7. Anwendungen

Automatische Zusammenfassungen von Videos

Videoanalyse

Dr. Stephan Kopf

Übersicht

- Motivation
- Analyse historischer Filme
- Automatische Erzeugung von Zusammenfassungen
  - Analyse des Videos
  - Auswahl von Kameraeinstellungen
- Ergebnisse
- Zusammenfassung

Motivation (I)

- Projekt: “European-Chronicles-Online”
- EU-Projekt mit mehreren Teilnehmern:
  4 europäische nationale Filmarchive
- In den Archiven werden zum Teil mehr als 100.000 Stunden historischer Filme gespeichert

Ziel

- Entwicklung eines Systems, um die umfangreichen Sammlungen mit historischen Filmen digital zu speichern und einen einfachen Zugriff zu ermöglichen.

Motivation (II)

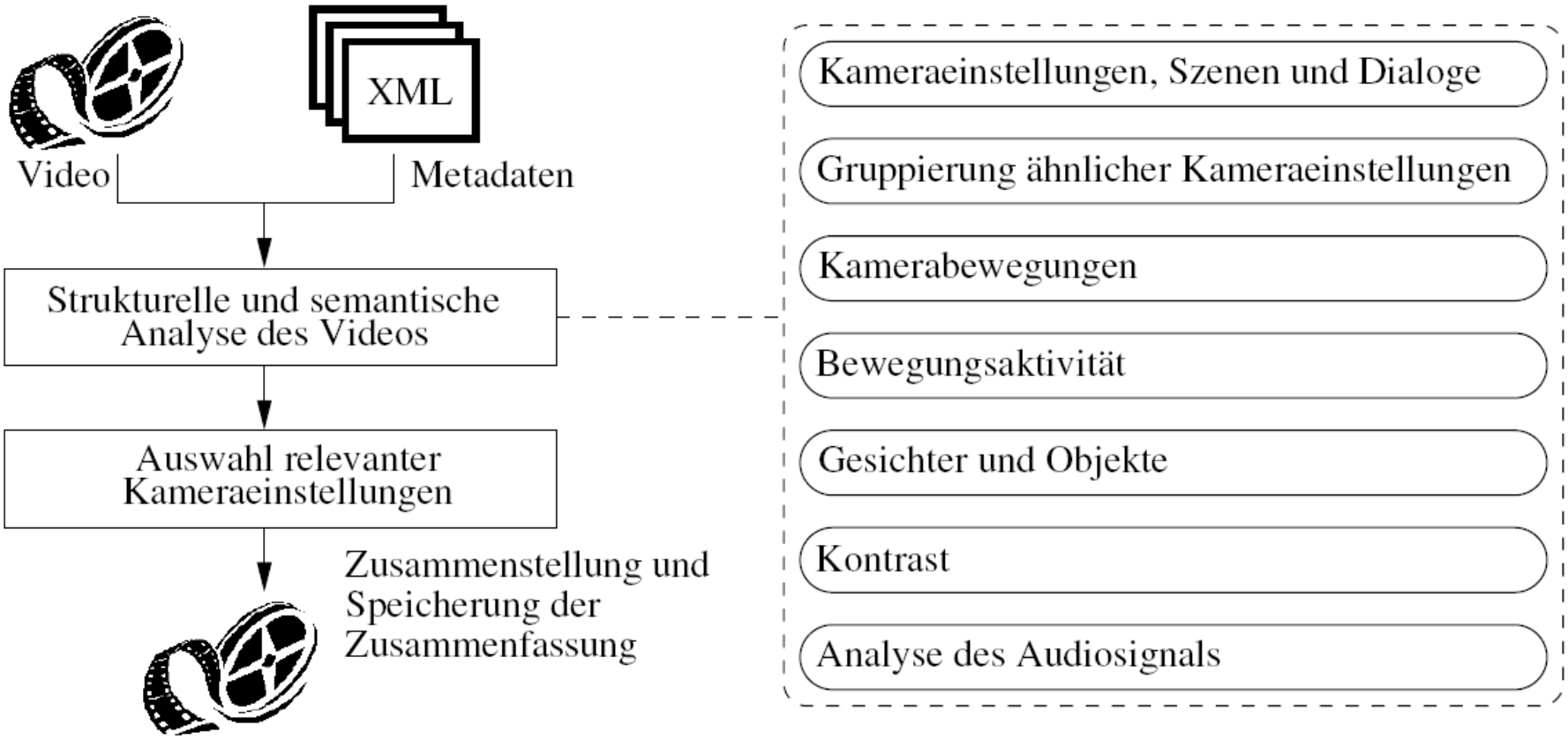
Zu lösende Probleme

- Wie können in umfangreichen Archiven Videos oder Videosegmente gefunden werden?
  → manuelle Annotationen
  → automatische Analyse von semantischen Inhalten
  → Text-basierte Suchanfragen liefern zum Teil mehrere hundert Videos
- Der Zeitaufwand ist sehr hoch, um den Inhalt von Videos zu erfassen
  → Automatische Zusammenfassungen von Videos

Besonderheiten historischer Filme

- Schwarz-Weiß
- Rauschen
- Schwankungen der Helligkeit
- Verwackelte Kamera
- Kratzer und Streifen
- Fehler bei der Aufnahme

Erzeugung von Zusammenfassungen



Analyse des Videos (I)

- Es werden Informationen über Schnitte, Kamerabewegungen, Gesichter, Objekte und Textregionen ermittelt.
- Die Informationen werden auf der Ebene der Kameraeinstellungen zusammengefasst und durch einen aggregierten Merkmalswert beschrieben.

Analyse des Videos (II)

Erkennung von ähnlichen Kameraeinstellungen

- Es werden *repräsentative Bilder* von allen Kameraeinstellungen benötigt. Zunächst wird das mittlere Bild einer Kameraeinstellung als repräsentatives Bild ausgewählt.
