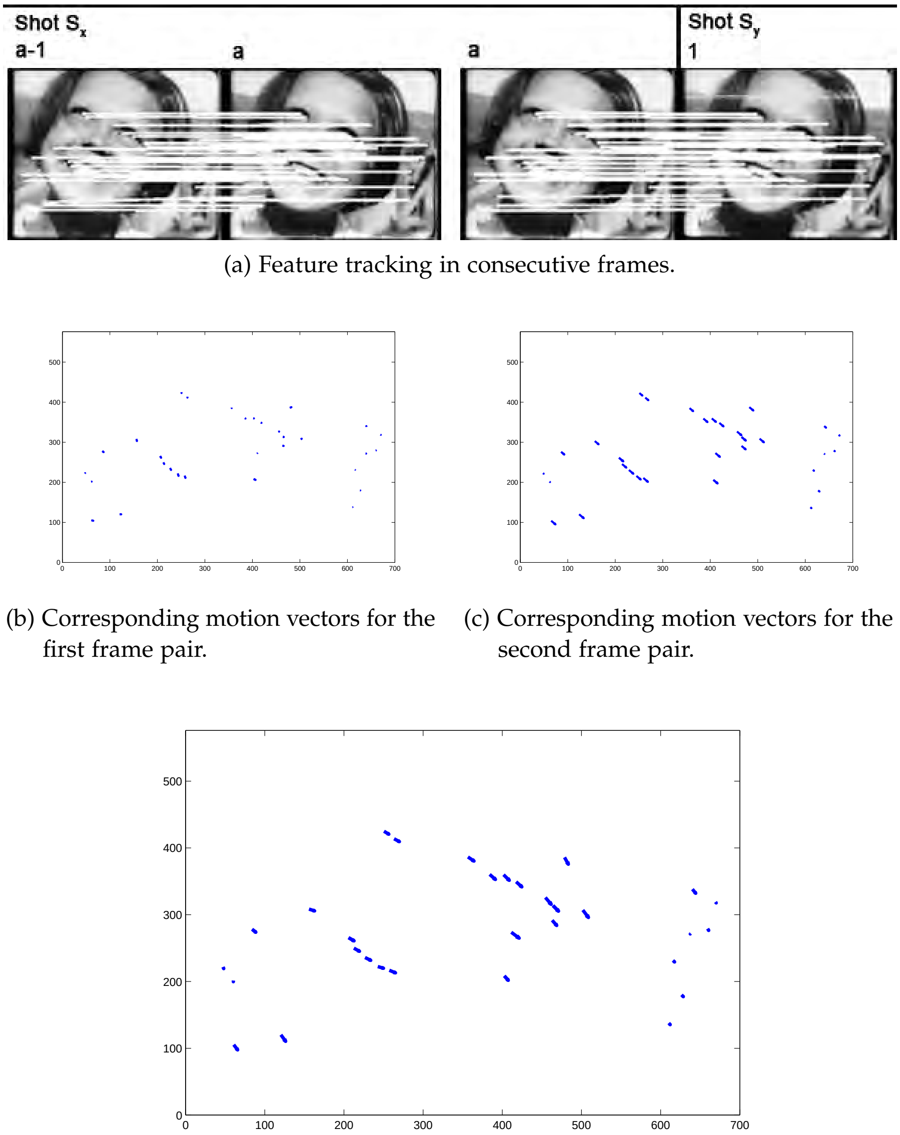
1. the majority frames from S_x vote for S_y, and
2. there is at least one reverse vote, i.e. at least one frame from S_y votes for S_x.

In case, S_y is a following shot of S_x, both are assigned to a new candidate camera take: $CT_i =< S_x, S_y >$. For every last shot of the current CT_i the process is repeated until there are no more following shots detected.

5.2.1.2 *Motion Smoothness Analysis*

Motion vector fields estimated for consecutive video frames are slowly varying over both space and time. Therefore, we measure the variations of the motion vectors along the temporal direction in the continuity region of each candidate camera take. Figure 5.6 shows an example for consecutive shots. The difference between the respective motion vectors is very low and, thus, indicates high motion smoothness. On the contrary, Figure 5.7 depicts frames that are visually similar but belong to different, temporally non consecutive shots. The slight move of the girl's head results in significantly larger differences in the motion vectors.

For motion detection and tracking we apply local feature tracking based on SIFT matching (see Section 3.2.2 for details on the feature).



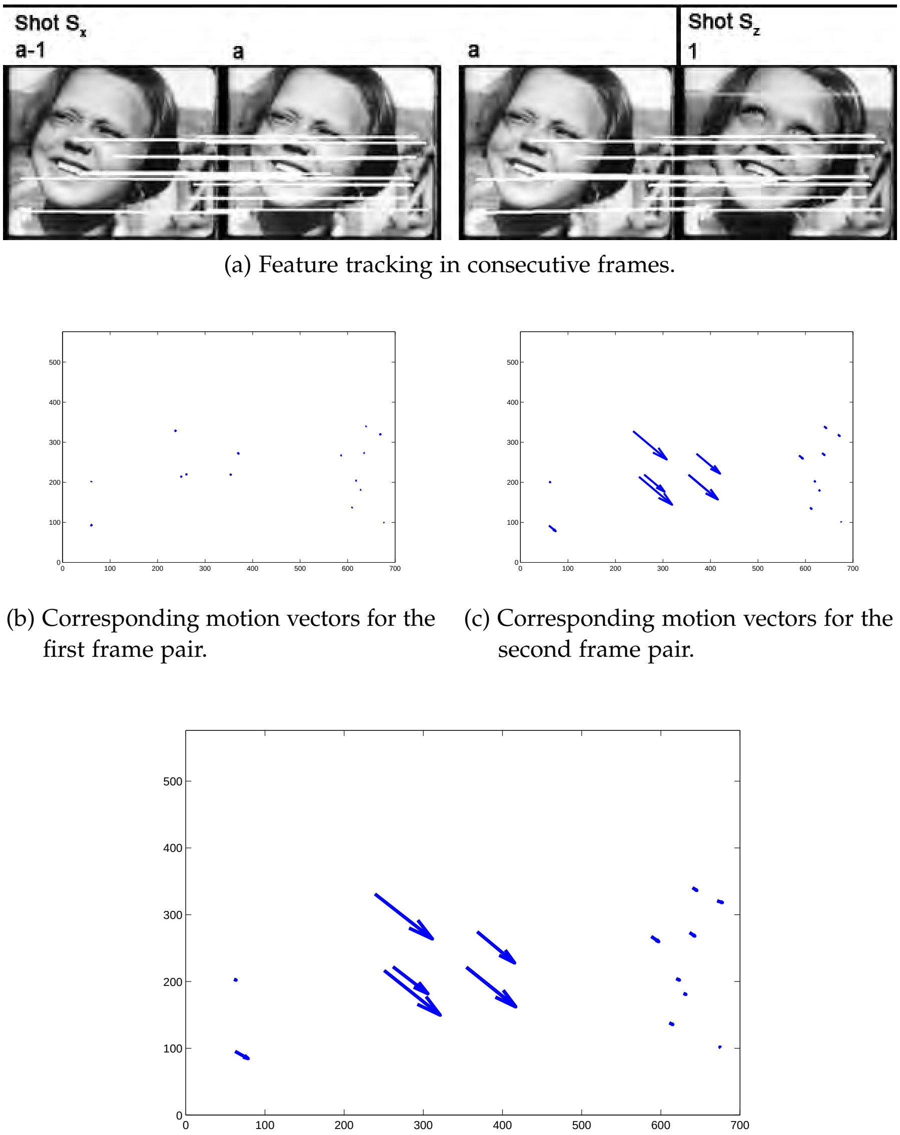
(a) Feature tracking in consecutive frames.

(b) Corresponding motion vectors for the first frame pair.

(c) Corresponding motion vectors for the second frame pair.

(d) Differences between the corresponding motion vectors.

Figure 5.6: Motion smoothness for frames of the same camera take from *The Eleventh Year* (1928).



(a) Feature tracking in consecutive frames.

(b) Corresponding motion vectors for the first frame pair.

(c) Corresponding motion vectors for the second frame pair.

(d) Differences between the corresponding motion vectors.

Figure 5.7: Motion smoothness for frames of similar but not consecutive shots from *The Eleventh Year* (1928).

However, other motion tracking methods can be applied as well. We limit the number of extracted SIFT features per frame to 500. The resulting feature descriptors are matched by identifying the first two nearest neighbors in terms of Euclidean distances. A descriptor is accepted if the nearest neighbor distance is below a predefined threshold. The value of 0.8 was determined experimentally and used through the evaluation tests described in Section 5.2.2. Finally, only camera takes with smooth motion vectors are accepted.

5.2.2 *Experiments*

In this section we describe the performed experiments on camera take detection and montage pattern reconstruction.

5.2.2.1 *Camera Take Detection*

The first experiment focusses on the evaluation of camera take detection. We explored two movies, *Man with a Movie Camera* (1929), consisting of 1,768 shots (95,678 frames) and *The Eleventh Year* (1928), consisting of 660 shots (63,123 frames). Our algorithm detected 186 camera takes of two and more shots. The results were evaluated manually. Approx. 93% of all detected camera shots were correct and false positive rate was less than 5% (see Table 5.1).

	MMC		EYE		Average	
True positives	110	92.44%	63	94.02%	173	93.01%
False positives	5	4.20%	4	5.98%	9	4.84%
Ambiguous	4	3.36%	0	0.00%	4	2.15%
Detected takes	119	100.00%	67	100.00%	186	100.00%

Table 5.1: Performance results on camera take detection. MMC: *Man with a Movie Camera* (1929), EYE: *The Eleventh Year* (1928).

The lack of motion and the same visual appearance of shots may cause false positive detection of camera takes for identical, static shots. Another reason for incorrect detected camera takes is the dissolve editing technique. The gradual replacement and high degree of similarity between shots can falsely assign them to the same camera take (see Figure 5.8 for an example).

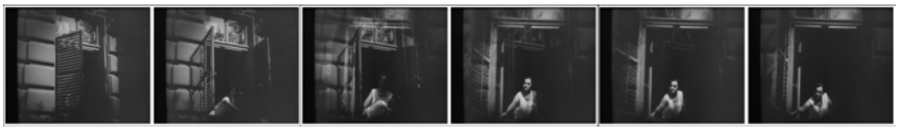


Figure 5.8: False positive camera take due dissolve (the same scene is shot from two different perspectives). *Man with a Movie Camera* (1929).

Additionally, four detected camera takes could not be verified due to ambiguity. An example for such shots is presented in Figure 5.9. The shot depicts a figure in a shooting gallery on a fair. Due to the repetitive movement of the timbal in the right hand it is often not possible to definitely determine if 1) multiple shots are part of the same camera take or 2) it is always the same shot on different positions.

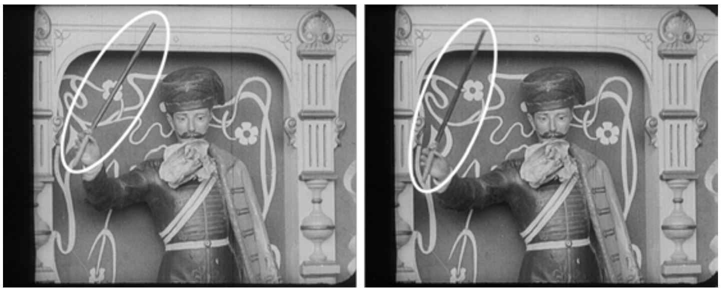


Figure 5.9: Ambiguous shots. *Man with a Movie Camera* (1929).

5.2.2.2 *Montage Reconstruction*

The next experiment addresses the reconstruction and the analysis of the original montage schemas. Montage schemas describe the assemblage of a film through editing. They allow for the analysis of editing techniques and montage patterns. Furthermore, the reconstruction of montage schemas is essential for the analysis of archive film material where the original versions (filmstrips) do often no longer exist. As previously discussed, the remaining copies are usually backup copies from film archives that are often incomplete due to bad storage, mold, and film tears.

We investigate different film sequences from three archive documentaries. We first detect camera takes. In the next step, we assign labels to the shots of the same camera take.

The first sequence from *The Eleventh Year* (1928) presents workers building a railway. The whole sequence of 19 shots (204 frames) originates from three cross-cut camera takes. Our algorithm successfully detected and assigned the respective shots (see Figure 5.10).

Since we investigate experimental video material, not all of the resulting montage schemas comply with conventional editing patterns. Figure 5.11 presents the detected montage pattern in a sequence of 66 shots (191 frames) from the *Man with a Movie Camera* (1929). It exhibits an unusual editing technique that is not reconstructable with other common scene detection algorithms. The interpretation of such patterns is a research subject for film experts.

The evaluation of the third sequence (42 shots, 773 frames) from *Kino Eye* (1924) was motivated by the discovery of the original montage schema. It shows the experiment of the filmmaker to graphically chart the montage of shots within a scene (see Figure 5.12b). The

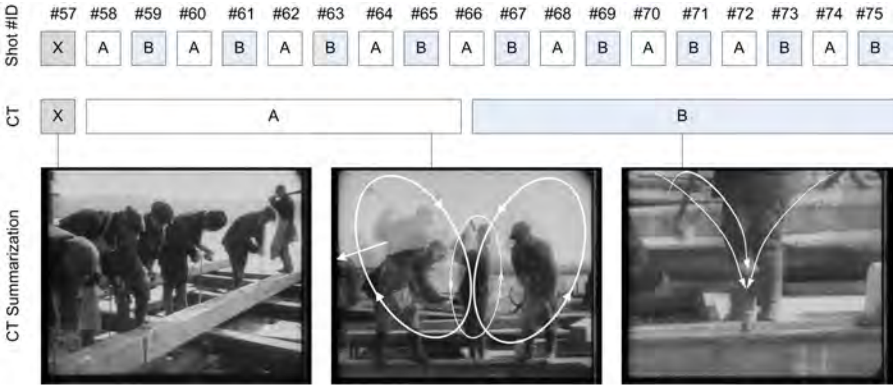


Figure 5.10: Detected camera takes in the sequence from *The Eleventh Year* (1928) (white arrows show dominant motion).

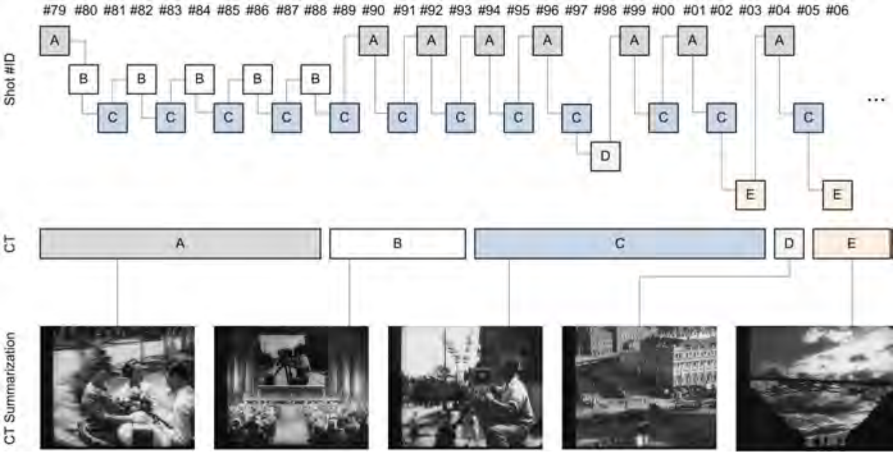
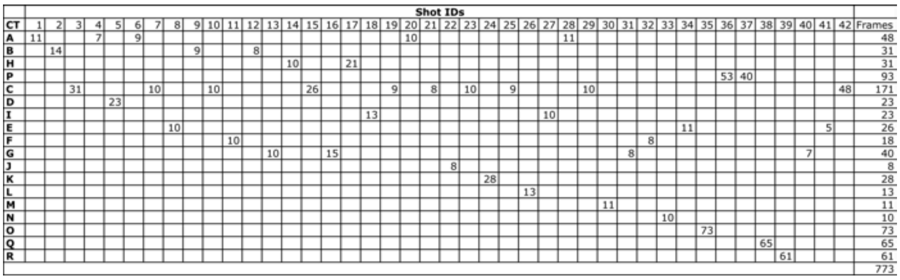


Figure 5.11: Detected camera takes in the sequence from *Man with a Movie Camera* (1929).

reconstructed montage schema indicates missing and rearranged shots (see Figure 5.12a). Currently, it is not clear whether the original film complied with the discovered schema and the nowadays available copy is a full version of the original film. Notwithstanding, the results demonstrate the reliability of the algorithm and its applicability in a scenario where the original montage schema is not available.

