HAW HAMBURG

BACHELORARBEIT

Entwicklung und Evaluation eines lokalen KI-gestützten Schnittassistenzsystems für die Filmpostproduktion

vorgelegt am 24. August 2026

Pascal Nikiema
Matrikelnummer [redacted]

Erstprüfer: Prof. Dr.-Ing. Marco Grimm
Zweitprüfer: Thorsten Wagener

HOCHSCHULE FÜR ANGEWANDTE WISSENSCHAFTEN HAMBURG
Fakultät Elektro-, Medien- und Informationstechnik
Department Medientechnik
Finkenau 35
22081 Hamburg

Zusammenfassung

Diese Arbeit untersucht CinAssist, ein KI-gestütztes Schnittassistenzsystem für die Filmpostproduktion. Die quelloffene Anwendung läuft im Browser und bringt eine vollständige Schnittoberfläche mit Vorschau, Spuren und Zeitleiste mit. CinAssist analysiert Rohmaterial, verankert Szenen und Sprache zeitlich, verbindet Bild- und Textmerkmale und erzeugt daraus einen Schnittvorschlag, der sich in der Zeitleiste weiterbearbeiten und in andere etablierte Schnittprogramme übertragen lässt. Über eine Assistenzschicht lassen sich Anweisungen in natürlicher Sprache geben. Die abschließende kreative Entscheidung bleibt beim Editor. Forschungsdesign, Anforderungen und Evaluationskriterien werden aus Grundlagen und Stand der Technik hergeleitet. Die Bewertung erfolgt zweifach: objektiv über die Kennzahlen einer Ablationsstudie und subjektiv über eine kriteriengeleitete Sichtung des Ergebnisses.

Abstract

This thesis examines CinAssist, an AI-assisted editing system for film post-production. The open-source application runs in the browser and provides a complete editing interface with preview, tracks and timeline. CinAssist analyses raw footage, anchors scenes and speech in time, combines image and text features, and generates an editing proposal that can be further edited in the timeline and transferred to other established editing programs. An assistance layer accepts instructions in natural language. Final creative responsibility remains with the editor. Research design, requirements and evaluation criteria are derived from the foundations and the state of the art. The assessment is twofold: objective through the metrics of an ablation study, and subjective through a criterion-based viewing of the result.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungen

AAC	Advanced Audio Coding (Audiocodec, Export)
API	Application Programming Interface (Programmierschnittstelle)
ASR	Automatic Speech Recognition (automatische Spracherkennung)
BCD	Binary Coded Decimal (Zifferncodierung der Zeitcode-Felder)
BM25	Best Match 25 (Rankingfunktion, Textsuche)
CLIP	Contrastive Language–Image Pre-training (Bild-Text-Modell, visuelle Suche)
CRF	Constant Rate Factor (Qualitätsmaß der Videokompression)
EBU	European Broadcasting Union (Normungsgremium des Rundfunks)
FA	funktionale Anforderung
FCPXML	Final Cut Pro XML (Timeline-Austauschformat)
H.264	Advanced Video Coding (MPEG-4-Videocodec, Export)
HTML	Hypertext Markup Language (Format des Prüfberichts)
IoU	Intersection over Union (Überdeckungsgrad zweier Bildbereiche)
KI	künstliche Intelligenz
LLM	Large Language Model (großes Sprachmodell)
LTC	Linear Timecode (Längszeitcode auf einer Tonspur)
ML	Machine Learning (maschinelles Lernen)
MP4	MPEG-4 Part 14 (Container, Exportformat)
MTCNN	Multi-task Cascaded Convolutional Networks (Verfahren der Gesichtssuche)
NFA	nichtfunktionale Anforderung
NLE	Non-Linear Editing (nichtlinearer Videoschnitt)
RIFF	Resource Interchange File Format (Dateistruktur der WAV-Datei)
SMPTE	Society of Motion Picture and Television Engineers (Normungsgremium)

Glossar

Die Arbeit verbindet zwei Fachsprachen, die des Filmschnitts und die der maschinellen Auswertung. Die folgenden Begriffe werden im Text vorausgesetzt.

Ablationsstudie	Vergleich des vollständigen Verfahrens mit Varianten, denen je ein Bestandteil fehlt. Die Differenz misst dessen Beitrag.
Alternativspur	Spur über der Hauptspur mit einer anderen Einstellung oder einem anderen Take.
Anspiel-Barriere	Grenze, vor der nur der zum Warmwerden vorgespielte Szenenanfang liegt.
Beat	Kleinste erzählerische Einheit einer Szene.
Einstellung	Kameraeinrichtung einer Szene. Zu einer Einstellung gehören mehrere Takes.
Embedding	Zahlenvektor, der Bild und Text rechnerisch vergleichbar macht.
Feinschnitt	Ausarbeitung des Rohschnitts mit Schnittpunkten und Übergängen.
Klappe	Ansage zu Beginn einer Aufnahme, die Szene, Einstellung und Take nennt.
Provenienzkette	Rückführung eines Eintrags bis zur Originaldatei.
Proxy	Leichte Vorschaufassung anstelle des Originals.
Prüfbericht	Dokument, das zu jeder Entscheidung Begründung und Beleg ausweist.
Reaktionsschnitt	Schnitt auf die zuhörende Figur, während die andere spricht.
Retrieval	Auffinden passender Aufnahmen über Bild- und Textmerkmale zugleich.
Rohschnitt	Erste durchgehende Abfolge in der Reihenfolge der Handlung, ohne Feinarbeit.
Szenen-Master	Durchlaufender Take, der eine Szene auf der Hauptspur trägt.
Take	Einzelne Aufzeichnung einer Einstellung.
Wellenform	Darstellung des Tonverlaufs über der Zeit, hier zur Bild-Ton-Zuordnung.
Zeitcode	Fortlaufende Zeitangabe, die Bild und Ton auf eine gemeinsame Zeitachse bringt.

1. Einleitung

1.1 Problemstellung

Die frühe Filmpostproduktion beginnt mit der Sichtung, Ordnung und inhaltlichen Erschließung des Rohmaterials. Der Editor muss relevante Einstellungen, Dialogpassagen und zeitliche Zusammenhänge finden, bevor ein belastbarer Rohschnitt entstehen kann. Der Schnitt ist deshalb nicht nur ein technischer Vorgang, sondern Teil einer Auswahl- und Strukturierungsleistung, durch die eine filmische Aussage entsteht [1, S. 149–150]. Bei umfangreichem Rohmaterial kann insbesondere die manuelle Verschlagwortung und Sichtung zeitaufwendig sein [2].

Diese Arbeit untersucht, ob ein lokales KI-gestütztes Assistenzsystem diese vorbereitenden Aufgaben unterstützen kann. CinAssist ist darauf ausgelegt, Bild, Ton, Sprache, Metadaten und Zeitinformationen zusammenzuführen, relevante Szenen auffindbar zu machen und einen ersten Schnittvorschlag zu erzeugen. Ob der Vorschlag tatsächlich überprüfbar und bearbeitbar ist, wird anhand der Implementierung und der Evaluation geprüft.

1.2 Motivation

Die Motivation ergibt sich aus dem hohen manuellen Aufwand bei der Sichtung umfangreicher oder unzureichend dokumentierter Medienbestände. Eine textbasierte Suche erfasst gesprochene Inhalte, beschreibt aber nicht automatisch das sichtbare Geschehen. Eine visuelle Suche kann Motive finden, berücksichtigt jedoch nicht zwingend Dialog, Handlung oder dramaturgischen Kontext. Eine multimodale Assistenz soll diese Zugänge verbinden.

Die lokale Ausführung ist ein zweiter Motivationsgrund. Rohmaterial kann unveröffentlichte Inhalte, Personenaufnahmen oder vertraglich geschützte Informationen enthalten. Eine lokale Verarbeitung kann den Transfer der Rohdaten an externe Dienste begrenzen, stellt für sich allein jedoch noch keinen vollständigen Datenschutz- oder Sicherheitsnachweis dar. Die Arbeit untersucht daher einen Workflow, in dem die Analyse der eigentlichen Bild- und Audiodaten in einer kontrollierten lokalen Umgebung erfolgt.

1.3 Zielsetzung der Arbeit

Ziel ist die Entwicklung und Evaluation eines lokalen KI-gestützten Schnittassistenzsystems. Der Prototyp soll Videodateien ingestieren, technische Metadaten erfassen, Audio transkribieren, Szenen segmentieren, visuelle Merkmale und Embeddings erzeugen, semantische Suche ermöglichen und daraus eine weiterbearbeitbare Timeline erstellen.

Nicht Ziel ist die Entwicklung eines neuen Videogenerationsmodells oder die autonome Erstellung eines fertigen Films. Die abschließende Entscheidung über Auswahl, Reihenfolge, Rhythmus, dramaturgische Wirkung und Freigabe verbleibt beim Editor.

1.4 Forschungsfrage und Unterforschungsfragen

Die Forschungsfrage lautet:

> Inwiefern kann ein lokales KI-gestütztes Assistenzsystem die Filmpostproduktion unterstützen, und wie können die generierten Schnittvorschläge objektiv und subjektiv bewertet werden?

Die Unterfragen lauten:

1. Welche Aufgaben der Filmpostproduktion können durch ein lokales KI-gestütztes Assistenzsystem unterstützt werden?
2. Wie lassen sich Schnittvorschläge anhand struktureller Schnittlogik, sprachlicher Kontinuität und visueller Kohärenz objektiv bewerten?
3. Wie lässt sich die subjektive Beurteilung als kriteriengeleitete Eigenbewertung nachvollziehbar anlegen, und was kann sie nicht leisten?
4. Welche Potenziale und Grenzen ergeben sich aus lokaler Architektur, multimodaler Analyse und menschlicher Kontrolle?

1.5 Aufbau der Arbeit

Kapitel 2 erläutert die Grundlagen der Filmpostproduktion, filmische Schnittprinzipien und die für CinAssist relevanten technischen Konzepte. Kapitel 3 ordnet etablierte Videoschnittsoftware und KI-gestützte Schnittlösungen ein und leitet die Forschungslücke ab. Kapitel 4 beschreibt Forschungsdesign, Anforderungen, Kriterien und die Anlage der qualitativen Eigenbewertung.

Kapitel 5 dokumentiert die Entwicklung von CinAssist, Kapitel 6 die Evaluation. Kapitel 7 diskutiert die Ergebnisse und die Grenzen des Systems. Kapitel 8 beantwortet die Forschungsfrage und formuliert den Ausblick. Abbildung 1.1 zeigt den Systemkontext und den Ablauf der Verarbeitung im Überblick.

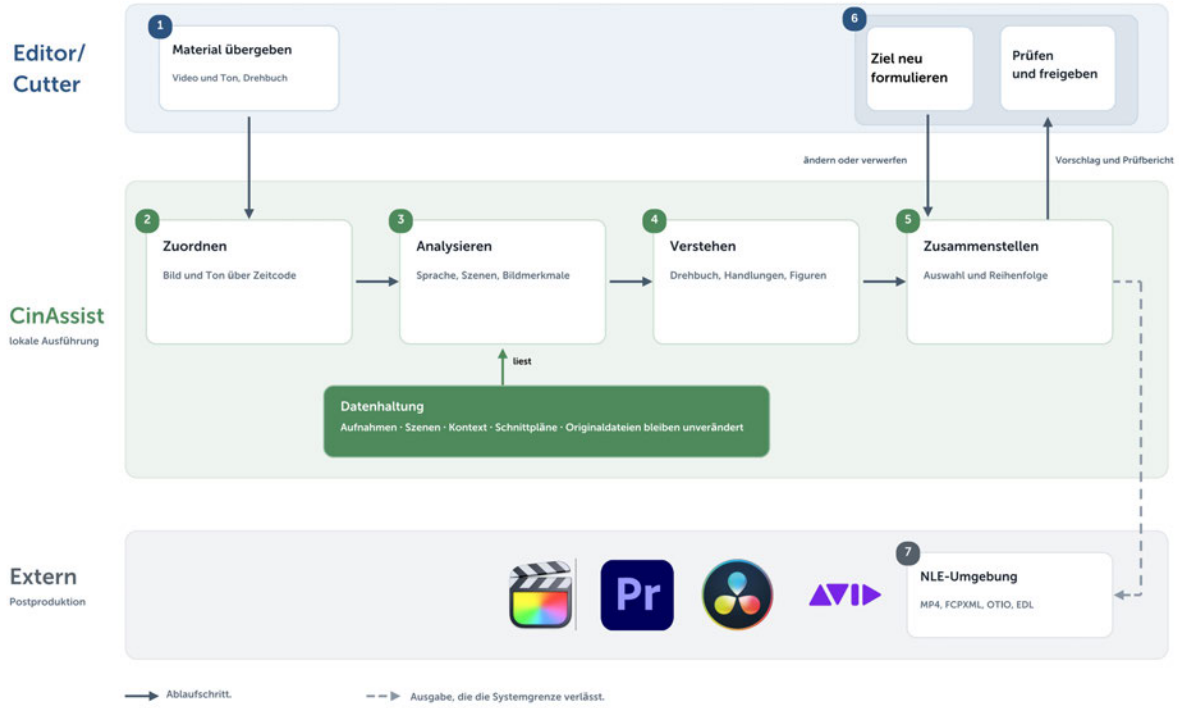

Abbildung 1.1: Systemkontext und Verarbeitungsablauf

2. Grundlagen

2.1 Filmpostproduktion

Nach den Dreharbeiten liegt ein Bestand aus einzelnen Bild- und Tonaufnahmen, Produktionsdaten und gegebenenfalls getrennten Audiospuren vor. Die Aufnahmen werden gesichtet, geordnet und zunächst zu einem Rohschnitt verbunden [3, S. 66, 212]. Im anschließenden Feinschnitt werden Schnittpunkte, Übergänge und das Verhältnis von Bild und Ton präzisiert [3, S. 210, 214].

Das zentrale Arbeitsmittel dieser Phase ist die Timeline: Sie zeigt alle Bild- und Tonspuren sowie die dort positionierten Sequenzen und macht Schnittpunkte, Übergänge und Dauern unmittelbar bearbeitbar [3, S. 210]. Wenn diese Arbeit von einem Schnittvorschlag spricht, ist damit eine solche Timeline gemeint und keine fertig gerenderte Filmdatei. Die zentrale kreative Verantwortung liegt beim Editor. Metadaten, Transkriptionen und automatische Tags sind Hilfsmittel für Sichtung und Suche, nicht Ersatz für die Bewertung von Kontext, Erzählabsicht und Wirkung.

2.2 Filmische Schnittprinzipien

Schnitt bezeichnet zunächst das technische Zerlegen, Auswählen und zeitliche Begrenzen von Einstellungen. Montage bezeichnet zusätzlich die ästhetische und erzählerische Organisation dieser Einstellungen zu einer zusammenhängenden Folge [1, S. 150]. Bedeutung entsteht durch Auswahl, Reihenfolge, Dauer und Beziehung der Einstellungen.

Continuity Editing strebt räumliche, zeitliche und kausale Verständlichkeit an. Räumliche und Bewegungskontinuität können dazu beitragen, dass harte Schnitte unauffällig bleiben [1, S. 151, 166–168]. Blickrichtung, Schuss-Gegenschuss und Handlungsachse werden in dieser Arbeit als mögliche Indikatoren berücksichtigt [1, S. 167, 172]. Diskontinuierliche Verfahren sind jedoch nicht automatisch Fehler. Ein bewusster Regelbruch kann einen Ortswechsel, eine Ellipse, Spannung oder einen Kontrast erzeugen.

Die Montage wird außerdem durch Wahrnehmung vervollständigt. Zuschauerinnen und Zuschauer verbinden Übergänge mental und erzeugen dadurch einen zusammenhängenden Erlebnisraum [4, S. 499–546, 603]. Für CinAssist folgt daraus, dass visuelle oder semantische Ähnlichkeit nur ein Hinweis auf mögliche Eignung ist.

Für die Implementierung kommen drei operative Arbeitskategorien hinzu: A-Roll bezeichnet die inhaltstragende Aufnahme, B-Roll ergänzt oder illustriert sie. Das Begriffspaar wird in der Bedeutung verwendet, die auch die Forschung zu transkriptbasierter Schnittunterstützung zugrunde legt [5]. Der Establishing Shot leitet eine Sequenz mit einer Eröffnungstotale ein [1, S. 168] und etabliert zugleich Ort und handelnde Personen, trägt also eine erzählerische und nicht nur eine räumliche Funktion [6, S. 157]. Feste Bedeutungen sind das nicht, eine B-Roll kann im konkreten Film eine zentrale narrative Funktion übernehmen. Die Zuordnung bleibt ein Vorschlag, den der Editor ändern können muss.

2.3 Dramaturgische Einordnung der Materialauswahl

Die bisher beschriebenen Prinzipien betreffen die Verbindung einzelner Einstellungen. Die Auswahl des Materials folgt darüber hinaus einer erzählerischen Ordnung, und diese Ordnung kann dem Dreh vorausgehen oder erst im Schnitt entstehen. Liegt ein Drehbuch vor, steht die Struktur bereits vor dem Material fest. Beim dokumentarischen Arbeiten ist das nicht so, und die Auswahl des Materials wird selbst zur strukturierenden Handlung. Diese Unterscheidung trägt die vorliegende Arbeit. Das entwickelte System hält für jeden der beiden Fälle einen eigenen Verarbeitungsweg bereit (Kapitel 5), und das in Kapitel 6 untersuchte Material gehört zum ersten.

Den Bezugsrahmen für den zweiten Fall liefert die dokumentarische Dramaturgie. Ihre Werkzeuge betreffen drei Arbeitsbereiche: die Strukturierung des Stoffes, die Darstellung der Inhalte durch Bild, Geräusche, Musik, O-Ton, Schrift und Sprache sowie die Gestaltung und Steuerung des Films [7, S. 34]. CinAssist arbeitet auf der Ebene der Darstellung: Es erfasst Bild, Ton, gesprochene Sprache und technische Metadaten und macht sie auffindbar. Strukturierung des Stoffes und Steuerung des Films bleiben beim Editor. Ein Suchtreffer beschreibt, welches Material vorhanden ist, nicht, welche Funktion es in der geplanten Erzählung übernehmen soll. Vom „Erzähl-Team“ aus Bild, Geräusch, Musik, O-Ton, Schrift und Filmtext [7, S. 38] wertet CinAssist zwei Mittel systematisch aus, Bild und O-Ton. Der Anspruch des Systems ist damit von vornherein auf einen Ausschnitt der erzählerischen Mittel begrenzt.

Wesentlich für die Anlage des Systems: Bei dokumentarischen Projekten weicht die endgültige Fassung regelmäßig von der geplanten dramaturgischen Struktur ab, weil die endgültige Story erst im Schnitt entsteht [7, S. 64]. Daraus folgt eine Anforderung über die reine Suchfunktion hinaus: Ein Schnittvorschlag darf kein Endzustand sein, sondern muss ein Zwischenstand bleiben, der sich verschieben, teilen, ersetzen und verwerfen lässt. Ein System, das eine fertige Fassung ausgibt, würde genau den Prozess unterbrechen, in dem die Struktur überhaupt erst entsteht.

2.4 Künstliche Intelligenz, ML und LLM

Drei häufig gleichgesetzte Begriffe bezeichnen verschiedene Ebenen. *Künstliche Intelligenz* (KI) dient als Oberbegriff. Die Praxisliteratur unterscheidet generative, analytische und reaktive Systeme, wobei die Einteilung nicht trennscharf ist [8, S. 7]. CinAssist arbeitet überwiegend analytisch und bringt keine neuen Bildinhalte hervor. Eine generative Komponente tritt allein bei der sprachlichen Strukturierung der Anfrage hinzu (Abschnitt 3.1).

Machine Learning (ML) lernt die Zuordnung aus Beispielen, statt sie vorab zu formulieren [8, S. 7]. Daraus folgt die für die Bewertung entscheidende Eigenschaft: Ein gelerntes Modell liefert Ähnlichkeitsaussagen und keine Begründungen. Vortrainierte Modelle haben auf großen Datenmengen gelernt und lassen sich an nachgelagerte Aufgaben anpassen [8, S. 15]. CinAssist trainiert keine eigenen Modelle, sondern verwendet solche unverändert.

Large Language Models (LLM) lassen sich ohne erneutes Training auf neue Aufgaben ausrichten, indem die Aufgabe sprachlich beschrieben wird. Dieses Vorgehen heißt Prompting [9]. Architektonisch beruhen sie auf Transformern mit Aufmerksamkeitsmechanismus [10]. Im System stehen sie an zwei Stellen, bei der Übersetzung der Anfrage in Suchkriterien und bei der Anordnung der Kandidaten. Sie sind damit eine Strukturierungskomponente und keine autonome kreative Instanz. Diese Festlegung folgt aus der in Abschnitt 2.3 begründeten Rollenverteilung und nicht aus einer technischen Grenze.

2.5 Whisper und automatische Spracherkennung

Automatic Speech Recognition (ASR) überführt gesprochene Sprache in Text. Moderne End-to-End-Systeme fassen die früher getrennten Verarbeitungsschritte in einer gemeinsam lernbaren Architektur zusammen [11]. Whisper ist ein solches System,

trainiert auf einem sehr großen mehrsprachigen Bestand mit schwacher Überwachung und ohne projektspezifisches Nachtraining einsetzbar [12]. Diese Eigenschaft ist hier entscheidend, denn Rohmaterial liegt nicht als annotierter Trainingsdatensatz vor, ein aufgabenspezifisches Training scheidet damit aus. Modell und Referenzimplementierung stehen unter der MIT-Lizenz [13], die Nutzung, Veränderung und Weitergabe erlaubt, ohne Gewährleistung und ohne Offenlegungspflicht für eigenen Quellcode [14].

Im System ist das Transkript die Brücke vom Ton zur textbasierten Verarbeitung: Es bildet das Korpus der lexikalischen Suche (Abschnitt 2.7) und die Grundlage des Abgleichs mit dem Drehbuch (Kapitel 5). Gespeichert wird es segmentweise, mit Zeitmarken zusätzlich auf Wortebene. Diese Wortzeiten sind Voraussetzung für die Bestimmung des Einstiegspunkts innerhalb einer Szene, denn ein Schnitt mitten in ein Wort lässt sich nur vermeiden, wenn die Wortgrenzen bekannt sind. Die Erkennungsqualität hängt von Sprache, Akzent, Hintergrundgeräuschen, Musik, Sprecherwechseln und Aufnahmequalität ab [12]. Das Transkript ist damit eine belastbare, aber keine fehlerfreie Grundlage. Welche Fehler das Modell im vorliegenden Material tatsächlich erzeugt hat, behandelt Kapitel 7.

2.6 Bildanalyse und Segmentierung

Ein Video ist eine zeitliche Folge digitaler Einzelbilder. Eine Datei von dreißig Minuten ist als Ganzes kein sinnvolles Suchergebnis. Die Zerlegung in Abschnitte erfolgt über einen Wechsel des Bildinhalts: Überschreitet die Veränderung zwischen aufeinanderfolgenden Bildern einen Schwellenwert, wird eine Grenze gesetzt. Das Ergebnis ist eine Folge von Abschnitten mit Anfangs- und Endzeit, die zeitliche Verankerung, auf der alle weiteren Analyseschritte aufsetzen. Ausgewertet werden darin nicht alle Bilder, sondern einzelne zeitlich verteilte Frames.

Dieser technische Abschnitt ist vom filmischen Begriff der Szene zu unterscheiden. Erkannt wird ein Wechsel des Bildinhalts, was in der Regel dem Wechsel der Einstellung entspricht. Eine Szene im dramaturgischen Sinn kann aus mehreren Einstellungen bestehen und lässt sich aus dem Bildsignal nicht ableiten.

Beschreibende Merkmale. Ein kompaktes Bild-Sprach-Modell erzeugt zu einem repräsentativen Frame eine kurze Beschreibung in natürlicher Sprache. Diese Beschreibung ist selbst wieder Text und geht damit in die in Abschnitt 2.7 beschriebene lexikalische Suche ein. Solche Modelle erzeugen jedoch gelegentlich Beschreibungen, die nicht zum Bild passen. Die Verarbeitung muss offensichtlich fehlerhafte Ausgaben erkennen und verwerfen.

Formale Merkmale. Die Einstellungsgröße gibt die Entfernungsspanne zwischen Kamerastandpunkt und abgebildetem Objekt an und beschreibt damit die Nähe oder Distanz, die das Publikum zu den Dargestellten einnimmt. Da Menschen meist im Zentrum der Inszenierung stehen, orientiert sich der Begriff am menschlichen Körper. Die Filmanalyse unterscheidet acht Abstufungen von der Panoramaaufnahme bis zum Detail [1, S. 10–11]. CinAssist nähert diese Kategorie an, indem es Gesichter im Frame erkennt und aus dem Flächenanteil des größten Gesichts eine von wenigen Stufen ableitet. Die Annäherung ist gröber als die filmische Einteilung. Sie versagt, sobald kein Gesicht sichtbar ist: Eine Landschaftstotale und eine Detailaufnahme eines Gegenstands sind für sie nicht unterscheidbar. Und ihre Schwellen sind gesetzt, nicht aus dem Material abgeleitet.

Diese Grenzen sind kein Mangel der Umsetzung, sondern folgen aus der Sache. Technische Merkmale bleiben beschreibende Hinweise auf mögliche Eignung. Ein Urteil über die filmische Qualität einer Aufnahme lässt sich aus ihnen nicht gewinnen.

2.7 Multimodale Repräsentationen und hybrides Retrieval

Unter *Retrieval* wird das Auffinden und Ordnen relevanter Einheiten aus einem Bestand nach einer Anfrage verstanden: hier die zeitlich verankerten Szenen des Rohmaterials, geordnet nach Relevanz zur Formulierung des Editors, eine Vorauswahl, keine Entscheidung. *Hybrid* heißt das Verfahren, wenn es mehrere unterschiedlich gebildete Signale zu einem Rangwert zusammenführt: Hier sind es ein visuelles und ein textbasiertes, ihre Verrechnung beschreibt Kapitel 5.

Visuelle Ähnlichkeit. CLIP verbindet Bild- und Textrepräsentationen in einem gemeinsamen Merkmalsraum und ermöglicht dadurch eine sprachbezogene visuelle Suche [15]. Eine Anfrage und ein Bild werden dazu jeweils in einen Zahlenvektor überführt, ein sogenanntes *Embedding*, der die inhaltlichen Eigenschaften des Ausgangsmaterials in verdichteter Form abbildet. Beide Embeddings liegen im selben Raum, sodass ihre Nähe als Maß der Übereinstimmung dienen kann. Der Vorteil liegt darin, dass die Anfrage frei formuliert sein kann und keine vorab vergebenen Schlagwörter voraussetzt. Für Videos ist zu berücksichtigen, dass ein Bewegtbild aus vielen Einzelbildern besteht. Die Repräsentation einer Szene beruht daher auf einer Auswahl von Frames und bildet den zeitlichen Verlauf nur eingeschränkt ab.

Lexikalische Relevanz. Die zweite Komponente arbeitet auf Text. BM25 ist eine Rangfolgefunktion aus dem probabilistischen Relevanzmodell des Information Retrieval. Sie bewertet ein Dokument danach, wie häufig die Begriffe der Anfrage darin vorkommen, und gewichtet dabei die Häufigkeit der Begriffe gegen die Länge des Dokuments [16]. Anders als bei CLIP entsteht kein semantischer Raum: Gesucht wird nach den Wörtern selbst. In CinAssist bildet der Bestand aus Szenenbeschreibung und Transkript das durchsuchte Textkorpus.

Warum beide. Die Verfahren versagen an verschiedenen Stellen: CLIP findet visuell passende Szenen auch ohne die gesuchten Wörter, kann aber einen nur gesprochenen Eigennamen nicht zuverlässig auffinden. BM25 findet solche Begriffe zuverlässig, weiß aber nichts über das Bild und bewertet eine wortlose Aufnahme mit null. Ein Assistenzsystem für dokumentarisches Material muss beide Fälle abdecken: Ein Interview trägt seine Information im Wort, eine B-Roll im Bild. Die Werte werden normalisiert und über eine offengelegte Gewichtung verrechnet. Diese Gewichtung ist eine Festlegung, Kapitel 5 benennt sie, und sie bleibt veränderbar. Abbildung 2.1 veranschaulicht diese Komplementarität am Material dieser Arbeit.