- In den historischen Videos treten häufig fehlerhafte Bildbereiche und zum Teil vollständig defekte Bilder auf.
- Durch einen Vergleich des Histogramms des festgelegten Bildes mit dem durchschnittlichen Histogramm aller Bilder der Kameraeinstellung kann verhindert werden, dass einzelne fehlerhafte Bilder verwendet werden.
- Bei einer großen Differenz beider Histogramme wird das repräsentative Bild durch das Bild der Kameraeinstellung ersetzt, dessen Histogramm möglichst ähnlich dem durchschnittlichen Histogramm ist.

Analyse des Videos (III)

Gruppierung ähnlicher Kameraeinstellungen

- *Größe einer Gruppe:* Summe der Länge der Kameraeinstellungen dieser Gruppe
- Größe gibt einen Hinweis auf die Bedeutung der Gruppe für das Video
- Große Gruppen erhalten eine hohe Priorität, so dass diese Gruppen durch mindestens eine Kameraeinstellung in der Zusammenfassung repräsentiert werden.
- Es existiert kein spezieller zeitlicher Bezug innerhalb einer Gruppe mit ähnlichen Kameraeinstellungen (im Gegensatz zu Szenen).

Analyse des Videos (IV)

Zuordnung zu Gruppen

- Vergleich der repräsentativen Bilder
- Die Summe der absoluten Differenzen von Graustufenhistogrammen liefert ein Ähnlichkeitsmaß für die Bilder.
- Identifikation von speziellen Zentren für jede Gruppe
- Die Zentren werden durch Graustufenhistogramme beschrieben (Punkt im mehrdimensionalen Raum).
- Während der Gruppierung werden neue Zentren festgelegt, bis der Abstand aller Bilder zum jeweils nächstgelegenen Zentrum einen Schwellwert unterschreitet.
- Falls der Abstand mindestens eines repräsentativen Bildes über dem Schwellwert liegt, wird ein zusätzliches Zentrum benötigt und hinzugefügt.