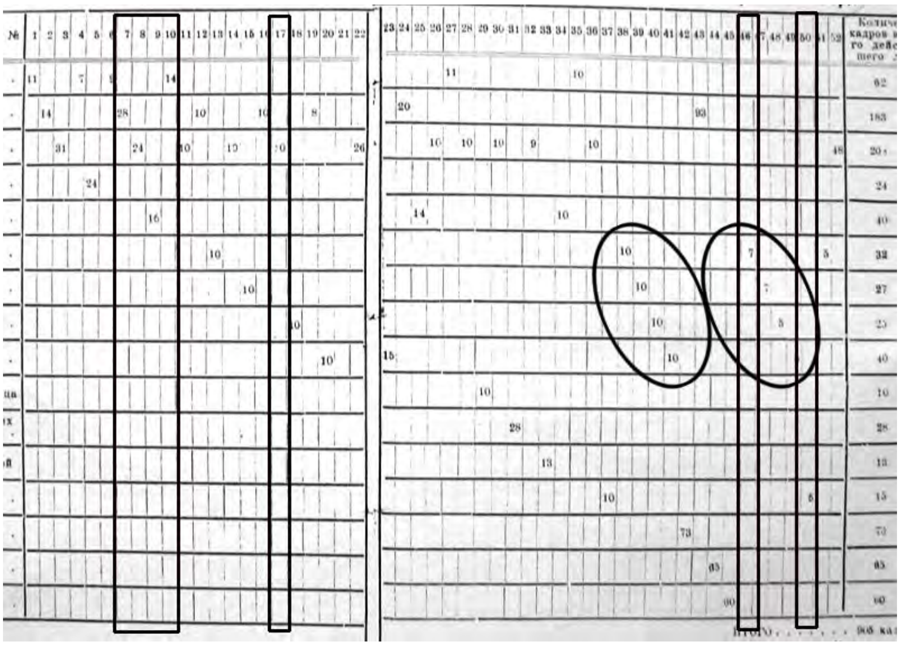
5.2.3 *Related Work*

A camera take reconstruction algorithm reveals information about the structure of a movie. This perspective on the film structure represents the film production point of view. In contrast, current work on video structure analysis focusses mainly on scene detection and classification. Recent approaches on scene detection and classification group shots into a scene if they are content-correlated and temporally close to each other [30, 113, 123, 124, 182]. Content correlation is usually determined based on color information. An essential disadvantage of



	Shot IDs																																										
CT	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	Frames
A	11			7		9														10								11															48
B		14							9			8																															31
H														10			21																										31
P																																				53	40						93
C			31				10			10					26				9		8		10		9				10													48	171
D					23																																						23
I																		13									10																23
E								10																										11							5		26
F											10																					8											18
G													10			15															8									7			40
J																						8																					8
K																								28																			28
L																										13																	13
M																														11													11
N																																	10										10
O																																			73								73
Q																																						65					65
R																																							61				61
																																											773

(a) Reproduced schema.



(b) Extract from the original schema from the mid 1920s (black box frames indicate some of the detected missing shots, oval frames: rearranged shots)

Figure 5.12: Montage schemas for the sequence from *Kino Eye* (1924).

this approach is that false color matches between shots of different scenes result in falsely combined shots. Dynamic scenes often possess different color information which impedes the process of keyframes selection for reliable shot representation. Motion information is often neglected within the process of scene detection. Ngo et al. use motion information for the selection and formation of keyframes as representative for the shot [113]. However, motion is no further used as a matching criterion. Rasheed et al. merge shots together that have high motion activity and small shot lengths to enable high scene dynamics [123]. However, the assumption that shots of the same scene follow the same dynamics holds only for very limited scenarios.

Recently, Truong et al. addressed the extraction of film takes [155]. The authors applied merge-and-split clustering techniques to group similar shots based on color histograms. A substantial assumption of the approach is that at most one shot is presented from a single camera take. This assumption holds for a great part of Hollywood-type movies but fails for most documentary and experimental films. A further limitation of the approach is that it cannot be applied to shots with extensive camera and/or object motion (e.g. action shots) due to the restrictions of the selected shot representation. Eventually, the task of camera take extraction is reduced to a shot similarity detection.

In contrast to existing approaches, we strongly rely on motion information. Motion smoothness between frames of the same camera take allows for the reliable recognition of consecutive shots. Thus, shots are linked together without the problem of appropriate keyframe selection or shot representation. Furthermore, since shots of the same camera take can be temporally apart from each other in the edited film, the reconstruction of camera takes captures information, which is lost by a scene detection algorithm.

5.2.4 *Conclusion and Discussion*

We presented a novel application for the reconstruction of camera takes. We applied the proposed algorithm on a test set of experimental archive documentaries. Presented results demonstrate the reliability of the algorithm and outline its applicability for manifold application scenarios such as montage pattern analysis or the comparison of different film cuts. This new topic in the field of video retrieval radically affects the analysis of archive documentaries for two reasons. *Firstly*, the evolution of the production process has changed essentially over the past decades. In the past, filmmaking was an expensive process where a scene was shot just a single or a few times which meant that acquired film material was used as completely as possible in the current final cut and possibly reused in further film compilations. On the contrary, today, a scene is shot until it fits the vision of the producer(s) which can result in raw film material up to 100 times the length of the final cut [110]. *Secondly*, documentaries often use unusual camera and editing techniques. Currently, the study of such patterns is a tedious manual process performed by film experts

Reliable camera take detection provides a new perspective to the domain of film analysis. From a *technical point of view*, it allows for the comparison of different film cuts and the analysis of montage patterns that do not follow conventional editing rules. Moreover, further analysis of the motion smoothness between two shots can provide information about missing frames from the original camera take.

From a *semantical point of view*, reconstructed camera takes capture information that can be missed by conventional scene detection algorithms. By the analysis of motion smoothness within a given continuity region, the proposed method does not require appropriate shot representations or keyframes and feature selection. Moreover, two shots to be grouped are not required to be visually similar for the whole shot length. Highly dynamical shots (e.g. action) or large camera motion often result in great dissimilarity in the visual perception. However, motion smoothness analysis can still detect consecutive shots due to the smooth transition present in a continuos camera take. This information can be further used to improve the process of video representation and retrieval.

5.3 FILM COMPARISON

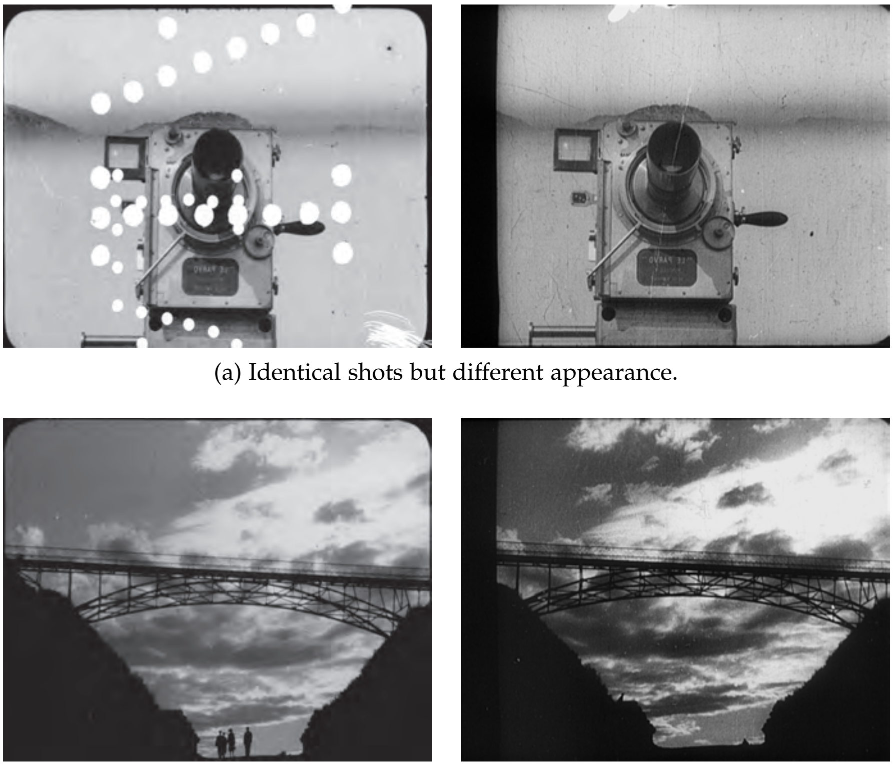
What is a video copy?

By definition, a video copy is a transformed video sequence [83]. The transformation can be of technical nature (e.g. change in video format, frame rate, resizing, shifting, etc.) or editorial modification (frame insertion/deletion, background change, etc.). Video copy detection is an active research area driven by ever-growing video collections. The detection of video duplicates allows for the efficient search and retrieval of video content. Existing applications for content-based video copy detection comprise e.g. clip identification in a given video set [85, 170], copyright protection [69, 72], identification of duplicated news stories [180], and TV broadcast monitoring and detection of commercials [88, 132]. Presented experiments are often limited to high quality video clips of pre-defined fixed length and synthetically generated transformations such as resizing, frame shifting, contrast and gamma modification, Gaussian noise additions, etc.

Challenges

In contrast, film and video comparison reaches beyond the boundaries of a single shot and aims at the identification of both reused and unique film material in two video versions. The compared videos can be two versions of the same feature film, e.g. director's cut and original cut, or two different movies that share a particular amount of film material, such as documentary films and compilation films. Archive film material additionally challenges existing approaches for video analysis by the state and the nature of the material. Different versions vary significantly not only by the actual content (e.g. loss of frames/shots due to censorship or re-editing) but also due to material-specific artifacts such as mold, film tears, flicker, and low contrast. Furthermore, existing algorithms often provide only limited robustness to illumination changes, affine transformation, cropping, and partial occlusions, which restricts their applicability for low-quality archive films. Archive film material is well-suited for the evaluation of film comparison techniques since it contains a large number of natural (not synthetically generated) transformations among different film versions and represents a complex real world scenario for film comparison and copy detection.

In practice, it is not always obvious whether or not two shots are identical. Figure 5.13 illustrates the challenge of identifying shot correspondences in film archives. Figure 5.13a depicts the first frames of two identical shots that possess different appearance due to visible cue marks and scratches in the first shot, frame shifting in the second shot, and contrast differences in both shots. On the contrary, Figure 5.13b shows the first frames of two similar and yet different shots with high perceptual similarity.



(a) Identical shots but different appearance.

(b) Different shots despite high perceptual similarity.

Figure 5.13: Identical vs. similar shots.

Approach

In general, a film comparison process passes well-defined steps from shot boundary detection to shot representation and matching. At each step different algorithms can be applied. The combination of and the interaction between the selected methods are crucial for the overall comparison process. In this section we present an approach for automated film comparison that accounts for the temporal and hierarchical structure of a video, i.e. the frame, shot, and video level. The approach allows for the selection of the appropriate hierarchy level for a given task and, thus, enables different application scenarios such as the identification of missing shots or the reconstruction of the original film cut. Using the proposed methodology we evaluate the performance of established shot boundary detection algorithms and investigate the influence of keyframe selection and feature representation on the film comparison process. The results show that the approach presented yields high recognition rates for the investigated application scenarios. Furthermore, the integration of knowledge about the hierarchical structure of a video allows for the outstanding performance of a simple, edge-based descriptor at much lower computational cost than state-of-the-art local feature-based approaches.

Section structure

This section is organized as follows. We describe the underlying methodology of our approach for automated film comparison in Section 5.3.1. Section 5.3.2 outlines the methods for shot boundary detection, keyframe extraction, and feature representation used for the experiments presented in Section 5.3.3. We present current related work and discuss its limitations in Section 5.3.4. Finally, we conclude in Section 5.3.5.

5.3.1 *Underlying Methodology*

From a technical point of view, a video consists of temporally aligned shots. Each video shot is a continuous sequence of frames recorded from a single camera. We present an approach which accounts for this logical structure of a video and does not require any additional information. Starting from a raw and unsegmented video, the first step is to determine shot boundaries automatically. Following, each shot is represented by a set of robust and distinctive features. Finally, our matching and decision process is applied to the segmented video stream. Figure 5.14 visualizes the workflow and information propagation within the framework.

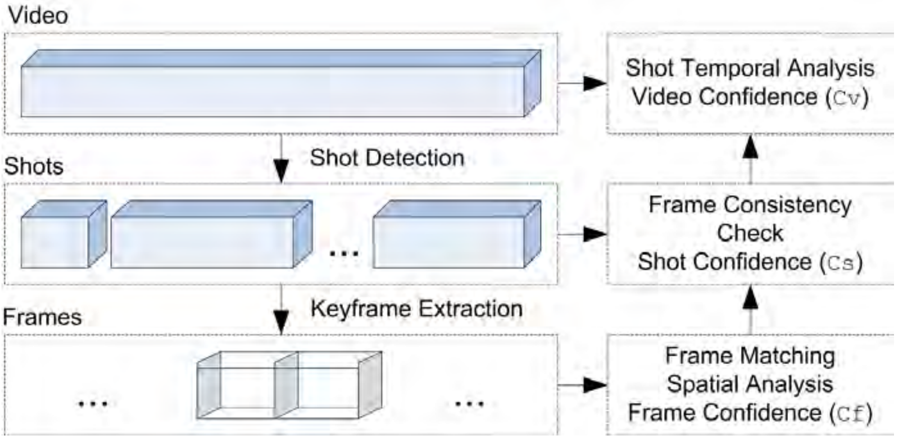


Figure 5.14: Workflow and information propagation.

5.3.1.1 *Frame Level*

Video frames are the basic building blocks of a video sequence. Since a single frame represents a still image, manifold global and local features can be applied to describe its content. For performance reasons, features are usually not extracted for each frame of the shot but only for selected keyframes. Each keyframe is represented by a set of features and compared to each frame of the second video. The similarity between frame features is used to assign a keyframe to a shot by means of frame voting. Dependent on the selected feature, various distance metrics can be applied to measure the visual similarity. In our work, we use nearest neighbor ratio matching based on Euclidean distances, i.e. two features are considered similar if the distance to the second most similar feature is above a predefined threshold. Furthermore, we introduce a frame confidence measure, c_f, based on the distance spreading of all matches, i.e. if all distances lie closely together, the corresponding match is considered less reliable. In contrast, an outlier suggests a high matching confidence:

$$c_f = 1 - \frac{d_m}{d}, \tag{5.2}$$

where d_m is the mean matching distance of the matched features, and d the mean matching distance of all descriptors. Finally, the comparison for each keyframe results in a quadruple holding the frame position in the current shot, the frame confidence, the frame voting and the positioning of the frame within the matched shot.

5.3.1.2 *Shot Level*

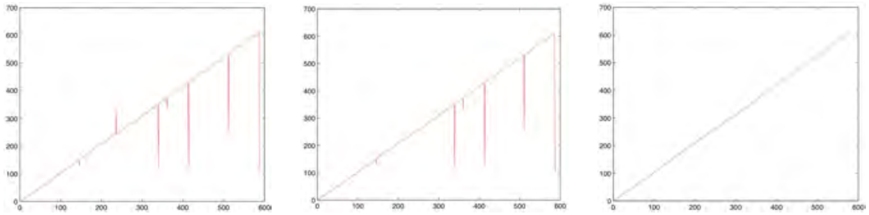
Since a video shot consists of frames, three factors may influence the shot matching decision: 1) frames' votes, 2) corresponding frame confidences and, optionally, 3) the temporal ordering of the frames. In our evaluation, at least two out of the three keyframes have to vote for the same shot, otherwise the shot is rejected and classified as unknown. The shot confidence, c_s, accounts for the confidence of the voting frames and is defined as the average of their scores:

$$c_s = \frac{\sum_{i=1}^{n} s_i \times c_{fi} \times w_{fi}}{\sum_{i=1}^{n} s_i} \begin{cases} s = 1 & \text{for a voting frame} \\ s = 0 & \text{otherwise} \end{cases} \tag{5.3}$$

where c_{fi} is the frame confidence of the i-th frame and n the number of keyframes in the shot, and w_{fi} the weight factor for the corresponding temporal position. If matched keyframes have corresponding temporal positions within the respective shots, $w_{fi} = 1$, otherwise $w_{fi} = 0.8$. The consideration of the temporal ordering of the frames increased the precision score by up to 10% in experimental tests. Additionally, we apply the majority rule, i.e. the majority of the keyframes have to vote for the same shot otherwise the shot is rejected and classified as unknown.

5.3.1.3 *Video Level*

The video level represents the highest layer in the framework. Given the domain of video comparison, the corresponding shots in different videos build a well-defined sequence. This additional knowledge is used to eliminate matched shots, which do not fit in the overall ordering sequence. To detect outliers we apply local minima and maxima suppression on the sequence formed by the respective shot ids (see Figure 5.15). Finally, the average confidence score of matched shots is defined as video confidence c_v.



(a) Shot id correspondences after shot-level analysis

(b) Maxima suppression

(c) Minima suppression and final correspondences

Figure 5.15: An example for minima/maxima suppression at video-level.

A further approach to increase the performance of film comparison is the investigation of shots that have no corresponding shots in the second video. Such shots can be either missed shots (due to failed comparison) or unique shots. Figure 5.16 visualizes a scenario with both matched and unknown shots in different video versions. Since the search field for corresponding shots is limited to a well defined area, the matching performance can be further improved while the additional computational costs are low.

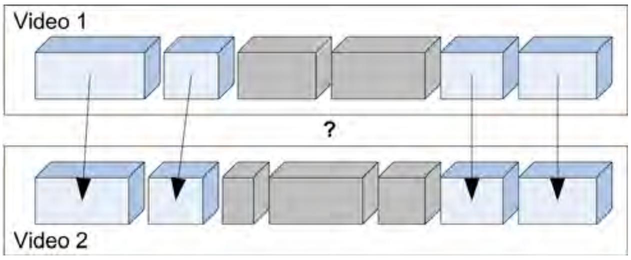


Figure 5.16: Video sequence with matched and unknown shots.

5.3.2 *Methods Compared*

The choice of underlying technology is crucial at each step of the film comparison process. The proposed approach defines the logic of the comparison process. It does not define the specific techniques used at each level. The user has the opportunity to select the adequate method at each step of the process. The three most important factors are:

1. the selection of shot detection algorithm;

2. the selection of keyframes as representatives for a given shot; and

3. the feature representation of the keyframes.

In this section we give a brief description of different approaches for shot boundary detection, keyframe selection, and feature extraction.