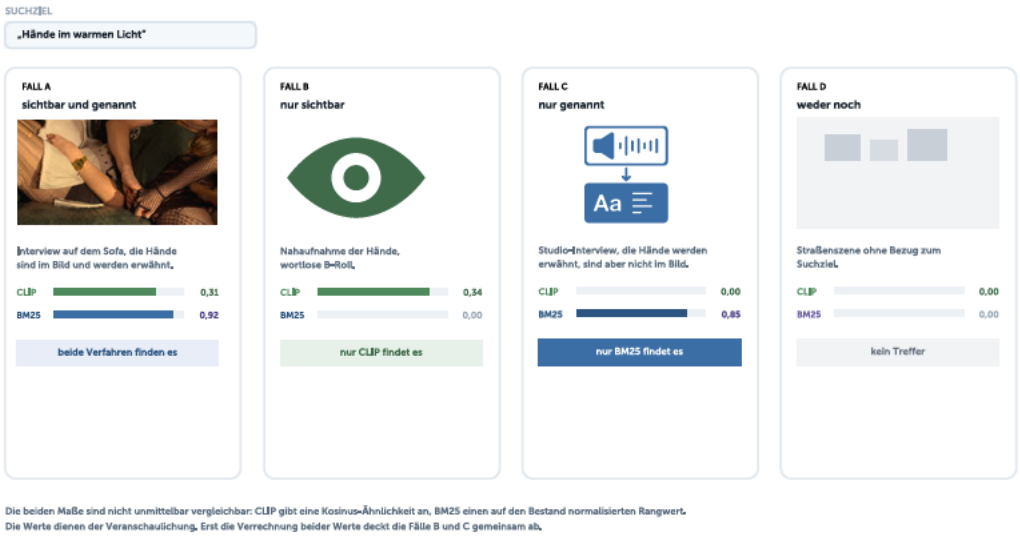

Abbildung 2.1: Komplementarität der beiden Retrieval-Signale am Beispiel „Hände im warmen Licht".

2.8 Lokale Ausführung als Architekturbedingung

Ein Teil der genannten Modelle lässt sich lokal betreiben. Dafür sprechen mehrere Gründe: Die verarbeiteten Informationen bleiben auf dem eigenen Rechner, es fallen keine Nutzungsgebühren an, der Betrieb funktioniert ohne Internetverbindung, und die Modelle lassen sich ohne Mengenbegrenzung durch einen Anbieter an eigene Anwendungen anpassen [8, S. 27].

Ausschlaggebend ist hier der erste Punkt. Rohmaterial enthält regelmäßig Personenaufnahmen, unveröffentlichte Inhalte und vertraglich geschützte Informationen. Die lokale Verarbeitung begrenzt den Transfer dieser Daten. Sie ersetzt jedoch keine vollständige Datenschutzprüfung, weil auch lokale Speicherung, Protokolldateien und optionale externe Komponenten zu betrachten bleiben.

2.9 Zusammenfassung der Grundlagen

Die Grundlagen führen zu vier Anforderungen: Analyseergebnisse müssen zeitlich verankert sein, Retrieval und Sequenzierung müssen getrennt untersucht werden, technische Merkmale dürfen nicht mit filmischer Qualität verwechselt werden, und der Editor muss Vorschläge prüfen und verändern können.

3. Stand der Technik

Als Bezugspunkt dienen zwei verbreitete Systeme des professionellen Schnitts: DaVinci Resolve von Blackmagic Design und Adobe Premiere. Beide verbinden Medienimport, Organisation, Metadaten, Transkription, Suche, Schnitt, Farb- und Audiobearbeitung sowie Export in einer Umgebung. Dieses Kapitel behandelt sie zuerst, weil sich an ihnen entscheidet, was an CinAssist überhaupt noch offen ist.

Die Angaben zu beiden Systemen stammen aus der jeweiligen Herstellerdokumentation. Eine unabhängige wissenschaftliche Quelle zum aktuellen Funktionsstand kommerzieller Schnittsysteme existiert nicht. Die Angaben sind daher als Herstellerangaben zu lesen und wurden nicht unabhängig überprüft.

Abschnitt 3.6 ordnet die Veränderungen ein, die während des Bearbeitungszeitraums eingetreten sind.

3.1 Einordnung von KI-Werkzeugen im Videoworkflow

In der Praxisliteratur werden vier Kategorien von KI-Werkzeugen im Videobereich unterschieden: generative Anwendungen, die aus einer Anweisung unmittelbar eine Videoausgabe erzeugen, Videoeditoren mit KI-Bearbeitungswerkzeugen, Produktivitätsanwendungen für die Aufbereitung über mehrere Kanäle sowie Werkzeuge, die vorhandene Videos durchsuchen und auswerten [8, S. 25].

CinAssist gehört zur vierten Kategorie: Es erzeugt kein Bildmaterial, sondern erschließt vorhandenes. Die beiden marktbestimmenden Systeme fallen zugleich in die zweite und die vierte Kategorie. Sie sind Schnittumgebungen und enthalten inzwischen auch Erschließungswerkzeuge. Genau dieser zweite Teil ist der Vergleichsgegenstand der folgenden Abschnitte.

3.2 DaVinci Resolve

Die Herstellerdokumentation weist die Version 21 aus [17], deren Neuerungen der Hersteller in einem auf April 2026 datierten Leitfaden dokumentiert [18]. Das Produkt ist zweigeteilt: Neben einer kostenlosen Fassung steht eine Studio-Fassung, die der Hersteller mit einem Einmalpreis von 295 US-Dollar ausweist [19].

Neuerungen. Für die Materialerschließung ist AI IntelliSearch einschlägig: Es analysiere das Material und ermögliche die Suche nach bestimmten Objekten oder nach Schlüsselwörtern im Dialog [17]. Bereits mit Version 20 im April 2025 kamen zwei Funktionen zur Sequenzbildung hinzu: AI IntelliScript erzeugt Timelines auf Grundlage eines Textskripts, AI Multicam SmartSwitch setzt eine Timeline mit Kamerawinkeln auf Basis einer Sprechererkennung zusammen [20]. Weitere KI-Werkzeuge betreffen Gestaltung und Bildbearbeitung und bleiben hier außer Betracht.

Gemeinsamkeiten mit CinAssist. Drei Fähigkeiten decken sich: die Verbindung von visueller und sprachlicher Suche, die Erzeugung einer Timeline aus dem Material und die Zuordnung getrennt aufgenommener Bild- und Tonspuren über Timecode oder Wellenform, die auf der Medienebene seit Langem etabliert ist [21]. Keine dieser Aufgaben ist demnach ein offenes Problem.

Unterschiede. Der Unterschied liegt nicht darin, *ob* aus einem Drehbuch geschnitten wird, sondern worauf sich der Abgleich stützt und was er offenlegt.

IntelliScript vergleicht den Text des Drehbuchs mit dem transkribierten Ton. CinAssist verfolgt denselben Weg, setzt aber an drei Stellen anders an: Der Abgleich läuft über Embeddings und trägt auch bei verschiedenen Sprachen von Drehbuch und Dreh. Die Zuordnung zur Drehbuchszene stützt sich auf die gesprochene Klappe statt auf den Dateinamen. Und für jede Drehbuchzeile weist das System aus, welche Einstellung sie abdeckt und welche Zeilen ohne Material bleiben: Jeder Eintrag des Schnittplans trägt Begründung und Beleg, nachvollziehbar im Prüfbericht (Abschnitt 5.5). Zudem setzen beide Blackmagic-Verfahren eine vorhandene Struktur zwingend voraus (Drehbuch bzw. Mehrkameramaterial). CinAssist deckt auch den Fall ab, in dem die Struktur erst im Schnitt entsteht (Abschnitt 2.3). Das Verfahren wird nicht offengelegt (die Dokumentation nennt nicht, auf welcher Repräsentation IntelliSearch beruht). Die KI-Funktionen setzen zudem die kostenpflichtige Studio-Fassung voraus: Das Referenzhandbuch führt IntelliSearch ausdrücklich als „Studio Version Only“ [22, S. 409], und die tragende DaVinci Neural Engine

ist nach Herstellerangabe exklusiver Bestandteil der Studio-Fassung [19]. Aufschlussreich ist der Umgang mit der Filmklappe: AI Slate ID liest die dort notierten Metadaten aus [17], während CinAssist Klappen-Szenen aus dem Kandidatenpool ausschließt, weil sie für die Sequenzbildung wertlos sind (Kapitel 5).

3.3 Adobe Premiere

Adobe Premiere[1] weist in der Fassung vom 7. August 2026 die Version 26.3 vom Juni 2026 aus [23].

Neuerungen. Maßgeblich ist die Funktion Media Intelligence: Sie erlaubt eine Suche nach visuellen Beschreibungen und eingebetteten Metadaten [24] sowie eine gesonderte Suche über gesprochene Inhalte [25]. Hinzu kamen 2026 ein KI-gestütztes Object Masking, ein durchsuchbarer Sequence Index und Content Credentials, die beim Export vermerken, ob generative KI beteiligt war. Seit Juni 2026 lässt sich Material aus Firefly Boards in eine Sequenz übernehmen. Deren Reihenfolge stammt aus der manuellen Auswahl und nicht aus einer Analyse [23].

Gemeinsamkeiten mit CinAssist. Hier ist die Überschneidung am größten, weil Adobe das Verfahren beschreibt. Die visuelle Suche beruhe nicht auf einer Verschlagwortung. Die Analyse bilde das Material in ein mehrdimensionales semantisches Verständnis ab, die Suchanfrage werde in denselben Raum abgebildet, und die nächstgelegenen Treffer bildeten das Ergebnis [26]. Beschrieben ist damit derselbe gemeinsame Merkmalsraum, auf dem auch der in Abschnitt 2.7 dargestellte Ansatz beruht.

Zwei weitere Annahmen dieser Arbeit finden sich wieder. Die Suche arbeite unterhalb der Clipebene: Ein Treffer könne einen Abschnitt eines Clips bezeichnen, der mit In- und Out-Punkten markiert werde [26]. Und Analyse wie Suche fänden vollständig lokal und ohne Internetverbindung statt [26]. Weder die zeitliche Verankerung von Treffern noch die lokale Verarbeitung sind damit ein Alleinstellungsmerkmal.

Unterschiede. Der deutlichste Unterschied betrifft den Verbleib der Analyseergebnisse. Adobe legt den Index der visuellen Analyse als eigene Datei neben der Projektdatei ab. Ob sich diese Ergebnisse außerhalb von Premiere nutzen lassen, beantwortet der Hersteller

[1] Adobe führt die Desktop-Anwendung inzwischen als „Adobe Premiere“. Die frühere und weiterhin verbreitete Bezeichnung lautet „Premiere Pro“. Beide bezeichnen dasselbe Produkt.

ausdrücklich mit Nein: Analyseergebnisse und Projektindex seien seinen Modellen eigen und für andere Anwendungen nicht verwendbar [26].

Der Aufwand der Materialerschließung ist damit an eine Anwendung gebunden. Hinzu kommt das Lizenzmodell: Premiere ist ausschließlich im Abonnement erhältlich, das Einzelprodukt weist der Hersteller mit 22,99 US-Dollar pro Monat im Jahresabo aus [27]. Da der Index nur in Premiere lesbar ist, bleibt auch der Zugriff auf einmal erzeugte Analyseergebnisse an ein laufendes Abonnement gebunden, während die Studio-Fassung von DaVinci Resolve mit einem Einmalpreis geführt wird (Abschnitt 3.2). Für eine wissenschaftliche Untersuchung wiegt schwerer, dass ein nicht auslesbarer Index auch nicht überprüfbar ist: Weder die abgelegten Repräsentationen noch die Rangfolge der Treffer lassen sich unabhängig nachvollziehen. CinAssist legt die Analyseergebnisse dagegen offen ab und gibt die Timeline in etablierten Austauschformaten aus (Kapitel 5).

Zur Sequenzbildung schließlich leistet Media Intelligence nichts: Sie findet Material, ordnet es aber nicht an. Der Import aus Firefly Boards ist eine Importfunktion, keine inhaltsgestützte Sequenzbildung. Eine anweisungsgesteuerte Erzeugung existiert bei Adobe erst mit dem Premiere AI Assistant, der nach Beginn dieser Arbeit erschien und in Abschnitt 3.6 behandelt wird.

3.4 Forschungsansätze

Neben den Produkten steht eine Forschungslinie, die dieselbe Aufgabe seit fast drei Jahrzehnten behandelt.

Butler und Parkes setzen mit dem System LIVE Sequenzen aus Fragmenten einer annotierten Videodatenbank zusammen, wobei die Anfrage eine einfache Geschichte ohne filmische Anweisungen beschreibt [28, S. 367]. Zentral ist ihre zweite Beobachtung: Der Film verfügt über keine abschließende Festlegung seiner Syntax und Semantik, sondern über eine lose Sammlung beschreibender und vorschreibender Aussagen [28, S. 368]. Daraus folgt die Evaluationslage dieser Arbeit: Es existiert keine eindeutig richtige Montage als Prüfmaßstab. Bocconi verschiebt den Schwerpunkt auf die Bedeutung: Eine Anwendung müsse wissen, womit und wie ein Datenelement kombiniert werden kann (Repurposing) [29]. Beide Arbeiten trennen damit dasselbe, was hier als Retrieval und Sequenzierung unterschieden wird.

Jüngere Arbeiten setzen an einzelnen Teilaufgaben an. AutoTag automatisiert die Vergabe von Metadaten für die Postproduktion [2]. LAVE verbindet Videobeschreibungen, ein Large

Language Model und Bearbeitungsaktionen, sodass Nutzer Ziele formulieren, Vorschläge erhalten und die Ergebnisse direkt nachbearbeiten können [9].

Parallel hat sich ein Feld zur automatischen Beschreibung von Videoinhalten entwickelt. Drei dort benannte Schwierigkeiten betreffen auch diese Arbeit. Der Beitrag der visuellen Merkmale lässt sich kaum von dem des Sprachmodells trennen. Die verfügbaren Datensätze sind in Vielfalt und Komplexität begrenzt. Und die gebräuchlichen automatischen Maße bilden nur unzureichend ab, wie weit maschinelle und menschliche Beschreibung übereinstimmen [30]. Diese dritte Schwierigkeit, die begrenzte Aussagekraft automatischer Maße, ist der Grund, warum diese Arbeit die automatisch prüfbaren Indikatoren nicht als Qualitätsurteil behandelt. Für die Vergleichbarkeit von Ergebnissen haben sich zudem gemeinsame Evaluationskampagnen etabliert, in denen Aufgaben, Testmaterial und Bewertungsverfahren vorab offengelegt werden [31]. Für die hier untersuchte Aufgabe existiert keine solche Kampagne, weshalb Kriterien und Anfragen in dieser Arbeit vor der Messung selbst festgelegt werden (Kapitel 4).

3.5 Vergleichende Übersicht

Tabelle 3.1 fasst die betrachteten Ansätze entlang jener Merkmale zusammen, die für die Forschungslücke tragend sind. Nicht aufgenommen ist das Merkmal der menschlichen Kontrolle: Alle Ansätze sehen eine nachträgliche Bearbeitung vor, es unterscheidet sie also nicht. Zu beachten ist der unterschiedliche Belegstatus, AutoTag und LAVE sind begutachtete Veröffentlichungen, die Produktangaben stammen vom Hersteller.

Tabelle 3.1: Gegenüberstellung ausgewählter Ansätze nach den jeweils zitierten Herstellerangaben

Ansatz	Belegstatus	Eingabe für die Sequenzbildung	Ausgewertete Signale	Verfahren offengelegt	Ergebnis außerhalb nutzbar
AutoTag	Publikation	keine Sequenzbildung	Bild, Metadaten	ja	nicht zutreffend
LAVE	Publikation	freie Anweisung	Videobeschreibung, Sprache	ja	Forschungsprototyp
DaVinci Resolve 21	Hersteller	Skript oder Multicam	Bild, Dialog	nein	Sequenz nur in Resolve
Adobe Premiere 26.3	Hersteller	manuelle Auswahl	Bild, Transkript, Metadaten	nein	nein, Index anwendungsgebunden
Premiere AI Assistant	Hersteller, Beta	freie Anweisung	nicht benannt	nein, nur nachträgliche Erläuterung	Sequenz nur in Premiere
CinAssist	diese Arbeit	freie Anweisung oder Drehbuch	Bild, Text, Sprecher, Zeit	ja, Parameter und Protokoll	FCPXML

3.6 Entwicklungen während des Bearbeitungszeitraums

Der Untersuchungsgegenstand hat sich während der Bearbeitung verändert. Diese Veränderung wird hier offengelegt, weil sie die Einordnung der Ergebnisse betrifft. Zu Beginn der Recherche im März 2026 existierten die verglichenen Einzelfunktionen bereits weitgehend: semantische Suche bei Adobe und automatische Timeline-Erzeugung bei Blackmagic Design jeweils seit April 2025, Letztere jedoch nur auf Grundlage eines Drehbuchs oder synchronen Mehrkameramaterials. Kein System verband eine frei formulierte Anweisung mit der Erzeugung einer Sequenz aus unstrukturiertem Rohmaterial. Neu an dieser Arbeit sind daher nicht die einzelnen Verfahren (Whisper, CLIP, BM25 und die Szenenerkennung waren sämtlich verfügbar), sondern ihre Zusammenführung unter den Bedingungen lokaler Ausführung, offengelegter Parametrisierung und exportierbarer Ergebnisse.

Die weitreichendste Veränderung datiert auf den 18. Juni 2026, mehrere Monate nach Beginn dieser Arbeit: Adobe veröffentlichte die Dokumentation zum Premiere AI Assistant als öffentliche Beta. Über natürlichsprachliche Eingaben lassen sich damit Medien ordnen, Rohschnitte aufbauen und geordnete Medien auf einer Timeline platzieren, das Ergebnis bleibt bearbeitbar, und zu jeder Antwort lässt sich eine Begründung aufklappen, die zeigt, wie die Anfrage verstanden wurde [32]. Damit ist auch die Verbindung von freier Anfrage und Rohschnitt-Erzeugung nicht mehr allein Gegenstand der Forschung. Drei Einschränkungen bleiben: Die Funktion ist eine Beta, sie setzt ein Abonnement und die Bindung an einen Anbieter voraus, und die aufklappbare Begründung bleibt eine nachträgliche Erläuterung ohne benannte, veränderbare Parametrisierung.

3.7 Ableitung der Forschungslücke

Vier Fähigkeiten, die zu Projektbeginn noch als Abgrenzungsmerkmale hätten gelten können, sind heute in den marktführenden Systemen vorhanden: die semantische Suche über Bild und gesprochene Sprache, die zeitlich verankerte Ausgabe von Teilclips, die lokale Ausführung der Analyse und (als Beta) die Erzeugung eines ersten Schnitts aus einer natürlichsprachlichen Anweisung. Keine davon ist für sich genommen noch eine Forschungslücke.

Die Lücke liegt eine Ebene tiefer und ist methodischer Natur: Sie betrifft die Frage, unter welchen Bedingungen ein solches Verfahren wissenschaftlich untersucht werden

kann. Drei Voraussetzungen fehlen bei den kommerziellen Systemen. Erstens ist das Verfahren nicht offengelegt: Eine nachträgliche Begründung in natürlicher Sprache erlaubt weder Wiederholung noch kontrollierten Vergleich. Zweitens sind die Zwischenergebnisse nach Herstellerangabe außerhalb der Anwendung nicht verwendbar [26], sodass die Grundlage einer Auswahl nicht unabhängig geprüft werden kann. Drittens fehlt ein Bewertungsmaßstab: Es existiert weder eine Evaluationskampagne mit vorab festgelegten Kriterien [31] noch eine eindeutig richtige Montage als Prüfmaßstab [28, S. 368].

Diese Arbeit untersucht deshalb nicht, ob sich Rohmaterial semantisch durchsuchen und daraus eine Timeline erzeugen lässt, das ist beantwortet. Sie untersucht, wie ein solcher Vorgang so gebaut werden kann, dass er offengelegt, wiederholbar und an nachvollziehbaren Kriterien bewertbar ist, und welche Kriterien sich dafür eignen. Dass zwei Hersteller im selben Zeitraum vergleichbare Funktionen eingeführt haben, spricht nicht gegen die Fragestellung, sondern für sie: Mit jedem weiteren nicht einsehbaren Verfahren gewinnt sie an Gewicht. Abzugrenzen ist die Arbeit schließlich von der generativen Videosynthese: Untersuchungen zu offenen generativen Modellen behandeln die Erzeugung neuen Bildmaterials [33], CinAssist erzeugt keines, sondern ordnet vorhandenes.

4. Methodik

4.1 Forschungsdesign

Die Arbeit folgt einem konstruktiven Forschungsdesign: Erkenntnis über einen Problembereich und seine Lösung wird durch das Erstellen und die Anwendung eines Artefakts gewonnen [34]. Dass dieses Vorgehen auch im deutschsprachigen Raum als wissenschaftliche Methode anerkannt ist, halten Österle et al. in einem Memorandum fest [35]. Das Artefakt ist in dieser Arbeit der Prototyp CinAssist. Für den Ablauf von der Problemidentifikation über Zielsetzung, Entwurf und Umsetzung bis zur Evaluation und zur Kommunikation der Ergebnisse liegt ein etabliertes Vorgehensmodell vor, an dem sich der Aufbau dieser Arbeit orientiert [36]. Ein konkreter Prototyp wird auf Grundlage theoretischer und technischer Anforderungen entwickelt. Anschließend wird untersucht, ob die vorgesehenen Funktionen unter dokumentierten Bedingungen ausgeführt werden können und welche Eigenschaften die erzeugten Schnittvorschläge besitzen.

Die Evaluation trennt technische Nachweise, objektive Kriterien und subjektive Wahrnehmung, denn erst die Kombination aus objektiven Metriken und menschlicher Beurteilung erlaubt im Bereich KI-gestützter Videoerzeugung ein belastbares Urteil [8, S. 21]. Ein erfolgreicher Upload beweist keine gute Sequenz, ein hoher Retrieval-Score keine visuelle Kohärenz. Diese Ebenen werden deshalb getrennt erhoben und erst in der Diskussion zusammengeführt.

4.2 Iteratives Vorgehen und Versionsstände

Der Prototyp liegt in mehreren Fassungen vor. Das ist kein Nebenprodukt unfertiger Arbeit, sondern das Vorgehen selbst: Das gewählte Forschungsdesign sieht ausdrücklich vor, nach einer Bewertung zur Gestaltung zurückzukehren, statt eine einmal gebaute Fassung als endgültig zu behandeln [36, S. 56]. Für die Beurteilung dieser Arbeit ist deshalb nicht allein die letzte Fassung maßgeblich, sondern die Abfolge der Fassungen und die Gründe für den jeweiligen Übergang.

Der Arbeitszyklus. Jeder Durchgang folgt demselben Ablauf: Aus Material und Drehbuch wird ein Schnitt erzeugt, dazu ein Prüfbericht, der jede Entscheidung mit ihrer Begründung ausweist. Zeigt sich beim Lesen eine Stelle, deren Begründung nicht trägt, wird nicht der Einzelfall behoben, sondern die Regel geändert, die ihn erzeugt hat. Anschließend wird der Schnitt neu erzeugt. Ein Mangel wird also nicht daran erkannt, dass das Ergebnis missfällt, sondern daran, dass die mitgelieferte Begründung ihn nicht deckt. Damit ist der Übergang von einer Fassung zur nächsten belegbar und nicht Geschmackssache. Eine Übersicht der auf diesem Weg entstandenen Regeländerungen mit ihren Auslösern führt Anhang A.3.

Prüfebenen des Verfahrens. Das Verfahren prüft an sechs Stellen gegen eine unabhängige Quelle. Die Übersicht beantwortet die methodische Frage, worauf sich die Ergebnisse in Kapitel 6 stützen und wo sie an Grenzen stoßen (die Prüfungen selbst beschreibt Kapitel 5). Entscheidend ist die dritte Spalte.

Die ersten fünf Ebenen laufen ohne Zutun ab und liefern je Entscheidung einen Beleg. Die sechste ist die einzige, die eine Person erfordert, und die einzige, die beurteilen kann, ob ein Beleg trägt.

Die Fassungen und ihre Auslöser. Am Referenzmaterial entstanden insgesamt 49 Fassungen, ihre chronologische Liste führt Anhang A.3. Vier davon markieren die methodischen Übergänge, an denen eine Regel geändert wurde, und sind hier gegenübergestellt. Sie wurden am selben Material erzeugt und sind dadurch vergleichbar.

Bemerkenswert ist der Übergang von v1 zu v2. Obwohl eine zusätzliche Prüfung hinzukam, sinkt die Zahl der Einträge, denn die neue Bildprüfung lässt alle Einschübe entfallen, für die sich im Drehbuch kein Anlass findet. Eine Prüfung kann ein Ergebnis also auch verkleinern. Die Fassungen sind nicht gegeneinander gemessen worden, das setzte ein Maß für Schnittqualität voraus, das hier nicht vorliegt. Was die Tabelle belegt, ist der Weg und sein Anlass, nicht eine Verbesserung im messbaren Sinn.

4.3 Anforderungsanalyse

Die Anforderungen sind für diese Arbeit in zwei Gruppen geführt: funktionale Anforderungen, die benennen, *was* das System leisten muss (Upload, Ingestion, Transkription, Szenenerkennung, Bildanalyse, Embeddings, Suche, Timeline-Generierung, manuelle Bearbeitung und Export), und nichtfunktionale Anforderungen, die festlegen,

Tabelle 4.1: Prüfebenen des Verfahrens

Was geprüft wird	Wogegen und womit	Verbleibende Unsicherheit
Zuordnung von Bild und Ton (Abschn. 5.3)	Zeitcode zweier unabhängiger Geräte, gegengeprüft über die Wellenform und bext gegen iXML	Auseinanderlaufen der Uhren nicht erkennbar, Wellenform braucht einen Referenzton
Zugehörigkeit einer Aufnahme zu einer Drehbuchszene (Abschn. 5.5)	gegen die gesprochene Klappe im Ton, regelbasiert	Aufnahmen ohne hörbare Klappe bleiben unzugeordnet
Gesprochener Satz zu Drehbuchzeile (Abschn. 5.5)	gegen den Drehbuchtext, Sequenz-Alignment mit Ähnlichkeitsschwelle	Abweichung durch Improvisation oder Fehler in der Transkription
Drehbuch-Handlung im Bild (Abschn. 5.5.4)	gegen Einzelbilder, geschlossene Fragen an ein Bildmodell mit Ähnlichkeit als Zweitsignal	kleine Gegenstände und innere Zustände unzuverlässig
Person im Bild (Abschn. 5.5.5)	gegen die Auftrittsverteilung der Figuren im Drehbuch	Figuren mit ähnlichem Auftrittsmuster verwechselbar
Der Schnitt als Ganzes	gegen den Prüfbericht, gelesen von einer Person	bislang nur durch den Verfasser gelesen

Tabelle 4.2: Fassungen des Schnittverfahrens am Korpus

Fassung	Änderung	Auslöser	Ergebnis
Rohschnitt	Zeilen den Aufnahmen zugeordnet, Reihenfolge nach Drehbuch	Ausgangsfassung	8:50
Feinschnitt v1	Bildwechsel in Sprechpausen mit durchlaufendem Ton, Handlung in Höhepunkten statt Blöcken, engere Übergänge	Ergebnis zu lang und blockhaft	42 Einträge, 6:00
Feinschnitt v2	stumme Aufnahmen nach im Bild bestätigten Drehbuch-Handlungen, Einschübe nur bei Anlass im Drehbuch	Einschub ohne tragfähige Begründung im Bericht	28 Einträge, 6:13
Feinschnitt v3	Wechsel auf die zuhörende Figur bei durchlaufendem Ton	fehlende Gegeneinstellung im Dialog	37 Einträge, 6:13

wie es das tun soll: lokale Verarbeitung, Reproduzierbarkeit, Robustheit, Wartbarkeit und Bedienbarkeit.

Die vollständige Zuordnung FA-1 bis FA-39 und NFA-1 bis NFA-19 wird im Repository geführt [37]. Jede zentrale Anforderung ist dort einer Komponente, einem Anwendungsfall und einem Testfall zugeordnet.

4.4 Material und Rolle des Verfassers

4.4.1 Wahl des Referenzmaterials

Als Grundlage dient der vollständige Bestand des Kurzfilms *Pinky Promise*, entstanden im Studienprojekt „Shortcut“ an der HAW Hamburg im Wintersemester 2023/24, gedreht an zwei Tagen im November 2023. Verwendet wird nicht ein ausgewählter Ausschnitt, sondern alles, was am Drehort aufgezeichnet wurde, einschließlich abgebrochener Aufnahmen und nicht verwendeter Einstellungen. Wo die Untersuchung darüber hinaus ergänzendes Material heranzieht, ist das an der betreffenden Stelle ausgewiesen (Abschnitt 4.5.3).