Analyse des Videos (V)

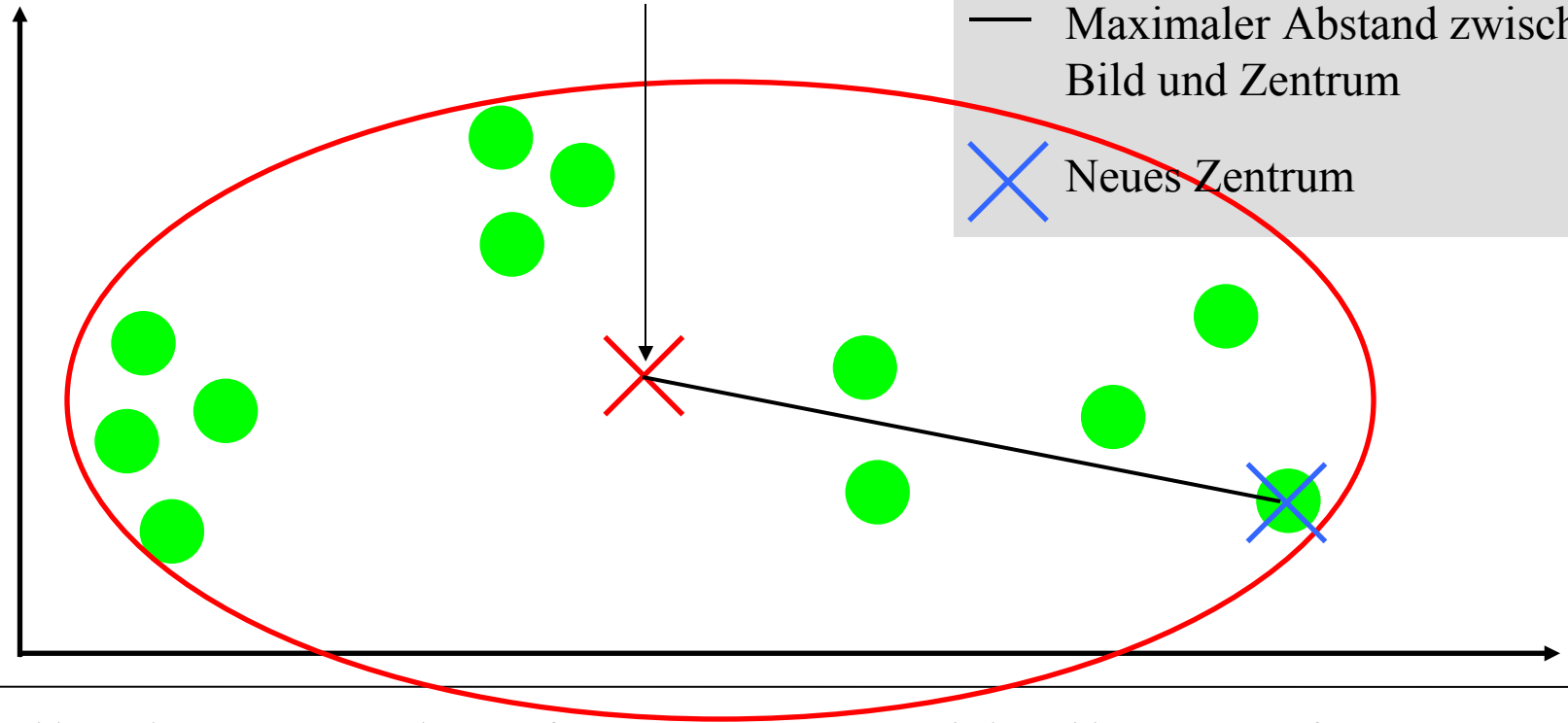
Zuordnung zu Gruppen (Modifizierter *K-Means*-Algorithmus)

2. Das erste Zentrum wird als durchschnittliches Histogramm aller repräsentativen Bilder initialisiert. Die Summe der Abstände zwischen dem Zentrum und allen Bildern ist für diesen Punkt minimal.
3. Für jedes repräsentative Bild wird das nächstgelegene Zentrum identifiziert. Jedes Bild wird dem nächstgelegenen Zentrum zugeordnet.
4. Die Positionen aller Zentren werden aktualisiert. Die neue Position eines Zentrums ist definiert als durchschnittlicher Histogrammwert aller Bilder, die diesem Zentrum zugeordnet sind.
5. Das Bild mit dem größten Abstand zu seinem Zentrum wird ausgewählt. Falls der Abstand über einem Schwellwert liegt, sind die Unterschiede innerhalb der Gruppe sehr hoch, und ein neues Zentrum wird an der Position dieses Bildes eingefügt. Gehe zu 2. bis alle repräsentativen Bilder innerhalb einer Gruppe eine große Ähnlichkeit besitzen.