5.3.2.1 *Shot Boundary Detection*

Shot boundary detection is a basic preprocessing step for most high-level video analysis tasks such as scene segmentation and video summarization. Many different techniques have been proposed in the last decades. The principle behind different approaches is similar. Usually, differences between consecutive frames are computed. If the differences exceed a certain threshold a shot cut is identified. In shot cut detection the single frames are usually represented by compact features, which are based on color, intensity, edges, motion, and frequency information. We evaluate three standard methods in cut detection, which are based on edges and intensity information. The fourth method (self-similarity matrix) is based on edge and frequency information and was specifically adapted to low-quality archive material.

INTENSITY HISTOGRAM. This method is based on the bin-wise differences of the intensity histograms between two consecutive frames [50]. To include spatial information, each frame is divided into M non-overlapping sub-images:

$$D^{IH}_{F_1,F_2} = \frac{1}{MN} \sum_{j=1}^{M} \sum_{i=1}^{N} |h_{1,j}[i] - h_{2,j}[i]|, \tag{5.4}$$

where h_j is the corresponding sub-image histogram and N the number of bins. A shot cut is detected if the difference exceeds a predefined threshold.

ADAPTIVE THRESHOLD. Instead of using a global threshold on the histogram differences, Truong et al. propose the use of a simple adaptive thresholding method to detect peaks in the histogram difference curve [153]. An adaptive threshold usually adapts better to local properties of the difference curve such as motion and flicker. The authors consider a sliding window along the temporal axis. They detect a shot cut if a given histogram difference 1) has the maximum value within the window, and 2) is α times greater than the mean of remaining histogram differences within the window.

EDGE CHANGE FRACTION. The basic idea of this approach is that the positions of edges change considerably at shot boundaries: existing edges disappear and new edges appear where there were no edges before [172]. Following, the authors count the entering, ρ_{in}, and vanishing, ρ_{out}, edge pixels among two frames and define the edge change fraction as

$$\rho = \max(\rho_{in}, \rho_{out}). \tag{5.5}$$

Peaks in the edge change fraction indicate shot cuts. By analysis of the spatial distribution and relative values of entering and exiting edge pixels the authors classify the shot cut type as a cut, dissolve, fade or wipe.

SELF-SIMILARITY MATRIX. This method is based on the self-similarity between adjacent video frames [176]. First, each frame is split uniformly into blocks and for each block an edge histogram and the low-frequency DCT coefficients are extracted. The edge histogram captures the orientations of the edges and is robust to frame displacements and flicker. The DCT feature represents the coarse spatial intensity distribution across a frame and is robust against dirt and scratches. The features are well-suited for combination, since they capture complementary information. Next, the similarity matrices for both features are computed separately. Sequences of similar frames produce bright squares along the diagonal of the matrix. Shot cuts are detected by moving a Gaussian weighted checkerboard kernel along the diagonal of the similarity matrix. The checkerboard kernel yields high correlation at the shot cuts and low correlation at other positions. Finally, the two similarity matrices result in two kernel correlation functions that are linearly combined. Shot boundaries are located by means of peak detection.

5.3.2.2 *Keyframe Selection*

There are different approaches for the selection of keyframes from a given video shot. Simple techniques do not account for the shot content but rather select the keyframes according to a predefined pattern. More sophisticated methods consider the visual dynamics of a shot and perform keyframe selection based on visual characteristics (e.g. color histograms) or motion information. Law-To et al. select keyframes corresponding to extrema of the global intensity of motion in a comparative study of video copy detection algorithms [83]. Originally, this approach was proposed by Eickeler and Müller [42]. To explore the influence of keyframe selection on the film comparison, we evaluate the following approaches:

1. KS1: always select the first frame as a keyframe [112],
2. KS2: the first and the last frames as keyframes [126],
3. KS3: the first, middle and last frames as keyframes [174], and

4. KS4: motion-based selection of keyframes [42].

Eickeler and Müller [42] define intensity of motion as:

$$i(t) = \frac{\sum_{x,y} d(x,y,t)}{XY}, \tag{5.6}$$

where $d(x,y,t)$ is the difference image of the gray values of adjacent frames. To overcome the problem of abrupt visual changes caused by e.g. flashes, the authors propose to use the smaller value of the motion intensities for the frames $(t, t+1)$ and $(t-1, t+2)$.

5.3.2.3 *Feature Extraction*

Different features can be used for the representation of keyframes. We compare three different types of features with a complementary structure. The MPEG-7 Edge Histogram is a global statistical descriptor, the SIFT features are local image descriptors and the differential-based descriptors capture representative information for an entire shot (for background on the features see Chapter 3).

MPEG-7 Edge Histogram

The MPEG-7 Edge Histogram descriptor is an effective feature for image similarity retrieval [97]. Furthermore, it possesses promising characteristics for the comparison of archive films. The feature captures global information within each block and, thus, is highly robust against frame displacements and invariant to flicker. Since it captures high-frequency information, it is prone to local artifacts (e.g. scratches, dirt) and reflects global artifacts such as tears across the entire frame.

SIFT

The SIFT descriptors are highly discriminative local features. They are invariant to changes in translation, scale, and rotation and partially invariant to changes in illumination and affine distortions. Thus, frame displacement and flicker have no influence on the features. Artifacts, which result in loss of visual information (scratches, dirt, tears), automatically lead to loss of potential keypoints. However, since there is a large number of keypoints per frame, their fraction does not impede the matching process significantly.

Differential-based descriptors

Finally, recently, several authors reported outstanding performance of differential-based descriptors in the context of video copy detection [70, 82, 83]. In this evaluation we follow the approach proposed by Law-To et al. in [82], which was reported as top-performing in a comparative study on video copy detection algorithms [83]. Unlike the original approach, we do not distinguish between motion and background trajectories. Bouncy and unsteady video sequences often exhibit high motion characteristics, which may lead to mislabeling and, thus, misclassification.

5.3.3 *Experiments*

In this section we present the performed evaluations. We selected three case studies that cover, next to the artistic challenges of the film material, different issues and challenges in the film comparison process from a technical point of view. Examples include different film source material, resultant artifacts, as well as various technical and editorial modifications. The case studies are described in Section 5.3.3.1. For the presented data set we evaluate the influence of the choice of underlying technology at each step of the film comparison process: shot boundary detection (Section 5.3.3.2), and keyframe selection and feature representation (Section 5.3.3.3). Eventually, our last experiment addresses a novel application that aims at the identification of unique shots (Section 5.3.3.4).

5.3.3.1 *Video Data and Case Studies*

We explore ten historical artistic documentaries (see Section 5.1) grouped into three case studies:

CS1 The first case study investigates two films by Dziga Vertov in two versions respectively: *Man with a Movie Camera* (1929) and *Enthusiasm* (1930). All films originate from tape-based analog sources. In one case, the copies were derived from the original source with several decades in between. They differ greatly in image quality and censored content. In the other case, the two versions originate from the same analog copy. One copy is the result of an effort to manually reconstruct the original film by manually adding and removing shots.

CS2 The second case study investigates again two films by Dziga Vertov – *Cinema Eye* (1924) and *Three Songs About Lenin* (1934) – in two different versions whereas the second copies originate from unknown sourced DVDs. In addition to the differences in image quality and content, digitization artifacts further impede the process of film comparison.

CS3 The last case study compares two different but related analog films: an original documentary by Dziga Vertov – *The Eleventh* (1928) – and a compilation film by A.V. Blum – *In The Shadow of The Machine* (1928) – where a number of shots from Vertov have been used.

The length of the films ranges from 20 to 90 mins. Similarly, the length of the shots is strongly varying from 1 to over 1500 frames/shot.

5.3.3.2 *Shot Boundary Detection*

The first step of an automated film comparison process is shot boundary detection. Shot boundary detection is widely seen as solved for contemporary film material. However, experiments with shot boundary detection demonstrate the task still being challenging in the context of archive film material (see Table 5.2 for a summary of the achieved

		R	P
CS1	Intensity Histogram (IH)	0.68	0.68
	Adaptive Threshold (AT)	0.67	0.65
	Edge Change Fraction (ECF)	0.67	0.67
	Self-Similarity Matrix (SSM)	**0.91**	**0.91**
CS2	Intensity Histogram (IH)	0.82	0.77
	Adaptive Threshold (AT)	0.88	0.92
	Edge Change Fraction (ECF)	0.77	0.76
	Self-Similarity Matrix (SSM)	**0.95**	**0.95**
CS3	Intensity Histogram (IH)	0.77	0.77
	Adaptive Threshold (AT)	0.80	0.81
	Edge Change Fraction (ECF)	0.75	0.76
	Self-Similarity Matrix (SSM)	**0.93**	**0.93**

Table 5.2: Recall (R) / Precision (P) results for shot boundary detection.

results). Unintended alterations of the content such as artifacts that generate abrupt visual changes (e.g. dirt, scratches, and film tears, see Figure 5.3) interfere with established algorithms that are based on pixel differences (*Intensity Histogram*, IH, and *Adaptive Threshold*, AT) and edge information (*Edge Change Fraction*, ECF). Furthermore, these methods are sensitive to global motion such as camera shaking, large object and camera motion. In such cases, preceding motion compensation can be performed. However, preliminary tests showed that prior motion compensation of archive film material introduces new artifacts that impede following feature detection and analysis. The *Self-Similarity Matrix* (SSM) outperforms significantly the other tested methods in terms of recall and precision and proves robustness for the complex spatio-temporal structure and manifold artifacts of archive film material. The use of robust image features and the larger analysis window (checkerboard window size of up to 8 frames) significantly increases the robustness of the method. Thus, all following experiments are performed based on the shots detected by the SSM method.

5.3.3.3 *Keyframe Selection and Feature Representation*

In this section we present the evaluation results of different keyframe selection methods (KS1-4) in combination with the MPEG-7 edge histogram descriptor (EHD) and the SIFT features. Since the differential descriptors (DD) are based on feature trajectories and, thus, process each frame of a shot, their performance is reported only at the shot- and video-level.

EHD features are matched using simple Euclidean distance. Local feature descriptors (SIFT and DD) additionally identify the first two nearest neighbors in terms of Euclidean distances. A descriptor is accepted if the nearest neighbor distance ratio is below a predefined threshold of 0.8. Since the local descriptors represent the characteristics of a small area around a point of interest, these approaches usually result in a high number of matching descriptors. Given the partially high similarity between different shots, the total number of matches is often misleading. To increase the reliability of detected matches we ignore all ambiguous matches, i.e. all descriptors are eliminated that match several features in the other frame. Additionally, the RANdom SAmple Consensus (RANSAC) algorithm is applied to remove outliers that do not fit a homography transformation [48]. Finally, a frame votes for the shot of the frame with most matches. However, if less than 5% of the descriptors are matches or if the frame confidence is below 50-60% (depending on the features extracted), the match is considered unreliable and is ignored, i.e. the frame is classified as unknown. At shot-level, the required shot confidence score is set initially to 60-70%. All shots with lower confidence score are rejected and classified as unknown. At video-level, we account for the temporal alignment of matched shots and discard shots that do not fit in the ordering sequence by applying peak (local minima and maxima) detection. Finally, all unknown shots are re-evaluated by reducing the required confidence score for a positive match by 10%.

CASE STUDY 1 (CS1). The first case study focusses on the comparison of analog sourced films. The different film versions share around 90% of all shots. In general, the remaining (unique) shots are the result of loss in the process of storage or copying during the years. However, corresponding shots bear also partially large differences due to e.g. film tears, contrast differences and removed frames (see for examples Figure 5.17).

Table 5.3 summarizes the experimental results in terms of recall-precision measures. In general, the SIFT features outperform the remaining descriptors independently of the keyframe selection method and on all three levels of the comparison. Surprisingly, the performance difference to MPEG-7 edge histogram (EHD) is very low. EHD proves to be a very competitive descriptor and as performant as the computationally more expensive SIFT algorithm. In terms of recall and precision, EHD scores 0.90 and 0.98 respectively whereas SIFT achieves 0.92 and 0.99. Although the differential descriptors (DD) build on information from each frame of a given shot, they show very low performance. An analysis of the extracted features shows very low variance, which results in low distinctiveness of the computed descriptors. Thus, such descriptors are only applicable for highly discriminative data. In a scenario of multiple low quality shots with high

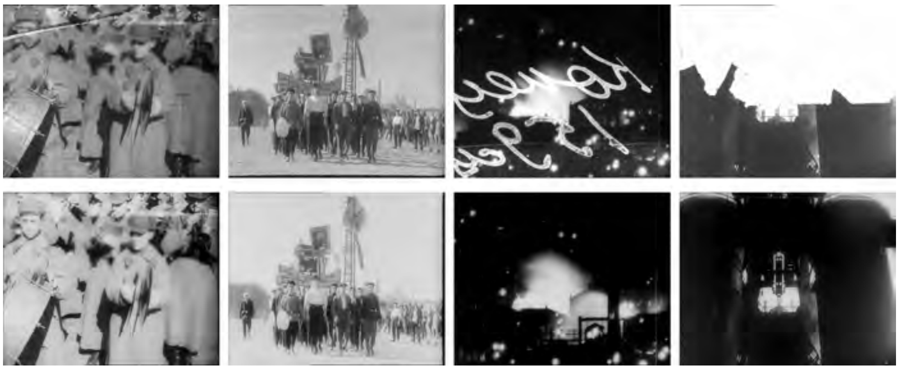


Figure 5.17: Examples for differences in corresponding shots in different film versions (each shot is represented by its first frame). First shot: film tear and illumination differences. Second shot: additional frame displacement (see the black lines on the corresponding frame borders. Third shot: frame mark removed in the second film version. Fourth shot: high contrast difference.

visual similarity, this approach fails to correctly assign corresponding shots.

		Frame-Level		Shot-Level		Video-Level	
		R	P	R	P	R	P
MPEG-7	KS-1	0.89	0.90	0.89	0.90	**0.90**	**0.98**
EHD	KS-2	0.85	0.87	0.84	0.95	0.87	0.97
	KS-3	0.88	0.87	0.89	0.93	0.90	0.96
	KS-4	0.83	0.85	0.89	0.90	0.90	0.96
SIFT	KS-1	0.92	0.90	0.92	0.90	**0.92**	**0.99**
	KS-2	0.86	0.90	0.82	0.96	0.85	0.98
	KS-3	0.86	0.91	0.89	0.96	0.90	0.97
	KS-4	0.87	0.83	0.89	0.91	0.90	0.94
DD	–	–	–	0.58	0.62	0.59	0.96

Table 5.3: Recall (R) / Precision (P) results for CS1.

The comparison of the keyframe selection methods shows that – for the given case study – KS1 (first frame is a representative for the shot) outperforms the remaining methods closely followed by KS3 (selection of the first, middle, and last frame for a shot) and KS4 (motion-based selection of keyframes). KS2 (first and last frames are keyframes for a shot) results in lower performance due to the majority decision rule on the frame-level of the framework, i.e. if both keyframes vote for different shots, the votes are discarded and the shot is classified as unknown even if one of the frames is assigned correctly. In general, the performances of all keyframe selection methods lie closely together.

However, the computational costs differ significantly. KS1 results in a single frame as representative for each shot. In contrast, within the given data set, KS4 results in the selection of 1 to 30 keyframes per shot, which increases the number of required comparisons significantly.

False positives

An analysis of the false positives reveals two facts. *First*, the investigated films contain a large number of static, repeating shots. In general, such shots are assigned correctly outside of the context of a complete film. Within the given context, they are often assigned to an identical shot that appears on a different position in the film sequence. Thus, a generally correct match is classified as a false positive. *Second*, the large number of shots with high perceptual similarity also increases the false positives at frame- and shot-level (see Figure 5.18 for examples). Since the video-level of the framework accounts for the temporal ordering of the shots, such false positives are easily identified and correctly re-assigned.

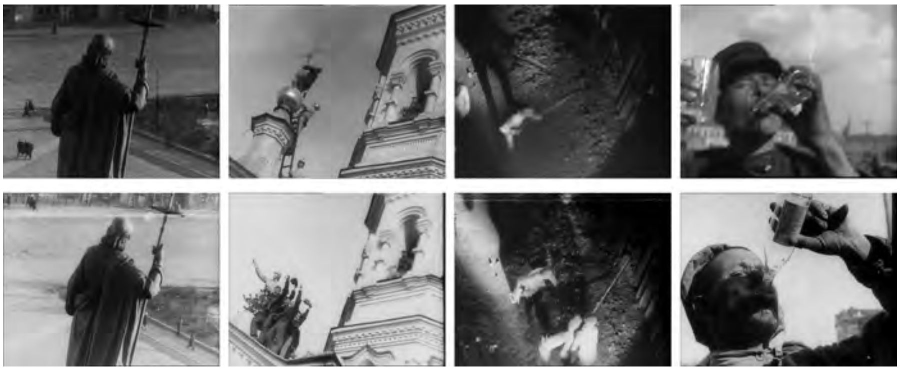


Figure 5.18: Examples for false positives. The assigned shots bear high visual similarity. In the first three examples the shots present the same scene settings with slightly different motives (e.g. people walking by or different workers). Despite the different subject in the last example, both shots have identical composition and action flow.

Figure 5.19 shows resulting correspondences from the film comparison for the *Man with a Movie Camera* (1929). The results show missing shots in Film A (shots 366 and 367 from Film B) which were filled with black frames (shot 386 from Film A). Especially noteworthy is the loss on information due to the introduction of an optical track at the left side of the film A as well as the contrast differences in the two sequences.

CASE STUDY 2 (CS2). The second case study compares different film versions of different origin: analog and digital. The film versions share around 70% of all shots. In addition to the already discussed artifacts in archive film material, artifacts that result from a preprocessing of the material (e.g. noise reduction, stabilization, contrast enhancement) as well as the coding technology lead to high differences in the visual perception of different shots (see Figure 5.20).