Tabelle 4.3: Komprimierte Anforderungsmatrix

ID	Anforderung	Priorität
FA-1	Unterstützte Videos hochladen und Quelle kennzeichnen.	MUSS
FA-2	Ingestion asynchron starten und Fortschritt anzeigen.	MUSS
FA-3	Metadaten, Audio, Szenen, Thumbnails und Merkmale erzeugen.	MUSS
FA-4	Transkript und Embeddings zeitlich verankert speichern.	MUSS
FA-5	Szenen semantisch und textbasiert suchen.	MUSS
FA-6	Eine geordnete Timeline mit Stil und Prompt erzeugen.	MUSS
FA-7	Timeline manuell verschieben und teilen können.	MUSS
FA-8	MP4 und optional NLE-Austauschdatei exportieren.	MUSS/KANN
NFA-1	Medienverarbeitung lokal und ohne zwingende Cloud-Abhängigkeit.	MUSS
NFA-2	Deterministische Grundkonfiguration dokumentieren.	MUSS
NFA-3	Fehler kontrolliert behandeln und sichtbar machen.	MUSS
NFA-4	Daten, Parameter und Entscheidungen nachvollziehbar speichern.	MUSS

Diese Wahl ist bewusst getroffen: Eine studentische Produktion arbeitet ohne Script Supervisor, ohne festes Datenmanagement und unter Zeitdruck, ihr Material trägt entsprechend Unregelmäßigkeiten. Für die Entwicklung eines Verfahrens, das Material erschließen soll, ist genau das der Vorteil, ein aufgeräumter Bestand hätte die Annahmen des Verfahrens bestätigt, ohne sie zu prüfen. Welche Unregelmäßigkeiten das sind, zeigte sich erst bei der Verarbeitung. Sie sind an den betreffenden Stellen belegt und seien hier zusammengefasst:

- Die Take-Nummern von Bild und Ton sind gegeneinander verschoben, in einem Fall widersprechen sich zusätzlich die Einstellungs-Nummern.
- Ein Teil der Tondateien wurde bereits am Aufnahmegerät umbenannt, sodass der Dateiname nicht mehr zur ursprünglichen Bezeichnung passt.

- Drei von 58 Bildaufnahmen tragen keinen Zeitcode. Alle 58 tragen zusätzlich einen im Gerät hinterlegten Zeitcode, der für jede Aufnahme denselben Wert zeigt und damit unbrauchbar ist.
- Die Kameraspuren enthalten überwiegend keinen verwertbaren Ton, sodass die Prüfung über die Wellenform nicht greift.
- Einzelne Drehbuchzeilen und Handlungen wurden gar nicht gedreht.

Jeder dieser Punkte hat eine Regel des Verfahrens hervorgebracht oder verändert: die Weigerung, allein nach dem Dateinamen zuzuordnen, das Verwerfen des geräteseitigen Zeitcodes, die Meldung „nicht anwendbar“ statt eines geratenen Wertes, das Ausweisen nicht gedrehter Zeilen als Lücke. Zwei Klarstellungen gehören dazu. Erstens werden die eingesetzten Modelle mit diesem Material *nicht trainiert*: Sie sind vortrainiert, das Material diente der Entwicklung, der Einstellung der Schwellen und der Prüfung. Zweitens gilt der Schluss nur in eine Richtung: Dass das Verfahren mit unregelmäßigem Material zurechtkommt, lässt erwarten, dass es mit sorgfältiger geführtem Material ebenfalls zurechtkommt. Ob es die zusätzlichen Angaben eines professionell dokumentierten Bestands (Script Supervisor, Drehbericht) ausschöpfen würde, ist damit nicht gesagt, es wertet sie derzeit nicht aus.

4.4.2 Rolle des Verfassers und Grenzen der Aussagekraft

Bei der in Abschnitt 4.4.1 genannten Produktion führte der Verfasser Regie. Der Vorteil: Die Zuordnung von Bild und Ton ist aus der Drehsituation bekannt, erst dadurch lässt sich beurteilen, ob eine automatische Zuordnung zutrifft, ohne diese Kenntnis wäre die Kaskade aus Abschnitt 5.3 nicht überprüfbar. Die Einschränkung: Entwicklung, Durchführung und Beurteilung fallen in einer Person zusammen. Die quantitativen Maße sind davon unberührt, weil sie ohne Zutun berechnet werden. Die Auswahl der Maße und die Deutung der Ergebnisse sind es nicht. Eine von der Entwicklung unabhängige Beurteilung liegt nicht vor, das ist die wesentliche Grenze dieser Arbeit und wird bei der Einordnung der Ergebnisse durchgehend mitgeführt.

4.5 Evaluationsdesign

Die Frage, ob ein Schnittvorschlag „gut“ ist, lässt sich nicht unmittelbar messen. Messbar ist, *was ein einzelner Bestandteil des Verfahrens beiträgt*. Die Untersuchung ist deshalb als Ablationsstudie angelegt: Dasselbe Verfahren wird mehrfach ausgeführt, wobei jeweils ein Bestandteil entfernt wird. Die gestalterische Qualität beantwortet das

nicht, sie bleibt der qualitativen Eigenbewertung vorbehalten. Beantwortet wird, ob die Entwurfsentscheidungen aus Kapitel 5 eine messbare Wirkung haben, was bei den kommerziellen Systemen aus Kapitel 3 nicht möglich ist, weil sich deren Verfahren nicht zerlegen lässt.

4.5.1 Vergleichsverfahren

Neben dem vollständigen Verfahren werden drei reduzierte Varianten ausgeführt, alle mit denselben Maßen bewertet.

Zufallsauswahl Die Anfrage wird ignoriert, Szenen werden zufällig gezogen, bis die Zieldauer erreicht ist: die untere Schranke.

Ohne harte Filter Gleiche Zerlegung, aber ohne die Bedingungen zu Einstellungsgröße, Sprecher und Dialog und ohne Vermeidung von Wiederholungen: isoliert den Beitrag der Filter.

Ohne Zerlegung Ein einziger Sprachmodell-Aufruf erhält Anfrage und Materialübersicht und gibt die Auswahl unmittelbar aus: isoliert den Beitrag der dreistufigen Architektur aus Planung, Abruf und Zusammenstellung.

4.5.2 Definition der erhobenen Maße

Je Durchlauf werden Maße aus fünf Gruppen erhoben. Sie sind im Auswertungsprogramm als reine Funktionen über den protokollierten Zwischenergebnissen definiert. Die folgenden Festlegungen geben diese Definitionen wieder, einschließlich der Schwellen, die über Treffer und Nichttreffer entscheiden.

Semantische Abdeckung. Für jeden belegten Slot des Plans wird dessen Absichtsformulierung mit der automatisch erzeugten Beschreibung der gewählten Szene verglichen. Herangezogen wird die englische Formulierung, ersatzweise die deutsche, wenn keine englische vorliegt. Beide Texte überführt derselbe CLIP-Textencoder in Vektoren, der auch das Retrieval trägt (Abschnitt 2.7). Die Abdeckung eines Durchlaufs ist der Mittelwert der Kosinusähnlichkeiten über die n belegten Slots,

$$C = \frac{1}{n} \sum_{i=1}^{n} \cos\Big(e(a_i),\, e(b_i)\Big), \tag{4.1}$$

mit der Absichtsformulierung a_i, der Szenenbeschreibung b_i und dem Textencoder e. Slots ohne Beschreibung oder ohne Absichtsformulierung gehen nicht in den Mittelwert ein. Minimum und Maximum je Durchlauf werden zusätzlich festgehalten.

Zur Gültigkeit: Das Maß vergleicht Text mit Text, gemessen wird die Nähe der Anfrage zur maschinellen *Beschreibung* der Szene, nicht zur Szene selbst, sodass Fehler des Bildmodells (Abschnitt 2.6) unmittelbar eingehen. Es taugt für den Vergleich von Verfahren auf demselben Bestand, nicht als absolute Güteaussage.

Bedingungstreue. Drei Anteile geben an, wie oft eine im Plan geforderte Eigenschaft in der gewählten Szene tatsächlich vorliegt. Die Treue der *Einstellungsgröße* zählt einen Treffer, wenn der Slot keine Vorgabe macht oder die erkannte Klasse der Szene der Vorgabe entspricht. Die *Sprecher*-Treue wird nur über die Slots gebildet, die einen Sprecher fordern. Ein Treffer liegt vor, wenn einem einzelnen Sprecher in der Szene mindestens eine halbe Sekunde Sprechzeit zugeordnet ist. Maßgeblich ist der Sprecher mit der längsten Sprechzeit, nicht die Summe über alle Sprecher. Die *Dialog*-Treue wird nur über die Slots gebildet, die Dialog fordern. Als Dialog gilt eine Transkription von mindestens 20 Zeichen. Fordert kein Slot eines Durchlaufs die jeweilige Eigenschaft, bleibt das Maß undefiniert und wird nicht als Null gewertet.

Zeittreue. Festgehalten werden die absolute Abweichung der erzeugten Gesamtdauer T von der Zieldauer T_{Ziel} und ihr Verhältnis

$$\Delta_T = \frac{|T - T_{\text{Ziel}}|}{T_{\text{Ziel}}}. \tag{4.2}$$

Vielfalt. Die Quellenvielfalt setzt die Zahl der verschiedenen Quelldateien ins Verhältnis zur Zahl der Segmente,

$$V = \frac{|\{\text{Quelldateien}\}|}{n_{\text{Segmente}}}, \tag{4.3}$$

und erreicht den Wert 1, wenn jedes Segment aus einer anderen Datei stammt. Die Verteilung der Einstellungsgrößen wird als Shannon-Entropie über die Klassen der gewählten Szenen erfasst,

$$H = -\sum_k p_k \log_2 p_k, \tag{4.4}$$

mit dem Anteil p_k der Klasse k. Der Wert 0 bedeutet, dass alle Segmente dieselbe Klasse tragen.

Laufzeit. Die Wanduhrzeit wird je Phase getrennt festgehalten: für die Zusammenfassung des Bestands, die Zerlegung der Anfrage, den Abruf der Kandidaten und die Zusammenstellung. Die Gesamtzeit eines Laufs wird eigenständig gemessen und ist keine Summe dieser Werte.

Zwei Einschränkungen binden die Auswertung. Erstens ist die Bedingungstreue nur dort aussagekräftig, wo eine Bedingung gestellt wurde. Ein Vergleich dieses Maßes ist deshalb nur zwischen dem vollständigen Verfahren und der Variante ohne harte Filter zulässig, da beide von derselben Zerlegung ausgehen. Zweitens ist die Zeittreue kein Qualitätsmaß: Die Zufallsauswahl füllt konstruktionsbedingt genau bis zur Zieldauer auf und erreicht darin den besten Wert, ohne inhaltlich etwas zu leisten.

4.5.3 Durchführung

Die Maße, die Vergleichsverfahren und der Satz an Anfragen wurden vor der Durchführung festgelegt. Die dabei zugrunde gelegte Anforderung an ein solches Vorgehen ist im folgenden Abschnitt ausgeführt. Die Anfragen sind nach Profilen gegliedert, die typische Aufgabenstellungen abbilden: reine Bildstrecken, Interviewpassagen, gemischte Formen sowie Grenzfälle.

Da für diese Aufgabe keine Evaluationskampagne mit vorgegebenem Anfragesatz existiert (Abschnitt 3.6), wurden die Anfragen für diese Arbeit selbst formuliert. Jede von ihnen trägt eine festgelegte Prüfabsicht: Eine Anfrage dient als Bezugspunkt, andere zwingen den Abruf auf einen bestimmten Inhalt, prüfen die Unterscheidung von Bewegung und Ruhe oder stellen eine harte Bedingung an Sprecher, Dialog oder Einstellungsgröße. Die Grenzfälle loten die Ränder aus, von acht Sekunden bis zu zwei Minuten und bis zu Bedingungen, die der Bestand kaum erfüllen kann. Die Zieldauern verteilen sich über diesen Bereich, damit sowohl der sehr kurze als auch der lange Schnitt vertreten ist. Prüfabsicht und Zieldauer je Anfrage führt Anhang A. Beide sind zusammen mit den Anfragen im Quelltext der Messung hinterlegt und seit dem ersten Lauf unverändert. Alle 18 Anfragen sind mit Kennung, Profil und Zieldauer im Wortlaut in Anhang A wiedergegeben. Sie wurden vor der Durchführung festgeschrieben und seither nicht verändert, damit spätere Läufe vergleichbar bleiben.

Der Bestand dieses Vergleichslaufs ist nicht mit dem Referenzmaterial aus Abschnitt 4.4.1 identisch: zwölf Szenen aus elf Dateien, davon fünf aus vier Aufnahmen des Kurzfilms und sieben aus ergänzenden Dateien (sechs frei verfügbare Archivaufnahmen, eine weitere Dialogaufnahme). Der Grund liegt in den Profilen, ein fiktionaler Kurzfilm enthält weder ruhige Natur-B-Roll noch Interviewmaterial. Die Ablationsstudie misst also das Verhalten

auf einem gemischten Bestand. Die Ergebnisse zur Zuordnung von Bild und Ton und zum Drehbuch-Abgleich (Kapitel 6) beruhen dagegen ausschließlich auf dem Material des Kurzfilms. Jeder Durchlauf schreibt seine vollständigen Zwischenergebnisse mit.

4.5.4 Vorkehrungen zur Objektivität der Messung

Ein Vergleich zwischen Verfahren ist nur dann aussagekräftig, wenn sich die Verfahren tatsächlich allein in dem Bestandteil unterscheiden, der untersucht werden soll. Fünf Vorkehrungen stellen das sicher. Sie sind im Auswertungsprogramm angelegt und nicht von der Person abhängig, die den Durchlauf startet.

Gleiches Material Der Bestand wird einmal ermittelt und an alle vier Verfahren unverändert übergeben.

Gleiche Bezugsgrundlage Die Sprecherzuordnung wird einmal ermittelt und für die Bewertung aller Verfahren herangezogen.

Gleiche Berechnung Dieselbe Programmfunktion berechnet die Maße für alle Verfahren, ohne verfahrensabhängige Verzweigung.

Festgelegter Zufall Die Zufallsauswahl arbeitet mit festem Startwert und ist reproduzierbar.

Fehler werden verzeichnet Abgebrochene Läufe gehen mit Fehlermeldung in die Ergebnisdatei ein, sodass ein Verfahren seinen Mittelwert nicht durch verschwindende schlechte Läufe verbessern kann.

Während der Durchläufe findet keine Beurteilung durch eine Person statt. Anfrage, Zieldauer und Verfahren stehen vorher fest, kein Lauf wird nachträglich verworfen oder wiederholt. Drei Grenzen sind zu benennen: Die Zerlegung arbeitet mit einem Sprachmodell bei Temperatur 0,3 und ist nicht vollständig wiederholbar. Jede Kombination aus Anfrage und Verfahren wird genau einmal ausgeführt, sodass diese Schwankung nicht durch Mittelung ausgeglichen wird. Und System und Auswertungsprogramm stammen von derselben Person, die Berechnung ist unbeeinflusst, die Auswahl dessen, was berechnet wird, nicht (Abschnitt 4.4.2).

4.6 Kriterien für die Beurteilung der Sequenzen

Die in Abschnitt 4.5 beschriebenen Maße erfassen, was sich zählen lässt. Sie sagen nichts darüber, ob eine Abfolge von Bildern als Erzählung trägt. Dafür werden drei Kriterien herangezogen, die am fertigen Ergebnis beurteilt werden.

Die drei Kriterien wurden als generische und systemunabhängige Dimensionen festgelegt. Dass Kriterien, Testmaterial und Bewertungsvorschriften vor der Messung festgelegt und offengelegt werden müssen, entspricht dem Vorgehen etablierter Evaluationskampagnen im Bereich der Videoanalyse [31].

Strukturelle Schnittlogik prüft, wo und wie Schnittpunkte gesetzt werden: Ein Schnitt soll an einer wahrnehmbaren Grenze zwischen Bildeinheit, Einstellung, Szene oder Handlung liegen. Betrachtet werden zudem Reihenfolge, Wiederholungen, A-Roll-/B-Roll-Beziehungen und Informationsaufbau.

Sprachliche Kontinuität prüft die Verbindung der Audiospuren über die Schnittgrenze hinweg: Unvollständige Wörter, fehlende Antworten, unklare Sprecherwechsel und sinnvolle Ellipsen werden getrennt dokumentiert.

Visuelle Kohärenz prüft, ob die Bildabfolge einen zusammenhängenden Erzählfluss erzeugt: Blickrichtung, Bewegungsanschluss, räumliche Orientierung, Einstellungsgröße, Licht und Farbe. Ein Unterschied gilt nur als Problem, wenn er im konkreten Kontext unbeabsichtigt oder störend wirkt.

4.7 Operationalisierung der Kriterien

Operationalisierung bezeichnet die Festlegung, anhand welcher beobachtbaren Größen ein theoretisches Konzept festgestellt wird, einschließlich der dafür gewählten Messinstrumente [38, S. 229]. Die Operationalisierung kombiniert hier automatisch prüfbare Indikatoren mit qualitativer Beurteilung. Automatisch prüfbar sind beispielsweise ungültige Zeitbereiche, doppelte Szenen, fehlende Quellen, Segmentdauer, Wechselraten und Exportstatus. Qualitativ werden Informationsaufbau, Dialogverständlichkeit, räumliche Orientierung und bewusste Regelbrüche beurteilt. Diese Zweiteilung ist notwendig, weil automatische Maße im Bereich der videobezogenen Beschreibung die wahrgenommene Qualität nur eingeschränkt abbilden und deshalb regelmäßig durch menschliche Beurteilung ergänzt werden [30].

4.8 Qualitative Eigenbewertung

Die Eigenbewertung nimmt der Verfasser selbst vor, unabhängige Betrachterinnen und Betrachter waren nicht vorgesehen. Beurteilt wird anhand der drei Kriterien aus Abschnitt 4.6: Notiert wird, wo ein Kriterium erfüllt oder verletzt ist und woran. Eine Bewertungsskala entfällt, weil ein Zahlenwert eine Messung nahelegte, wo eine begründete Einschätzung vorliegt. Ein Schluss auf die Wirkung bei Dritten ist nicht gedeckt (Abschnitt 4.4.2).

5. Entwicklung von CinAssist

5.1 Systemkonzept und Architektur

CinAssist ist eine lokal ausführbare, modulare Assistenzarchitektur. Grundlage der folgenden Darstellung sind der Quellcode des Projekts [37] und die zugehörigen technischen Aufzeichnungen [39]. Beide liegen der Arbeit bei (Anhang A). Beschrieben wird ausschließlich, was im Repository vorliegt und in einem Testlauf ausgeführt wurde.

Abbildung 5.1 zeigt den Aufbau in vier Ebenen: die Bereiche, ihre Bausteine und die jeweils tragenden Bibliotheken. Die vollständige Aufstellung der eingesetzten Bestandteile mit ihrer Aufgabe folgt in Abschnitt 5.1. Die folgenden Abschnitte gehen deshalb nicht die Bereiche der Reihe nach durch, sondern behandeln die Stellen, an denen eine Entwurfsentscheidung zu treffen war.

Eingesetzte Komponenten. Das System ist eine Eigenentwicklung, die durchgängig auf quelloffenen Komponenten aufsetzt. Die Medienverarbeitung übernimmt das extern aufgerufene FFmpeg. Die lokal ausgeführten Modelle sind Whisper (large-v3-turbo) für das Transkript mit Wortzeiten, CLIP (ViT-L-14) für die Bild- und Text-Embeddings der visuellen Suche, moondream für die Szenenbeschreibung, bge-m3 für den Abgleich von Drehbuch und gesprochenem Text, qwen2.5 für Zerlegung, Zusammenfassungen und Assistenzschicht sowie pyannote für die Sprecherdiarisierung, ausgeführt über Ollama beziehungsweise PyTorch. Szenengrenzen erkennt PySceneDetect, die Sprechaktivität silero-vad, die lexikalische Rangfolge rank-bm25. Dienst und Datenhaltung beruhen auf FastAPI, PostgreSQL, Celery und Redis, die Oberfläche auf Next.js und React. Sämtliche Bibliotheken stehen unter freizügigen Lizenzen (überwiegend MIT, BSD-3-Clause, Apache 2.0). Die vollständige Aufstellung mit Version und Lizenz je Komponente führt Anhang A. Sämtliche Modelle werden vortrainiert bezogen und lokal ausgeführt, ein eigenes Training findet nicht statt.

Wie sich dieser Aufbau der Nutzerin und dem Nutzer darstellt, zeigt Abbildung 5.2: Die Oberfläche gliedert sich in sieben Bereiche, von der Medienablage mit der semantischen Suche über den Programmmonitor bis zum Timeline-Editor und der Modulleiste, die

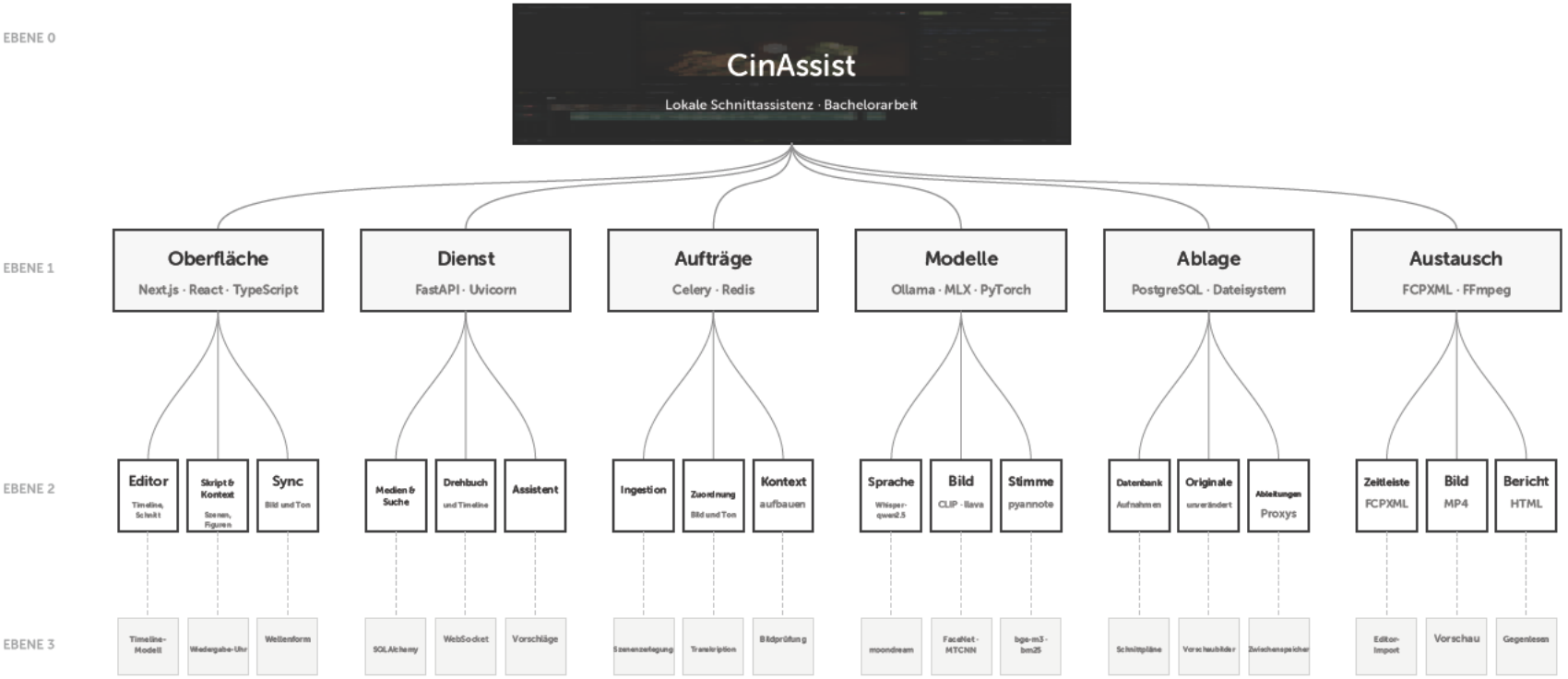

Abbildung 5.1: Aufbau des Systems in vier Ebenen: Bereiche, Bausteine und die jeweils tragenden Bibliotheken beziehungsweise Erzeugnisse.

zwischen Synchronisation, Drehbuch-Abgleich, Schnitt, Bearbeitung, Farbe und Ton wechselt.

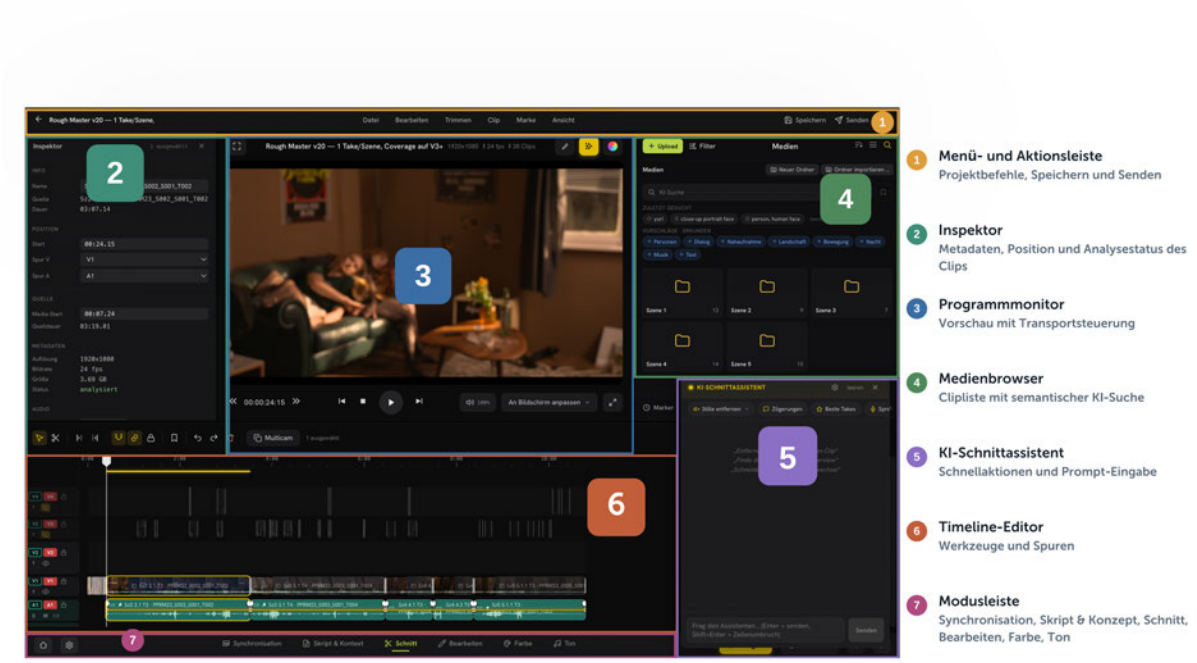

Abbildung 5.2: Funktionsbereiche der CinAssist-Oberfläche.

Ergänzende Systemdokumentation. Zwei Begleitdokumente vertiefen diesen Abschnitt und liegen den Materialien im Ordner `07_Systemdokumentation` bei. Das erste beschreibt den Aufbau Datei für Datei mit Aufgabe und Umfang je Modul, die 19 Tabellen des Datenmodells, die 92 Endpunkte der Schnittstelle, die vier tragenden Verfahren mit ihren Grenzfällen, die 97 Testfälle mit ihrem Gegenstand sowie die Bedienung Schritt für Schritt. Das zweite führt die eingesetzten Bibliotheken mit ihren Verfahren und Rechenvorschriften auf.

5.2 Videotechnische Grundlagen des Verfahrens

Die Zuordnung von Bild und Ton, die Bestimmung der Einstellungsgröße und die Ausgabe beruhen auf etablierten Verfahren der Videotechnik. Da diese den Kern der maschinellen Auswertung tragen, werden sie hier mit ihren Größen und Rechenvorschriften angegeben.

Zeitcode nach SMPTE 12M. Der auf einer Tonspur aufgezeichnete Längszeitcode (LTC) überträgt je Bild ein Wort von 80 Bit in Biphase-Mark-Codierung. Ein Wort schließt mit dem Synchronwort 0011 1111 1111 1101 ab, an dem sich die Wortgrenze eindeutig bestimmen lässt. Die Zeitfelder sind BCD-codiert (Binary Coded Decimal), jede Dezimalziffer belegt also eine eigene Vierergruppe von Bits.

Aus Stunden, Minuten, Sekunden und Bildnummer ergibt sich die fortlaufende Bildnummer

$$N = \Big((hh \cdot 60 + mm) \cdot 60 + ss\Big) \cdot f + ff, \tag{5.1}$$

und daraus die Zeit in Sekunden seit Mitternacht als $t = N/f$, wobei f die Bildfrequenz bezeichnet.

Die Bildfrequenz wird nicht angenommen, sondern aus dem Signal geschätzt: Da ein Wort 80 Bit umfasst, folgt aus der Abtastrate f_s und der Bitdauer T in Abtastwerten $f = f_s/(80\,T)$. Die Bitdauer wird robust über eine Trennung der Flankenabstände in zwei Häufungen geschätzt (ein festes Perzentil versagte bei vier von sechs Aufnahmen). Hinzu kommt eine Plausibilitätsprüfung der BCD-Felder samt Fortsetzung aufeinanderfolgender Bilder um genau eins, ohne die das Synchronwort auf Rauschkanälen zufällige Treffer liefert.