Analyse des Videos (VI)

Bsp.: Zuordnung zu Gruppen

Initialisierung:
Schwerpunkt aller Bilder

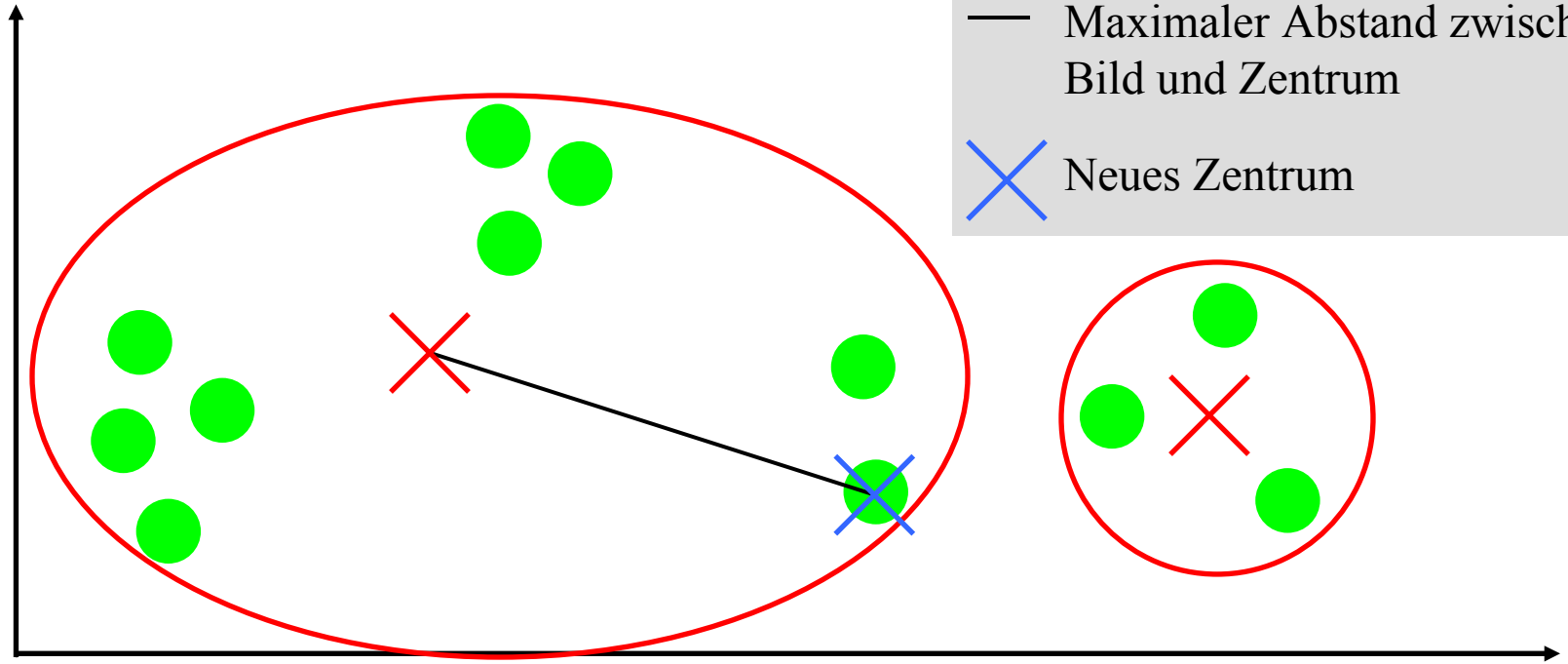


Iteration 1

- Am nächsten gelegenes Zentrum
- Aktualisierte Position der Zentren
- Maximaler Abstand zwischen Bild und Zentrum
- Neues Zentrum