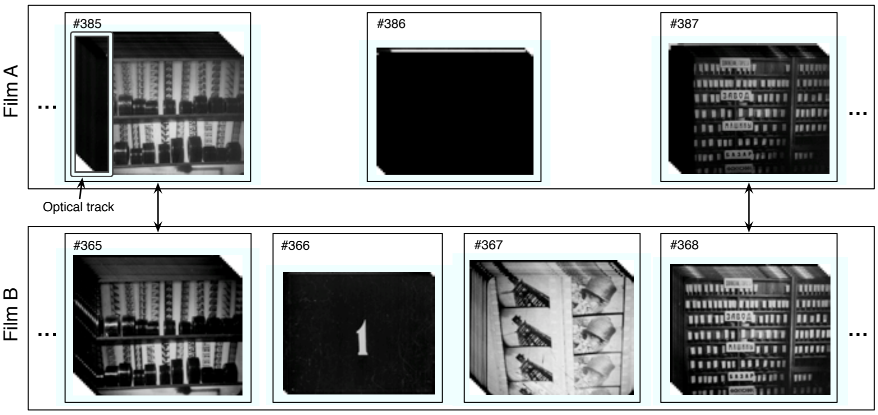


Figure 5.19: Experimental results from the automated film comparison.

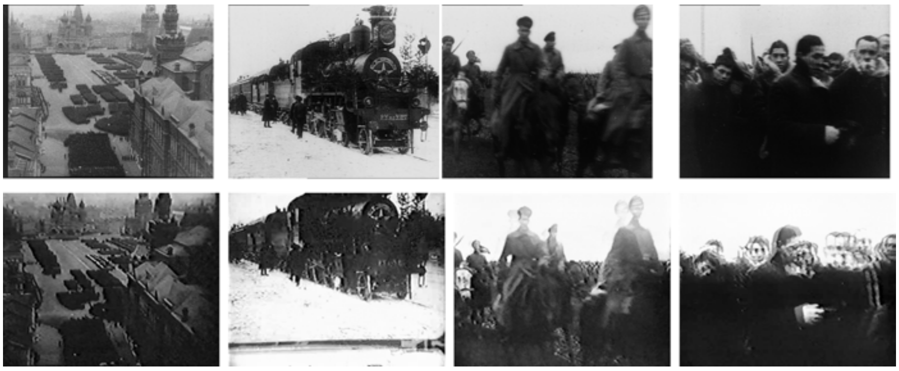


Figure 5.20: Examples for differences in corresponding shots resulting from a preprocessing step (e.g. noise reduction and contrast enhancement) and coding and compression technology.

The results of the experiments are presented in Table 5.4. Again, SIFT features are the top performing approach. In contrast to the first case study, the selection of three keyframes as representatives for the shot proves to be the best keyframe selection method. The manifold artifacts presented in this case study require for a robust decision rule at shot- and video-level of the framework. In general, all recall-precision scores are slightly lower than those achieved in the first case study because of the intensification of the presented artifacts as well as the introduction of new artifacts due to preprocessing and video coding. MPEG-7 edge histogram bears higher sensitivity to motion artifacts in shots (see for examples the last two shots pictured in Figure 5.20) and, thus, often fails to classify shots with large motion. The differential-based descriptors completely fail to achieve any reasonable results.The low performance of the method on the shot-level of the framework does not allow for further evaluation on video-level. The video-level involves peak detection in the ordering sequence of shots and requires a precision of at least 51% for the detection of a reliable sequence.

		Frame-Level		Shot-Level		Video-Level	
		R	P	R	P	R	P
MPEG-7 EHD	KS-1	0.65	0.56	0.65	0.56	0.81	0.82
	KS-2	0.57	0.57	0.58	0.62	0.81	0.78
	KS-3	0.69	0.60	0.69	0.75	**0.85**	**0.82**
	KS-4	0.63	0.57	0.68	0.58	0.79	0.79
SIFT	KS-1	0.81	0.88	0.81	0.88	0.87	0.85
	KS-2	0.76	0.75	0.76	0.77	0.76	0.88
	KS-3	0.79	0.74	0.83	0.84	**0.91**	**0.88**
	KS-4	0.79	0.80	0.81	0.83	0.86	0.84
DD	–	–	–	0.03	0.02	–	–

Table 5.4: Recall (R) / Precision (P) results for CS2.

CASE STUDY 3 (CS3). The last case study compares an original documentary and a compilation film that uses less than 5% of the shots. This case study clearly outlines the limitations of the MPEG-7 edge histogram and the differential-based descriptors: both approaches fail to find sufficient corresponding shots. However, SIFT features also achieve very low performance. Best recall and precision scores are 40% and 70% respectively using the KS1 keyframe selection method.

5.3.3.4 *Unique Shot Detection*

Our last experiment investigates the identification of shots that are unique. The SIFT descriptor correctly identifies 83% of all unique shots followed by the MPEG-7 Edge Histogram (EHD) with 67% and the differential-based descriptors with 29%. The large differences on a percentage basis are due to the low number of unique shots in our data set. Out of 44 unique shots only 24 are longer than three frames which is a basic requirement in our framework. Thus, the absolute difference of 4 shots between the performance of the MPEG-7 Edge Histogram and the SIFT descriptors is relatively low (see Table 5.5 for details).

	Unique shots	MPEG-7 EHD	SIFT	DD
Man with a Movie Camera (1929)	2	2	2	1
Enthusiasm (1930)	22	14	18	6
Aver.		0.67	**0.83**	0.29

Table 5.5: Unique shot detection.

5.3.4 *Related Work*

Global feature-based approaches

Existing approaches for video copy detection usually rely on the extraction of local and/or global features that are matched against a video reference set. In general, algorithms based on global features allow for efficient computation, search, and indexing. Typical features include color, edge, and motion information. Lie et al. propose a compact binary signature based on color histogram for the recognition of TV commercials [88]. Zhang et al. use color moments and a stochastic attributed relational graph matching to detect duplicated news videos [180]. Kim et al. apply color and motion information to describe video content [73]. Video similarity is measured using a group-based record linkage technique. Leon et al. apply video tomography to create spatio-temporal signatures (see Section 3.2.2) [85]. Bertini et al. propose video fingerprints based on MPEG-7 color and edge descriptors and use edit distance, defined as the minimal cost of insertions, deletions, and substitutions of symbols to make two fingerprints equal, for measuring video similarity [16]. However, such methods align two videos for the entire length and, thus, are inefficient if only a small part of the reference or query video is a copy or if there is a single sequence appearing multiple times in the reference video. This limitation is improved by the approach proposed by Yeh et al. [167]. The authors extend the MSF as proposed in [86] by implementing color histograms. Following, they extend the edit distance to find local alignments of two videos based on the Smith-Waterman algorithm [140]. In general such global feature based methods result in a compact feature representation and, thus, enable efficient search and indexing. However, they are not robust to illumination changes, cropping, and partial occlusions.

Local feature-based approaches

Local feature-based methods overcome these limitations and often achieve better performance. Sivic et al. use a combination of affine covariant regions and SIFT for object and scene matching and retrieval [137]. Laptev et al. propose spatio-temporal fingerprints for event detection [81]. Joly et al. apply the Harris corner detector and a differential description of the local region around each interest point [69]. This approach was shown to be superior over further methods employed in the literature such as ordinal intensity signatures or space-time interest points [83]. Zhou et al. present a video similarity model which combines a normalized chromaticity histogram and shot duration [184]. The proposed model requires color information and uses each frame of a shot to build its visual feature. A further limitation of the approach is that it is only robust to low-level transformations such as frame rate conversion and compression format. Sand and Teller propose an image-registration method for aligning two videos recorded at different times into the spatio-temporal domain [127]. The authors combine interest point based matching and local motion estimation (based on Kanade-Lucas-Tomasi (KLT) frame tracker) for frame alignment. The proposed method has low invariance to affine transformation and high computational costs of several minutes per second of video. Recently, Douze et al. applied a combination of Hessian-Affine detector and SIFT descriptor and integrate it into a bag-of-features framework [40]. The authors report best results on the TRECVID 2008 copy detection task providing manifold video modifications such as contrast change, blur and noise introduction, occlusions and cropping.

Current limitations and scope of the work

Our work shows some similarities to the approach by Ng et al. [112]. The authors propose a tree matching algorithm based on the hierarchical structure of a video. Unlike our work, the authors define the video shot as the lowest hierarchical level of a video structure whereas each shot is represented by its first frame. Following, similarity is computed by a combination of color histogram and shot style (camera motion and the length of a shot). A significant limitation of this work is that it is only useful for comparison of videos which exhibit high distinctive patterns of motions and have not undergone strong modification (e.g. illumination changes or frame deletion/insertion). Furthermore, recent evaluations bear mainly two limitations: *First*, the used data sets comprise video clips of high quality and pre-defined fixed length between 5 and 60 sec. *Second*, the experiments are usually performed using synthetic transformations such as resizing, frame shifting, contrast and gamma modification, gaussian noise additions, etc. Our work differs from previous research in the area of video copy detection in several aspects:

1. We aim at the comparison of complete film versions. The additional knowledge about the video structure allows for the easy integration of temporal constraints of matched video frames and shots and increases the overall matching performance.

2. We evaluate the combination and influence of different state-of-the-art algorithms for shot boundary detection, keyframe selection and feature representation.

3. We perform the evaluation on a real-world video data set of archive film material exposing challenging artifacts such as:
   - artifacts originating from the analog filmstrips, e.g. contrast and exposure changes, blurring, frame shift, dirt, film tears;
   - digitization artifacts, e.g. coding transformations;
   - technical transformations, e.g. changes in video format, resizing, cropping; and
   - editorial operations such as frame/shot insertion and frame/shot deletion.

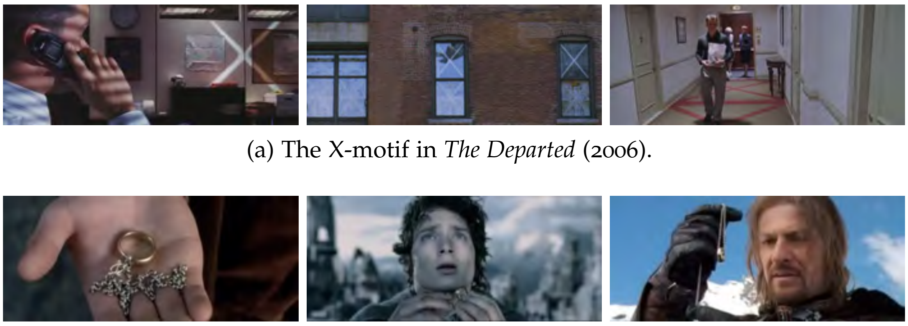
5.3.5 *Conclusion*

We presented an approach for film comparison, which accounts for the overall video structure. Within this framework we compared state-of-the-art methods for shot boundary detection as well as feature representation and investigated the influence of keyframe selection on the performance of the film comparison process. We presented the results of the evaluation based on a real world scenario on challenging archive film material. The results of the performed experiments put the competition between global and local descriptors into perspective. SIFT features are very discriminative and reliable and thus the amount of data to be explored can be reduced significantly. Despite the low quality and partially large differences between corresponding shots, just three frames per shot are sufficient to correctly assign them. However, MPEG-7 Edge Histogram is more competitive than expected. Where SIFT is starting to be more and more the universal weapon with which to attack such problems, MPEG-7 edge histogram proves to be almost as performant as the computationally much more expensive SIFT. Despite the low video quality and partially large differences between corresponding shots, MPEG-7 edge histogram descriptors achieve outstanding performance in terms of recall and precision that is only marginally lower than those of SIFT features .

5.4 recurring element detection

Near-duplicate detection is a rapid emerging research field focused at the identification of identical or near-identical video sequences. Its vast development is mostly driven by requirements of large media providers, advertising agencies, and commercial companies. Near-duplicate detection facilitates application scenarios such as improved search and retrieval of videos by reducing the number of duplicated videos, the monitoring of commercials broadcastings, and copyright protection [131]. While currently near-duplicate detection explores video sequences as a whole, it does not allow for search on more detailed level such as the investigation of duplicated characters or objects. The detection of such recurring elements is a new requirement for automated film analysis stated by art and film experts.

Recurring elements are a common tool in visual arts such as painting, photography, and filmmaking. Examples for such elements can be found in the paintings by Salvador Dali (the piano is typical for his Surrealist compositions) or the films by Dziga Vertov (rails, spinning wheels, etc.) or Alfred Hitchcock (birds, cameo appearances, etc.), etc. Recurring elements are often applied to convey a certain message, theme, or mood. They are usually called motifs and take very different shapes such as the use of a specific color or sound in a given context, a particular movement of an object or character, camera position, composition, or even a story line [14, 18]. Motifs are often highly symbolic. Thus, their detection requires for semantic understanding and relies on the experience and attentiveness of the audience. An example for such a visually highly varying motif is the X-motif in *The Departed* (2006) by Martin Scorsese (see Figure 5.21a). The X appears whenever a character is in mortal danger and takes very different shapes by means of lighting, color, and material. However, motifs can also be easily recognizable such as the ring in the *Lord of the Rings* (2001-2003) by Peter Jackson (see Figure 5.21b).



(a) The X-motif in *The Departed* (2006).

(b) The easily recognizable ring-motif in the *Lord of the Rings* (2001-2003).

Figure 5.21: Examples for motifs in movies.

In this section we explore the feasibility of the detection of recurring elements by means of automated computer vision methods. A clear restriction for such an approach is the requirement for certain similarities in the visual appearance of present recurring elements. We propose a method based on local features that are robust to changes in illumination, rotation, and scaling. Furthermore, the system automatically learns recurring regions and creates links between related views of one and the same object. Following, detected elements may differ significantly in their position, orientation, and scale. In our approach, a region (or element) can be an object, a part of it, or a recurring character (usually the main actors). Finally, the proposed method allows for the detection of recurring elements not only in a single movie but also among different works from the same author and, thus, can support a high-level film analysis currently performed tediously and manually by film experts. In summary, the main contributions of this section are:

- We define a new research task motivated by the requirements of film experts.
- We propose an automated method to detect recurring elements in movies independently of their position, orientation, and scale.
- The output of the proposed system allows for a variety of summarizing visualizations of semantically related information.

This section is organized as follows. Section 5.4.1 describes the algorithm for recurring region detection. Section 5.4.2 presents experiments we performed as proof-of-concept for the evaluation of the proposed algorithm. In Section 5.4.3 we give an overview over related research. We conclude in Section 5.4.4 and give an outlook for further research.

5.4.1 *Approach*

The aim of the proposed system is to detect recurring regions within a video sequence. A video sequence can be a shot, a scene, or a whole movie. Detected regions have to meet two essential requirements. First, they have to be distinctive and not homogeneous regions such as the sky, or a wall. Second, detected regions should allow for a multiple view representation of the captured element. Thus, the proposed system includes two critical components, region detection and region representation, which will be discussed in the following sections (see Figure 5.22).

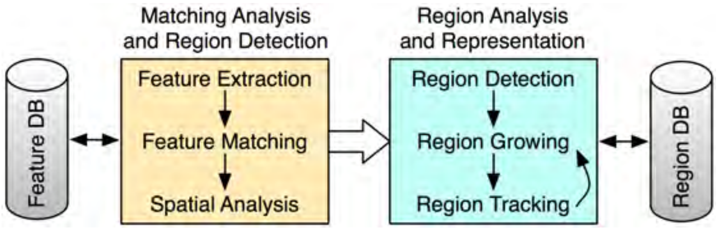


Figure 5.22: Algorithm workflow.

Given a video sequence for recurring object detection, the first step is, as in any general video analysis approach, the detection of shot cuts and the extraction of keyframes as shot representation. Both topics are well-investigated research areas resulting in numerous existing methods. For shot boundary detection we employed the method proposed by Truong et al. [153]. It is a simple adaptive thresholding technique detecting peaks in the histogram difference curve of consecutive frames. For each detected shot, we extract the first, middle, and last frames as keyframes. Despite the simplicity of both methods, they proved to work efficiently with the involved data set and achieved satisfactory results in the performed experiments. Since the input for the proposed system is a sequence of keyframes, they can be easily replaced by more sophisticated methods if needed.

5.4.1.1 *Region Detection*

For each keyframe K^{S_j}, where S_j is the corresponding shot, we detect distinct interest points and extract local features based on SIFT [92]. SIFT features are invariant to changes in translation, scale, and rotation and partially invariant to changes in illumination and affine distortions and, thus, allow for matching across different viewing conditions[1]. Each feature F is described by a quadruple $\{K_i^{S_j}, x, y, D\}$, where K_i is the associated keyframe id, x and y the corresponding coordinates, and D the local feature descriptor.

Following, we perform initial, coarse region detection based on feature matching. Each keyframe is compared to each following keyframe in the input video sequence. Feature descriptors are matched by identifying the first two nearest neighbors in terms of Euclidean distances. A descriptor is accepted if the nearest neighbor distance is below a predefined threshold. The value of 0.8 was determined experimentally and used through the evaluation tests described in Section 5.4.2. To reduce the number of false matches we introduce a loose spatial constraint. Each match is considered within the cluster of its three nearest neighboring feature points. A match is accepted if there is at least one further match present in the cluster. Finally, all accepted clusters are set as initial regions.