Zeitbezug der Tonaufnahmen. Die Tonaufnahmen tragen ihren Zeitbezug im *bext*-Abschnitt des Broadcast-Wave-Formats nach EBU Tech 3285. Dessen Feld *time_reference* enthält die Zahl n_{ref} der Abtastwerte seit Mitternacht, woraus sich der Beginn der Aufnahme ergibt als

$$t_0 = \frac{n_{\mathrm{ref}}}{f_s}. \tag{5.2}$$

Gelesen werden die Abschnitte der RIFF-Struktur unmittelbar, ohne den Datenteil anzutasten. Ergänzend führt der iXML-Abschnitt Bildfrequenz und Spurbelegung, gegen die der aus (5.2) gewonnene Wert geprüft wird.

Kreuzkorrelation der Wellenformen. Liegt auf der Kameraspur ein hörbarer Ton vor, lässt sich der Versatz zwischen Kameraton v und Tonaufnahme a über die Kreuzkorrelation bestimmen. Berechnet wird sie über den Korrelationssatz im Frequenzbereich,

$$(a \star v)[m] \;=\; \mathcal{F}^{-1}\Big\{\, \mathcal{F}\{a\}(k) \cdot \overline{\mathcal{F}\{v\}(k)} \,\Big\}[m], \tag{5.3}$$

mit auf 8 kHz herabgesetzter Abtastrate und einer Transformationslänge als nächster Zweierpotenz. Der gesuchte Versatz ist

$$\Delta = \frac{1}{f_s} \arg\max_m \left| (a \star v)[m] \right|. \tag{5.4}$$

Ein Ergebnis wird nur dann übernommen, wenn sich der Hauptwert deutlich abhebt: Er muss mindestens das Dreifache des größten Wertes außerhalb eines Fensters von ±0,5 s um ihn herum betragen. Andernfalls gilt die Stufe als nicht anwendbar. Das Vorzeichen ist so festgelegt, dass ein negativer Wert bedeutet, die Tonaufnahme sei vor dem Bild gestartet.

Einstellungsgröße aus dem Bildanteil. Die Einstellungsgröße wird aus dem Flächenanteil des größten gefundenen Gesichts am Bild bestimmt,

$$r = \frac{w_{\mathrm{G}} \cdot h_{\mathrm{G}}}{W \cdot H}, \tag{5.5}$$

und über feste Schwellen fünf Klassen zugeordnet: $r > 0{,}30$ Detail, $r > 0{,}15$ Nah, $r > 0{,}05$ Halbnah, darunter Halbtotale mit Person, ohne gefundenes Gesicht Totale. Mehrfachfunde derselben Person werden zuvor über den Überdeckungsgrad (IoU, Schwelle 0,30) zusammengefasst. Die fünf Klassen vergröbern die in Abschnitt 2.6 genannten acht Abstufungen. Die Zuordnung ist eine Näherung, die von der Gesichtsgröße auf die Einstellung schließt.

Ausgabe und Austauschformate. Die fertige Zeitleiste verlässt das System auf zwei Wegen, als gerendertes Bild und als Projekt für ein Schnittprogramm.

Die Bildausgabe erfolgt als H.264-Datenstrom im MP4-Behälter mit dem Qualitätsmaß CRF 18, der Ton als AAC mit 192 kbit/s. Weicht das Seitenverhältnis der Quelle vom Zielformat ab, wird das Bild eingepasst und der Rest mit schwarzen Balken gefüllt, sodass kein Bildinhalt verlorengeht.

Für die Weitergabe an ein Schnittprogramm schreibt das System die Zeitleiste als FCPXML nach Version 1.8 des Formats. Die Datei beschreibt Segmente, Quellen und Zeitbereiche, ohne Bilddaten zu übertragen, die Weiterarbeit erfolgt am unveränderten Originalmaterial. Sie trägt dabei mehr als die reine Abfolge: Zu jedem Segment der ersten Spur hängt sie die zugeordnete Tondatei mit dem im Sync ermittelten Versatz als eigene Tonlage an, dazu die Marken der Beats und die Kennzeichnung nach Szene, Einstellung und Take. Die in Abschnitt 5.3 beschriebene Zuordnung von Bild und Ton reist damit mit.

Für DaVinci Resolve baut das System die Zeitleiste zusätzlich unmittelbar über die Skript-Schnittstelle des Programms auf: Ein mitgeschriebenes Projektverzeichnis führt je Segment Quelle, Zeitlage, Spur und Freischaltung sowie die zugeordnete Tondatei mit Versatz, dazu die Ablage nach Szenen, die Marken der Beats und den berichtigten Startzeitcode der Tondateien. Erprobt ist dieser Weg mit DaVinci Resolve in Version 20. Abbildung 5.3 zeigt das Ergebnis.

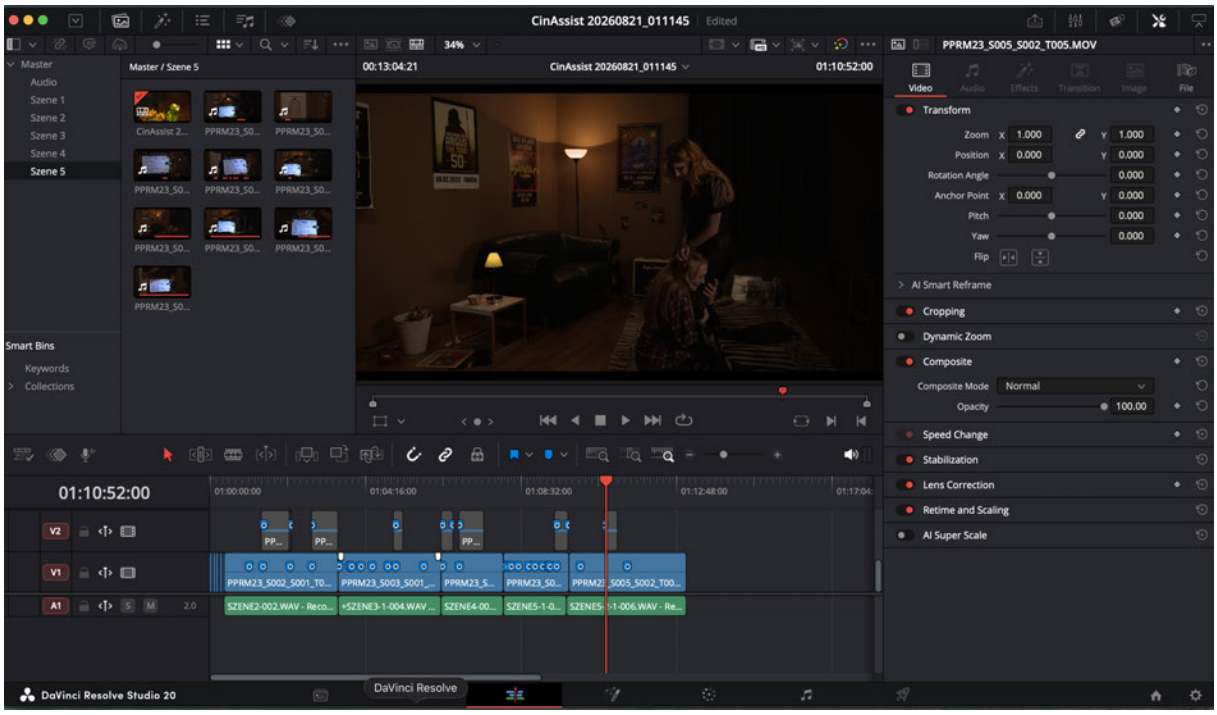

Abbildung 5.3: Über die Skript-Schnittstelle aufgebaute Zeitleiste in DaVinci Resolve. Die Medienablage links ist nach Szenen gegliedert. Spur V1 trägt die Abfolge, Spur V2 die abgeschalteten Auflagen, Spur A1 die zugeordneten Tonaufnahmen aus der Synchronisation. Die Marken auf den Segmenten der ersten Spur bezeichnen die Beats.

5.3 Synchronisation von Bild und Ton

Vor jeder inhaltlichen Analyse steht ein Problem, das in der Praxis die Regel ist: Bild und Ton werden getrennt aufgezeichnet, die Kamera hält oft nur einen Referenzton oder ein Timecode-Signal fest. So auch im Referenzkorpus: Eine Transkription der Kameraspur liefert nur Bruchstücke, erst die des verknüpften Tons den vollständigen Wortlaut, und damit die Grundlage der textbasierten Suche aus Abschnitt 2.7.

Vierstufige Zuordnungskaskade. CinAssist ordnet daher vor der Analyse jedem Ton sein Bild zu. Das Verfahren ist deterministisch und kommt ohne Sprachmodell aus:

Dieselben Dateien führen zu denselben Verknüpfungen. Es arbeitet in vier Stufen, die in dieser Reihenfolge greifen.

Stufe 1, Timecode. Der Timecode der Tondateien wird aus dem bext-Chunk gelesen und gegen den iXML-Chunk geprüft. Weichen beide ab, gilt er als nicht bestimmbar. Auf der Bildseite wird der als Audiosignal aufgezeichnete Timecode dekodiert. Überlappen sich die Zeitintervalle zweier Dateien um mindestens ein Fünftel der kürzeren Dauer, gelten sie als Kandidatenpaar, ab vier Fünfteln als gesichert.

Stufe 2, Wellenform. Liegt hörbarer Referenzton vor, wird der Versatz zusätzlich über die Kreuzkorrelation der Wellenformen bestimmt. Sie bestätigt die erste Stufe oder stuft die Zuordnung mit Warnung herab. Fehlt der Referenzton, meldet die Stufe ausdrücklich „nicht anwendbar“.

Stufe 3, Klappe. Der Klappenschlag als Transiente ist auffindbar, doch die nötige Szenenbeschreibung liegt bei der Ingestion noch nicht vor, sodass die Stufe in der Grundkonfiguration nicht aktiv ist.

Stufe 4, Dateiname. Namensbestandteile dienen nur der Gruppierung und der Erzeugung von Warnungen. Eine Zuordnung allein nach dem Namen erfolgt nur auf ausdrücklichen Wunsch und gilt als unsicher. Der Grund ist am Korpus zu sehen: Die Take-Nummern von Ton und Bild sind systematisch um eins verschoben: Ein Abgleich über den Namen wäre falsch und sähe zugleich richtig aus.

Entwurfsentscheidungen. Drei Festlegungen betreffen unmittelbar die in Kapitel 3 formulierte Anforderung an die Nachvollziehbarkeit. Erstens trägt jede Verknüpfung ihre Herkunft mit sich: Stufe, Konfidenzwert und eine lesbare Begründung. Zweitens wird Mehrdeutigkeit nicht aufgelöst, sondern ausgewiesen: Deckt ein Ton zwei überlappende Aufnahmen ab, entsteht der Status *unklar* mit Kandidatenliste, und die Verarbeitung ist blockiert, bis eine Person entschieden hat, denn ein System, das hier rät, erzeugte einen Fehler, der später kaum noch auffällt. Drittens werden unplausible Angaben verworfen statt übernommen: Der Container-Timecode wird ignoriert, sobald er auf mehreren Dateien identisch ist. Eine vorgeschaltete Klassifikation der Tonkanäle überspringt die Transkription von Stille, Timecode oder Rauschen und vermerkt dies. Ein ausgewiesener Verzicht ist einem stillschweigend erzeugten Fehlergebnis vorzuziehen.

Grenzen der Zuordnungskaskade. Jede Stufe benennt, was sie nicht leistet: Die Timecode-Stufe kann ein Auseinanderlaufen der Uhren nicht erkennen und deckt weder Aufnahmen über Mitternacht noch Drop-Frame-Arithmetik ab. Die Wellenform-Stufe setzt einen brauchbaren Referenzton voraus und ist bei Abstand oder Nachhall unzuverlässig. Die Klappen-Stufe erreicht nur den Status *plausibel*. Diese Aufstellung ist Teil des Verfahrens: Ein Zuordnungsvorschlag ohne Angabe seiner Reichweite wäre für die anschließende Bewertung wertlos.

5.4 Analysepipeline

Die Aufbereitung durchläuft je Aufnahme nacheinander Metadaten, Tonextraktion, Transkription, Szenenzerlegung, Bildbeschreibung und Merkmalsberechnung (die beteiligten Bibliotheken nennt Abschnitt 5.1). Hier ist allein festzuhalten, wie mit unvollständigen Ergebnissen umgegangen wird: der Punkt, an dem sich eine Aufbereitung bewährt oder unbrauchbare Daten in alles Weitere trägt.

Jedes Ergebnis bleibt an eine Aufnahme und einen Zeitbereich gebunden. Ohne diese Bindung wäre es später nicht mehr zuzuordnen. Der Zustand *fertig* wird erst vergeben, wenn die dafür erforderlichen Erzeugnisse tatsächlich vorliegen, ein Schritt, der ausfällt, macht den Auftrag unvollständig und nicht etwa fertig. Wo ein nachrangiges Modell nicht verfügbar ist, greift ein gröberes Verfahren, und dieser Rückfall wird gesondert vermerkt, statt im Ergebnis unsichtbar zu bleiben.

Umgang mit fehlender Sprechaktivität. Die Spracherkennung läuft nicht über die gesamte Datei, sondern über die Abschnitte, in denen ein vorgeschalteter Schritt Sprechaktivität feststellt. Das spart Zeit und verhindert, dass das Modell in Stille hineinhalluziniert.

Diese Vorauswahl hat eine Schwäche, die erst am Material auffiel: Weit entfernte oder stark verhallte Sprache, Rufe quer durch einen Raum in der Totalen, wird nicht als Sprechaktivität erkannt. In einem Fall blieb von 112 Sekunden einer Aufnahme nur die Klappenansage übrig. Die betroffene Szene verlor ihre Beats, und der erzeugte Schnitt kippte.

Die Antwort ist ein zweiter Durchlauf: Deckt die erkannte Sprechaktivität weniger als zehn Sekunden oder 15 Prozent der Dauer ab, läuft die Spracherkennung über die ganze Datei. Übernommen wird nur, was keinen erkannten Bereich überlappt, den Filter gegen Leerlauf-Ausgaben übersteht und eine mittlere Wortsicherheit von mindestens 0,35

erreicht, streng, weil diese Bedingungen genau jenen Schutz ersetzen, dessentwegen die Vorauswahl eingeführt wurde. Auf dem Referenzbestand betraf das 19 Aufnahmen, alle mit Zugewinn, einzelne von einem auf neun oder fünfzehn Abschnitte. Der Vorgang zeigt ein wiederkehrendes Muster: Eine Maßnahme gegen einen Fehler erzeugt einen zweiten, und dessen Behebung muss den ursprünglichen Schutz erhalten.

5.5 Drehbuchgestützte Schnittplanung

Liegt zu einer Produktion ein Drehbuch vor, steht eine zweite Strukturquelle zur Verfügung, die dem Rohmaterial vorausgeht. CinAssist nutzt sie in einem eigenen Verarbeitungsweg, der neben der in Abschnitt 5.6 beschriebenen anfragegesteuerten Erzeugung steht. Beide Wege münden in dieselbe Timeline und lassen sich unabhängig voneinander verwenden.

5.5.1 Vom Drehbuch zum geordneten Material

Zerlegung des Textes. Das Drehbuch wird zunächst in Szenen und Zeilen zerlegt, wobei Dialog, Aktion und Übergang unterschieden werden. Die Zerlegung ist regelbasiert und toleriert die uneinheitlichen Formate, die in nicht-industriellen Produktionen üblich sind. Liegt das Drehbuch in einer anderen Sprache vor als der Dreh, werden die Dialogzeilen einmalig übersetzt. Die Übersetzung ist editierbar und wird verworfen, wenn die Zeilenzahl nicht mehr übereinstimmt, ein Sprachmodell darf hier weder Zeilen zusammenfassen noch hinzufügen.

Die gesprochene Klappe. Für die Zuordnung eines Takes zur Drehbuchszene erweist sich die *gesprochene* Klappe als das verlässlichste Merkmal. Die Spracherkennung gibt sie in schwankender Form wieder („Scene 2.1, Take 3“, „Szene 3.2, Teil 3“), was die Auswertung berücksichtigt. Entscheidend ist der zugrunde liegende Befund: Die genannte Szenennummer verweist auf das Drehbuch, die Kamera-Dateinamen weichen davon regelmäßig ab, eine Zuordnung über den Namen wäre verlockend und falsch, derselbe Schluss wie bei der Tonzuordnung in Abschnitt 5.3.

Beginn des Spiels. Der Anfang einer Aufnahme ist nicht der Anfang des Spiels. Am Drehort wird vor der eigentlich gewollten Passage häufig der Szenenanfang angespielt. Die Regie setzt den maßgeblichen Teil dann mitten in der laufenden Aufnahme mit einem gesprochenen Zeichen in Gang. Wer die Aufnahme von vorn verwendet, zeigt eine Probe.

Diese Praxis war dem Verfahren zunächst nicht bekannt, und das Ergebnis zeigte es: Eine Einstellung begann mit dem Betreten des Raums, obwohl die maßgebliche Handlung erst zwanzig Sekunden später einsetzt, aufgefallen beim Gegenlesen, nicht dem System. Daraus wurde eine Regel: Die letzte Produktionsäußerung *vor* der ersten Spieläußerung bildet eine harte Grenze, vor der kein Fenster beginnen darf. Bemerkenswert ist der Fall, weil die Regel nicht aus den Daten ableitbar war: Sie stammt aus der Kenntnis des Drehs und musste dem System mitgeteilt werden, ein Beleg dafür, dass die Auswertung von Rohmaterial nicht allein aus dem Material heraus gelingt.

5.5.2 Zuordnung gesprochener Sätze zu Drehbuchzeilen

Der Kern der Schicht ist ein Abgleich zwischen den tatsächlich gesprochenen Sätzen eines Takes und den Dialogzeilen der zugehörigen Drehbuchszene. Der Abgleich erfolgt über Embeddings und funktioniert dadurch auch dann, wenn Drehbuch und Dreh in verschiedenen Sprachen vorliegen. Ergänzend wird die lexikalische Ähnlichkeit gegen die Übersetzung herangezogen, wobei der jeweils höhere Wert zählt.

Die Zuordnung ist bewusst *nicht* als streng aufsteigende Folge modelliert. Schauspielerinnen und Schauspieler wiederholen Zeilen, und die Spracherkennung zerteilt lange Sätze. Ein Verfahren, das eine monotone Reihenfolge erzwingt, verliert unter diesen Bedingungen zutreffende Treffer. Stattdessen wird je Satz die ähnlichste Zeile oberhalb einer Schwelle gewählt, ergänzt um einen geringen Bonus für die erwartete Reihenfolge.

5.5.3 Erzeugung des Rohschnitts

Aus Drehbuch und Takes entsteht anschließend ein Rohschnitt nach festen Regeln, ohne Beteiligung eines Sprachmodells: Die Reihenfolge folgt dem Drehbuch, je Szene dient die Einstellung mit der größten Zeilenabdeckung als Rückgrat, ein Wechsel findet nur an Sprecherwechseln statt und nur, wenn die andere Einstellung die Zeile ebenfalls abdeckt, die klassische Dialogauflösung. Aufeinanderfolgende Zeilen desselben Takes bilden ein einziges Segment. Drehbuchzeilen ohne Material werden als Lücke ausgewiesen, und jeder Eintrag des Schnittplans trägt Begründung und Beleg.

Wahl des Takes je Einstellung. Liegen zu einer Einstellung mehrere Takes vor, ist zu entscheiden, welcher in den Vorschlag gelangt. Die erste Fassung nahm den Take mit der höchsten Skript-Abdeckung. Das erwies sich als zu grob, weil ein abgebrochener Take die Zeilen bis zum Abbruch vollständig abdeckt und deshalb gewinnen konnte.

Heute bewertet ein Punktwert jeden Take aus fünf gemessenen Größen und einer Nutzerangabe: Anteil der gedeckten Drehbuchzeilen (Gewicht 0,5), Abbruchstatus (+0,2 ohne, −0,3 mit Abbruch, −0,2 für sehr kurze Takes), Länge des gespielten Abschnitts relativ zur Szene (0,2) und im Bild bestätigte Regieanweisungen (je 0,25 bis 0,75, nur bei Bestätigung in mindestens zwei Einzelbildern). Die Nutzerbewertung, das Gegenstück zum eingekreisten Take des Script Supervisors, den es bei dieser Produktion nicht gab (Abschnitt 4.4.1), verschiebt den Wert um ±1,0 beziehungsweise +0,3. Sie ist das stärkste Einzelgewicht, überstimmt die gemessenen Größen aber nicht in jedem Fall. Als abgebrochen gilt ein Take, in dem nach Spielbeginn Produktions-Sprech einsetzt, der eine Wiederholung verlangt („nochmal“, „von vorne“, „sorry“). Der Befund wird mit Zeitpunkt und Wortlaut festgehalten.

Der Punktwert wirkt zweistufig: Bei der Wahl des besten Takes je Einstellung scheiden abgebrochene ganz aus und kehren nur ohne Alternative zurück, denn ein abgebrochener Take ist besser als eine Lücke. In der anschließenden beatweisen Planung treten alle Takes erneut an. Der Abbruch zieht dort 0,8 ab, statt auszuschließen, sodass ein abgebrochener Take einen einzelnen Beat gewinnen kann, wenn er ihn als einziger abdeckt. Wer einen Take verbindlich setzen will, verschiebt ihn im Schnitt selbst, die in Abschnitt 5.7 beschriebene Aufgabenteilung.

5.5.4 Abgleich der Drehbuch-Handlungen mit dem Bild

Die bis hierher beschriebenen Schritte prüfen das Gesprochene gegen das Drehbuch. Für stumme Passagen gab es diese Prüfung zunächst nicht, mit sichtbarer Folge: In einer Dialogpause stand eine Detailaufnahme, für die sich im Drehbuch kein Anlass fand, das Bildmaterial war nie *gegen* das Drehbuch geprüft worden.

Dafür wurde ein eigener Schritt ergänzt. Aus den Regieanweisungen einer Szene formuliert das Sprachmodell je Handlung zwei bis drei geschlossene Fragen sowie eine kurze Bezeichnung. Diese Fragen werden einem Bildmodell an dicht gesetzten Einzelbildern aller Aufnahmen der Szene gestellt (alle fünf Sekunden in stummen, alle zehn in dialoghaltigen Passagen). Als zweites, unabhängiges Signal dient die Ähnlichkeit zwischen Einzelbild und Bezeichnung im gemeinsamen Merkmalsraum.

Der Zuschnitt der Fragen beruht auf einer eigenen Messung: Fragen nach *Körperhandlungen und Posen* werden zuverlässig beantwortet, Fragen nach *kleinen Gegenständen* häufig fälschlich verneint, nach Stimmungen oder Absichten gar nicht belastbar. Solche Fragen sind ausgeschlossen. Berücksichtigt wird zudem, dass ein Drehbuch nicht wörtlich verfilmt wird: Steht dort „einen Song schreiben“, kann im Bild jemand Gitarre spielen. Das Ergebnis

ist je Drehbuch-Handlung einer von drei Zuständen, *gedreht*, *unsicher* oder *fehlt*, mit den belegten Zeitfenstern. Es ist in der Oberfläche sichtbar und geht in den Prüfbericht ein. Abbildung 5.4 zeigt einen Ausschnitt des erzeugten Prüfberichts mit dem Ergebnis dieses Abgleichs.

✓	A	Skript-Aktion	Status	Belegt in (Take · Einstellung · Zeit)
☐	A0	Orpheus and Eurydice lie on the sofa, their little fingers are intertwined. The alarm rings on Orpheus's phone. He wakes up and turns it off. He looks to Eurydi	gedreht	S002_S001_T004 2.1 1:10.0–2:30.0 · S002_S001_T007 2.2 1:20.0–2:20.0 · S002_S004_T001 2.4 0:05.0–0:30.0 · S002_S001_T006 2.2 1:10.0–2:30.0 · S002_S001_T003 2.1 0:10.0–0:50.0, 2:40.0–3:20.0, 3:40.0–3:50.0 · S002_S001_T002 2.1 2:20.0–2:40.0, 3:00.0–3:10.0
☐	A2	Orpheus gets up, kisses her on the forehead and goes to the kitchen. We see Eurydice still laying motionless on the sofa.	unsicher	S002_S001_T007 2.2 (CLIP +0.04) · S002_S001_T006 2.2 (CLIP +0.05) · S002_S001_T003 2.1 (CLIP +0.04) · S002_S001_T001 2.1 (CLIP +0.03)
☐	A3	Shortly after Orpheus comes back with two cups of tea. He puts them on the table in front of the sofa and looks to Eurydice.	gedreht	S002_S001_T004 2.1 0:20.0–0:30.0 · S002_S001_T007 2.2 1:10.0–1:30.0 · S002_S001_T005 2.2 0:20.0–0:30.0
☐	A5	As he gets no response, Orpheus becomes worried and starts touching and shaking Eurydice to wake her up.	gedreht	S002_S001_T004 2.1 0:00.0–0:20.0, 0:50.0–2:20.0 · S002_S001_T007 2.2 0:00.0–0:20.0, 1:30.0–2:20.0 · S002_S001_T006 2.2 2:10.0–2:20.0 · S002_S001_T005 2.2 0:00.0–0:05.0, 0:20.0–0:25.0 · S002_S001_T003 2.1 0:00.0–0:50.0, 1:10.0–1:20.0, 2:50.0–3:10.0, 3:40.0–4:00.0 · S002_S001_T002 2.1 0:10.0–0:50.0, 1:10.0–1:40.0, 2:10.0–2:40.0, 3:00.0–3:20.0
☐	A7	Orpheus realizes that Eurydice is dead and looses it completely.	gedreht	S002_S001_T004 2.1 2:10.0–2:20.0 · S002_S001_T007 2.2 1:50.0–2:20.0 · S002_S001_T006 2.2 2:10.0–2:20.0 · S002_S001_T003 2.1 0:10.0–0:40.0, 2:50.0–3:10.0, 3:40.0–3:50.0 · S002_S001_T002 2.1 2:20.0–2:30.0, 3:00.0–3:10.0 · S002_S001_T001 2.1 0:10.0–0:50.0
☐	A9	The scene ends by Orpheus crying and hugging the dead body of Eurydice.	fehlt	–

Abbildung 5.4: Abgleich der Drehbuch-Handlungen mit dem Bild, Ausschnitt aus einem erzeugten Prüfbericht mit den sechs Handlungen der Szene 2. Die Nummern beziehen sich auf alle Drehbuchzeilen, weshalb sie Lücken aufweisen. Handlung A2 bleibt *unsicher*, weil allein das nachrangige Ähnlichkeitssignal anspricht und keine der geschlossenen Fragen bejaht wurde.

Sie ändert zugleich die Anordnung des Rohschnitts. Stumme Aufnahmen werden nun nach den *bestätigten* Handlungen in der Reihenfolge des Drehbuchs gesetzt statt nach Klappennummer und Bewegung. Eingeschobene Aufnahmen sind nur noch dann zulässig, wenn das Drehbuch zwischen zwei Dialogzeilen eine Handlung nennt, diese im Bild belegt ist und aus derselben Szene stammt. Eine bestätigte Handlung wird höchstens einmal je Szene verwendet, und ein Treffer in einem einzelnen Einzelbild zählt nicht.

5.5.5 Gesichter, Figuren und der Reaktionsschnitt

Der Abgleich der Handlungen sagt, *was* in einer Aufnahme geschieht, nicht *wer* darin zu sehen ist, für die Dialogmontage mit ihrem Wechsel auf die zuhörende Figur eine wesentliche Lücke. Ein weiterer Schritt schließt sie ohne hinterlegte Gesichtsdatenbank: Aus allen Clips werden regelmäßig Einzelbilder entnommen, gefundene Gesichter als Merkmalsvektoren beschrieben und bestandsweit zu Gruppen zusammengefasst, jede Gruppe eine Person, zunächst ohne Namen.

Die Benennung erfolgt über das Drehbuch: Für jede Figur ist ablesbar, in welchen Szenen sie vorkommt, für jede Gesichtsgruppe, in welchen Szenen sie erscheint. Der Abgleich beider Verteilungen ergibt die Zuordnung. Eine Person wird also nicht daran erkannt, wer sie ist, sondern daran, *wann sie auftritt*. Abweichende Zuordnungen kann die Nutzerin überschreiben.

Damit wird der Reaktionsschnitt möglich: Innerhalb eines Dialogabschnitts wird eine Aufnahme gesucht, in der die *zuhörende* Figur mindestens zweieinhalb Sekunden im Bild ist und die sprechende weitgehend nicht. Bemerkenswert ist die Benennung des Ergebnisses: Zeigt die Aufnahme zwei Personen, heißt es „engere Gegeneinstellung (Zweier, Zuhörer im Bild)“ statt Reaktion, das Verfahren beansprucht keinen echten Gegenschuss, wenn ihm nur eine engere Einstellung auf beide Figuren vorliegt.