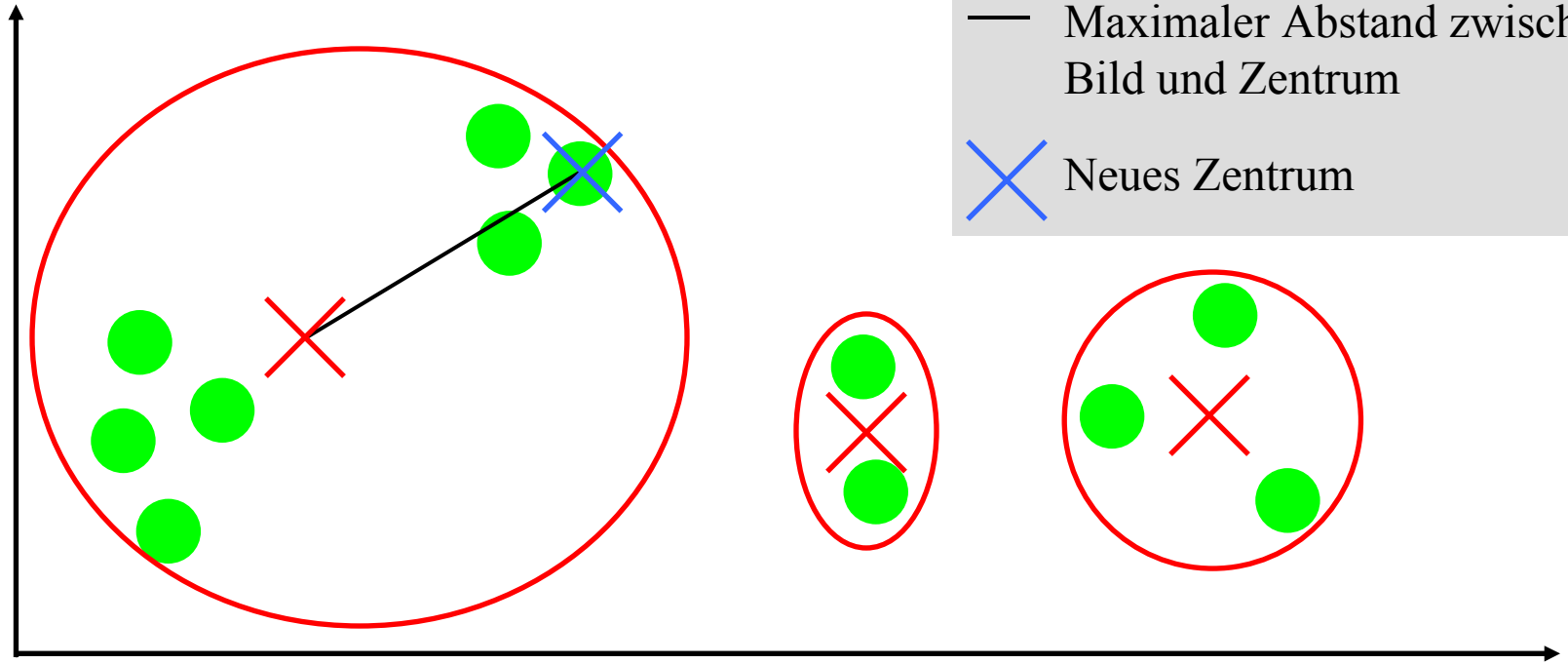
Analyse des Videos (VII)