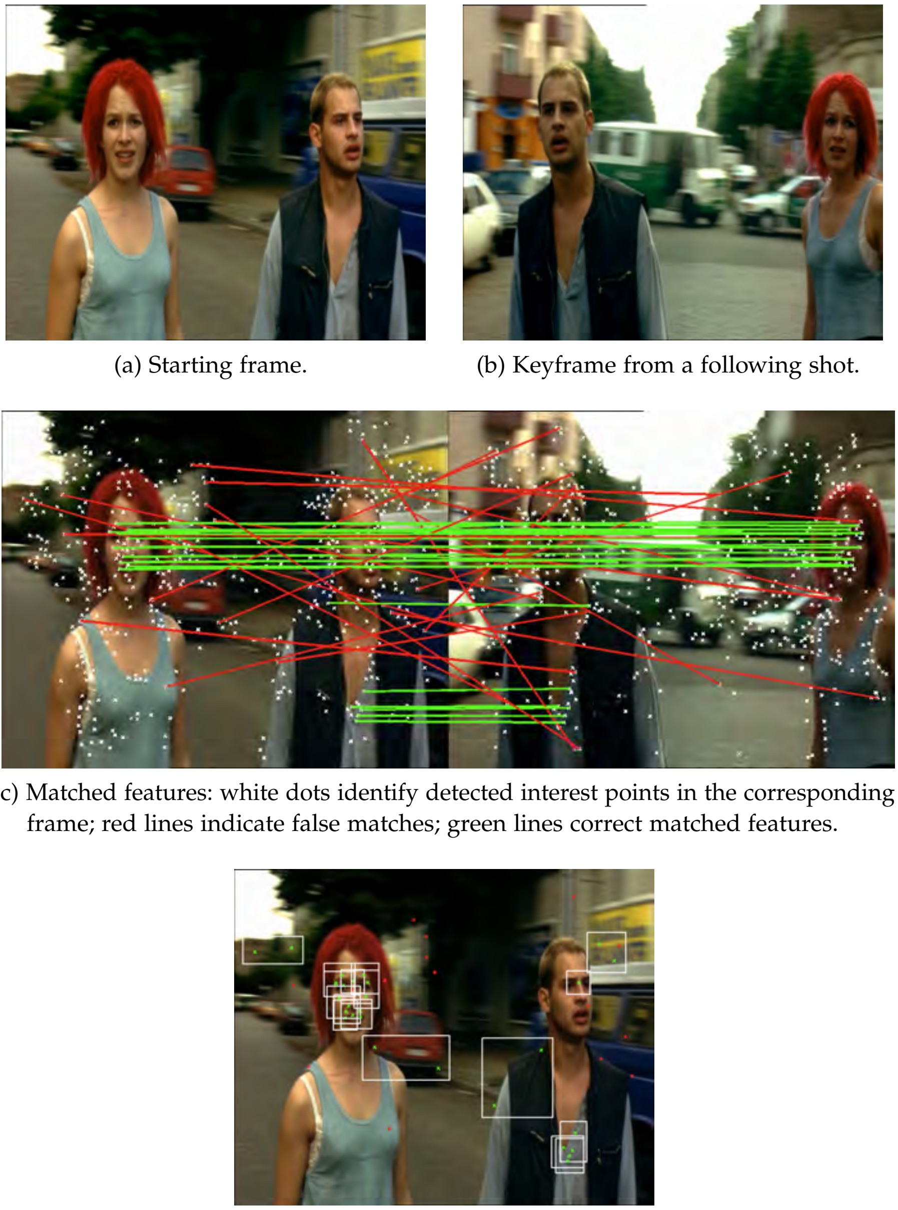
Figure 5.23 shows an example for an initial region detection from the movie *Run Lola Run* (1998) by Tom Tykwer. Compared are two frames from two different shots showing Lola and her boyfriend on the run from the police. The scene is shot from two different viewpoints (see Figures 5.23a-5.23b). Although the matching process produces a number of false positives (see the red lines in Figure 5.23c), most of the false matches are dropped due to the spatial constraint on the next stage of the algorithm (see Figure 5.23d).

5.4.1.2 *Region Analysis and Representation*

The first stage of the algorithm, coarse region detection, results in numerous regions. To reduce their number we first remove all regions with dimensions and area below a given threshold. Following, we perform region growing by detecting and merging all overlapping regions. Finally, each region R is defined as $\{K_i^{S_j}, x, y, w, h, R_M\}$, where w is the width of the region, h its height, and R_M is a set of links to matched regions $\{R_1, R_2, ..., R_N\}$.

Figure 5.24 visualizes the process of region dropping and merging for the previous example from the movie *Run Lola Run*. From 28 initially detected regions, more than 50% were dropped due to the dimensional restriction (see Figure 5.24a). In our experiments we set the minimum for both width and height of detected region to 5 px. In such way a region can be visually perceived and interpreted by the viewer even if it only depicts a small part of an object. Following, all overlapping regions are merged together building preliminary final regions (see Figure 5.24b). Since the whole process of region detection for the starting frame is repeated for all following keyframes,

[1] For details on SIFT see Section 3.2.2



(a) Starting frame.

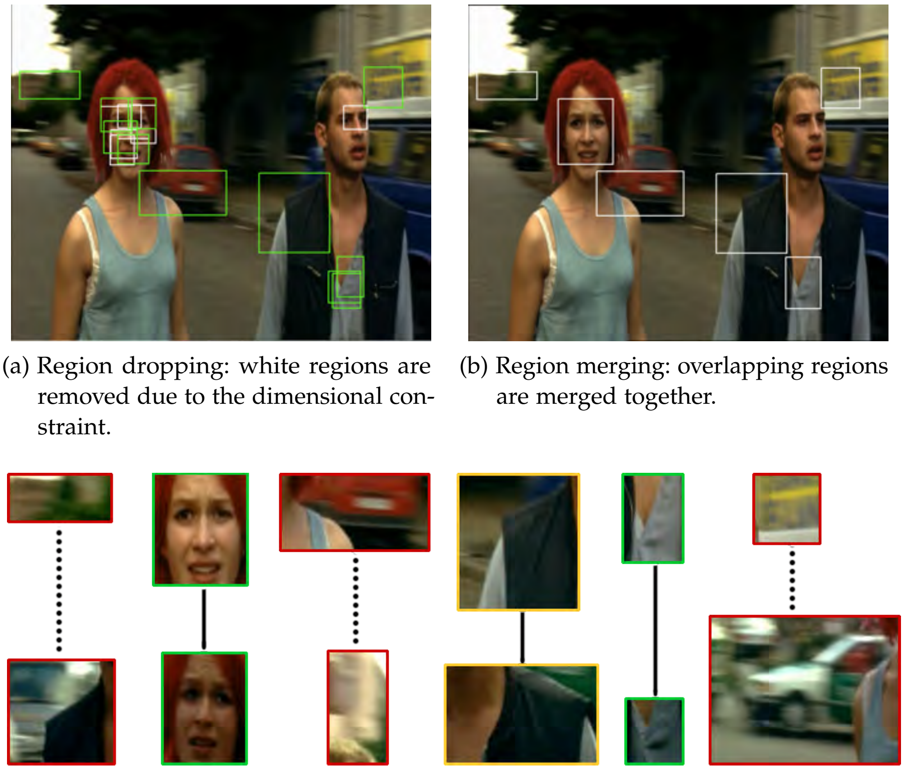
(b) Keyframe from a following shot.

(c) Matched features: white dots identify detected interest points in the corresponding frame; red lines indicate false matches; green lines correct matched features.

(d) Detected initial regions in the starting frame. Red dots indicate dropped features due to the spatial constraint.

Figure 5.23: Example for initial region detection (for better visualization some spacing is introduced within the detected regions).

detected regions are constantly updated in size, quantity, and the set of linked regions. For the detection of final recurring elements all linked regions can be recursively traversed. Eventually, some regions have few repetitions for the whole video sequence while others indicate recurring elements (see Figure 5.24c).



(a) Region dropping: white regions are removed due to the dimensional constraint.

(b) Region merging: overlapping regions are merged together.

(c) Final region linking: red borders indicate false positive linking; yellow borders show templates with similar parts of the same object; green borders indicate correct linked templates. Dotted lines shows elements with very few repetitions for the whole video sequence. Solid lines indicate detected recurring elements for the investigated video sequence.

Figure 5.24: Example for region dropping and merging.

5.4.2 *Experiments*

As proof-of-concept for the proposed algorithm we perform two experiments. The first one focusses on the detection of recurring elements in a single, contemporary movie, and the second one explores the reuse of elements in and among several archived documentaries by the same filmmaker.

5.4.2.1 *Contemporary Movie*

For the first experiment we employed the German movie *Run Lola Run* (1998). The story follows Lola who has 20 minutes to raise 100,000

German marks and save her boyfriend's life. The film presents sequentially three possible scenarios about the story development and its outcome. All three scenarios share the same locations and characters. Following, the film involves many recurring elements (objects as well as characters) which makes it extremely suitable for our experimental tests.

Figure 5.25 depicts the decreasing distribution of the amount of linked shots per detected region. For a better visualization we only show the top 2% regions that have been linked to 7 or more shots. Approx. 98% of all detected regions are linked to less than 7 shots and, thus, considered as insignificant for our application scenario.

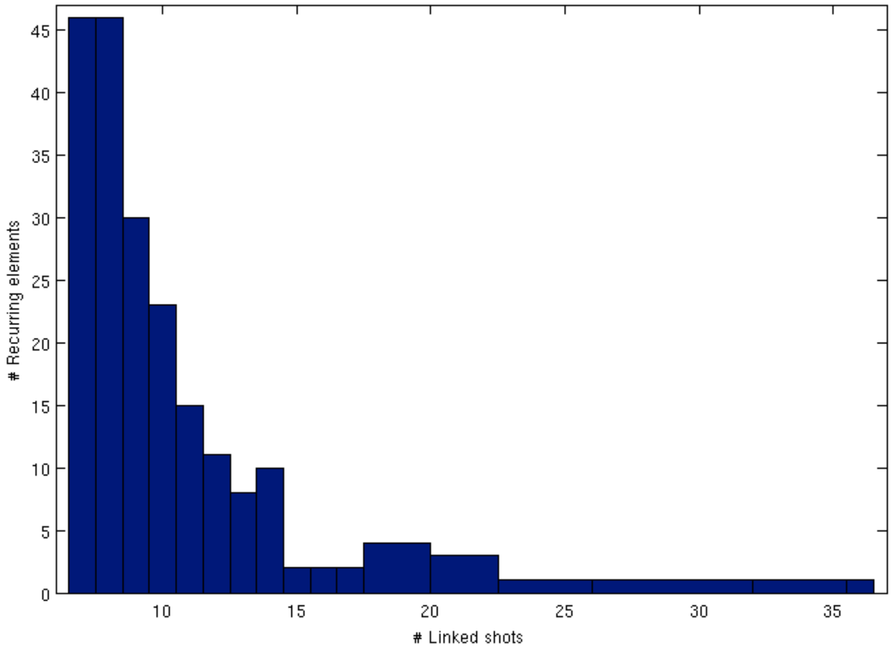


Figure 5.25: Amount of detected recurring elements vs. corresponding amount of linked shots for the top 2% detected regions.

The definition of ground truth for recurring objects in a movie is a tedious process feasible probably by the filmmaker only. Therefore, we focus on the precision performance of the conducted experiments. In our evaluations, we define precision by the ratio of correct linked regions vs. all linked regions. The precision for the top 2% of all detected regions is approx. 75% which confirms the potential of the algorithm. In summary, we investigated over 200 regions with the corresponding associated regions. In average, for each detected region 10 shots have been linked (or 17 regions since multiple keyframes per shot are possible). The average area per region is 38% of the frame size (see Figure 5.26). It turns out that detected regions should not be too small. Anything bellow 5% is not really a meaningful region but rather a part of an object such as a skin section, a shirt detail, a wall texture, etc. Following, the region cannot be tracked reliably since it is found in a large number of frames in spite of their non semantical relation.

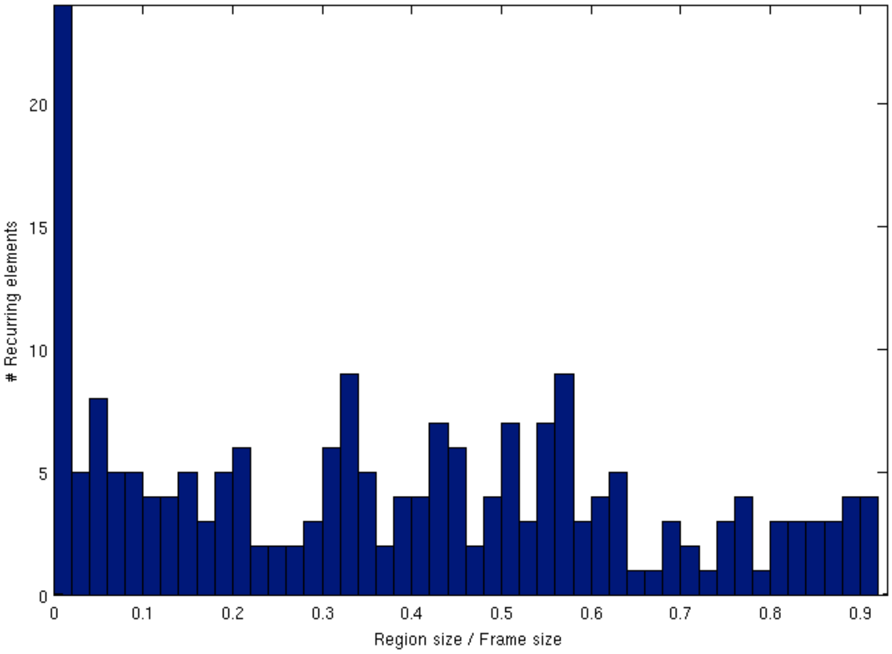


Figure 5.26: Distribution of the size of detected recurring elements.

Currently, few falsely linked regions reduce the overall performance. It is an implication of the approach that if a newly detected region is matched to an existing one, the new region inherits all established links of the second region. Figure 5.27 shows four examples for detected recurring regions. The first example depicts two main characters, Lola and her father, from various scenes in the movie. The remaining examples show recurring objects: a huge dollar bill on the wall of the office of Lola's father, a phone, and a flying bag. The last example also demonstrates a false linking with Lola's hair since the texture of the bag and Lola's hair exhibit high similarities in their texture. Horizontal lines illustrate the level of linked elements. Especially noteworthy is the visual variance within the same level. While in the example with the dollar bill there is a high degree of visual similarity on the level below the top region, in the first example, Lola and her father are matched separately and the linked regions do not have any common visual information although they share the same semantical topic.

Figure 5.28 shows a detected recurring element with its complete set of linked regions. The example shows Lola, running to raise money, her boyfriend looking at the clock on the wall, and a close up of the clock, which is an often occurring scene in the movie. All three objects (the two characters and the clock) are repeatedly found and linked together in various degrees of detail: from the initial close up, via a long shot of Lola, to an extreme close up of her trousers. Based

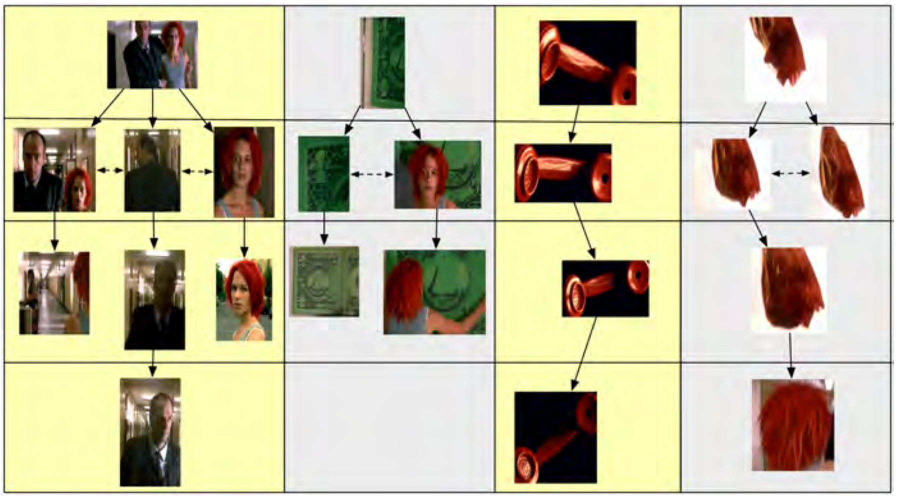


Figure 5.27: Examples for detected recurring elements: solid lines depict directly linked regions; dashed lines shows indirectly associated successors of the same region.

on the region characteristics, the next task could be the classification of regions according to the shooting length into e.g. a close up, a medium, and a long shot. The example illustrates the two main characteristics of the approach. *First*, due to the applied local features in the one-to-one keyframe comparison, directly linked regions (depicted by directional solid lines) share some common visual information. This does not necessarily hold true for the successors of a given region. Note, the green highlighted regions in Figure 5.28. Although, they are both successors of the same region, they do not share common visual information. However, they are both involved into the same semantical topic. *Second*, linked regions can be distributed over the entire movie. Next to the tree representations we used in the discussed examples, a variety of visualization methods can be applied to represent semantically related information based on detected recurring elements such as MPEG-7 collections, hierarchical and sequential summaries, etc.

5.4.2.2 *Archived Documentaries*

The second experiment we performed in the context of recurring element detection investigates three archived documentaries by Dziga Vertov: *Man with a Movie Camera* (1929), *Enthusiasm* (1931), and *Three Songs about Lenin* (1934). The reason to choose the three documentaries is a suggested motif by film experts shared in all three movies: the rails. Hence, we first explore the movies separately and compare the results to those of the contemporary material. Following, we verify whether or not the proposed algorithm is able to detect recurring elements among different works of the same filmmaker.