In der Oberfläche sind je Gesichtsgruppe Aufnahmenzahl, Gesichterzahl und Übereinstimmung mit der Auftrittsverteilung ausgewiesen, samt der Grenzen des Verfahrens: Dieselbe Person kann als mehrere Gruppen erscheinen, kleine Gruppen bleiben ohne Namen, einzelne Zuordnungen beruhen auf schwachen Übereinstimmungswerten.

5.5.6 Spurordnung und Prüfbericht

Die Ordnung der Spuren. Grundprinzip der Anordnung ist der *Szenen-Master*: Je Szene trägt die erste Spur eine möglichst durchlaufende Aufnahme, gewählt nach dem Punktwert aus Abschnitt 5.5.3, sodass die Erzählung innerhalb einer Szene in einem Fluss bleibt, statt aus vielen kurzen Stücken zusammengesetzt zu werden. Die übrigen Aufnahmen derselben Szene, darunter *Geschwister*, also parallel gedrehte Einstellungen derselben Passage in anderer Auflösung, werden an den Beats des Masters ausgerichtet und darüber angeordnet. Zunächst zerschnitten Reaktionen und Einschübe diese durchlaufende Aufnahme. Das Ergebnis war richtig, aber nicht rückgängig zu machen: Wer den Einschub verwarf, hatte ein Loch. Heute liegen die Ebenen übereinander: Die erste Spur trägt den Master in der Reihenfolge des Drehbuchs, die zweite Reaktionen und Einschübe als

stumme Auflagen, deren Ton von unten weiterläuft, die weiteren Spuren die an den Beats ausgerichteten Alternativen derselben Stelle.

Die Wiedergabe läuft von der obersten Spur nach unten und nimmt das erste sichtbare Segment. Trägt es keinen Ton, wird dieser auf der nächsten tonführenden Spur darunter gesucht. Darauf beruht die Umkehrbarkeit: Die Alternativspuren sind nach dem Laden *ausgeblendet*, nicht stummgeschaltet, ein bloßes Stummschalten nähme ihnen das Bild nicht. Eine Auflage lässt sich damit ausblenden, ohne dass eine Lücke entsteht, und Alternativen lassen sich durch Ein- und Ausblenden vergleichen. Der Vorschlag ist keine fertige Abfolge mehr, die man annimmt oder verwirft, sondern eine geordnete Auswahl, aus der geschnitten wird, was der in Abschnitt 5.7 beschriebenen Aufgabenteilung genauer entspricht.

Beim Export nach FCPXML wird die Ordnung bewusst abgeflacht. Alle Segmente außerhalb der ersten Spur werden als abgeschaltet übergeben, damit die Zeitleiste im Schnittprogramm zunächst die durchlaufende Fassung zeigt und der Editor die Auflagen einzeln zuschaltet. Sie sind vorhanden und zeitlich richtig gelegen, aber nicht aktiv.

Der Prüfbericht. Zu jedem erzeugten Schnitt entsteht ein eigenständiges Dokument, das das Ergebnis zum Gegenlesen aufbereitet, mit drei Darstellungen je Szene. Die erste führt die Dialogzeilen des Drehbuchs auf: Original und Übersetzung, gewählte Aufnahme, Position, Ein- und Ausstieg, den tatsächlich gesprochenen Satz, Ähnlichkeitswert und Begründung. Nicht gefundene Zeilen erscheinen als Lücke. Die zweite trägt alle Aufnahmen der Szene nach Beats auf, jede Zeile ein Beat, jede Spalte eine Aufnahme, in den Zellen Zeitbereich und Art des Belegs, damit ist an einer Stelle ablesbar, welche Aufnahmen dieselbe Stelle der Geschichte tragen und warum eine den Vorzug erhielt. Abbildung 5.5 gibt die zweite dieser Darstellungen wieder.

Die dritte führt die Handlungsanweisungen mit ihrem Prüfergebnis auf (Abschnitt 5.5.4). Stumme Abschnitte erscheinen mit ihrer Bildbeschreibung, Lücken gesondert. Der Bericht schließt mit einer Bilanz und einer Ankreuzspalte für das Gegenlesen. Er löst damit ein, was Kapitel 3 als fehlende Eigenschaft der kommerziellen Systeme benannt hat: Der Vorschlag ist Zeile für Zeile daraufhin prüfbar, warum eine bestimmte Aufnahme an einer bestimmten Stelle steht.

Rolle des Sprachmodells. Die Arbeitsteilung innerhalb dieser Schicht ist festgelegt. Zerlegung, Klappenauswertung, Abgleich, Abdeckungsberechnung und Schnittplanung arbeiten regelbasiert und liefern bei gleicher Eingabe dasselbe Ergebnis. Ein Sprachmodell wird nur für Übersetzungen und für zusammenfassende Beschreibungen herangezogen, und auch dort unter der Auflage, jede Aussage zu belegen. Liefert es kein verwertbares Ergebnis,

Beat	2.1 T1 S002_S001_T001	2.1 T2 S002_S001_T002	2.1 T3 S002_S001_T003	2.1 T4 S002_S001_T004	2.2 T2 S002_S001_T005	2.2 T3 S002_S001_T006	2.2 S0
B0 *eroeffnung*: Orpheus and Eurydice lie on the sofa, their little fingers are intertw	–	★ E 0:35.0–1:09.7	VE 0:35.5–1:05.7	–	–	–	E 0:1
B1 Z1 ORPHEUS: „Guten Morgen Babe, ich mach uns gleich mal Tee.“	–	★ A 1:09.4–1:18.2	A 1:03.8–1:18.2	0:38.2–1:20.2	–	–	–
B2 Z4 ORPHEUS: „Babe, es ist Zeit aufzustehen, hast du heute nicht Vorlesung“	–	–	–	1:19.8–1:20.2	V 0:27.5–0:28.5	A 0:26.3–0:57.3	★ 0:1
B3 Z6 ORPHEUS: „Das ist nicht witzig Eury, wach auf!“	V 0:13.1–0:15.2	★ AV 1:24.6–2:31.1	V 1:13.1–2:50.8	AV 1:20.2–1:50.0	–	A 0:56.2–1:01.6	V 1:3
B4 Z8 ORPHEUS: „Nein Eury, bitte, ich kann das nicht... Komm zurück!“	AV 0:39.6–1:05.5	★ A 2:29.0–3:00.8	AV 2:50.6–3:04.7	A 1:49.4–2:01.6	–	1:01.5–2:05.0	AV 1:3
B5 *schluss*: The scene ends by Orpheus crying and hugging the dead body of Eurydice	–	–	–	2:01.2–2:06.1	–	–	–

★ = gewählte Quelle · A = alignierte Skriptzeile (Anker) · s = eindeutige Improvisation (semantisch) · V = Skript-Aktion im Bild bestätigt · E =

Abbildung 5.5: Übersicht der Beats einer Szene über die verfügbaren Aufnahmen. Aus Platzgründen sind nicht alle Spalten abgebildet. Die Markierung weist die gewählte Quelle aus.

wird der betreffende Eintrag als unsicher gekennzeichnet, statt eine Ersatzannahme zu treffen.

5.6 Schnittlogik und Timeline-Generierung

Bewertung einer Szene. Damit eine Abfolge bewertet werden kann, braucht jede Szene eine Kennzahl für ihre Bewegtheit. Eine frühere Fassung aus fest gewichteten Bildmaßen wurde ersetzt, weil die Gewichte willkürlich blieben. An ihre Stelle trat eine Bewertung im gemeinsamen Bild-Text-Raum. Die Szene wird den Beschreibungsgruppen für Bewegung und für Ruhe gegenübergestellt, aus dem Abstand beider mittlerer Ähnlichkeiten ergibt sich

$$E = \mathrm{clamp}\Big(0{,}5 + 2\Big(\overline{\cos}(e, P_{\mathrm{akt}}) - \overline{\cos}(e, P_{\mathrm{ruhe}})\Big),\ 0,\ 1\Big), \tag{5.6}$$

mit dem Bild-Embedding e der Szene und den Embeddings der beiden Beschreibungsgruppen. Der Operator $\mathrm{clamp}(x, 0, 1)$ begrenzt das Ergebnis auf das Intervall $[0, 1]$. Der Rohabstand liegt erfahrungsgemäß zwischen $-0{,}10$ und $+0{,}15$, Faktor und Verschiebung bilden ihn auf das Einheitsintervall ab. Das Verfahren folgt der Anwendung von Bild-Text-Modellen ohne aufgabenspezifisches Nachtrainieren [15].

Der Gewinn liegt weniger in der Genauigkeit als in der Begründbarkeit: Es sind keine frei gewählten Koeffizienten mehr zu rechtfertigen, und alle Werte stammen aus demselben Merkmalsraum wie die Suche.

Fehlt das Bild-Embedding, greift die frühere Formel als Rückfall, und fehlt auch dieses, eine Abschätzung über die Dauer. Diese Staffelung ist eine durchgehende Eigenschaft des Systems: Ein fehlendes Signal führt zu einem gröberen, aber benannten Ergebnis statt zum Abbruch.

Visueller Abstand zweier Szenen. Für die Vermeidung gleichförmiger Abfolgen wird der Abstand zweier Szenen benötigt. Liegen Embeddings vor, ist er der Kosinusabstand

$$d(a, b) = 1 - \cos(e_a, e_b). \tag{5.7}$$

Andernfalls tritt eine Näherung aus Bildmaßen an seine Stelle, die Unterschiede in Helligkeit, Kontrast, Farbtemperatur und Bewegung mit den Anteilen 0,30, 0,25, 0,30 und 0,15 verrechnet. Auch hier ist die Näherung als solche gekennzeichnet.

Suche nach der Abfolge. Die anfragegesteuerte Timeline-Generierung folgt dem Muster Plan, Retrieve, Assemble: Der Prompt wird in Slots übersetzt, noch unbesetzte

Positionen mit feststehender Dauer, Bildausschnitt und Anforderung. Kandidaten werden über CLIP-Ähnlichkeit und BM25-Relevanz gefunden, und die Anordnung erfolgt als Beam Search mit Strahlbreite drei, die Bewertung berücksichtigt Energie, visuelle Diversität, Dialogbezug sowie A-Roll-/B-Roll- und Clip-Wechsel. Eine schrittweise Wahl des jeweils besten nächsten Segments unterbleibt, weil sie früh in eine ungünstige Folge führen kann. Der Aufwand von $\mathcal{O}(b \cdot |C|^2)$ ist für die vorkommenden Größenordnungen unerheblich. Die Schnittstellen werden anschließend auf Sprechpausen gelegt (Lücke über 300 ms zwischen Transkriptabschnitten), damit kein Schnitt mitten in einen Satz fällt. Die Timeline bleibt ein Vorschlag. Abbildung 5.6 zeigt die Oberfläche mit einem erzeugten Schnitt.

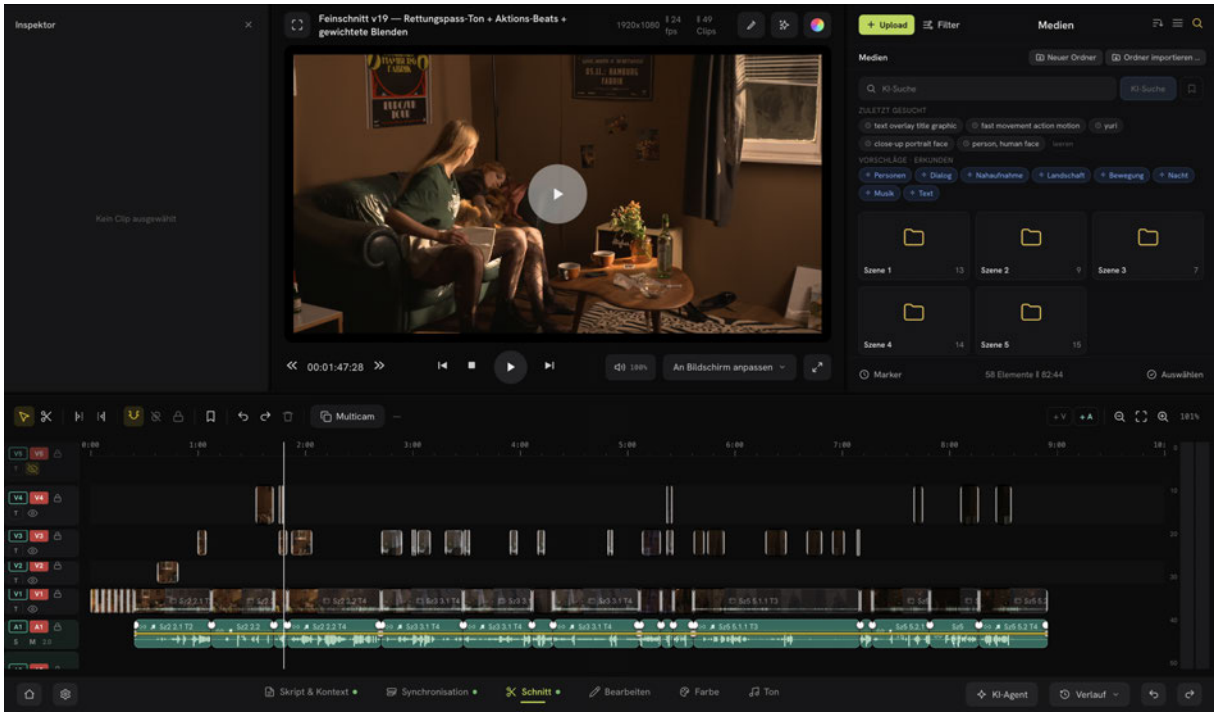

Abbildung 5.6: Editor mit erzeugtem Schnitt. Oben die Vorschau, rechts der nach Drehbuchszenen geordnete Materialbestand mit der hybriden Suche, unten die Timeline mit den Bildspuren (V1 trägt die Abfolge, V2 bis V5 die Alternativen) und der Tonspur. Die Beschriftung der Segmente nennt Szene, Einstellung und Take der Quelle.

5.7 Assistenzschicht

Die beiden bisher beschriebenen Erzeugungswege setzen voraus, dass der Editor weiß, welches Verfahren er anstoßen möchte. Darüber liegt eine Schicht, die eine in natürlicher Sprache formulierte Absicht entgegennimmt und in eine Folge von Arbeitsschritten übersetzt.

Aufbau. Die Schicht folgt einem etablierten Muster: Das lokal laufende Sprachmodell formuliert einen Gedanken, wählt eine Aktion, erhält deren Ergebnis als Beobachtung zurück und entscheidet über den nächsten Schritt, bis eine Antwort vorliegt oder eine Höchstzahl an Durchläufen erreicht ist. Die verfügbaren Aktionen sind eine begrenzte Menge benannter Funktionen für drei Aufgaben: Material auffinden, die Aufbereitung nachbessern und eine Timeline erzeugen. Das Sprachmodell kann ausschließlich diese Funktionen aufrufen. Auf Dateisystem, Datenbank oder Timeline hat es keinen Zugriff. Jeder Zwischenschritt wird während der Ausführung an die Oberfläche übertragen, der Ablauf ist damit während seiner Entstehung sichtbar.

Vorschlag statt Ausführung. Die Schicht verändert die Timeline nicht, sondern erzeugt einen *Vorschlag*, ein Bündel von Bearbeitungsbefehlen, das dem Editor zur Annahme oder Ablehnung vorgelegt wird. Diese Trennung ist im Aufbau verankert: Die Vorschlagsverwaltung besitzt selbst keine Möglichkeit, einen Vorschlag anzuwenden, die Ausführung liegt allein beim Timeline-Editor. Ein Vorschlag kann die Zustimmung folglich nicht umgehen, auch nicht durch einen Fehler im Sprachmodell. Vor der Entscheidung erscheint der Vorschlag als gestrichelte Vorschau über den bestehenden Segmenten. Derselbe Mechanismus gilt auch für die regelbasierten Wege: Rohschnitt (Abschnitt 5.5) und anfragegesteuerte Timeline (Abschnitt 5.6) durchlaufen dieselbe Zustimmung. Damit ist die in Abschnitt 2.3 begründete Anforderung baulich eingelöst: Ein Schnittvorschlag tritt als Angebot auf und nicht als Ergebnis.

5.8 Datenmodell und Schnittstellen

Das Datenmodell ist in drei Schichten gegliedert, die dem Weg des Materials durch das System folgen. Entscheidend ist dabei weniger die Zahl der Entitäten als die Eigenschaft, dass jedes Analyseergebnis bis auf seine Quelle zurückführbar bleibt.

Import und Zuordnung. Jede Datei wird als eigenständiges Medienobjekt geführt, mit technischen Kennwerten, Timecode samt Herkunft, den aus dem Dateinamen gelesenen Angaben und einem Fingerabdruck gegen Doppel-Importe, wobei die Originaldateien unverändert an ihrem Ort bleiben. Darüber liegt der Take als Zusammenfassung eines Bildes mit den zugeordneten Tönen. Jede Zuordnung trägt Versatz, Verfahren, Konfidenz und Begründung, und ein Take ohne Gegenstück bleibt sichtbar.

Analyse. Der Clip ist der Bezugspunkt aller inhaltlichen Ergebnisse: An ihm hängen die erkannten Szenen mit Zeitbereich, Beschreibung, Einstellungsgröße und Repräsentation für die visuelle Suche. Sprecher werden getrennt geführt und über eine Zwischentabelle mit ihren Szenen verbunden, so bleibt dieselbe Person über Clipgrenzen hinweg dieselbe Person.

Drehbuch und Kontext. Das zerlegte Drehbuch (Szenen, Zeilen, Original und Übersetzung) trägt drei Kontextebenen: Take-Kontext (was in einer Aufnahme geschieht), Szenen-Kontext (die Aufnahmen einer Drehbuchszene) und Story-Kontext. Der erzeugte Schnittplan behält für jeden Eintrag Begründung und Beleg.

Zeitliche Provenienz. Quellzeit und Timelinezeit werden durchgängig getrennt geführt: Ein Segment verweist auf einen Zeitbereich einer Szene, diese auf einen Clip, dieser auf einen Take, dieser auf die unveränderte Originaldatei, und die Tonzuordnung dazwischen trägt Versatz und Begründung. Diese Kette ist die Voraussetzung der in Kapitel 3 als fehlend beschriebenen Prüfbarkeit: Zu jedem Segment lässt sich beantworten, aus welcher Aufnahme es stammt, mit welchem Ton, aufgrund welcher Zuordnung und mit welcher Sicherheit. Die vollständige Aufstellung der Tabellen führt Anhang A.

5.9 Entwicklung mit generativer KI

Claude und Visual Studio Code dienten als Entwicklungswerkzeuge, nicht als Bestandteile der ausgelieferten Anwendung. Keines der im System verwendeten Modelle stammt aus dieser Assistenz. Verwendet wurde eine Entwicklungsassistenz, die unmittelbar im Projektverzeichnis arbeitet: Sie liest und schreibt Dateien und führt Kommandos aus, statt nur Textvorschläge auszugeben. Dieser Unterschied erklärt, warum sie das Ergebnis ihrer eigenen Änderungen unmittelbar prüfen konnte.

Die folgenden Angaben beruhen nicht auf Erinnerung. Der Verlauf einer zusammenhängenden Arbeitsphase von rund 25 Stunden, von der Zuordnung von Bild und Ton bis zu den Reaktionsschnitten, wurde vollständig aufgezeichnet und liegt der Arbeit bei [40]. In ihm stehen 53 Eingaben des Entwicklers 939 Werkzeugaufrufen der Assistenz gegenüber.

Die Eingaben sind kurz, 20 Wörter im Median, und enthalten überwiegend keine Lösungsvorgabe, sondern die Beobachtung einer Fehlfunktion. Knapp ein Drittel beanstandet ein beobachtetes Verhalten. Die Iteration bestand also nicht darin, eine Anweisung zu verfeinern, sondern das Ergebnis zu benutzen und den nächsten Mangel

zu benennen, derselbe Zyklus wie in Abschnitt 4.2, nur mit dem Entwickler anstelle des Prüfberichts als Beobachter. Auf eine Anweisung folgen im Mittel etwa achtzehn selbstständige Schritte, davon rund 71 Prozent das Ausführen von Kommandos und nur etwa sieben Prozent das Schreiben von Dateien (Abbildung 5.7): Die Assistenz hat überwiegend ausgeführt und nachgesehen, nicht Code verfasst. Diese Arbeitsteilung folgt derselben Regel wie das entwickelte System: Die Maschine bereitet vor und legt offen, die Entscheidung bleibt bei der Person (Abschnitt 5.7).

Wirksamkeit der Eingabearten. Die 53 Eingaben lassen sich nachträglich fünf Arten zuordnen. Die Einteilung ist keine Messung, sondern eine Auswertung des eigenen Vorgehens am aufgezeichneten Wortlaut. Sie ist in der beiliegenden Übersetzung je Eintrag ausgewiesen und damit nachprüfbar. Tabelle 5.1 führt die fünf Arten mit ihrer jeweiligen Wirkung auf.

Tabelle 5.1: Arten der Eingaben und ihre Wirkung auf die Entwicklung

Art	Anzahl	Wirkung
Beobachtung einer Fehlfunktion	16	Trug am weitesten, sobald ein konkreter Fall genannt war. „Diese Löffel-Szene läuft zweimal und zeigt genau dasselbe“ führte zur Erkennung paralleler Aufnahmen derselben Passage
Fachwissen aus der Drehsituation	8	Lieferte Regeln, die aus dem Material nicht ableitbar waren, etwa: stumme Einstellungen nur aus dem Ordner derselben Szene (Abschnitt 5.5.4)
Überlegung vor der Umsetzung	8	„Lass uns das erst überlegen“ verhinderte mehrfach, in eine falsche Richtung zu bauen
Angehängter Kontext	3	Drehbuch und Architekturregeln als Datei statt als Beschreibung
Unscharfer Wunsch	4	Trug am wenigsten. „Ich hätte es gern feiner“ und „geh an die Grenzen“ nennen kein Kriterium, an dem sich ein Ergebnis prüfen ließe, und führten zu beliebigen Änderungen

Der Befund lässt sich knapp fassen: Wirksam waren Eingaben, die einen *Mangel* oder eine *Randbedingung* benannten, unwirksam solche, die eine *Verbesserung* wünschten. Das entspricht dem in Abschnitt 4.2 beschriebenen Arbeitszyklus, in dem ein Mangel nicht am Missfallen erkannt wird, sondern an der Begründung, die ihn nicht deckt. Die vollständige Liste der Eingaben liegt in deutscher Übersetzung mit dieser Zuordnung bei, daneben der unveränderte Wortlaut und der gesamte Sitzungsverlauf (Anhang B).

Das wiederkehrende Problem war nicht fehlerhafter, sondern *plausibler* Code: Eine Änderung lief fehlerfrei durch, wirkte richtig und war es nicht. Der in Abschnitt 5.5.4 geschilderte Einschub ohne Anlass im Drehbuch ist das deutlichste Beispiel. Jede Änderung

wurde deshalb an einem erzeugten Artefakt geprüft, an einer Timeline, einem Prüfbericht, einem Messwert, und nicht daran, ob sie fehlerfrei durchlief.

Vier Einschränkungen sind zu nennen. Die Aufzeichnung umfasst zwei Tage, frühere Abschnitte der Entwicklung liegen nicht in dieser Form vor. Die gezählten Aufrufe schließen fehlgeschlagene Versuche ein und sind deshalb kein Maß für Ergiebigkeit. Der Verlauf wurde einmal automatisch zusammengefasst, sodass ein Teil des Zwischenstands nur verdichtet vorliegt. Und die Verantwortung für die Integration bleibt in jedem Fall beim Entwickler, weil eine plausibel aussehende Lösung nicht zur tatsächlichen Laufzeitumgebung passen muss.

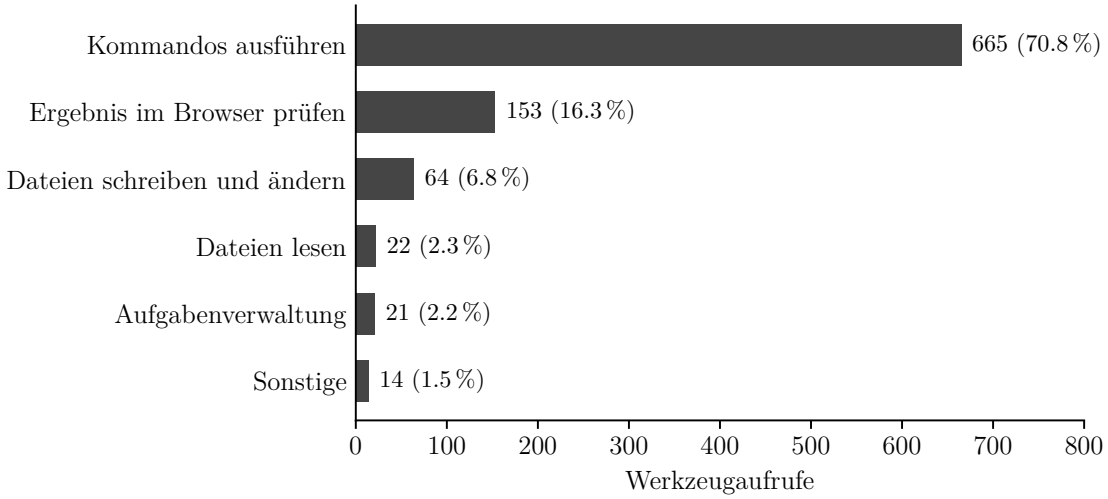

Abbildung 5.7: Verteilung der 939 Werkzeugaufrufe der aufgezeichneten Arbeitsphase nach Art, ausgezählt aus dem Sitzungsjournal (Anhang B). „Ergebnis im Browser prüfen“ bezeichnet das Kontrollieren der laufenden Anwendung nach einer Änderung.

6. Evaluation

6.1 Versuchsaufbau

Die Versuche verwenden dokumentierte Clips, Prompts und Konfigurationen. Festgehalten werden Hardware, Softwareversionen, Modelle, Embedding-Dimension, Stil, Schwellenwerte, LLM-Status, Laufzeit, Szenenanzahl, Fehler und erzeugte Artefakte. Die deterministische Grundkonfiguration wird bei identischer Eingabe wiederholt.

6.2 Quantitative Ergebnisse der Ablationsstudie

6.2.1 Durchführung und Datengrundlage

Achtzehn Anfragen wurden über vier Verfahren geführt, was 72 Einzelläufe ergibt. Die Anfragen verteilen sich auf vier Profile: fünf reine Bildstrecken, fünf Interviewpassagen, fünf gemischte Formen und drei Grenzfälle. Die vorgegebenen Zieldauern reichen von 8 bis 120 Sekunden.

Der Bestand umfasst zwölf Szenen aus elf Dateien (Zusammensetzung: Abschnitt 4.5.3, Anfragen: Anhang A). Diese Zahl begrenzt die Aussagekraft aller folgenden Werte erheblich: Bei langen Zieldauern bleibt kaum Auswahlspielraum, die Ergebnisse zeigen das Verhalten der Verfahren *unter knapper Materiallage* und sind auf große Bestände nicht ohne Weiteres übertragbar. Eine zweite Einschränkung betrifft die Vergleichbarkeit: Die Zerlegung durch das Sprachmodell ist nicht vollständig deterministisch, erkennbar daran, dass die Zahl der Fälle mit Sprecherbedingung neun gegenüber acht beträgt. Die Paarung ist weitgehend, aber nicht exakt.

6.2.2 Gültigkeitsbereich der einzelnen Maße

Vor der Auswertung ist festzulegen, welche Vergleiche die Maße überhaupt tragen. Die Einschränkungen aus Abschnitt 4.5 bestätigen sich, und eine dritte kommt hinzu: Die

Zufallsauswahl weist eine semantische Abdeckung von exakt 0,000 auf, dieser Wert ist nicht gemessen, sondern nicht messbar, weil sie keine Zerlegung erzeugt. Gleiches gilt für den mittleren Rangwert der beiden Verfahren ohne Abruf.

Tabelle 6.1: Zulässiger Vergleichsbereich der erhobenen Maße

Maß	Zulässiger Vergleich	Grund der Begrenzung
Bedingungstreue (Einstellungsgröße, Sprecher, Dialog)	nur vollständiges Verfahren gegen Variante ohne Filter	Verfahren ohne Bedingung erfüllen sie rechnerisch vollständig
Semantische Abdeckung	alle außer Zufallsauswahl	ohne Zerlegung kein Bezugstext
Mittlerer Rangwert	nur die beiden abrufenden Verfahren	ohne Abruf kein Rangwert
Vielfalt, Segmentzahl, Laufzeit	alle vier Verfahren	am Ergebnis unmittelbar bestimmbar
Zeittreue	keiner als Qualitätsaussage	Zufallsauswahl füllt konstruktionsbedingt exakt auf

6.2.3 Einhaltung der inhaltlichen Bedingungen

Das deutlichste Ergebnis betrifft die Bedingungen an Sprecher und Dialog. Wo die Zerlegung einen bestimmten Sprecher verlangt, trifft das vollständige Verfahren diese Bedingung in allen Fällen. Ohne die harten Filter wird sie in keinem Fall getroffen. Für die Bedingung, dass tatsächlich gesprochen wird, gilt dasselbe.