Bsp.: Zuordnung zu Gruppen



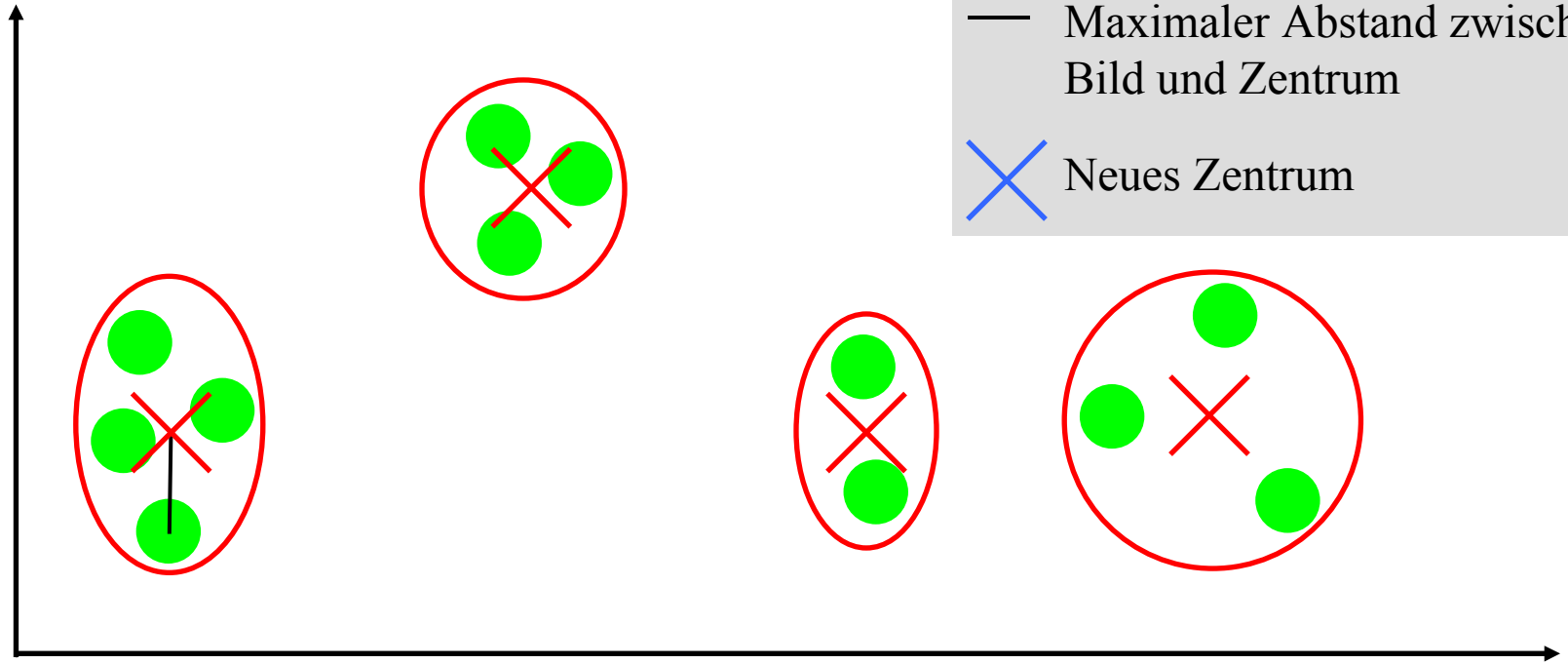
Analyse des Videos (VIII)

Bsp.: Zuordnung zu Gruppen



Analyse des Videos (IX)

Bsp.: Zuordnung zu Gruppen



Analyse des Videos (X)