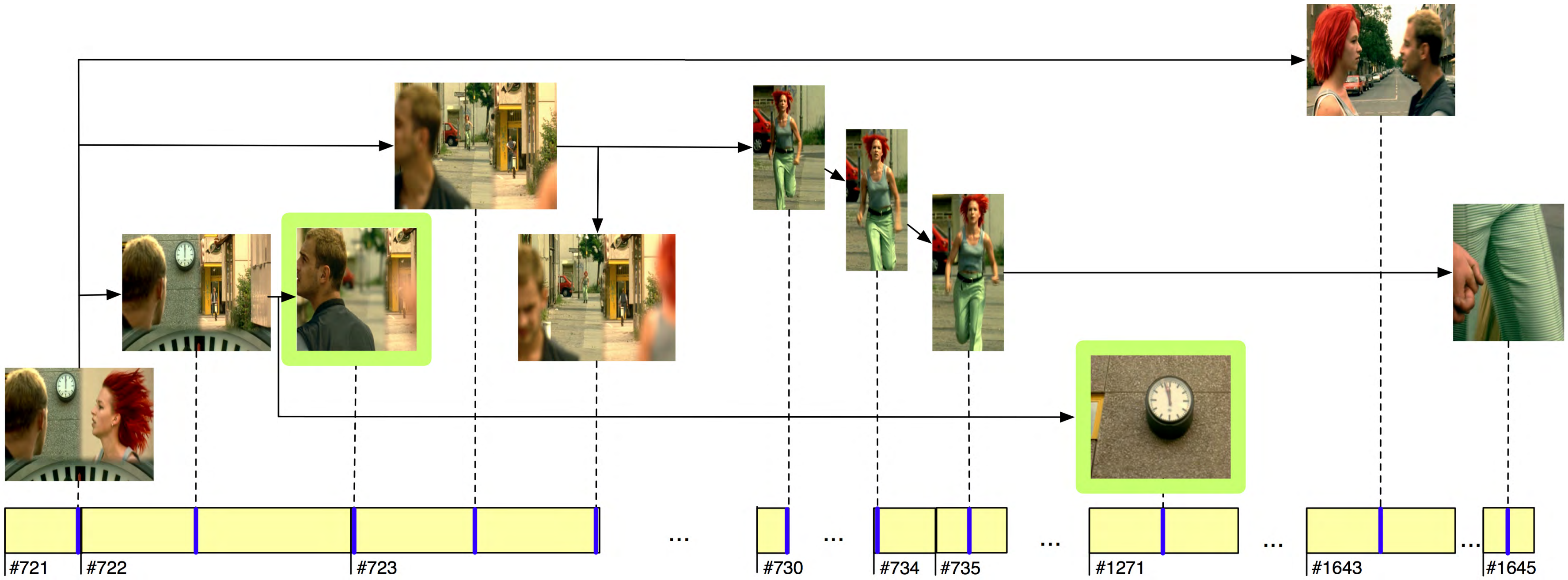


Figure 5.28: A complete example for a detected region and the corresponding linked elements. Yellow boxes depict corresponding shots, blue lines indicate keyframe positions within the shots. Solid lines between the regions show a direct linkage between regions. The two green highlighted regions are an example for siblings of the same region.

In contrast to the contemporary movie from the previous experiment, the explored archived documentaries exhibit less recurring elements with much lower amount of linked shots per detected region (see Figure 5.29). This is mainly due to the fact that most documentaries care less about narration and actors, they often change locations, and characters do not necessarily recur. As a result, detected recurring elements are mostly a long camera take that was cross-cut with a second scene. Hence, such shots exhibit a high visual similarity. Figure 5.30 shows an example for detected recurring elements in the movie *Enthusiasm* (1931)[2].

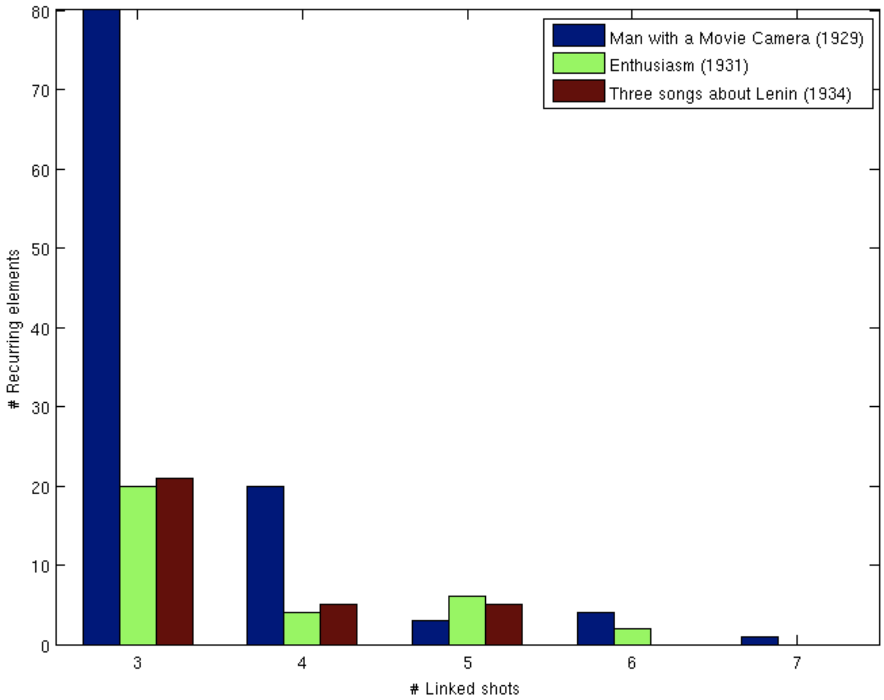


Figure 5.29: Amount of detected recurring elements vs. corresponding amount of linked shots.

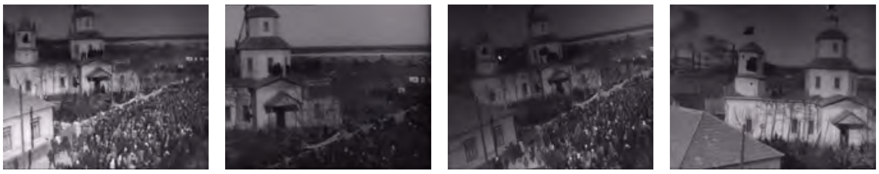


Figure 5.30: An example for detected recurring element in *Enthusiasm* (1931).

In general, documentaries turn out to exhibit, to a greater extent, recurring scenes or sets rather that recurring elements (objects or characters). Following, detected regions occupy predominantly (nearly) the full frame size (see Figure 5.31).

[2] Some examples from *Man with a Movie Camera* (1929) are depicted in Figure 4.32.

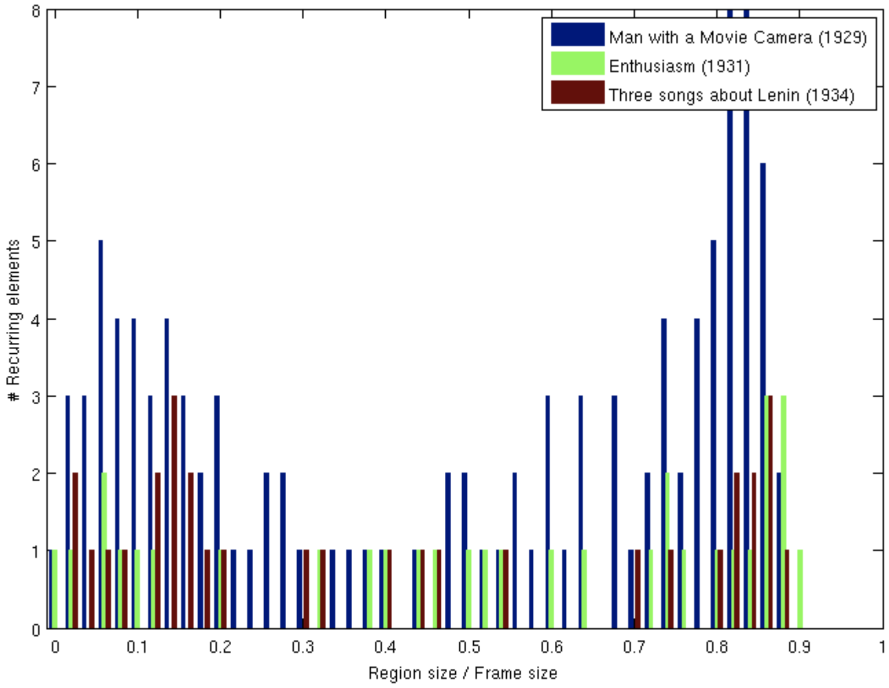


Figure 5.31: Distribution of the size of detected recurring elements.

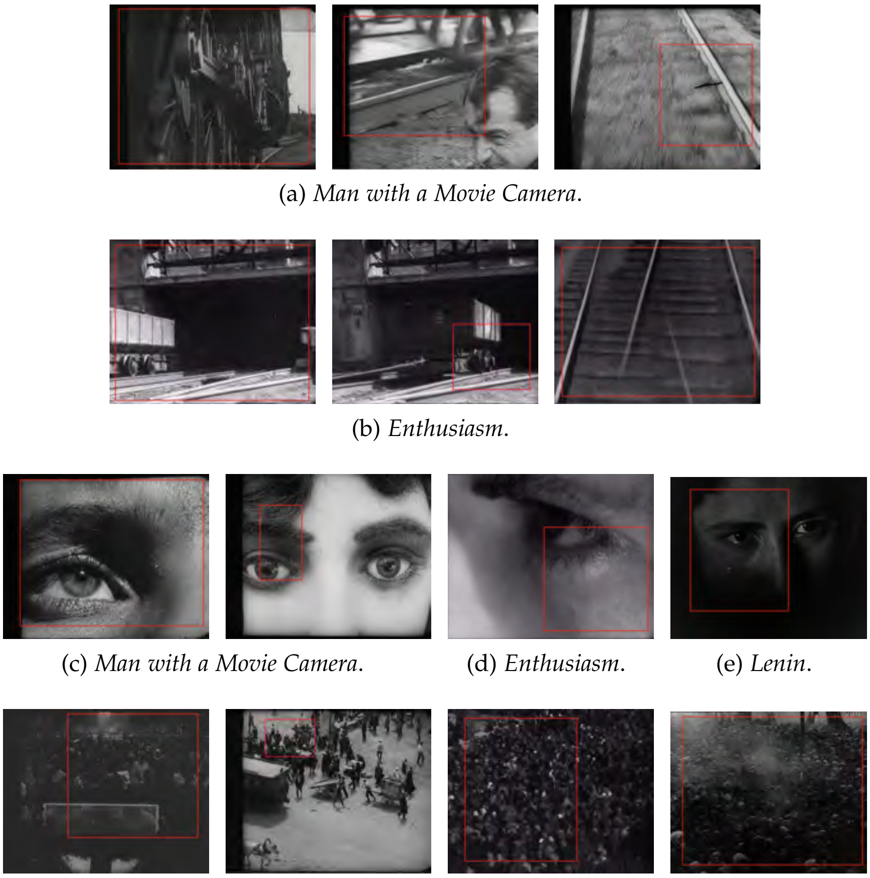
Finally, the precision performance is comparable to the results achieved in the first experiment. The average precision performance for the archived documentaries is approx. 70%.

Starting point for the cross-movie analysis are previously detected recurring regions in each movie. Similar to region tracking within a single movie, corresponding regions are matched using local features and a nearest neighbor ratio matching strategy. For our evaluations at least five matches per region are required to define a reasonable match. Figure 5.32 shows the top three elements detected in the explored movies: rails, eye, and crowd. Despite the partially high visual dissimilarities, all three detected elements represent typical Vertov motifs applied across different works.

5.4.3 *Related Work*

To the best of our knowledge, recurring elements detection has not been subject to research so far. Related research areas comprise near-duplicate detection and object detection and tracking.

Near-duplicate detection aims at identifying images or video sequences showing slight variance due to editing or changes in lighting, viewpoint, motion, etc. [40, 70, 183]. This research area has emerged in recent years for a variety of applications such as the recognition of TV commercials, detection of duplicated news videos, media linking, and copyright infringement detection. Recently, Huang et al. proposed a method for scene recognition based on near-duplicates object detec-



(a) *Man with a Movie Camera*.

(b) *Enthusiasm*.

(c) *Man with a Movie Camera*.

(d) *Enthusiasm*.

(e) *Lenin*.

(f) *Man with a Movie Camera*.

(g) *Enthusiasm*.

(h) *Lenin*.

Figure 5.32: Cross-movie analysis. 5.32a-5.32b: rails-motif. 5.32c-5.32e: eye-motif. 5.32f-5.32h: crowd-motif. All detected regions are embedded into the original frame for better visualization.

tion [64]. The authors argue that shots of the same scene most probably share a large number of similar objects or background. However, the authors do not perform any object but simple keypoint detection and tracking. Following, a shot is represented by an average space-time feature, called imprecisely object key feature. In contrast to recent research in near-duplicate detection, our work performs on more detailed level. While existing approaches detect duplicated or reused media (images or video) as a whole we aim at the identification of recurring elements within a given medium and their reuse among different media.

Object detection and tracking usually requires a predefined appearance model of the salient object or a priori information about the scene for reliable background subtraction and motion tracking [44, 58, 84]. The application scenarios are manifold ranging from traffic control and surveillance to sport video analysis and the recognition of human action. Recently, Celik et al. proposed a method for unsupervised object detection in unlabeled surveillance video data [28, 29]. The authors first detect salient objects based on motion information and simple dimensional features (e.g. height). In the next step, similarity-based clustering allows for the grouping of objects according to the category they belong to. The approach is only applicable in a restricted scene with a static camera. Salient objects have to be moving and within a certain degree of perspective deformation due to the dimensional features in the initial step. Our approach differs significantly from existing methods for object detection and tracking in respect of available knowledge about both object and scene and in respect of the degree of detection, i.e. general category (a person, a car, etc.) vs. a specific subject.

5.4.4 *Conclusion*

In this section we presented a new approach for the detection of recurring elements in movies. Since detected regions can be an object, a part of it, or a character, the system allows for the detection of visually similar motifs and recurring characters. The linking between detected regions shows possible different views of the recurring elements and facilitates the quick retrieval of relevant sequences. Performed experiments with different works by the same filmmaker demonstrate the potential of the proposed algorithm to assist experts in film analysis and film studies.

Part III

SUMMARY

6 SUMMARY AND CONCLUSION

The best way out is always through.

— Robert Frost

Existing approaches for automated film and video analysis bear, for the most part, two essential characteristics:

1. Relevant features are identified in comparison with other movies, e.g. in the context of genre recognition: a horror movie exhibits darker color distribution than a comedy, and

2. Consumer-driven applications aim at an improved retrieval and handling of media data, e.g. video summarization, genre and event-recognition, copy detection, etc.

In contrast, this work addressed the task of automated film understanding from a filmmaking point of view. Instead of the final product, we explored the process of film creation, editing, and presentation as a source for relevant features. Every choice for a given film technique or setting has a purpose. While the automated detection of some well-established film techniques requires for additional knowledge and is not feasible at the current state of research, many other techniques can be analyzed fully automated using computer vision methods. Identified research tasks address mainly the requirements of film experts and improve the process of film understanding and film studies. However, applications for the broader audience can also make use of the acquired knowledge. An example is the use of recurring element detection. The proposed method can reveal a common semantic topic between visually almost dissimilar regions. In a next step, this information can be applied for enriched visualization and summarization methods (e.g. in the context of MPEG-7).

6.1 ACHIEVEMENTS

This work investigated the possibilities for building a common ground between the requirements of film experts and existing computer vision methods for automated film analysis and understanding. Within the scope of this study we presented a mapping between media aesthetic elements and concepts that influence the production, presentation, and perceptions of films, their application by means of well-established film techniques, and existing methods and features in computer vision. This new view on film analysis allowed for the exploration of the boundaries of current research in computer vision and for the identification of open research tasks. Finally, we presented three novel research questions and their solutions in the context of automated film analysis: camera take reconstruction, film comparison, and recurring element detection. Performed experiments bear two significant potentials:

1. The proposed algorithms can assist film experts by providing support for tasks that are currently manually performed (e.g. for film archives and museums, for film studies, and for the filmmaker looking for a specific footage).

2. The proposed algorithms provide a roadmap and pave the way for further application scenarios such as montage pattern analysis, the comparison of different film cuts, the identification of missing shots, the reconstruction of the original film cut, or the detection of recurring elements in the works of the same filmmaker.

6.2 FUTURE DEVELOPMENT

- This thesis presents an initial fundamental research in the context of automated film understanding. It is an attempt to provide a mapping between features that are of particular interest for the film study community and computer vision approaches. Further research, and in particular, intensified communications and a common vocabulary between film experts and computer vision experts is required to complement the proposed mapping and to identify further research questions in the context of automated film analysis and understanding.

- Another major issue is missing ground truth. To enable further research and comparison between different approaches a common set of data is required. However, the definition of ground truth in this context is a notably tedious process feasible probably by film experts only. Furthermore, the resulting ground truth is often shaped by individual judgements and subjectivity.

- Proposed algorithms in this thesis are designed as proof-of-concept. Further work is required to improve the overall performance and computation time in order to provide practical tools for the film study community.

- The performed study identified a large set of achievable research tasks that have not been subject to research so far, e.g. the analysis of film compositions and continuity techniques (see Table 4.1).