Tabelle 6.2: Vollständiges Verfahren gegen Variante ohne harte Filter (Mittelwerte über 18 Anfragen)

Maß	vollständig	ohne Filter	Differenz
Sprecherbedingung erfüllt	1,000	0,000	$+1,000$
Dialogbedingung erfüllt	1,000	0,000	$+1,000$
Einstellungsgröße erfüllt	0,528	0,381	$+0,147$
Vielfalt der Quellen	0,723	0,376	$+0,347$
Semantische Abdeckung	0,180	0,325	$-0,145$
Mittlerer Rangwert	0,325	0,467	$-0,142$
Laufzeit in Sekunden	93,2	92,2	$+1,0$

Der Unterschied von 1,000 gegenüber 0,000 beruht nicht auf einer besseren Bewertung, sondern auf einem Ausschluss: Das vollständige Verfahren entfernt Szenen, die die Bedingung verletzen, vor der Bewertung aus der Kandidatenmenge. Die ähnlichste Szene enthält in diesem Bestand offenbar regelmäßig nicht die geforderte Person. Gemessen wird die Wirksamkeit eines Ausschlusses, nicht eine graduelle Verbesserung. Bei der Einstellungsgröße fällt der Vorsprung mit 0,147 geringer aus, weil bei reinen Bildstrecken selten eine bestimmte Größe verlangt wird (0,900 gegenüber 0,917). Er entsteht in den gemischten Formen (0,533 gegenüber 0,217) und bei Interviews (0,200 gegenüber 0,050).

6.2.4 Vielfalt des verwendeten Materials

Die Vermeidung von Wiederholungen wirkt durchgängig: Die Vielfalt der Quellen steigt von 0,376 auf 0,723, bei reinen Bildstrecken von 0,410 auf 0,933. Ohne diesen Schritt greift das Verfahren wiederholt auf dieselbe Aufnahme zurück, weil diese für mehrere Abschnitte zugleich die höchste Ähnlichkeit aufweist, und für eine Schnittfolge ist das unbrauchbar.

6.2.5 Semantische Nähe und der auftretende Zielkonflikt

Bei zwei Maßen liegt das vollständige Verfahren zurück: Die semantische Abdeckung fällt von 0,325 auf 0,180, der mittlere Rangwert von 0,467 auf 0,325, die unmittelbare Kehrseite des Ausschlusses. Der Rückstand konzentriert sich fast vollständig auf die Interviewpassagen (0,355 auf 0,091), also gerade dort, wo die Bedingungen an Sprecher und Dialog greifen. Bei reinen Bildstrecken ist der Abstand deutlich kleiner (0,337 gegenüber 0,218). Die Ursache: Scheidet die dem Text ähnlichste Aufnahme aus, weil sie die geforderte Person nicht zeigt, wählt das Verfahren die ähnlichste *unter den zulässigen*, und dieser Wert liegt naturgemäß niedriger.

Für die Interviewpassagen bezeichnet das keinen Mangel: Eine Einstellung, die den Sprechenden nicht zeigt, ist an dieser Stelle fachlich falsch, auch wenn ein Ähnlichkeitsmaß sie bevorzugt. Für die übrigen Profile bleibt ein tatsächlicher Verlust an inhaltlicher Nähe bestehen. Festzuhalten ist ein Zielkonflikt zwischen der Einhaltung von Bedingungen und der inhaltlichen Nähe: Beides zugleich zu maximieren ist bei knappem Bestand nicht möglich, und welche Seite den Vorrang erhält, ist eine gestalterische Entscheidung, weshalb die Möglichkeit, einen Vorschlag im Editor zu verwerfen oder zu ersetzen (Kapitel 5), hier ihre Berechtigung hat.

Die Laufzeiten der Verfahren werden zusammenhängend in Abschnitt 6.5 behandelt. Abbildung 6.1 stellt die vier Maße für beide Verfahren gegenüber.

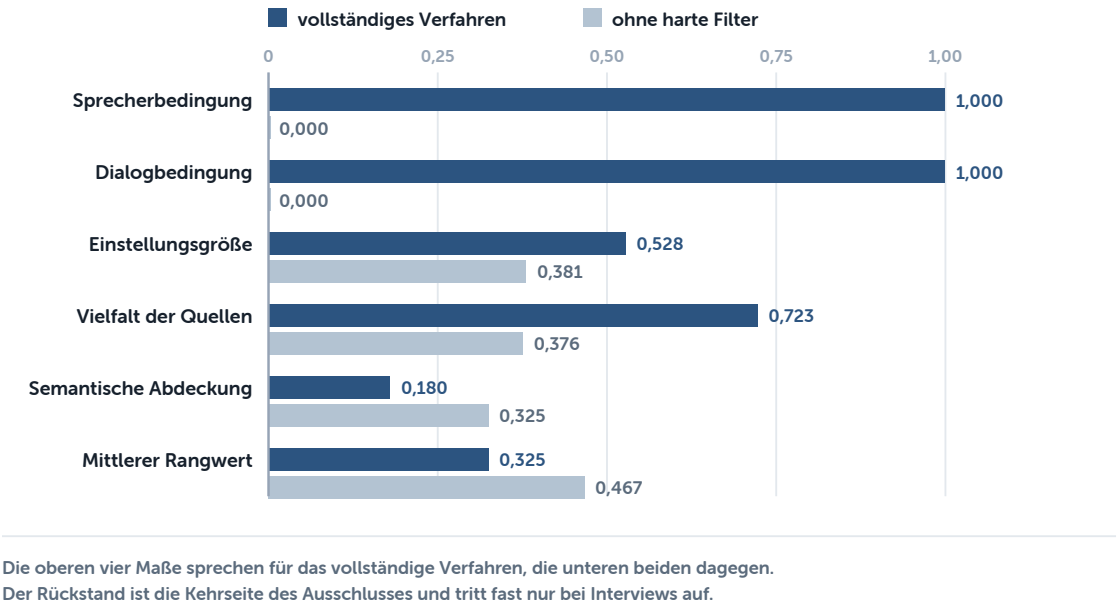

Abbildung 6.1: Vollständiges Verfahren gegenüber der Variante ohne harte Filter, Mittelwerte über 18 Anfragen.

6.2.6 Zusammenfassung der quantitativen Befunde

Drei Aussagen lassen sich aus dem Durchlauf begründen. Erstens setzen die harten Filter die Bedingungen an Sprecher und Dialog vollständig durch, während sie ohne diese Filter durchgängig verletzt werden. Zweitens verhindert die Vermeidung von Wiederholungen, dass dieselbe Aufnahme mehrfach herangezogen wird. Die Vielfalt der Quellen steigt um 0,347. Drittens geht beides zulasten der gemessenen inhaltlichen Nähe, und zwar konzentriert in den Interviewpassagen.

Nicht begründen lässt sich aus diesen Zahlen, ob die erzeugten Abfolgen als Schnitt überzeugen. Dazu treffen die erhobenen Maße keine Aussage. Das ist Gegenstand der folgenden Abschnitte.

6.3 Zuordnung von Bild und Ton

Die in Abschnitt 5.3 beschriebene Kaskade ist der einzige Teil des Systems, für den eine von ihm unabhängige Wahrheit vorliegt: Welche Tonaufnahme zu welcher Bildaufnahme gehört, ist aus der Drehsituation bekannt. Die folgenden Werte stammen aus einem Prüflauf über den Referenzbestand.

Erfassung. Von 58 Bilddateien tragen 55 einen im Ton eingebetteten Zeitcode, drei keinen. Alle 58 tragen zusätzlich einen im Container hinterlegten Zeitcode mit durchgehend demselben Wert. Da 58 Aufnahmen nicht zur selben Uhrzeit entstanden sein können, wurde er als Artefakt verworfen. Die 30 Tonaufnahmen tragen sämtlich einen Zeitstempel im Dateikopf. Ein zweiter Import derselben Ordner erzeugte keine Dubletten.

Tabelle 6.3: Zuordnung von 58 Bild- zu 30 Tonaufnahmen

Einstufung	Anzahl	Bedeutung
sicher	27	Zeitcode beider Seiten deckt sich
plausibel	0	Stufen widersprechen einander
unklar	2	mehrere Aufnahmen kommen in Betracht
ohne Gegenstück	33	keine Zuordnung gefunden

Ergebnis der Zuordnung. Für eine Einstellung mit fünf Aufnahmen liegen die Werte im Einzelnen vor: alle fünf über den Zeitcode zugeordnet, Konfidenz mindestens 0,986, Versatz zwischen −2,24 und −3,12 Sekunden. Zugleich meldete das System bei allen fünf eine Warnung, weil die Take-Nummern von Bild und Ton um eins verschoben sind: Die Zuordnung wurde getroffen *und* der Widerspruch benannt. Zwei aufeinanderfolgende Läufe über denselben Bestand lieferten byte-identische Ergebnisse, zu erwarten, weil kein Modell beteiligt ist.

Wirkung auf die weitere Verarbeitung. Dass die Zuordnung nicht nur formal stimmt, zeigt der nachgelagerte Schritt: Auf der Kameraspur einer Aufnahme lag der Pegel bei etwa −97 dB, die Spracherkennung ergab dort die Ausgabe „Thank you.“, ein bekanntes Verhalten bei nahezu stummem Eingang. Nach der Zuordnung (Versatz −2,7 Sekunden, Pegel −38 dB) lieferte dieselbe Aufnahme den tatsächlich gesprochenen Text mit Klappenansage. Die Zuordnung ist damit Voraussetzung: Ohne sie arbeitet die gesamte sprachbezogene Auswertung auf einem Signal, das keine Sprache enthält.

Nicht wirksame Stufen der Kaskade. Zwei ihrer vier Stufen tragen im vorliegenden Bestand nichts bei: Die Wellenform-Prüfung setzt hörbaren Kameraton voraus und meldet sonst „nicht anwendbar“. Die Klappen-Auswertung benötigt Szenenbeschreibungen, die beim Import noch nicht vorliegen. Der Dateiname entscheidet grundsätzlich nicht allein, auf diesem Bestand widersprechen sich Datei- und Ton-Nummerierung mehrfach.

Die 33 Aufnahmen ohne Gegenstück sind überwiegend Bildaufnahmen ohne zugehörige Tonaufnahme und werden als solche geführt, nicht mit einer schwachen Zuordnung versehen.

An drei Stellen wich das Ergebnis von der vorliegenden Aufstellung des Materials ab, etwa weil sich zwei dort als „ohne Gegenstück“ geführte Aufnahmen über den Zeitcode eindeutig zuordnen ließen. In allen Fällen wurde den Daten und nicht der Aufstellung gefolgt und die Abweichung vermerkt. Die Metadaten zeigen, dass ein Teil der Tonaufnahmen bereits am Aufnahmegerät umbenannt worden war, was die Abweichungen erklärt und begründet, warum der Dateiname als alleinige Quelle ungeeignet ist.

Reichweite der Ergebnisse. Die Werte gelten für einen Bestand und eine Drehsituation mit durchgehend synchronisierten Zeitcode-Geräten. Wo diese Voraussetzung fehlt, trägt die erste Stufe nicht. Die Zahlen sind daher kein allgemeiner Gütewert, sondern der Nachweis, dass das Verfahren unter den dokumentierten Bedingungen zutreffend und wiederholbar arbeitet. Nach dem Prüflauf wurde der Bestand um einen zweiten Ordner mit Tonaufnahmen erweitert (seither 41 sichere und zwei plausible Zuordnungen bei 29 ohne Gegenstück). Die berichteten Werte beziehen sich auf den dokumentierten Lauf.

6.4 Abdeckung des Drehbuchs im erzeugten Schnitt

Neben der Ablationsstudie liegt ein zweiter Befund vor: die Auswertung des Prüfberichts zum jüngsten erzeugten Schnitt: fünf Szenen des Drehbuchs, 47 Einträge, Gesamtdauer 12:49 Minuten (die Länge der Master auf der ersten Spur. Die übrigen Einträge sind Auflagen und Geschwister darüber, Abschnitt 5.5.6). Sie beantwortet die Frage, wie viel des Drehbuchs im Ergebnis wiederzufinden ist.

Dialogzeilen. Von den 24 Dialogzeilen des Drehbuchs sind 22 mit einer konkreten Aufnahme und einem konkreten Zeitbereich verbunden. Die beiden verbleibenden wurden in keiner Aufnahme gefunden und werden mit ihrer Ursache ausgewiesen. Diese Lücken sind kein Fehler des Verfahrens: Eine Zeile, die nicht gedreht wurde, kann nicht geschnitten werden, und ein System, das an dieser Stelle irgendeine Aufnahme einsetzte, wäre nicht besser, sondern unehrlicher.

Drehbuch-Handlungen. Für die 28 Handlungsanweisungen des Drehbuchs weist der Bericht drei Zustände aus. Die Verteilung ist ungleich über die Szenen.

Tabelle 6.4: Abgleich der Drehbuch-Handlungen mit dem Bildmaterial

Szene	gedreht	unsicher	fehlt	gesamt
1 Wohnzimmer (stumme Inserts)	1	0	0	1
2 Wohnzimmer	4	1	1	6
3 Korridor	3	4	1	8
4 Wohnzimmer	4	0	0	4
5 Korridor	3	0	6	9
Gesamt	15	5	8	28

Gut die Hälfte der Handlungen ist im Bild belegt, knapp ein Fünftel bleibt unsicher, gut ein Viertel wird nicht gefunden.

Verteilung der unsicheren Fälle. Aufschlussreich ist weniger die Gesamtzahl als ihre Verteilung: Vier der fünf unsicheren Fälle entfallen auf Szene 3. Deren bestätigte Handlungen beschreiben Körperhaltungen (eine Gitarre abstellen, an der Wand lehnen), die unsicheren dagegen einen kleinen Beutel in der Hand, einen Mülleimer am Ende des Korridors und innere Zustände (unbeteiligt wirken, besorgt schauen). Die Unsicherheit trifft damit genau die beiden Kategorien, die Abschnitt 5.5.4 vorab als unzuverlässig ausgewiesen hat: kleine Gegenstände und nicht sichtbare Absichten. Für die Beurteilung ist das günstig, weil die unsicheren Fälle vorhersagbar sind und gezielt von Hand geprüft werden können. Szene 5 verhält sich anders, dort sind sechs Handlungen *nicht gefunden* ohne unsichere Fälle: Das deutet nicht auf ein Erkennungsproblem, sondern darauf, dass diese Handlungen nicht gedreht wurden.

Grenzen dieses Befundes. Drei Einschränkungen: Erstens beruhen die Werte auf einem einzigen Drehbuch und Schnitt und begründen keine allgemeine Trefferquote. Zweitens, und das wiegt schwerer, geben die drei Zustände die *Einschätzung des Systems* wieder, nicht eine geprüfte Wahrheit. Ob eine als „gedreht“ ausgewiesene Handlung tatsächlich zu sehen ist, wurde nicht unabhängig überprüft (der Bericht sieht dafür eine Ankreuzspalte vor). Drittens beanspruchte die Auswertung aller fünf Szenen rund drei Stunden Rechenzeit. Die Ergebnisse werden deshalb zwischengespeichert.

6.5 Zeitbedarf

Für ein Werkzeug, das während der Sichtung eingesetzt werden soll, ist der Zeitbedarf keine Nebengröße. Er zerfällt in zwei Teile, die getrennt zu betrachten sind: eine einmalige Aufbereitung des Materials und die Erzeugung eines einzelnen Vorschlags.

Erzeugung eines Vorschlags. Der vollständige Durchlauf einer Anfrage dauerte im Median 94 Sekunden, im günstigsten Fall 18 und im ungünstigsten 146 Sekunden. Die Aufschlüsselung nach Phasen zeigt eine deutliche Verteilung.

Tabelle 6.5: Zeitbedarf je Anfrage nach Phasen, Mittelwerte über 18 Anfragen in Sekunden

Verfahren	Zerlegung	Abruf	gesamt
vollständiges Verfahren	92,5	0,7	93,2
ohne harte Filter	91,8	0,4	92,2
ohne Zerlegung		35,7 (ein Aufruf)	35,7
Zufallsauswahl	0,0	0,0	0,02

Nahezu die gesamte Zeit entfällt auf den Aufruf des Sprachmodells, das die Anfrage zerlegt, während der Abruf der Kandidaten weniger als eine Sekunde beansprucht. Zwei Schlüsse: Die harten Filter, deren Wirkung Abschnitt 6.2 nachweist, kosten praktisch keine Zeit: Die Verbesserung ist nicht mit einem Laufzeitnachteil erkauft. Und der Mehraufwand der dreistufigen Abfolge liegt allein in der Zerlegung. Wer die Laufzeit senken will, muss dort ansetzen, nicht an der Suche. Der gesamte Vergleichslauf mit 72 Durchgängen beanspruchte 66 Minuten. Nach Profilen reicht die Laufzeit von 70,0 Sekunden bei reinen Bildstrecken bis 123,4 Sekunden bei gemischten Formen. Alle Werte gelten für die in Anhang A dokumentierte Hardware.

Einmalige Aufbereitung. Der Bestand umfasst 58 Aufnahmen mit zusammen 82,7 Minuten Material. Dessen Aufbereitung (Bildwandlung, Transkription, Szenenzerlegung, Bildbeschreibung, Merkmalsberechnung) ist der weitaus größere Posten, fällt aber je Aufnahme nur einmal an und wird zwischengespeichert. Der Drehbuch-Bild-Abgleich schlägt mit rund einer halben Stunde je Durchlauf zu Buche. Die Angaben sind eine obere Schranke: Sie stammen aus den Zeitstempeln der nacheinander abgearbeiteten Aufträge und trennen nicht zwischen Rechen- und Wartezeit. Für die Einordnung ist die Verteilung wichtiger als der Betrag: Die aufwendigen Schritte fallen einmal an, die Erzeugung eines Vorschlags bleibt bei ein bis zwei Minuten. Ein Wechsel der Formulierung ist ohne erneute

Aufbereitung möglich und damit Voraussetzung dafür, das Werkzeug ausprobierend zu verwenden.

6.6 Qualitative Eigenbewertung

Die bisherigen Abschnitte berichten, was sich zählen lässt. Ob die erzeugte Abfolge als Schnitt trägt, entscheidet sich beim Ansehen. Grundlage dieses Abschnitts ist die vollständige Sichtung des Rohschnitts in der Fassung 29, Länge 12:49 min, getrennt nach Spuren und entlang der drei in Abschnitt 4.6 festgelegten Kriterien. Die Beurteilung stammt vom Verfasser, also von derselben Person, die das System gebaut hat. Sie ist eine begründete Einzelmeinung und ersetzt keine unabhängige Fremdbewertung. Damit sie überprüfbar bleibt, liegen die Spuren den Materialien als einzelne Videodateien gleicher Länge bei, dazu die Schnittliste mit Zeitpunkt, Einstellung und Take je Clip.

6.6.1 Aufbau des Vorschlags

Der Vorschlag ist kein flacher Schnitt, sondern ein Stapel aus fünf Spuren (Tabelle 6.6). Die Hauptspur trägt je Szene die Ankereinstellung als durchlaufenden Master. Darüber liegen vier Spuren mit Alternativen, wobei jede Spur genau eine Einstellung führt und diese über die Szene hinweg beibehält. Spur V4 ist ein Sonderfall, sie führt dieselbe Einstellung wie der Master, aber andere Takes. Der Stapel trennt damit die beiden Fragen, die im Schnitt verschieden sind: eine andere Perspektive auf denselben Moment und derselbe Blick in einer anderen Ausführung.

Tabelle 6.6: Spuren des Rohschnitts in der Fassung 29

Spur	Rolle	Clips	Material	Einstellungen je Szene
V1	Master, durchlaufend	12	769,5 s	1.1 · 2.1 · 3.1 · 4.1 · 5.1.1 und 5.2.1
V2	andere Einstellung	8	216,4 s	2.2 · 3.2 · 4.3 · 5.2.3
V3	andere Einstellung	3	76,7 s	2.4 · 4.2 · 5.2
V4	andere Takes derselben Einstellung	21	177,3 s	2.1 · 3.1 · 4.1 · 5.1.1
V5	andere Einstellung	3	21,2 s	4.4 · 5.2.1

6.6.2 Hauptspur

Die strukturelle Schnittlogik trägt. Die Reihenfolge folgt dem Drehbuch, jede Szene erscheint genau einmal, und keine Szene wird wiederholt. Klappen sind nirgends im Schnitt sichtbar oder hörbar. Szene 1 ist als Folge von sieben kurzen Inserts aus derselben Einstellung aufgebaut und liest sich als Auftakt, die übrigen Szenen laufen je als ein Master durch. Auch die sprachliche Kontinuität ist gewahrt, der Ton sitzt durchgehend am zugehörigen Bild, eine falsch platzierte Tonstelle wurde nicht beobachtet.

Ein Mangel tritt wiederholt auf. Das Ende einzelner Einstellungen läuft zu lang, bis das Kommando der Regie zu hören ist. Die Ursache ist im Verfahren auffindbar. Der Ausstieg wird aus dem Transkript bestimmt: Erkennt die Spracherkennung nach der letzten Spielreplik ein Produktionskommando, endet die Einstellung drei Zehntelsekunden davor, andernfalls greift der Rückfall auf das Ende der Aufnahme abzüglich einer halben Sekunde. Bei drei der fünf Szenen-Master beträgt der Abstand zum Ende der Aufnahme genau diese halbe Sekunde, dort hat also der Rückfall gegriffen. In den beiden übrigen Fällen wurde das Kommando erkannt und der Ausstieg 3,8 s beziehungsweise 20,7 s vor dem Ende der Aufnahme gesetzt. Der Schutz vor Produktionssprech wirkt am Anfang der Aufnahme zuverlässig und am Ende nur dann, wenn das Kommando laut genug gesprochen wurde, um transkribiert zu werden. Die Beobachtung des Auges und der Befund im Datenbestand decken sich hier, und die Korrektur ist ein Nachziehen des Ausstiegspunkts, keine Umarbeitung.

Wie weit dieser Nachlauf reichen kann, zeigt das zweite Glied von Szene 5. Nach der letzten Replik laufen dort 120,8 s weiter, bei den übrigen Szenen-Mastern sind es 1,2 bis 25,6 s. In diesen zwei Minuten tritt die Regie ins Bild und führt eine Darstellerin heraus, weil die Kamera für die spätere Bildbearbeitung weiterlief und die Szene danach erneut angesetzt wurde (Abbildung 6.2). Kein gesprochenes Wort weist diese Unterbrechung aus, sie verläuft stumm. Die Regel, die den Ausstieg setzt, befragt Sprache und Bewegung, nicht den Bildinhalt, und kann eine stumme Unterbrechung deshalb nicht erkennen.

Ein Ansatz zur Abhilfe liegt im System bereits vor: Es gruppiert Gesichter über die Aufnahmen hinweg und ordnet die Gruppen den Figuren des Drehbuchs zu (Abschnitt 5.5.5). Eine Gesichtsgruppe ohne Rolle im Drehbuch, die im Nachlauf auftaucht, wäre ein brauchbares Warnzeichen. Verdrahtet ist das nicht. Bemerkenswert ist dabei, wie unterschiedlich schwer die beiden Wege zu diesem Befund sind. Beim Ansehen fällt die Stelle sofort auf, für das Verfahren ist sie unsichtbar, weil sie in keiner der Größen erscheint, die es erhebt.

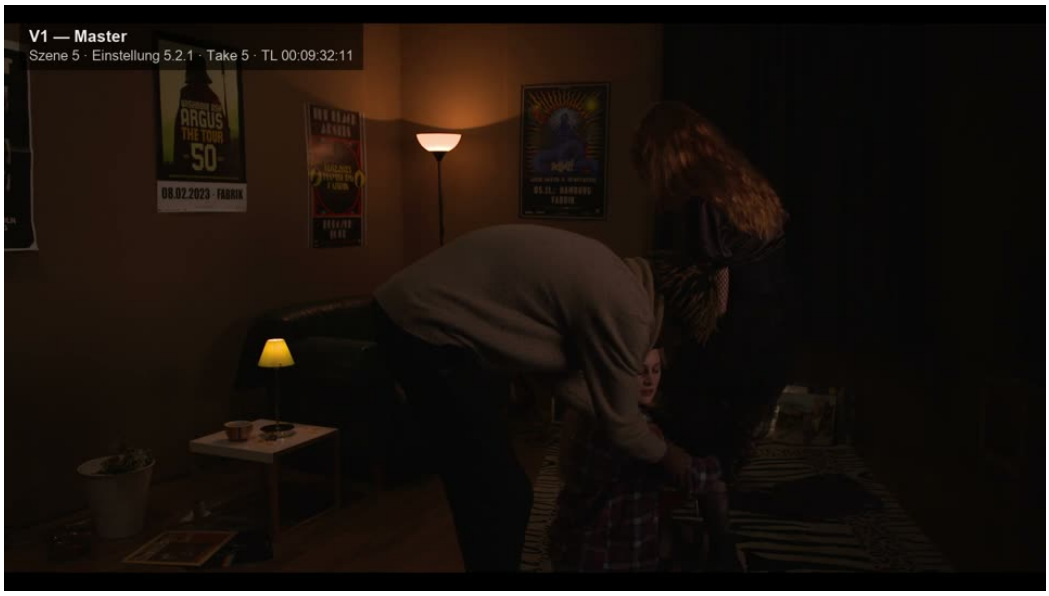

Abbildung 6.2: Nachlauf im zweiten Glied von Szene 5, rund 37 s nach der letzten Replik. Die Regie steht im Bild und führt eine Darstellerin heraus, bevor die Szene erneut angesetzt wird. Die eingeblendete Kennung stammt aus dem Spur-Export und nennt Spur, Szene, Einstellung, Take sowie den Beginn des Clips in der Timeline.

6.6.3 Alternativspuren

Die Alternativen liegen an der richtigen Stelle über dem Master, in der Reihenfolge der Takes und beim richtigen Moment der Handlung. Genau ist die Platzierung nicht. Der Grund ist beim Sichten erkennbar: Die Darstellenden spielen von Take zu Take unterschiedlich und sprechen nicht dieselben Worte, weshalb ein rein textlicher Anker keine exakte Deckung herstellen kann. Die Zuordnung stützt sich dann auf Bild und Handlung, und diese setzt den Ausschnitt auf den zutreffenden Beat, nicht auf das zutreffende Einzelbild. Für die drei näher betrachteten Spuren gilt derselbe Befund mit denselben Abweichungen.

Praktisch ist das brauchbar. Was eine Alternativspur liefert, ist die Vorauswahl, also welcher Take, welche Einstellung und welche Stelle in Frage kommen. Das Feinstellen des Ein- und Ausstiegs bleibt Handarbeit, hält sich aber in Grenzen, weil der Ausschnitt bereits am richtigen Beat liegt. Gemessen an dem, was diese Suche von Hand kostet, ist der verbleibende Aufwand gering.

6.6.4 Einordnung

Nach den drei Kriterien fällt die Beurteilung unterschiedlich aus. Die strukturelle Schnittlogik ist auf der Hauptspur erfüllt, mit dem benannten Vorbehalt am Ende einzelner Einstellungen. Die sprachliche Kontinuität ist dort ebenfalls erfüllt, auf

den Alternativspuren nur auf der Ebene des Beats, nicht auf der Ebene der Silbe. Die Bildabfolge liest zusammenhängend, die Wahl der Ankereinstellung je Szene ist nachvollziehbar, und ein störender Bildsprung wurde nicht beobachtet. Nicht beantworten kann diese Sichtung, ob die Sequenz auf Dritte wirkt. Dafür fehlt weiterhin eine unabhängige Beurteilung.

6.7 Technische Testfälle und offene Messungen

Die in Kapitel 4 vorgesehenen technischen Testfälle wurden am laufenden System durchgeführt, ergänzt um vier Fälle zu Rückführbarkeit, Wiederherstellbarkeit, Antwortzeit und Spur-Export. Tabelle 6.7 fasst Vorgehen und Ergebnis zusammen, das Testprotokoll liegt bei.