- In this work, the differentiation between feasible and not feasible tasks has been made based on the current state-of-the-art in computer vision and under the assumption that there is no a priori knowledge available (e.g. about characters, objects, shapes, etc.). However, tasks, such as rhythm analysis or motif detection, are of particular interest to the film study community (even within predefined constrictions) and bear high potential for further research.

- This thesis focussed on the study of the formal features of film style. We explored films in relation to their filmmakers and in terms of their construction and applied technologies. Film analysis can be further approached from a very different direction, e.g. how films are perceived and responded by the audience. Such approaches are closely dependent on users' experiences and users' preferences. Furthermore, they assume that certain film techniques trigger well-defined, distinct emotions and that filmmakers apply such film techniques according to the predefined association. However, as already discussed in Chapter 2, the psychological interpretation of film techniques is not fully explored yet. Recent research in affective content analysis focuses mainly on the detection and classification of emotions in distinct movie types, such as horror and comedy, using a limited number of film techniques (e.g. color distribution, key lighting, shot length, motion, and audio features). This thesis broadens the horizon for affective content analysis by discussing a wide set of fundamental film techniques and their intended purpose. Further research is required for the identification of the most effective film techniques for affective content analysis. Notwithstanding, such research is mostly applicable for mainstream cinema and not for avant-garde and art movies where the artistry of the filmmaker plays a central role in the applied film techniques and in the shaping of the movie.

- The development and, primarily, the establishment of new technologies involves changes in the filmmaking process: new film techniques appear and existing ones may not be applicable any more. An example for such a technology is stereoscopic or 3D cinema. Next to the strong perception of depth in 3D, notable adaptations include fine-tuned compositions due to the limitations of stereoscopic perception and slower pace due to increased visual complexity. In general, research in the context of automated film analyses stand to substantially benefit from such technology developments. Fundamental film techniques, such as those discussed in this thesis, are still present and essential in filmmaking. However, advanced research tasks, such as recurring element detection that require for object tracking and modeling, can make use of additional information available in stereoscopic images.

BIBLIOGRAPHY

[1] *American Standard Acoustical Terminology*. American Standards Association, 1960.

[2] A. E. Abdel-Hakim and A. A. Farag. Csift: A sift descriptor with color invariant characteristics. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1978–1983, 2006.

[3] B. Adams, C. Dorai, and S. Venkatesh. Automated film rhythm extraction for scene analysis. In *IEEE International Conference on Multimedia and Expo*, pages 1056–1059, 2001.

[4] B. Adams, C. Dorai, and S. Venkatesh. Finding the beat: An analysis of the rhythmic elements of motion pictures. In *Asian Conference on Computer Vision*, 2002.

[5] B. Adams, C. Dorai, and S. Venkatesh. *Media computing: computational media aesthetics*, chapter Formulating Film Tempo: The Computational Media Aesthetics Methodology in Practice, pages 57–79. Kluwer Academic Publishers, 2002.

[6] B. Adams, C. Dorai, and S. Venkatesh. Toward automatic extraction of expressive elements from motion pictures: tempo. *IEEE Transactions on Multimedia*, 4(4):472–481, 2002.

[7] P. Aigrain, P. Joly, and V. Longueville. Medium knowledge-based macro-segmentation of video into sequences. *Intelligent Multimedia Information Retrieval*, pages 159–173, 1997.

[8] J. Anderson and B. Anderson. The myth of persistence of vision revisited. *Journal of film and video*, 45(1):3–12, 1993.

[9] J. Annesley, J. Orwell, and J.-P. Renno. Evaluation of mpeg7 color descriptors for visual surveillance retrieval. In *Second Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pages 105–112, 2005.

[10] R. Arnheim. *Art and Visual Perception: A Psychology of the Creative Eye*. University of California Press, 1974.

[11] D. Ballard. Generalizing the hough transform to detect arbitrary shapes. *Pattern Recognition*, 13(2):111–122, 1981.

[12] A. Baumberg. Reliable feature matching across widely separated views. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 774–781, 2000.

[13] H. Bay, T. Tuytelaars, and L. V. Gool. Surf: Speeded up robust features. In *European Conference on Computer Vision*, volume 3951/2006 of *LNCS*, pages 404–417. Springer, 2006.

[14] F. E. Beaver. *Dictionary of film terms: the aesthetic companion to film art*. Peter Lang Publishing, 2009.

[15] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(4):509–522, 2002.

[16] M. Bertini, A. D. Bimbo, and W. Nunziati. Video clip matching using mpeg-7 descriptors and edit distance. In *International Conference on Image and Video Retrieval*, volume 4071 of *LNCS*, pages 133–142, 2006.

[17] R. A. Block. *Cognitive models of psychological time*, chapter Models of psychological time, pages 1–35. Lawrence Erlbaum Associates, 1990.

[18] D. Bordwell and K. Thompson. *Film art: an introduction*. McGraw-Hill, 8th edition, 2008.

[19] A. Bosch, A. Zisserman, and X. Mu noz. Scene classification using a hybrid generative/discriminative approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(4):712–727, 2008.

[20] E. Britannica. motion picture. Encyclopædia Britannica Online. http://www.britannica.com/EBchecked/topic/394107/motion-picture (last checked: 2011-09-15), 2011.

[21] E. Britannica. time perception. Encyclopædia Britannica Online. http://www.britannica.com/EBchecked/topic/596177/time-perception (last checked: 2011-09-15), 2011.

[22] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert. High accuracy optical flow estimation based on a theory for warping. In *European Conference on Computer Vision*, volume 4 of *LNCS*, pages 25–36. Springer, 2003.

[23] G. J. Burghouts and J.-M. Geusebroek. Performance evaluation of local colour invariants. *Computer Vision and Image Understanding*, 113(1):48–62, 2009.

[24] J. B. Burns, A. R. Hanson, and E. M. Riseman. Extracting straight lines. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(4):425–455, july 1986.

[25] J. Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(6):679–698, 1986.

[26] P. Cavanagh. Short-range vs long-range motion: Not a valid distinction. *Spatial Vision*, 5(4):303–309, 1991.

[27] P. Cavanagh and G. Mather. Motion: The long and short of it. *Spatial Vision*, pages 103–129, 1989.

[28] H. Celik, A. Hanjalic, and E. A. Hendriks. Unsupervised and simultaneous training of multiple object detectors from unlabeled surveillance video. *Computer Vision and Image Understanding*, 113(10):1076–1094, 2009.

[29] H. Celik, A. Hanjalic, E. A. Hendriks, and S. Boughorbel. Online training of object detectors from unlabeled surveillance video. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–7, 2008.

[30] V. T. Chasanis, A. C. Likas, and N. P. Galatsanos. Scene detection in videos using shot clustering and sequence alignment. *IEEE Transactions on Multimedia*, 11(1):89–100, 2009.

[31] L. Chen and M. T. Özsu. Rule-based scene extraction from video. In *International Conference on Image Processing*, volume 2, pages 737–740, 2001.

[32] L. Chen, S. J. Rizvi, and M. T. Özsu. Incorporating audio cues into dialog and action scene extraction. In *SPIE Storage and Retrieval for Multimedia Databases*, pages 252–264, 2003.

[33] I. Cherif, V. Solachidis, and I. Pitas. Shot type identification of movie content. In *International Symposium on Signal Processing and Its Applications*, pages 1–4, 2007.

[34] Y. Cui, J. S. Jin, S. Zhang, S. Luo, and Q. Tian. Music video affective understanding using feature importance analysis. In *ACM International Conference on Image and Video Retrieval*, pages 213–219, 2010.

[35] K. Dancynger. *The technique of film and video editing: history, theory, and practice*. Focal Press, 4th edition, 2007.

[36] M. De Santo, G. Percannella, C. Sansone, and M. Vento. Dialogue scenes detection in mpeg movies: A multi-expert approach. In *Multimedia Databases and Image Communication*, pages 192–201, 2001.

[37] R. Deriche. Recursively implementing the gaussian and its derivatives. In *International Conference on Image Processing*, pages 263–267, 1992.

[38] A. Desolneux, L. Moisan, and J.-M. Morel. Meaningful alignments. *International Journal of Computer Vision*, 40:7–23, 2000.

[39] C. Dorai and S. Venkatesh, editors. *Media computing: computational media aesthetics*. Kluwer Academic Publishers, 2002.

[40] M. Douze, H. Jégou, and C. Schmid. An image-based approach to video copy detection with spatio-temporal post-filtering. *IEEE Transactions on Multimedia*, pages 257–266, 2010.

[41] R. Dyer. *Film studies: critical approaches*, chapter Intorduction to film studies. Oxford University Press, 2000.

[42] S. Eickeler and S. Müller. Content-based video indexing of tv broadcast news using hidden markov models. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 2997–3000, 1999.

[43] A. Ekin, A. Tekalp, and R. Mehrotra. Automatic soccer video analysis and summarization. *IEEE Transactions on Image Processing*, 12(7):796–807, 2003.

[44] A. Ess, K. Schindler, B. Leibe, and L. V. Gool. Object detection and tracking for autonomous navigation in dynamic environments. *International Journal of Robotics Research*, 29(14):1707–1725, 2010.

[45] H. Fastl and E. Zwicker. *Psychoacoustics: Facts and Models*. Springer, 3 edition, 2007.

[46] C. Feng. Code for vanishing point detection using jlinkage and lsd. http://code.google.com/p/vpdetection (last checked: 2011-09-15), 2011.

[47] V. Ferrari, T. Tuytelaars, and L. V. Gool. Simultaneous object recognition and segmentation from single or multiple model views. *International Journal of Computer Vision*, 67(2):159–188, 2006.

[48] 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*, 24(6):381–395, 1981.

[49] S. C. Gaddam. Code for calculating color clouds. http://cns.bu.edu/~gsc/ColorHistograms.html (last checked: 2011-09-15), 2011.

[50] U. Gargi, R. Kasturi, and S. H. Strayer. Performance characterization of video-shot-change detection methods. *IEEE Transactions on Circuits and Systems for Video Technology*, 10(1):1–13, 2000.

[51] J. M. Gauch and A. Shivadas. Identification of new commercials using repeated video sequence detection. In *IEEE International*

Conference on Image Processing, volume 3, pages II–1252–1255, 2005.

[52] M. Gavrielides, E. Sikudova, and I. Pitas. Color-based descriptors for image fingerprinting. *IEEE Transactions on Multimedia*, 8(4):740–748, 2006.

[53] Y. Geng, D. Xu, and A. Wu. Effective video scene detection approach based on cinematic rules. In *Knowledge-Based Intelligent Information and Engineering Systems*, pages 165–165, 2005.

[54] J.-M. Geusebroek, R. van den Boomgaard, A. Smeulders, and H. Geerts. Color invariance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(12):1338 –1350, 2001.

[55] T. Gevers. Image segmentation and similarity of color-texture objects. *IEEE Transactions on Multimedia*, 4(4):509–516, 2002.

[56] X. Giro-i Nieto, R. Salla, and X. Vives. Digimatge, a rich internet application for video retrieval from a multimedia asset management system. In *International Conference on Multimedia Information Retrieval*, pages 425–428, 2010.

[57] M. Grabner, H. Grabner, and H. Bischof. Fast approximated sift. In *Asian Conference on Computer Vision*, volume 3851/2006 of *LNCS*, pages 918–927. Springer, 2006.

[58] W. Guo, C. Xu, S. Ma, and M. Xu. Visual attention based motion object detection and trajectory tracking. In *PCM 2010*, pages 462–470, 2011.

[59] A. Hampapur and R. Bolle. Feature based indexing for media tracking. In *IEEE International Conference on Multimedia and Expo*, pages 67–70, 2000.

[60] A. Hanjalic. Shot-boundary detection: unraveled and resolved? *IEEE Transactions on Circuits and Systems for Video Technology*, 12(2):90–105, 2002.

[61] C. Harris and M. Stephens. A combined corner and edge detector. In *Alvey Conference*, pages 147–152, 1988.

[62] M. Hershenson. *Visual space perception: a primer*. MIT Press, 2000.

[63] B. K. Horn and B. G. Schunck. Determining optical flow. *Artificial Intelligence*, 17(1–3):185–203, 1981.

[64] C.-R. Huang and C.-S. Chen. Video scene detection by link-constrained affinity-propagation. In *IEEE International Symposium on Circuits and Systems*, pages 2834–2837, 2009.

[65] J. Huang, S. R. Kumar, M. Mitra, W.-J. Zhu, and R. Zabih. Image indexing using color correlograms. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 762–768, 1997.

[66] ISO/IEC. *Information Technology - Multimedia Content Description Interface - part 3: Visual*. Number 15938-3. ISO/IEC. Moving Pictures Expert Group, 2002.

[67] L. Itti. Automatic foveation for video compression using a neurobiological model of visual attention. *IEEE Transactions on Image Processing*, 13(10):1304–1318, oct. 2004.

[68] A. Jacquot, P. Sturm, and O. Ruch. Adaptive tracking of non-rigid objects based on color histograms and automatic parameter selection. In *IEEE Workshop on Motion and Video Computing*, volume 2, pages 103–109, 2005.

[69] A. Joly, C. Frélicot, and O. Buisson. Robust content-based video copy identification in a large reference database. In *International Conference on Image and Video Retrieval*, volume 2728/2003, pages 414–424. LNCS, 2003.

[70] A. Joly, C. Frélicot, and O. Buisson. Content-based copy detection using distortion-based probabilistic similarity search. *IEEE Transactions on Multimedia*, 9(2):293–306, 2007.

[71] M. Kampel and M. Zaharieva. Recognizing ancient coins based on local features. In *Advances in Visual Computing*, volume 5358/2008 of *LNCS*, pages 11–22. Springer, 2008.

[72] Y. Ke and R. Sukthankar. Pca-sift: A more distinctive representation for local image descriptors. In *Computer Vision and Pattern Recognition*, volume 2, pages 506–513, 2004.

[73] H.-s. Kim, J. Lee, H. Liu, and D. Lee. Video linkage: group based copied video detection. In *ACM International Conference on Image and Video Retrieval*, pages 397–406, 2008.

[74] H. Kobayashi, Y. Okouchi, and S. Ota. Image retrieval system using kansei features. In *5th Pacific Rim International Conference on Artificial Intelligence: Topics in Artificial Intelligence*, pages 626–635, 1998.

[75] M. Kotti, D. Ververidis, G. Evangelopoulos, I. Panagakis, C. Kotropoulos, P. Maragos, and I. Pitas. Audio-assisted movie dialogue detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(11):1618–1627, 2008.

[76] P. Kovesi. Image features from phase congruency. *Videre: Journal of Computer Vision Research*, 1(3):2–26, 1999.

[77] P. Kovesi. Code for calculating phase congruency and phase symmetry/asymmetry. http://www.csse.uwa.edu.au/~pk/Research/research.html (last checked: 2011-09-15), 2011.

[78] B. Kroon, J. Nesvadba, and A. Hanjalic. Dialog detection in narrative video by shot and face analysis. In *SPIE Proceedings. Multimedia Content Access: Algorithms and Systems*, volume 6506, 2007.

[79] R. Laganière, R. Bacco, A. Hocevar, P. Lambert, G. Païs, and B. E. Ionescu. Video summarization from spatio-temporal features. In *Proceedings of the 2nd ACM TRECVid Video Summarization Workshop*, pages 144–148, 2008.

[80] I. Laptev. On space-time interest points. *International Journal of Computer Vision*, 64(2-3):107–123, 2005.

[81] I. Laptev and T. Lindeberg. Space-time interest points. In *IEEE International Conference on Computer Vision*, pages 432–439, 2003.

[82] J. Law-To, O. Buisson, V. Gouet-Brunet, and N. Boujemaa. Robust voting algorithm based on labels of behavior for video copy detection. In *ACM International Conference on Multimedia*, pages 835–844, 2006.

[83] J. Law-To, L. Chen, A. Joly, I. Laptev, O. Buisson, V. Gouet-Brunet, N. Boujemaa, and F. Stentiford. Video copy detection: a comparative study. In *ACM International Conference on Image and Video Retrieval*, pages 371–378, 2007.

[84] B. Leibe, K. Schindler, N. Cornelis, and L. Van Gool. Coupled object detection and tracking from static cameras and moving vehicles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(10):1683–1698, 2008.

[85] G. Leon, H. Kalva, and B. Furht. Video identification using video tomography. In *IEEE International Conference on Multimedia and Expo*, pages 1030–1033, 2009.

[86] J. Li, W. Wu, T. Wang, and Y. Zhang. One step beyond histograms: Image representation using markov stationary features. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[87] S. Li, L. Zhu, Z. Zhang, A. Blake, H. Zhang, and H. Shum. Statistical learning of multi-view face detection. In *Computer Vision – ECCV 2002*, volume 2353 of *LNCS*, pages 117–121, 2006.

[88] Y. Li, J. Jin, and X. Zhou. Video matching using binary signature. In *International Symposium on Intelligent Signal Processing and Communication Systems*, pages 317–320, 2005.

[89] T. Lin and H.-J. Zhang. Automatic video scene extraction by shot grouping. In *International Conference on Pattern Recognition*, volume 4, pages 39–42, 2000.

[90] D. Lowe. Demo software: Sift keypoint detector. `http://www.cs.ubc.ca/~lowe/keypoints/` (last checked: 2011-09-15), 2011.

[91] D. G. Lowe. Object recognition from local schale-invariant features. In *International Conference on Computer Vision*, volume 2, pages 1150–1157, 1999.

[92] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[93] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *International Joint Conference on Artificial Intelligence*, volume 2, pages 674–679, 1981.

[94] Y.-F. Ma, X.-S. Hua, L. Lu, and H.-J. Zhang. A generic framework of user attention model and its application in video summarization. *IEEE Transactions on Multimedia*, 7(5):907–919, 2005.

[95] R. Maltsy. *Hollywood Cinema*. Wiley-Blackwell, 2 edition, 2003.

[96] B. Mamer. *Film Production Technique: Creating the Accomplished Image*. Number 5. Wadsworth Publishing, 2008.

[97] B. Manjunath, J.-R. Ohm, V. Vasudevan, and A. Yamada. Color and texture descriptors. *IEEE Transactions on Circuits and Systems for Video Technology*, 11(6):703–715, 2001.

[98] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In *British Machine Vision Conference*, volume 1, pages 384–393, 2002.

[99] E. McKean, editor. *The New Oxford Amerdican Dictionary*. Oxford University Press, 2nd edition, 2005.

[100] K. Mikolajczyk and C. Schmid. Indexing based on scale invariant interest points. In *International Conference on Computer Vision*, pages 525–531, 2001.

[101] K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *International Journal of Computer Vision*, 60(1):63–86, 2004.

[102] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10):1615–1630, 2005.

[103] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. V. Gool. A comparison of affine region detectors. *International Journal of Computer Vision*, 65(1–2):43–72, 2005.

[104] D. Mitrović, M. Zeppelzauer, M. Zaharieva, and C. Breiteneder. Retrieval of visual composition in film analysis. In *International Workshop on Image Analysis for Multimedia Interactive Services*, 2011.

[105] J. Mitry. *The aesthetics and psychology of the cinema.* Indiana University Press, 2000.

[106] S. Moncrieff, C. Dorai, and S. Venkatesh. Affect computing in film through sound energy dynamics. In *ACM international conference on Multimedia*, pages 525–527, 2001.

[107] S. Moncrieff, C. Dorai, and S. Venkatesh. Detecting indexical signs in film audio for scene interpretation. In *IEEE International Conference on Multimedia and Expo*, pages 989–992, 2001.

[108] S. Moncrieff, C. Dorai, and S. Venkatesh. *Media computing: computational media aesthetics*, chapter Determining Affective Events through Film Audio, pages 131–155. Kluwer Academic Publishers, 2002.

[109] B. C. J. Moore. *An Introduction to the psychology of hearing*. Academic Press, 5 edition, 2004.

[110] W. Murch. *In the blink of an eye: a perspective on film editing*. Silman-James Press, 2001.

[111] F. Nack, C. Dorai, and S. Venkatesh. Computational media aesthetics: finding meaning beautiful. *IEEE Multimedia*, 8(4):10–12, 2001.

[112] C. W. Ng, I. King, and M. R. Lyu. Video comparison using tree matching algorithm. In *Proceedings of the International Conference on Imaging Science, Systems and Technology*, pages 184–190, 2001.

[113] C.-W. Ngo, T.-C. Pong, and H.-J. Zhang. Motion-based video representation for scene change classification. *International Journal of Computer Vision*, 50(2):127–142, 2002.

[114] P. Obrador, L. Schmidt-Hackenberg, and N. Oliver. The role of image composition in image aesthetics. In *IEEE International Conference on Image Processing*, pages 3185–3188, 2010.

[115] A. Oikonomopoulos, I. Patras, and M. Pantic. Spatiotemporal salient points for visual recognition of human actions. *IEEE Transactions on Systems, Man, and Cybernetics*, 36(3), 2006.

[116] T. Ojala, M. Aittola, and E. Matinmikko. Empirical evaluation of mpeg-7 xm color descriptors in content-based retrieval of semantic image categories. In *International Conference on Pattern Recognition*, volume 2, pages 1021–1024, 2002.

[117] G. Oldham. *First Cut: Conversations with Film Editors*. University of California Press, 1995.

[118] M. Park, S. Leey, P.-C. Cheny, S. Kashyap, A. Butty, and Y. Liuy. Performance evaluation of state-of-the-art discrete symmetry detection algorithms. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[119] G. Pass, R. Zabih, and J. Miller. Comparing images using color coherence vectors. In *ACM International Conference on Multimedia*, pages 65–73, 1996.

[120] S. Pfeiffer and U. Srinivasan. *Media computing: computational media aesthetics*, chapter Scene Determination Using Auditive Segmentation Models of Edited Video, pages 105–123. Kluwer Academic Publishers, 2002.

[121] L. Pickup and A. Zisserman. Automatic retrieval of visual continuity errors in movies. In *Proceeding of the ACM International Conference on Image and Video Retrieval*, pages 7:1–7:8, 2009.

[122] I. Radev, N. Pissinou, and K. Makki. Film video modeling. In *Workshop on Knowledge and Data Engineering Exchange*, page 122, 1999.

[123] Z. Rasheed and M. Shah. Scene detection in holywood movies and tv shows. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 343–348, 2003.

[124] Z. Rasheed and M. Shah. Detection and representation of scenes in videos. *IEEE Transactions on Multimedia*, 7(6):1097–1105, 2005.

[125] Z. Rasheed, Y. Sheikh, and M. Shah. On the use of computable features for film classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 15(1):52–64, 2005.

[126] Y. Rui, T. S. Huang, and S. Mehrotra. Constructing table-of-content for videos. *Multimedia Systems*, 7(5):359–368, 1999.

[127] P. Sand and S. Teller. Video matching. *ACM Transactions on Graphics*, pages 592–599, 2004.

[128] F. Schaffalitzky and A. Zisserman. Automated location matching in movies. *Computer Vision and Image Understanding*, 92(2–3):236–264, 2003.

[129] C. Schmid, R. Mohr, and C. Baukhage. Evaluation of interest point detectors. *International Journal of Computer Vision*, 2(37):151–172, 2000.

[130] R. Sekuler, S. N. Watamaniuk, and R. Blake. *Steven's handbook of experimental psychology: Sensation and Perception*, volume 1, chapter Perception of visual motion, pages 121–176. John Wiley & Sons, Inc., 3 edition, 2002.

[131] H. T. Shen, J. Liu, Z. Huang, C. W. Ngo, and W. Wang. Near-duplicate video retrieval: Current research and future trends. *IEEE Multimedia*, 2011.

[132] H. T. Shen, X. Zhou, Z. Huang, J. Shao, and X. Zhou. Uqlips: a real-time near-duplicate video clip detection system. In *International Conference on Very Large Data Bases*, pages 1374–1377, 2007.

[133] A. Shivadas and J. Gauch. Real-time commercial recognition using color moments and hashing. In *Fourth Canadian Conference on Computer and Robot Vision*, pages 465–472, 2007.

[134] P. Shivakumara, W. Huang, and C. L. Tan. Efficient video text detection using edge features. In *International Conference on Pattern Recognition*, pages 1–4, 2008.

[135] E. Sikov. *Film studies: an introduction*. Film and culture. Columbia University Press, 2009.

[136] J. Sivic, F. Schaffalitzky, and A. Zisserman. Object level grouping for video shots. *International Journal of Computer Vision*, 67(2):189–210, 2006.

[137] J. Sivic and A. Zisserman. Video google: a text retrieval approach to object matching in videos. In *Ninth IEEE International Conference on Computer Vision (ICCV'03)*, volume 2, pages 1470–1477, 2003.

[138] A. F. Smeaton, P. Over, and A. R. Doherty. Video shot boundary detection: Seven years of trecvid activity. *Computer Vision and Image Understanding*, 114(4):411–418, 2010.

[139] A. W. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, 2000.

[140] T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 1(147):195–197, 1981.

[141] M. Soleymani, G. Chanel, J. J. Kierkels, and T. Pun. Affective ranking of movie scenes using physiological signals and content analysis. In *ACM Workshop on Multimedia semantics*, pages 32–39, 2008.

[142] M. Sonka, V. Hlavac, and R. Boyle. *Image Processing, Analysis, and Machine Vision*. Thomson Learning, 3rd edition, 2007.

[143] M. Stark and B. Schiele. How good are local features for classes of geometric objects. In *11th International Conference on Computer Vision*, pages 1–8, 2007.

[144] A. N. Stein and M. Hebert. Local detection of occlusion boundaries in video. *Image and Vision Computing*, 27(5):514–522, 2009.

[145] M. Stricker and M. Orengo. Similarity of color images. In *SPIE Conference on Storage and Retrieval for Image and Video Databases III*, volume 2, pages 381–392, 1995.

[146] M. J. Swain and D. H. Ballard. Indexing via color histograms. In *International Conference on Computer Vision*, pages 390–393, 1990.

[147] J.-P. Tardif. Non-iterative approach for fast and accurate vanishing point detection. In *International Conference on Computer Vision*, pages 1250–1257, 2009.

[148] T. Tayama. The minimum temporal thresholds for motion detection of grading patterns. *Perception*, 29(7):761–769, 2000.

[149] R. Teixeira, T. Yamasaki, and K. Aizawa. Comparative analysis of low-level visual features for affective determination of video clips. In *International Conference on Future Information Technology (FutureTech)*, pages 1–6, 2010.

[150] K. Terasawa, T. Nagasaki, and T. Kawashima. Robust matching method for scale and rotation invariant local descriptors and its application to image indexing. In *Information Retrieval Technology*, volume 3689/2005 of *LNCS*, pages 601–615. Springer, 2005.

[151] D. W. Tjondronegoro and Y.-P. P. Chen. Knowledge-discounted event detection in sports video. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, 40(5):1009–1024, 2010.

[152] B. T. Truong and C. Dorai. Automatic genre identification for content-based video categorization. In *15th International Conference on Pattern Recognition*, volume 4, pages 230–233, 2000.

[153] B. T. Truong, C. Dorai, and S. Venkatesh. New enhancements to cut, fade, and dissolve detection processes in video segmentation.

In *ACM international Conference on Multimedia*, pages 219–227, 2000.

[154] B. T. Truong, S. Venkatesh, and C. Dorai. Application of computational media aesthetics methodology to extracting color semantics in film. In *ACM International Conference on Multimedia*, pages 339–342, 2002.

[155] B. T. Truong, S. Venkatesh, and C. Dorai. Extraction of film takes for cinematic analysis. *Multimedia Tools and Applications*, 26(3):277–298, 2005.

[156] T. Tuytelaars and L. V. Gool. Matching widely separated views based on affine invariant regions. *International Journal of Computer Vision*, 59(1):61–85, 2004.

[157] T. Tuytelaars and K. Mikolajczyk. Local invariant feature detectors: a survey. *Foundations and Trends in Computer Graphics and Vision*, 3(3):177–280, 2008.

[158] K. E. van de Sande, T. Gevers, and C. G. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1582–1596, 2010.

[159] J. van de Weijer, T. Gevers, and A. Bagdanov. Boosting color saliency in image feature detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(1):150 –156, 2006.

[160] R. G. von Gioi, J. Jakubowicz, J.-M. Morel, and G. Randall. Lsd: A fast line segment detector with a false detection control. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(4):722–732, 2010.

[161] H. L. Wang and L.-F. Cheong. Affective understanding in film. *IEEE Transactions on Circuits and Systems for Video Technology*, 16(6):689–704, 2006.

[162] C.-Y. Wei, N. Dimitrova, and S.-F. Chang. Color-mood analysis of films based on syntactic and psychological models. In *IEEE International Conference on Multimedia and Expo*, volume 2, pages 831–834, 2004.

[163] G. Willems, T. Tuytelaars, and L. V. Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. In *European Conference on Computer Vision*, pages 650–663, 2008.

[164] G. Willems, T. Tuytelaars, and L. V. Gool. Spatio-temporal features for robust content-based video copy detection. In *ACM International Conference on Multimedia Information Retrieval*, pages 183–190, 2008.

[165] C. Xu, J. Wang, H. Lu, and Y. Zhang. A novel framework for semantic annotation and personalized retrieval of sports video. *IEEE Transactions on Multimedia*, 10(3):421–436, 2008.

[166] M. Xu, X. He, J. Jin, Y. Peng, C. Xu, and W. Guo. Using scripts for affective content retrieval. In *Advances in Multimedia Information Processing – PCM 2010*, pages 43–51. 2011.

[167] M.-C. Yeh and K.-T. Cheng. Video copy detection by fast sequence matching. In *ACM International Conference on Image and Video Retrieval*, pages 1–7, 2009.

[168] B.-J. Yi, J.-T. Lee, H.-W. Woo, and H.-C. Rim. Contextual video advertising system using scene information inferred from video scripts. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 771–772, 2010.

[169] A. Yoshitaka, T. Ishii, M. Hirakawa, and T. Ichikawa. Content-based retrieval of video data by the grammar of film. In *IEEE Symposium on Visual Languages*, page 310, 1997.

[170] J. Yuan, L.-Y. Duan, Q. Tian, and C. Xu. Fast and robust short video clip search using an index structure. In *ACM SIGMM International Workshop on Multimedia Information Retrieval*, pages 61–68, 2004.

[171] J. Yuan, B. Wei, W. Lu, and L. Wang. A new video text detection method. In *ACM/IEEE Joint Conference on Digital Libraries*, pages 359–362, 2011.

[172] R. Zabih, J. Miller, and K. Mai. A feature-based algorithm for detecting and classifying scene breaks. In *ACM International Conference on Multimedia*, pages 189–200, 1995.

[173] M. Zaharieva, D. Mitrović, M. Zeppelzauer, and C. Breiteneder. Film analysis in archive documentaries. *IEEE Mutlimedia*, 18:38–47, 2011.

[174] M. Zaharieva, M. Zeppelzauer, D. Mitrović, and C. Breiteneder. Finding the missing piece: Content-based video comparison. In *IEEE International Symposium on Multimedia*, pages 330–335, 2009.

[175] X. Zeng, X. Zhang, W. Hu, and W. Li. Video scene segmentation using time constraint dominant-set clustering. In *International Multimedia Modeling Conference*, pages 637–643, 2010.

[176] M. Zeppelzauer, D. Mitrović, and C. Breiteneder. Analysis of historical artistic documentaries. In *International Workshop on Image Analysis for Multimedia Interactive Services*, pages 201–106, 2008.

[177] M. Zeppelzauer, M. Zaharieva, D. Mitrović, and C. Breiteneder. Retrieval of motion composition in film. *Digital Creativity*, 2011.

[178] H. Zettl. *Media computing: computational media aesthetics*, chapter Essentials of Applied Media Aesthetics, pages 11–38. Kluwer Academic Publishers, 2002.

[179] H. Zettl. *Sight, Sound, Motion: Applied Media Aesthetics*. Wadsworth Publishing Co Inc, 6th edition, 2010.

[180] D.-Q. Zhang and S.-F. Chang. Detecting image near-duplicate by stochastic attributed relational graph matching with learning. In *ACM International Conference on Multimedia*, pages 877–884, 2004.

[181] S. Zhang, W. Hu, T. Wang, J. Liu, and Y. Zhang. Speaker clustering aided by visual dialogue analysis. In *Advances in Multimedia Information Processing – PCM 2008*, pages 693–702, 2008.

[182] L. Zhao, S.-Q. Yang, and B. Feng. Video scene detection using slide windows method based on temporal constrain shot similarity. In *IEEE International Conference on Multimedia and Expo*, pages 1171–1174, 2001.

[183] W.-L. Zhao, C.-W. Ngo, H.-K. Tan, and X. Wu. Near-duplicate keyframe identification with interest point matching and pattern learning. *IEEE Transactions on Multimedia*, 9(5):1037–1048, 2007.

[184] J. Zhou and X.-P. Zhang. Automatic identification of digital video based on shot-level sequence matching. In *ACM International Conference on Multimedia*, pages 515–518, 2005.

SHORT CURRICULUM VITAE

Contact Information

Name:	Maia ZAHARIEVA
Address:	Vienna University of Technology, IMS Favoritenstr. 9-11/188-2, A-1040 Vienna, Austria
Phone:	+43-1-58801-18857
eMail:	zaharieva@ims.tuwien.ac.at

Education

2007 – 2011	Vienna University of Technology, Austria PhD in business informatics (Dr.rer.soc.oec.) Thesis title: Features in visual media analysis
1998 – 2003	University of Vienna, Austria MSc. in business informatics (Mag.rer.soc.oec.) Thesis title: Efficient description of multimedia learning objects
1995 – 1996	University of National and World Economy, Sofia, Bulgaria Study of business informatics

Work experience (selection)

2008 – present	Interactive Media Systems (IMS) Group, Institute of Software Technology and Interactive Systems, Vienna University of Technology, Austria Teaching and research assistant
2007 – 2008	Pattern Recognition and Image Processing (PRIP) Group, Institute of Computer Aided Automation, Vienna University of Technology, Austria Project assistant
2003 – 2007	Multimedia Information Systems (MIS) Group, Institute of Distributed and Multimedia Systems, University of Vienna, Austria Teaching and research assistant

Project experience

08.2008 – 01.2010 VERTOV | Digital Formalism: The Vienna Vertov Collection (WWTF project, "5 senses" call 2006), http://www.digitalformalism.org

02.2007 – 08.2008 COINS | COmbatting Illicit Numismatic Sales (6th EU Framework Programme, STREP), http://www.coins-project.eu

12.2005 – 01.2007 PROLIX | Process Oriented Learning and Information Exchange (6th EU Framework Programme, 5th call, IST), http//www.prolixproject.org

01.2005 – 09.2006 BRICKS | Building Resources for Integrated Cultural Knowledge Services (6th EU Framework Programme, 1st call, IST), http://www.brickscommunity.org

05.2005 – 03.2007 ebInterface, ebInvoice, ebTransfer http://www.ebinterface.at

02.2003 – 12.2005 MobiLearn | Media Informatics Any-Time Any-Where (NML 2 Programme), http://www.mobilearn.at

08.2002 – 12.2003 LaMedica, http://www.lamedica.de