Tabelle 6.7: Technische Testfälle und ihre Ergebnisse

Test	Vorgehen	Ergebnis
T-1 Ingestion	Zählung der Analyse-Artefakte in der Projektdatenbank	58 von 58 Aufnahmen mit Vorschaubildern. 66 von 66 Szenen mit Beschreibung, Embedding und Einstellungsgröße. 49 Szenen tragen eine Transkription, die übrigen 17 enthalten keine erkannte Sprache
T-2 Fehlerbehandlung	ungültiger Modus und unbekannte Plan-Kennung an die Schnittstelle	kontrollierte Fehlermeldungen mit klarem Text, kein Absturz
T-3 Determinismus	zweifache Erzeugung des Rohschnitts mit identischen Parametern (Testläufe anschließend verworfen)	beide Läufe liefern 24 Einträge, Einträge und Statistik sind byte-identisch
T-4 Hybrid-Retrieval	dieselbe Anfrage („Hände im warmen Licht“) rein visuell, rein textbasiert und kombiniert	alle drei Modi liefern geordnete Treffer, die Spitzenpositionen von Bild- und Textsuche unterscheiden sich
T-5 MP4-Export	Export der Master-Spur des jüngsten Schnitts, Prüfung mit ffprobe	abspielbare Datei, H.264 in 1920×1080, Ton AAC mit 48 kHz, Dauer 769,5 s entspricht der Planstatistik. Die Datei trägt 25 Bilder je Sekunde und weicht damit von den 24 der Quelle ab, die Laufzeit stimmt dennoch. Der für die Abgabe erstellte Spur-Export (T-10) hält die Bildrate der Quelle ein
T-6 NLE-Export	Strukturprüfung der Austauschdatei, Import in DaVinci Resolve	gültiges FCPXML 1.8 mit 52 Assets, 59 Beat-Marken und 35 deaktivierten Auflagen (Abbildung 5.3). Drei unabhängige Exporte desselben Plans sind SHA-256-identisch
T-7 Rückführbarkeit	Prüfung der Herkunftskette in der Projektdatenbank	alle 66 Szenen lösen sich auf eine Aufnahme auf, alle 58 Aufnahmen auf einen Take und dieser auf die unveränderte Originaldatei. Die in Abschnitt 5.8 beschriebene Kette ist damit lückenlos
T-8 Wiederherstellbarkeit	Einspielen des beiliegenden Datenbank-Abzugs in eine leere Datenbank	fehlerfrei eingespielt, mit 58 Aufnahmen, 66 Szenen, 49 Schnittfassungen und 57 Drehbuchzeilen im erwarteten Umfang
T-9 Antwortzeit der Suche	fünf Anfragen an die hybride Suche, Wanduhrzeit je Aufruf	0,10 bis 0,47 s, im Mittel 0,21 s. Der Abruf selbst ist damit nicht der zeitbestimmende Schritt, sondern die Zerlegung der Anfrage (Abschnitt 6.5)
T-10 Spur-Export	Ausgabe aller fünf Spuren als je eine Videodatei, Prüfung mit ffprobe	fünf abspielbare Dateien, H.264 in 1920×1080 mit 24 Bildern je Sekunde wie die Quelle, Ton AAC mit 48 kHz aus der synchronisierten Tonaufnahme. Alle fünf zählen 18 468 Bilder und sind damit bildgenau deckungsgleich (Anhang A.4)

Neben diesen Abnahmefällen, die das System als Ganzes prüfen, liegt eine automatisierte Testreihe für die einzelnen Bausteine vor: 97 Testfälle für den Zeitcode-Dekoder, das Lesen der bext- und iXML-Abschnitte, die Zuordnungskaskade, die Tonkanal-Klassifikation, den Drehbuch-Abgleich, die Bildprüfung und die Befehle der Assistenzschicht. Von ihnen laufen 81 durch, 16 werden übersprungen, weil sie den Referenzbestand auf der externen Festplatte oder ein geladenes Sprachmodell voraussetzen. Die Reihe liegt dem Quelltext bei.

Offen bleiben eine unabhängige Beurteilung durch Dritte (Anlage: Abschnitt 4.8) und die Prüfung der Analysepipeline je Schritt. Erfundene Zahlen wären auch dann unbrauchbar, wenn sie plausibel wirkten.

7. Diskussion

7.1 Deutung des gemessenen Zielkonflikts

Der auffälligste Befund der Ablationsstudie ist kein einzelner Wert, sondern ein Gegensatz: Dieselben Bestandteile, die die Bedingungen an Sprecher und Dialog durchsetzen, senken die gemessene inhaltliche Nähe zur Anfrage. Für die Bewertung des Systems ist entscheidend, wie dieser Gegensatz zu verstehen ist.

Er ist nicht die Folge einer ungeschickten Umsetzung, sondern strukturell: Jede zusätzliche Bedingung verkleinert die Auswahlmenge, der erreichbare Höchstwert kann nur sinken. Bei zwölf verfügbaren Szenen fällt das besonders ins Gewicht. Der gemessene Abstand beschreibt daher auch die Enge des Materials. Durch bessere Gewichtung lässt sich der Gegensatz nicht auflösen: Wer die Bedingungen lockert, gewinnt Nähe und verliert Verlässlichkeit, und welche Seite vorzuziehen ist, hängt vom Vorhaben ab. Ein Verfahren, das diese Abwägung selbstständig und unsichtbar träfe, wäre nicht besser, sondern unkontrollierbar. Sinnvoll ist, die Entscheidung sichtbar und umkehrbar zu halten, was der Editor durch das Verwerfen und der Prüfbericht durch die Angabe des Grundes leisten. Der gemessene Nachteil ist damit weniger ein Mangel als der Preis einer offengelegten Festlegung.

Eine Nachmessung an einem größeren Bestand. Die eben genannte Vermutung, der gemessene Abstand beschreibe auch die Enge des Materials, lässt sich prüfen. Dazu wurde der Vergleichslauf wiederholt, mit denselben 18 Anfragen, denselben vier Verfahren und denselben Maßen, aber auf einem anderen Bestand: 63 statt zwölf Szenen, und ausschließlich Material des Kurzfilms, ohne die ergänzenden Aufnahmen aus Abschnitt 4.5.3. Tabelle 7.1 stellt beide Läufe gegenüber.

Der Gegensatz kehrt sich um: Auf dem größeren Bestand liegt das vollständige Verfahren in der semantischen Abdeckung knapp *vor* der Variante ohne Filter, die Bedingungen werden weiterhin vollständig eingehalten, der Vorsprung bei der Einstellungsgröße wächst von 0,147 auf 0,565, die Abweichung von der Dauer sinkt von 0,427 auf 0,007. Die Vermutung ist damit bestätigt: Der berichtete Abstand beschreibt zu einem erheblichen Teil die Enge

Tabelle 7.1: Vergleichslauf auf zwölf gemischten Szenen und auf 63 Szenen des Kurzfilms

	vollständiges Verfahren		ohne harte Filter	
Maß	12 Szenen	63 Szenen	12 Szenen	63 Szenen
Semantische Abdeckung	0,180	0,228	0,325	0,223
Sprecher-Treue	1,000	1,000	0,000	0,655
Dialog-Treue	1,000	1,000	0,000	0,613
Einstellungsgröße	0,528	0,717	0,381	0,152
Quellenvielfalt	0,723	0,944	0,376	0,609
Abweichung der Dauer	0,427	0,007	0,438	0,275

des Materials. Die strukturelle Überlegung bleibt gültig, fällt aber nicht mehr ins Gewicht, sobald mehrere Aufnahmen zugleich Bedingung und Inhalt erfüllen.

Drei Einschränkungen: Es änderten sich Umfang *und* Zusammensetzung des Bestands zugleich (eine einzige Produktion ist einheitlicher als ein gemischter Bestand). Die Variante ohne Filter erreicht die Bedingungen nun in 65 bzw. 61 Prozent der Fälle, sodass der Beitrag der Filter kleiner erscheint. Und drei der 72 Durchläufe brachen ohne Messwerte ab. Für die Arbeit bleibt es bei den in Kapitel 6 berichteten Zahlen. Die Nachmessung tritt daneben und beantwortet eine Frage, die die Hauptmessung offenlassen musste.

7.2 Beitrag des Prüfberichts

Kapitel 3 hat die Forschungslücke nicht in der Erzeugbarkeit eines Schnittvorschlags verortet, sondern in seiner Prüfbarkeit. Kapitel 5 beschreibt das Mittel dazu. Ob dieses Mittel etwas taugt, lässt sich an einem Vorgang während der Entwicklung zeigen.

Beim Gegenlesen eines erzeugten Schnitts fiel eine kurze Detailaufnahme einer Hand auf, die in einer Dialogpause stand und für die sich im Drehbuch kein Anlass fand. Der Bericht wies als Begründung lediglich die Länge der Sprechpause aus. Damit war der Mangel benennbar: Das Bildmaterial war nie gegen das Drehbuch geprüft worden. Die Folge war keine Korrektur am Einzelfall, sondern eine Änderung der Regel: die Bildprüfung aus Abschnitt 5.5.4 und die Neuordnung stummer Aufnahmen nach bestätigten Handlungen. Das ist eine einzelne Beobachtung, kein Nachweis der Nützlichkeit. Bemerkenswert bleibt, dass ein System ohne diese Offenlegung sie gar nicht ermöglicht hätte. Der Befund aus Abschnitt 6.4 stützt denselben Punkt: Dass die unsicheren Fälle genau die vorab benannten Kategorien betreffen, ist nur feststellbar, weil die Einschätzung je Handlung mit ihrem Beleg ausgewiesen wird. Eine bloße Gesamtquote hätte den Zusammenhang verdeckt.

7.3 Grenzen der eingesetzten offenen Modelle

Die Entscheidung, ausschließlich frei verfügbare Modelle auf dem eigenen Rechner zu verwenden, ist für den Gegenstand dieser Arbeit begründet: Rohmaterial einer Produktion soll den Rechner nicht verlassen müssen, und ein Verfahren, das sich prüfen lassen soll, darf nicht von einem Dienst abhängen, dessen Verhalten sich jederzeit ändern kann.

Diese Entscheidung hat einen Preis, der sich an den Messungen dieser Arbeit ablesen lässt. Vier Beobachtungen sind zu nennen.

Bildverstehen bleibt an Körperhandlungen gebunden. Das eingesetzte Bildmodell beantwortet Fragen nach Haltungen und Tätigkeiten zuverlässig, versagt aber bei kleinen Gegenständen. Diese Grenze war vorab gemessen und wurde durch die Auswertung bestätigt: Fünf der sechs unsicheren Fälle betrafen genau diese Kategorie (Abschnitt 6.4). Die Fragen mussten deshalb von vornherein auf das eingeschränkt werden, was das Modell leisten kann, die Prüfung richtet sich nach dem Werkzeug und nicht allein nach der Sache.

Die Spracherkennung meldet Sprache, wo keine ist. Auf einer nahezu stummen Spur gab das Modell den Satz „Thank you.“ aus, statt das Fehlen von Sprache anzuzeigen. Dass dieser Fehler auffiel, lag an der Zuordnung des richtigen Tons. Ohne sie wäre er als gültiges Transkript in die weitere Verarbeitung eingegangen.

Die Zerlegung der Anfrage bestimmt die Laufzeit. Von 93,2 Sekunden je Anfrage entfallen 92,5 auf den Aufruf des Sprachmodells (Abschnitt 6.5). Ein größeres Modell würde die Wartezeit weiter erhöhen, ein kleineres die Zerlegung verschlechtern. Zugleich arbeitet das Modell nicht vollständig wiederholbar, was die Vergleichbarkeit der Läufe einschränkt.

Die Zuordnung von Gesichtern bleibt unvollständig. Dieselbe Person erscheint als mehrere Gruppen, kleine Gruppen bleiben ohne Namen, und einzelne Zuordnungen stützen sich auf einen negativen Übereinstimmungswert (Abschnitt 5.5.5). Die manuelle Bestätigung ist deshalb nicht Bequemlichkeit, sondern notwendiger Bestandteil.

Möglicher Vergleich mit einem kommerziellen Modell. Gemeint ist nicht ein Vergleich mit den Schnittsystemen aus Kapitel 3, der aus den in Abschnitt 7.5 genannten Gründen nicht möglich ist, sondern ein Austausch der *Modelle* im eigenen Verfahren:

dieselbe Pipeline, dieselben Daten, dieselben Maße, nur mit einem entfernt betriebenen Modell an der Stelle des lokalen.

Anders als die Frage nach der Schnittqualität wäre dieser Vergleich *zählbar*, denn an zwei Stellen liegt eine geprüfte Bezugsgröße vor. Bei der Bildprüfung (Abschnitt 5.5.4) sind es dieselben geschlossenen Fragen an dieselben Einzelbilder, deren Antworten sich gegen die im Prüfbericht ausgewiesenen Zustände halten lassen. Ein Unterschied wäre vor allem bei kleinen Gegenständen zu erwarten, bei der hier gemessenen Schwäche. Bei der Anfragezerlegung sind es die 18 festgeschriebenen Anfragen aus Anhang A, deren Slots sich in Zahl und Bedingungen unmittelbar gegenüberstellen lassen. Beides ergäbe Zahlen, keine Eindrücke.

Durchgeführt wurde der Vergleich nicht. Der Aufwand dafür ist überschaubar, aber nicht null: Modellname und Adresse des lokalen Dienstes sind an zentraler Stelle über Umgebungsvariablen einstellbar, in der Bildbeschreibung und in der Anfragezerlegung stehen sie dagegen fest im Quelltext. Wesentlich ist, wie ein Ergebnis zu lesen wäre. Fiele es zugunsten des entfernten Modells aus, spräche das nicht gegen den hier gewählten Weg, sondern bezifferte seinen Preis: Es wäre bekannt, wie viel Genauigkeit für die Unabhängigkeit von einem Dienst aufgewendet wird. Diese Zahl fehlt der Arbeit und wäre der nächste Schritt.

7.4 Übertragbarkeit auf andere Produktionen

Die Untersuchung fand an einem einzigen Film statt. Damit stellt sich die Frage, was von dem Verfahren an dieses Material gebunden ist und was nicht. Die folgende Betrachtung beantwortet sie durch Prüfung des Quellcodes. Sie ist *keine* empirische Aussage, denn ein zweiter Bestand wurde nicht verarbeitet. Sie benennt, was bei einer Übertragung zu erwarten und was vorher zu klären wäre.

Materialunabhängige Bestandteile. Der weitaus größte Teil des Verfahrens enthält keinen Bezug auf diese Produktion: die gesamte Aufbereitung, die Zuordnung von Bild und Ton über Zeitcode und Wellenform sowie der Editor. Namen aus dem Film sind nicht im Programm hinterlegt (die Eigennamen-Wortliste steht in einer Einstellungsdatei), der Drehbuch-Parser erwartet die übliche Form eines Drehbuchs, die Zerlegung einer Szene in Beats ist rein strukturell bestimmt, und der Abgleich zwischen Drehbuch und gesprochenem Text erfolgt über ein mehrsprachiges Modell. Dass ein englisches Drehbuch auf einen deutschen Dreh trifft, ist kein Sonderfall. Die Fragen der Bildprüfung werden je Projekt aus dessen Regieanweisungen erzeugt, die Zuordnung der Gesichter beruht auf

Auftrittsverteilungen statt hinterlegten Personen, und die Schnittregeln sind allgemeine Regeln der Montage.

Zwei Voraussetzungen an die Produktion. Zwei Annahmen betreffen nicht das Programm, sondern die Art, wie gedreht wurde.

Die erste und wichtigere ist die *Klappenkonvention*. Das Verfahren gewinnt die Szenenstruktur aus der gesprochenen Klappe, weil diese im vorliegenden Fall die Nummerierung des Drehbuchs trägt und ihre zweite Stelle die Einstellung bezeichnet. Wird nach anderer Übung geklappt, etwa mit fortlaufender Klappennummer statt der Drehbuchnummer, lässt sich diese Zuordnung nicht mehr aus dem Ton ableiten. Aufgefangen wird das durch die Ablage der Medien nach Szenen und durch die Möglichkeit, eine Klappe von Hand zu berichtigen. Als Voraussetzung ist es jedoch zu benennen.

Die zweite betrifft die *Ablage*: Die Zuordnung der Ordner zu Drehbuchszenen setzt voraus, dass das Material entsprechend abgelegt oder einmalig sortiert wird.

Eingestellte Werte. Die Verfahren sind allgemein, ihre Schwellen sind es nicht. Die Grenze, ab der ein gesprochener Satz als Treffer im Drehbuch gilt, liegt bei 0,55. Ab 0,66 gilt er als sicher, ab 0,72 als zweifelsfrei. Ein improvisierter Satz wird ab 0,48 einem Beat zugerechnet. Die Zerlegung eines Takes in Beats bestraft übersprungene Beats mit einem festen Wert. Ein einzelner bejahter Bildtreffer zählt nicht.

Diese Werte wurden an 58 Aufnahmen eines Films eingestellt. Bei einem anderen Bestand ist ein Nachziehen zu erwarten. Wesentlich ist, dass dieses Nachziehen *prüfbar* bleibt: Der Prüfbericht weist je Zeile den erreichten Wert aus, sodass sich zu hoch oder zu niedrig gesetzte Schwellen an den Lücken und an den unsicheren Zuordnungen ablesen lassen. Er ist damit nicht nur Ergebnis, sondern auch das Werkzeug für die Übertragung auf einen neuen Bestand.

Hinzu kommt eine sprachliche Bindung geringeren Gewichts: Die Erkennung der Klappe im Transkript kennt deutsche und englische Bezeichnungen. Ein Dreh in einer weiteren Sprache erforderte deren Aufnahme in die Erkennung.

Die strukturelle Grenze. Eine Einschränkung wiegt schwerer als alle genannten, weil sie sich nicht durch Einstellungen beheben lässt: Das Datenmodell kennt keinen Projektbezug. Es gibt genau ein aktives Drehbuch, und die Aufnahmen stehen ohne Zuordnung zu einem Projekt nebeneinander. Eine Installation trägt daher genau einen Film. Ein zweiter erfordert einen eigenen Datenbestand.

Für eine Arbeit, die ein Verfahren an einem Beispiel zeigt, ist das vertretbar. Für den Einsatz über mehrere Produktionen hinweg wäre es die erste vorzunehmende Änderung.

7.5 Einordnung gegenüber den betrachteten Systemen

Ein Vergleich mit den in Kapitel 3 beschriebenen Systemen ist nur eingeschränkt möglich, und die Grenze ist deutlich zu ziehen.

Nicht behauptet werden kann, dass das hier entwickelte Verfahren bessere Schnitte erzeugt. Die Untersuchung fand an zwölf Szenen eines einzigen Projekts statt, während die kommerziellen Systeme auf große Bestände und über viele Produktionen hinweg ausgelegt sind. Eine Aussage über Schnittqualität im Vergleich wäre durch nichts gedeckt.

Behaupten lässt sich etwas anderes: Die in dieser Arbeit durchgeführte Untersuchung ist an jenen Systemen nicht durchführbar. Eine Ablationsstudie setzt voraus, dass sich das Verfahren in Bestandteile zerlegen und einzeln abschalten lässt. Ein Prüfbericht setzt voraus, dass zu jeder Entscheidung eine Begründung vorliegt. Beides ist eine Eigenschaft der Bauweise, nicht der Ergebnisgüte. Der Beitrag dieser Arbeit liegt damit weniger in einem besseren Schnittvorschlag als in einem Vorschlag, über den sich überhaupt begründet streiten lässt.

7.6 Trennung der drei Bewertungsebenen

Technische Nachweise zeigen, ob Zeitbereiche, Segmente und Ausgaben formal gültig sind. Die drei Kriterien zeigen, ob eine Folge strukturell, sprachlich und visuell trägt. Wie eine Sequenz auf Dritte wirkt, ist mit beidem nicht erfasst. Gerade die Abweichungen sind aufschlussreich: Eine Timeline kann formal einwandfrei sein und dennoch verwirren, ein Bildsprung kann messbar auffallen und als bewusster Kontrast gelesen werden. Deshalb wird kein einzelner Wert als Gesamturteil geführt. Für die Reichweite des Systems folgt eine nüchterne Bestimmung: Es unterstützt Sichtung, Ordnung und Suche. Emotionale Wirkung, Ironie und dramaturgische Absicht kann seine Bewertungslogik nicht berücksichtigen. Der Anspruch ist nicht, diese zu erfassen, sondern die Arbeit bis zu der Stelle vorzubereiten, an der sie beurteilt werden müssen.

7.7 Limitationen der Arbeit

Neben den allgemeinen Einschränkungen sind vier Grenzen zu nennen, die sich unmittelbar aus den durchgeführten Messungen ergeben. Der Bestand umfasste zwölf Szenen, sodass die Ergebnisse das Verhalten bei knappem Material beschreiben. Jede Kombination aus Anfrage und Verfahren wurde einmal ausgeführt, wodurch die Schwankung des nicht vollständig wiederholbaren Sprachmodells nicht ausgeglichen ist. Die drei Zustände des Drehbuch-Abgleichs wurden nicht unabhängig überprüft. Und die Bewertung der Ergebnisse erfolgte durch dieselbe Person, die das System und die Auswertung erstellt hat. Eine unabhängige Fremdbewertung liegt nicht vor.

Hinzu kommen die Grenzen jedes Prototyps: ein einzelner Testbestand, eine Maschine, ein Satz Modelle, eine Konfiguration. Ohne unabhängige Ground Truth sind klassische Retrievalwerte nur eingeschränkt deutbar.

8. Fazit und Ausblick

8.1 Beantwortung der Forschungsfrage

Die Forschungsfrage lässt sich nach den durchgeführten Messungen in einem Teil beantworten und in einem anderen nicht. Beides ist im Folgenden getrennt auszuweisen. Ausgangspunkt war die Annahme, dass ein lokales KI-gestütztes Assistenzsystem die Filmpostproduktion unterstützen kann, wenn es Rohmaterial analysiert, Szenen und Sprachinformationen zeitlich auffindbar macht, multimodale Suche ermöglicht und einen bearbeitbaren Schnittvorschlag erzeugt. Der Nutzen liegt in vorbereitender Sichtung, Organisation, Suche und Vorauswahl. CinAssist ersetzt nicht die kreative, redaktionelle oder rechtliche Verantwortung des Editors.

8.1.1 Belegte Ergebnisse

Das Verfahren ist gebaut und läuft vollständig auf dem eigenen Rechner. Es führt Rohmaterial vom Einlesen über die Zuordnung von Bild und Ton, die Analyse, den Abgleich mit dem Drehbuch und die Erzeugung eines Schnittplans bis zu einer Timeline, die im Editor weiterbearbeitet werden kann. Sämtliche eingesetzten Bestandteile sind quelloffen. Es besteht keine Abhängigkeit von einem kostenpflichtigen Dienst und keine Notwendigkeit, Material an Dritte zu übertragen.

Die getroffenen Entwurfsentscheidungen haben eine messbare Wirkung. Die Bedingungen an Sprecher und Dialog werden mit den harten Filtern vollständig eingehalten und ohne sie durchgängig verletzt. Die Vielfalt der herangezogenen Aufnahmen steigt um 0,347. Dass diese Wirkung nachweisbar ist, beruht auf der Zerlegbarkeit des Verfahrens, eine Eigenschaft, die die betrachteten kommerziellen Systeme nicht bieten.

Der Preis dieser Entscheidungen ist beziffert und nicht verschwiegen: Die gemessene inhaltliche Nähe zur Anfrage sinkt, konzentriert auf Interviewpassagen, weil dort die geforderte Person häufig nicht in der ähnlichsten Aufnahme zu sehen ist.

Die vollständige Sichtung des Ergebnisses bestätigt den Aufbau und benennt eine Schwäche. Die Hauptspur gibt die Handlung in der Reihenfolge des Drehbuchs wieder,

ohne Wiederholung und ohne sichtbare Klappe, und der Ton sitzt durchgehend am zugehörigen Bild. Zu lang läuft dagegen das Ende einzelner Einstellungen. Im zweiten Glied von Szene 5 sind es 120,8 s nach der letzten Replik, in denen die Regie ins Bild tritt. Diese Beobachtung deckt sich mit dem Datenbestand: Bei drei der fünf Szenen-Master endet die Einstellung genau eine halbe Sekunde vor dem Ende der Aufnahme, dort hat also der Rückfall gegriffen, weil kein Produktionskommando erkannt wurde.

Schließlich ist jeder Vorschlag im Einzelnen prüfbar. Von 24 Dialogzeilen des Drehbuchs sind 22 mit einer bestimmten Aufnahme, einem Zeitbereich und dem tatsächlich gesprochenen Satz verbunden. Die beiden übrigen erscheinen als Lücke mit Angabe des Grundes. Für die 28 Handlungsanweisungen liegt je ein Zustand mit Beleg vor. Dass diese Offenlegung nicht nur formal besteht, zeigte sich während der Entwicklung, als sie eine fehlerhafte Anordnungsregel sichtbar machte und deren Änderung veranlasste.

8.1.2 Offene Punkte

Unbeantwortet bleibt die Frage, ob die erzeugten Abfolgen auf Dritte als Schnitt überzeugen. Die erhobenen Maße sagen dazu nichts: Sie erfassen Einhaltung, Vielfalt, Nähe und Zeit, nicht Wirkung. Die Beurteilung nach den drei Kriterien liegt mit Abschnitt 6.6 vor, stammt aber vom Verfasser und ist damit eine begründete Einzelmeinung. Eine unabhängige Fremdbewertung war nicht Bestandteil des Vorhabens und bleibt der nächste Schritt.

Ebenso wenig gedeckt sind allgemeine Aussagen. Alle Zahlen stammen aus einem Projekt mit zwölf verfügbaren Szenen, einem Drehbuch und einer Sprachfassung. Ob die drei Zustände des Drehbuch-Abgleichs zutreffen, wurde nicht unabhängig geprüft.

Für die Ausgangsannahme heißt das: Der Teil, der die Auffindbarkeit, die multimodale Suche und den bearbeitbaren Vorschlag betrifft, ist eingelöst. Der Teil, der die Unterstützung im Sinne eines gestalterischen Nutzens betrifft, ist begonnen, aber nicht belegt.

Die Ergebnisse sind deshalb getrennt nach technischen Nachweisen, objektiven Kriterien und subjektiven Beobachtungen berichtet. Eine noch nicht durchgeführte Evaluation darf nicht als positives Ergebnis formuliert werden. Die ausstehenden Teile sind als ausstehend gekennzeichnet.

8.2 Ausblick

Aus den Befunden dieser Arbeit ergeben sich vier Schritte unmittelbar. Der erste betrifft den Bestand: Zwölf Szenen begrenzen jede quantitative Aussage, und ein größerer Korpus würde

vor allem zeigen, ob der gemessene Zielkonflikt bei reichlicherem Material kleiner ausfällt. Der zweite betrifft die Anordnung des Versuchs: Mehrere Durchläufe je Kombination würden die Schwankung des Sprachmodells ausgleichen. Der dritte betrifft die Maße selbst: Die semantische Abdeckung bestraft in Interviewpassagen ein Verhalten, das handwerklich richtig ist, und müsste die gestellten Bedingungen berücksichtigen. Der vierte betrifft die Prüfung der Prüfung, die Zustände des Drehbuch-Abgleichs sollten von einer unabhängigen Person gegengelesen werden, wofür der Bericht bereits eine Spalte vorsieht.

Für den Einsatz über mehrere Produktionen hinweg ist zuerst der fehlende Projektbezug im Datenmodell zu ergänzen (Abschnitt 7.4). Alles Weitere, Klappenkonvention, Ablage, Schwellen, ist eine Frage der Absprache, nicht der Umarbeitung.

Auf der Seite des Verfahrens sind die Grenzen benannt: Der Rhythmus folgt Sprechpausen und Bildbewegung, nicht dem Bildinhalt, was sich am Ausstieg einzelner Einstellungen zeigt (Abschnitt 6.6), und die Zuordnung von Gesichtern zu Figuren ist bei sehr ähnlichem Auftrittsmuster nicht gesichert, weshalb sie überschreibbar bleiben muss. Darüber hinaus sind ein annotierter Datensatz, weitere Genres und Sprachen sowie robuste NLE-Importtests vorgesehen. Vorrangig bleiben Reproduzierbarkeit, transparente Unsicherheiten und eine klare Trennung zwischen technischer Assistenz und kreativer Montageentscheidung.

Über den Schnitt hinaus ist das hier gebaute Gerüst nicht festgelegt. Es besteht aus einer Aufbereitung, die Material lesbar macht, einer Entscheidungsschicht, die daraus einen Vorschlag ableitet, und einem Bericht, der jede Entscheidung offenlegt. Dieselbe Abfolge trägt auch andere Gewerke der Postproduktion. Naheliegend sind die Farbgebung, bei der sich Referenzbilder und gemessene Bildwerte zu einem begründeten Vorschlag verbinden lassen, sowie Übergänge und Effekte, die im jetzigen Verfahren nur als Blende mit fester Regel vorkommen. Jedes wäre eine eigene Arbeit mit eigenem Maßstab, könnte aber auf der bestehenden Aufbereitung, dem Drehbuch-Abgleich und der Provenienzkette aufsetzen.

Am Ende einer solchen Reihe stünde eine durchgehende Assistenz, die Material liest, transkribiert, ordnet, einen Schnitt vorschlägt und ihn farblich und in den Übergängen ausarbeitet, ohne dass Material das Haus verlässt und ohne kostenpflichtigen Dienst. Ausdrücklich nicht gemeint ist die Erzeugung von Bildern. Der Reiz liegt im Gegenteil: Alles Gezeigte stammt aus dem gedrehten Material und bleibt auf seine Quelle zurückführbar. Für eine Hochschule ist ein solcher Aufbau doppelt brauchbar, als Werkzeug in der Lehre und als Prüfstand, an dem sich einzelne Verfahren austauschen und gegeneinander messen lassen, weil der Quelltext offen und die Ausführung örtlich ist.

A. Anhang

A.1 Anhang A: Konfiguration und API-Referenz

Dieser Anhang führt die Ausstattung, die eingesetzten Modelle und die Kennzahlen des Referenzbestands auf, auf die sich die Angaben in Kapitel 5 und Kapitel 6 beziehen.

Rechner

Alle in dieser Arbeit berichteten Laufzeiten und Ergebnisse wurden auf einem Apple-Rechner mit M1-Pro-Prozessor, 32 GB Arbeitsspeicher und macOS 26.5 erhoben. Sämtliche Modelle laufen auf diesem Rechner. Es besteht keine Verbindung zu einem externen Dienst.

Eingesetzte Modelle

Versionsstände und Lizenzen

Die folgenden Angaben sind der Projektumgebung entnommen, in der die berichteten Läufe ausgeführt wurden. Sämtliche Bestandteile stehen unter freizügigen Lizenzen, überwiegend MIT, BSD-3-Clause und Apache-2.0. Eine Verpflichtung zur Offenlegung des eigenen Quellcodes entsteht daraus nicht.

FFmpeg wird als externes Programm aufgerufen und nicht mitgeliefert. Die maßgebliche Lizenz hängt von den beim Übersetzen aktivierten Bestandteilen ab und ist der jeweiligen Installation zu entnehmen.

Bestand und Kennzahlen

Diese Angaben beziehen sich auf den Stand des Quellcodes [37], wie er dieser Arbeit beiliegt. Der Arbeit liegen zudem die technischen Aufzeichnungen des Projekts bei [39].

Tabelle A.1: Modelle und ihre Aufgabe im System

Aufgabe	Modell	Anmerkung
Spracherkennung	whisper-large-v3-turbo	über die MLX-Fassung, auf Apple-Silicon beschleunigt
Bild- und Texteinbettung	open_clip ViT-L-14, vortrainiert datacomp_xl_s13b_b90k	768 Dimensionen. Gegenüber ViT-B-32 deutlich bessere Trefferlage
Zerlegung der Anfrage, Zusammenfassungen	qwen2.5:14b	9,0 GB, Rückfall auf llama3, wenn nicht vorhanden
Geschlossene Fragen zum Bild	llava:7b	4,7 GB, Grundlage der Bildprüfung
Bildbeschreibung	moondream	1,7 GB
Texteinbettung	bge-m3	1,2 GB
Gesichtsmerkmale	MTCNN und FaceNet (VGGFace2)	Erkennungsschwelle 0,9

Tabelle A.2: Tragende Komponenten mit Version und Lizenz

Backend	Version	Lizenz	Oberfläche	Version	Lizenz
FastAPI	0.136.1	MIT	Next.js	16.2.3	MIT
Uvicorn	0.46.0	BSD-3-Clause	React	19.2.4	MIT
SQLAlchemy	2.0.49	MIT	React-DOM	19.2.4	MIT
asyncpg	0.31.0	Apache-2.0	Wavesurfer.js	7.12.11	BSD-3-Clause
Celery	5.6.3	BSD-3-Clause	Zustand	5.0.12	MIT
Redis (Client)	7.4.0	MIT	Framer Motion	12.38.0	MIT
mlx-whisper	0.4.3	MIT	react-resizable-panels	2.1.9	MIT
open_clip_torch	3.3.0	MIT	Tailwind CSS	4.2.2	MIT
PyTorch	2.12.0	BSD-3-Clause	TypeScript	5.9.3	Apache-2.0
torchvision	0.27.0	BSD-3-Clause			
PySceneDetect	0.7	BSD-3-Clause			
OpenCV (headless)	4.13.0.92	Apache-2.0			
NumPy	2.4.4	BSD-3-Clause			
Pillow	12.2.0	MIT-CMU			
HTTPX	0.28.1	BSD-3-Clause			
silero-vad	6.2.1	MIT			
rank-bm25	0.2.2	Apache-2.0			
pyannote.audio	4.0.7	MIT			

Tabelle A.3: Referenzbestand und wesentliche Kennzahlen

Größe	Wert
Bildaufnahmen	58 Dateien, zusammen 82,7 Minuten
Tonaufnahmen	57 Dateien in zwei Ordnern
Drehbuch	fünf Szenen, 24 Dialogzeilen, 28 Handlungsanweisungen
Szenen im Auswahlbestand des Vergleichslaufs	12
Bildabstand der Bildprüfung	5 s in stummen, 10 s in dialoghaltigen Passagen, 448 Bildpunkte Breite
Bildabstand der Gesichtssuche	5 s, 896 Bildpunkte Breite
Temperatur bei der Zerlegung der Anfrage	0,3
Startwert der Zufallsauswahl im Vergleichslauf	42

Anfragen des Vergleichslaufs

Die 18 Anfragen der Ablationsstudie sind hier im unveränderten Wortlaut wiedergegeben, einschließlich der ursprünglichen Schreibweisen und Hervorhebungen. Sie wurden vor der Durchführung festgeschrieben (Abschnitt 4.5.3). Die letzte Spalte nennt die Prüfabsicht, mit der die jeweilige Anfrage aufgenommen wurde. Sie ist im Quelltext der Messung hinterlegt (`prompts_benchmark.py`) und wurde ebenfalls vor dem Lauf festgelegt.

A.2 Anhang B: Protokolle der KI-gestützten Entwicklung

Die Aufzeichnung der Arbeitsphase liegt den Materialien bei [40] und umfasst drei Dateien.

Die Entwicklung wurde auf Französisch geführt, der Arbeitssprache des Verfassers. Die lesbare Fassung der Eingaben liegt deshalb übersetzt vor, während das Journal den unveränderten Wortlaut wiedergibt, sodass sich die Übersetzung überprüfen lässt.

Am Journal ist inhaltlich nichts verändert. Ersetzt sind allein Dateipfade des Arbeitsrechners, die weder zur Sache gehören noch der Nachvollziehbarkeit dienen. Zeitstempel und Reihenfolge sind unangetastet, weil die ausgezählten Kennzahlen auf ihnen beruhen. Der Verlauf wurde einmal zusammengefasst, die Zusammenfassung

Tabelle A.4: Die 18 Anfragen des Vergleichslaufs im Wortlaut

Kennung	Profil	Ziel	Anfrage	Prüfabsicht
broll_nature_ short	Bildstrecke	15	15 Sekunden ruhige Naturaufnahmen, Wasser und Landschaft, keine Menschen	kurze Dauer, muss ohne Personen filtern und zufälligen Dialog vermeiden
broll_nature_60s	Bildstrecke	60	60 Sekunden ruhige Naturstimmung, Weitwinkel-Landschaften mit sanftem Übergang, keine Menschen	Grundfall, Bezugspunkt der Vergleiche
broll_urban	Bildstrecke	30	30 Sekunden Stadtbilder, Straßen und Gebäude aus der Distanz, kein Fokus auf Personen	städtisch statt Natur, im Bestand wenig vorhanden
broll_water_calm	Bildstrecke	20	20 Sekunden ruhige Wasseraufnahmen: Wellen, Fluss, See – meditativ, keine Menschen	zwingt den Abruf auf Wasser-Szenen
broll_energy_ dynamic	Bildstrecke	25	25 Sekunden dynamische Actionszenen, schnelle Bewegung, energische Weitwinkel-Bilder	prüft, ob CLIP Bewegung von Ruhe unterscheidet
interview_40s_ closeup	Interview	40	40 Sekunden Interview-Ausschnitt: Sprecher redet in Nahaufnahmen und Halbtotalen, kein B-Roll	durchgehend Dialog, Schnitt an Wortgrenzen
interview_60s_ serious	Interview	60	60 Sekunden ernstes Gespräch eines Sprechers direkt in die Kamera, ausschliesslich Nahaufnahmen	harte Bedingung auf Nahaufnahme
interview_broll_ mix	Interview	45	45 Sekunden: Sprecher erklärt seine Arbeit (Nahaufnahmen), dazwischen kurze B-Roll seiner Umgebung als Illustration	Wechsel zwischen A-Roll und B-Roll
interview_short	Interview	15	15 Sekunden Statement einer Person, Nahaufnahme, klare Aussage	kurzer Schnitt mit starkem Dialogbedarf
interview_multi_ speaker	Interview	50	50 Sekunden Diskussion mit mehreren Sprechern im Wechsel, Halbtotalen bevorzugt	mehrere Sprecher, prüft den Quellenwechsel
narrative_lonely_ cook	Gemischt	90	90 Sekunden über die Einsamkeit des Kochs am frühen Morgen in der leeren Küche, Wechsel zwischen Weitwinkel und Nahaufnahmen, endet auf einem stillen Close-up seines Gesichts	ursprüngliche Anfrage der Arbeit, stark erzählend
narrative_ documentary_ open	Gemischt	30	30 Sekunden Doku-Intro: Weitwinkel-Establisher, dann Nahaufnahme eines menschlichen Details, endet auf einem Halbtotal einer aktiven Person	klarer dramaturgischer Aufbau
narrative_music_ video	Gemischt	45	45 Sekunden Musikclip: rhythmische Wechsel zwischen Nahaufnahmen von Bewegung und Weitwinkel-Umgebung	schneller Rhythmus, prüft die Dauergrenzen der Slots
narrative_teaser_ energetic	Gemischt	20	20 Sekunden energischer Teaser mit schnellen Schnitten, wechselnde Framings, aufsteigende Intensität	dichter Schnitt, mindestens acht bis zehn Slots
narrative_slow_ atmospheric	Gemischt	120	120 Sekunden langsamer, atmosphärischer Cut: lange Weitwinkel, gelegentliche Nahaufnahmen als Akzent	langer Schnitt, prüft Zeittreue und Vielfalt
edge_very_short	Grenzfall	8	8 Sekunden Micro-Cut: 2 Weitwinkel + 1 Closeup, sehr schnell	prüft die untere Dauergrenze
edge_no_person_ hard	Grenzfall	30	30 Sekunden Landschaftsimpressionen, ABSOLUT KEINE MENSCHEN im Bild, nur Natur	harte Bedingung, prüft die bestandsbewusste Planung
edge_all_closeup	Grenzfall	45	45 Sekunden ausschliesslich Nahaufnahmen von Gesichtern oder Details, NIE Weitwinkel	harte Bedingung auf die Einstellungsgröße, im Bestand wenige Nahaufnahmen

Tabelle A.5: Bestandteile der Aufzeichnung der Entwicklungsphase

Datei	Inhalt
`nutzerprompts.md`	die 53 Eingaben des Verfassers in deutscher Übersetzung, in ihrer Reihenfolge, je mit der Kategorie aus Abschnitt 5.9
`sitzungsjournal_RAW.jsonl`	vollständiges maschinenlesbares Journal: alle Nachrichten, Werkzeugaufrufe, Ergebnisse und Zeitstempel, die Primärquelle, aus der die Kennzahlen in Abschnitt 5.9 ausgezählt sind
`sitzungsjournal_lesbar_de.html`	derselbe Verlauf im Browser lesbar: alle 229 Prosablöcke

steht im Journal als Nachricht und ist als solche erkennbar. Gesprächsinhalte früherer Entwicklungsabschnitte, die nicht aufgezeichnet wurden, wurden nicht nachträglich rekonstruiert.

A.3 Anhang C: Aus Fehlern abgeleitete Regeln

Der in Abschnitt 4.2 beschriebene Arbeitszyklus ändert bei einem Mangel nicht den Einzelfall, sondern die Regel, die ihn erzeugt hat. Insgesamt entstanden für den Referenzfilm auf diesem Weg 49 Schnittfassungen. Eine chronologische Liste aller Fassungen mit Name, Datum und Eintragszahl liegt dem Repository bei (`docs/fassungschronik_pinky_promise.md` [37]). Jede Fassung ist zudem in der Projektdatenbank vollständig erhalten, einschließlich der Begründung je Eintrag. Die folgende Übersicht ist eine Auswahl: Sie führt die Regeländerungen, die in dieser Arbeit im Zusammenhang beschrieben sind, kleinere Anpassungen zwischen den Fassungen sind nicht einzeln ausgewiesen. Jeder Eintrag ist an der genannten Stelle der Arbeit belegt. Tabelle A.6 führt die Mängel und die daraus abgeleiteten Regeln auf.

Zugang zum laufenden System

Für die Begutachtung läuft das System auf einem Rechner des Verfassers und ist unter https://macmini.tailef3707.ts.net/ erreichbar. Die Zugangsdaten werden den Prüfern gesondert übermittelt. Damit lassen sich die Oberfläche, der Materialbestand und der erzeugte Schnitt ohne eigene Installation ansehen.

Tabelle A.6: Beobachtete Mängel und die daraus abgeleiteten Regeln

Beobachteter Mangel	Abgeleitete Regel	Abschnitt
Detailaufnahme in einer Dialogpause ohne Anlass im Drehbuch, Begründung nannte nur die Pausenlänge	Bildprüfung: Regieanweisungen werden als geschlossene Fragen gegen Einzelbilder geprüft, stumme Aufnahmen nach *bestätigten* Handlungen geordnet, Einschübe nur bei Anlass im Drehbuch	5.5.4
Rohschnitt zu lang und blockhaft	Bildwechsel in Sprechpausen bei durchlaufendem Ton, Handlung in Höhepunkten statt Blöcken	4.2
Fehlende Gegeneinstellung im Dialog	Reaktionsschnitt: Wechsel auf die zuhörende Figur bei durchlaufendem Ton	5.5.5
Entfernte, verhallte Sprache nicht als Sprechaktivität erkannt, von 112 s blieb nur die Klappenansage	Rettungspass: zweiter Erkennungslauf über die ganze Datei bei Abdeckung unter zehn Sekunden oder 15 Prozent, mit strengen Übernahmebedingungen	5.4
Einstellung begann mit dem Betreten des Raums statt mit dem Spielbeginn	Anspiel-Barriere: die letzte Produktionsäußerung vor der ersten Spieläußerung ist eine harte Grenze	5.5.1
Abgebrochener Take gewann die Take-Wahl, weil er die Zeilen bis zum Abbruch vollständig abdeckte	Punktwert aus mehreren Größen inklusive Abbruchstatus, Abbruch-Erkennung über Produktions-Sprech nach Spielbeginn	5.5.3
Container-Timecode auf allen Dateien identisch	Unplausible Angaben werden verworfen statt übernommen	5.3
Take-Nummern von Bild und Ton systematisch um eins verschoben	Keine Zuordnung allein nach dem Dateinamen	5.3
Einschübe zerschnitten die durchlaufende Aufnahme und ihr Verwerfen hinterließ ein Loch	Spur-Ordnung: Auflagen liegen ausgeblendet über dem Master statt in ihn hineingeschnitten	5.5.6
Spracherkennung lieferte „Thank you.“ auf nahezu stummer Spur	Tonkanal-Klassifikation: Stille, Timecode und Rauschen werden nicht transkribiert, der Verzicht wird vermerkt	5.3
Festes Perzentil zur Schätzung der LTC-Bitdauer versagte bei vier von sechs Aufnahmen	Robuste Schätzung über die Trennung der Flankenabstände in zwei Häufungen, mit Plausibilitätsprüfung der BCD-Felder	5.2

Zugang zum Rohmaterial

Das Rohmaterial des Referenzfilms, also die Kamera- und Tonaufnahmen des Kurzfilms, umfasst rund 100 Gigabyte und liegt den Materialien nicht bei. Für die Wiedergabe der beiliegenden Schnittfassungen wird es nicht benötigt, wohl aber für eigene Läufe an der Quelle. Dafür ist es unter https://drive.google.com/drive/folders/1eQXTYJE6nx-j5XpPDvvDiO-AA_-NUknz abrufbar.

A.4 Anhang D: Sequenz-Export nach Spuren

Der in Abschnitt 6.6 beurteilte Schnittvorschlag liegt den Materialien als Videodatei je Spur bei, dazu eine Schnittliste mit Zeitpunkt, Szene, Einstellung, Take und Quellbereich für jeden der 47 Clips. Die fünf Dateien sind gleich lang und bildgenau deckungsgleich, sie zählen je 18 468 Bilder bei 24 Bildern je Sekunde, was einer Laufzeit von 00:12:49:12 entspricht. Ein Zeitpunkt in der Hauptspur ist damit derselbe Zeitpunkt in jeder Alternativspur, sodass sich die Spuren nebeneinander betrachten lassen. Wo für eine Einstellung keine Alternative vorliegt, steht auf der betreffenden Spur Schwarzbild mit Stille. Das bildet die Lücke in der Timeline ab und ist kein Ausfall.

Das Bild ist H.264 in 1920×1080 und behält die Bildrate der Quelle bei. Der Ton stammt aus der synchronisierten Aufnahme des Tonrekorders, zeitlich versetzt um den in Abschnitt 5.3 bestimmten Wert, übernommen sind dessen Kanäle 1 und 2. Der Kameraton wird nicht verwendet. Links oben ist eine Kennung eingeblendet, die Spur, Szene, Einstellung, Take und den Startzeitpunkt in der Timeline nennt. Sie dient dem Nachvollziehen und ist nicht Bestandteil des Schnitts.

Erklärung zur Verwendung KI-basierter Werkzeuge

Bei der Erstellung dieser Arbeit wurden die folgenden KI-basierten Werkzeuge eingesetzt. Inhalt, Aufbau, Argumentation und die Auswahl der Quellen stammen vom Verfasser.

Werkzeug	Einsatzbereich	Art der Nutzung
DeepL Write	sprachliche Überarbeitung	Vorschläge zu Formulierung und Satzbau für bereits verfasste Abschnitte
Claude Code in Visual Studio Code	Programmierung	Entwicklung des Prototyps und Hilfe beim Debugging. Näheres in Abschnitt 5.9, die Aufzeichnung liegt bei (Anhang B)
Google-Suche, Perplexity	Recherche	Auffinden von Fachliteratur und Herstellerdokumentation

Für die grafische Gestaltung einzelner Abbildungen wurde Canva Pro verwendet, also für Anordnung, Beschriftung und Farbgebung. Satz und Formatierung der Arbeit erfolgten in LaTeX über Overleaf (https://www.overleaf.com/).

Erklärung zur selbständigen Bearbeitung

Hiermit versichere ich, dass ich die vorliegende Arbeit ohne fremde Hilfe selbständig verfasst und nur die angegebenen Hilfsmittel benutzt habe. Wörtlich oder dem Sinn nach aus anderen Werken entnommene Stellen sind unter Angabe der Quellen kenntlich gemacht.

Hamburg	24. August 2026	
Ort	Datum	Unterschrift im Original

Literaturverzeichnis

[1] Oliver Keutzer, Sebastian Lauritz, Claudia Mehlinger und Peter Moormann. *Filmanalyse*. Film, Fernsehen, Neue Medien. Wiesbaden: Springer VS, 2014. ISBN: 978-3-658-02099-6.

[2] Marcelo Sandoval-Castañeda, Scandar Copti und Dennis Shasha. "AutoTag: automated metadata tagging for film post-production". In: *Multimedia Tools and Applications* 83.3 (2023), S. 6731–6753. DOI: 10.1007/s11042-023-15565-w.

[3] Ulrich Schmidt. *Digitale Film- und Videotechnik*. 3., erweiterte Auflage. München: Carl Hanser Verlag, 2011. ISBN: 978-3-446-42477-7.

[4] Peter Wuss. *Künstlerische Verfahren des Films aus psychologischer Sicht. Zum Wirkungspotenzial des Spielfilms*. Film, Fernsehen, Medienkultur. Wiesbaden: Springer VS, 2020. ISBN: 978-3-658-32051-5.

[5] Bernd Huber, Hijung Valentina Shin, Bryan Russell, Oliver Wang und Gautham J. Mysore. "B-Script: Transcript-based B-roll Video Editing with Recommendations". In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 2019, S. 1–11. DOI: 10.1145/3290605.3300311.

[6] Albert Heiser. *Das Drehbuch zum Drehbuch. Storytelling, Konzeption und Produktion für Werbefilme, Trailer, Imagefilme und Viral-Videos*. 3. Auflage. Wiesbaden: Springer Gabler, 2020. ISBN: 978-3-658-29733-6.

[7] Gregor Alexander Heussen. *Fakten · Bilder · Töne · Story. Dokumentarische Filmdramaturgie: TV · Video · Netz · Kino*. Journalistische Praxis. Wiesbaden: Springer VS, 2023. ISBN: 978-3-658-37794-6.

[8] Lothar Riedl. *Videos mit künstlicher Intelligenz gestalten. Kreative Tools und professionelle Workflows für die Videoproduktion*. X.media.press. Wiesbaden: Springer Vieweg, 2025. ISBN: 978-3-658-46662-6.

[9] Bryan Wang, Yuliang Li, Zhaoyang Lv, Haijun Xia, Yan Xu und Raj Sodhi. "LAVE: LLM-Powered Agent Assistance and Language Augmentation for Video Editing". In: *Proceedings of the 29th International Conference on Intelligent User Interfaces*. ACM, 2024, S. 699–714. DOI: 10.1145/3640543.3645143.

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser und Illia Polosukhin. “Attention Is All You Need”. In: *Advances in Neural Information Processing Systems*. Bd. 30. Curran Associates, Inc., 2017, S. 5998–6008. arXiv: 1706.03762. URL: https://arxiv.org/abs/1706.03762.

[11] Rohit Prabhavalkar, Takaaki Hori, Tara N. Sainath, Ralf Schlüter und Shinji Watanabe. “End-to-End Speech Recognition: A Survey”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 32 (2024), S. 325–351. DOI: 10.1109/TASLP.2023.3328283.

[12] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey und Ilya Sutskever. “Robust Speech Recognition via Large-Scale Weak Supervision”. In: *Proceedings of the 40th International Conference on Machine Learning*. Bd. 202. Proceedings of Machine Learning Research. PMLR, 2023, S. 28492–28518. URL: https://proceedings.mlr.press/v202/radford23a.html.

[13] OpenAI. *Whisper: Robust Speech Recognition via Large-Scale Weak Supervision*. Quelloffene Referenzimplementierung, MIT-Lizenz. URL: https://github.com/openai/whisper.

[14] Open Source Initiative. *The MIT License*. URL: https://opensource.org/license/mit.

[15] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger und Ilya Sutskever. “Learning Transferable Visual Models From Natural Language Supervision”. In: *Proceedings of the 38th International Conference on Machine Learning*. Bd. 139. Proceedings of Machine Learning Research. PMLR, 2021, S. 8748–8763. URL: https://proceedings.mlr.press/v139/radford21a.html.

[16] Stephen Robertson und Hugo Zaragoza. “The Probabilistic Relevance Framework: BM25 and Beyond”. In: *Foundations and Trends in Information Retrieval* 4.1–2 (2009), S. 1–174. DOI: 10.1561/1500000019.

[17] Blackmagic Design. *DaVinci Resolve 21 – What’s New*. Herstellerangabe, die Seite nennt kein Veröffentlichungsdatum. URL: https://www.blackmagicdesign.com/products/davinciresolve/whatsnew.

[18] Blackmagic Design. *DaVinci Resolve 21 – New Features Guide*. Herstellerangabe, Titelblatt datiert April 2026. 2026. URL: https://documents.blackmagicdesign.com/SupportNotes/DaVinci_Resolve_21_New_Features_Guide.pdf.

[19] Blackmagic Design. *DaVinci Resolve Studio*. Herstellerangabe. URL: https://www.blackmagicdesign.com/products/davinciresolve/studio.

[20] Blackmagic Design. *Blackmagic Design Announces DaVinci Resolve 20*. Herstellerangabe. 4. Apr. 2025. URL: https://www.blackmagicdesign.com/media/release/20250404-02.

[21] Blackmagic Design. *DaVinci Resolve: Media*. URL: https://www.blackmagicdesign.com/products/davinciresolve/media/.

[22] Blackmagic Design. *DaVinci Resolve 21 – Reference Manual*. Herstellerangabe. 2026. URL: https://documents.blackmagicdesign.com/UserManuals/DaVinci_Resolve_21_Reference_Manual.pdf.

[23] Adobe. *What's new in Adobe Premiere on desktop*. Herstellerangabe. 7. Aug. 2026. URL: https://helpx.adobe.com/premiere/desktop/whats-new/whats-new.html.

[24] Adobe. *Search media using AI in Premiere*. URL: https://helpx.adobe.com/premiere/desktop/organize-media/file-organization/search-for-media-using-ai-powered-media-intelligence.html.

[25] Adobe. *Suchen nach Audio mit Media Intelligence in Premiere*. URL: https://helpx.adobe.com/de/premiere/desktop/organize-media/file-organization/search-for-audio-using-media-intelligence.html.

[26] Adobe. *Media intelligence and Search panel FAQ in Adobe Premiere*. Herstellerangabe. 8. Apr. 2026. URL: https://helpx.adobe.com/premiere/desktop/organize-media/file-organization/media-intelligence-and-search-panel.html.

[27] Adobe. *Compare Premiere plans*. Herstellerangabe. URL: https://www.adobe.com/products/premiere/plans.html.

[28] Sean Butler und Alan Parkes. "Film sequence generation strategies for automatic intelligent video editing". In: *Applied Artificial Intelligence* 11.4 (1997), S. 367–388. DOI: 10.1080/088395197118190.

[29] Stefano Bocconi. "Semantic-Aware Automatic Video Editing". In: *Proceedings of the 12th Annual ACM International Conference on Multimedia*. New York: ACM, 2004. DOI: 10.1145/1027527.1027753.

[30] Nayyer Aafaq, Ajmal Mian, Wei Liu, Syed Zulqarnain Gilani und Mubarak Shah. "Video Description: A Survey of Methods, Datasets, and Evaluation Metrics". In: *ACM Computing Surveys* 52.6 (2019), S. 1–37. DOI: 10.1145/3355390.

[31] Alan F. Smeaton, Paul Over und Wessel Kraaij. "Evaluation campaigns and TRECVid". In: *Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*. ACM, 2006, S. 321–330. DOI: 10.1145/1178677.1178722.

[32] Adobe. *Premiere AI Assistant (beta) overview*. Herstellerangabe, öffentliche Beta. 18. Juni 2026. URL: https://helpx.adobe.com/premiere/desktop/premiere-ai-assistant/overview.html.

[33] Yannick Fuchs. “Vergleich generativer Video-KI und klassischer Videoproduktion. Eine Analyse von Effizienz und Qualität aktueller Open-Source-Modelle”. Ergänzende studentische Vorarbeit, nicht als zentrale wissenschaftliche Grundlage verwendet. Bachelorarbeit. Hochschule für Angewandte Wissenschaften Hamburg, 2025. URL: https://hdl.handle.net/20.500.12738/19130.

[34] Alan R. Hevner, Salvatore T. March, Jinsoo Park und Sudha Ram. “Design Science in Information Systems Research”. In: *MIS Quarterly* 28.1 (2004), S. 75–106. DOI: 10.2307/25148625.

[35] Hubert Österle, Jörg Becker, Ulrich Frank, Thomas Hess, Dimitris Karagiannis, Helmut Krcmar, Peter Loos, Peter Mertens, Andreas Oberweis und Elmar J. Sinz. “Memorandum zur gestaltungsorientierten Wirtschaftsinformatik”. In: *Schmalenbachs Zeitschrift für betriebswirtschaftliche Forschung* 62.6 (2010), S. 664–672. DOI: 10.1007/BF03372838.

[36] Ken Peffers, Tuure Tuunanen, Marcus A. Rothenberger und Samir Chatterjee. “A Design Science Research Methodology for Information Systems Research”. In: *Journal of Management Information Systems* 24.3 (2007), S. 45–77. DOI: 10.2753/MIS0742-1222240302.

[37] Pascal Nikiema. *CinAssist: Projekt-Repository*. URL: https://github.com/papy231/cinassist.

[38] Nicola Döring. *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften*. Berlin, Heidelberg: Springer, 2023. ISBN: 978-3-662-64761-5.

[39] Pascal Nikiema. “Interne technische Projektspezifikationen und Arbeitsnotizen”. Dem Repository beiliegend. 2026.

[40] Pascal Nikiema. *CinAssist: Aufzeichnung der KI-gestützten Entwicklung*. Beiliegende Materialien, Ordner 04_Prompting. 2026.