Zuordnung zu Gruppen

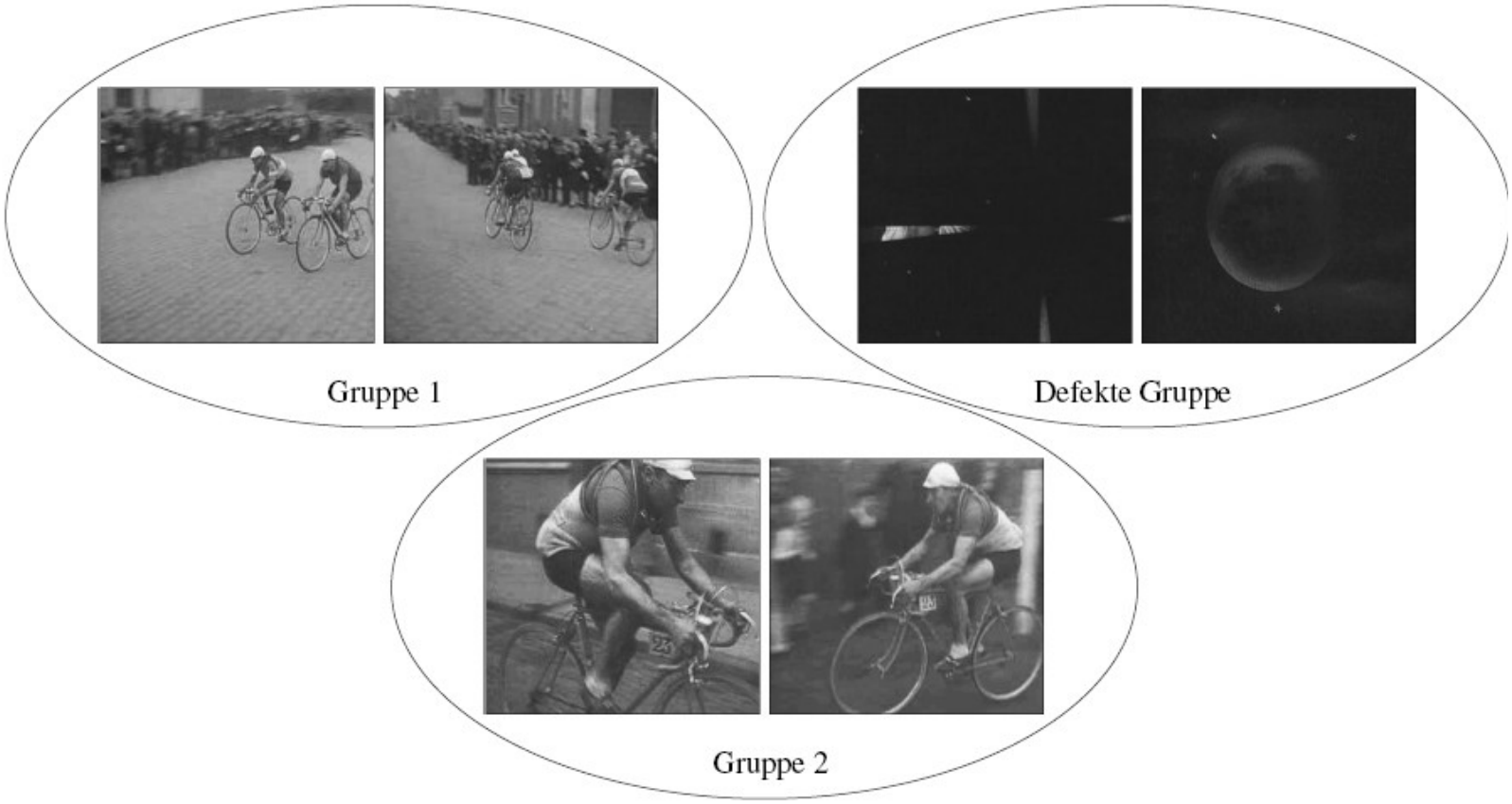
- In sehr kurzen Videos mit wenigen Kameraeinstellungen ist es möglich, dass die Anzahl der Gruppen und Kameraeinstellungen einander entsprechen.
- In Serien, Nachrichtensendungen und Sportveranstaltungen gibt es im Allgemeinen sehr große Gruppen mit vielen Kameraeinstellungen.

Erweiterung des Algorithmus zur Gruppierung von Kameraeinstellungen

- In vielen historischen Videos sind einzelne Bilder und längere Segmente des Videos beschädigt. Fehlerhafte Bereiche sollen nicht in der Zusammenfassung enthalten sein.
- Zur Identifikation fehlerhafter Kameraeinstellungen werden einzelne Zentren festgelegt, die *auf keinen Fall* in der Zusammenfassung enthalten sein sollen.
- Wird ein Bild solch einer Gruppe (*defekte Gruppe*) zugeordnet, so bleibt die entsprechende Kameraeinstellung für die Zusammenfassung unberücksichtigt.

Analyse des Videos (XI)

Erkennung fehlerhafter Gruppen



Analyse des Videos (XII)

Beispiel: Segmentierung und Erkennung von Objekten

Analyse des Videos (XIII)

Kamerabewegung

- Kamerabewegungen und Kameraoperationen geben Hinweise auf besonders wichtige Segmente des Videos.
  - Bei einem eingehenden Zoomeffekt ist häufig das Objekt im Bildzentrum von zentraler Bedeutung.
  - Vertikale Schwenks werden sehr selten eingesetzt und lenken die Aufmerksamkeit auf die Umgebung bzw. den Bildhintergrund.
- Semantische Beschreibung der Kamerabewegung innerhalb einer Kameraeinstellung wird ermittelt (Schwenks, Zoomeffekte, Rotationen).
- Verwackelte Aufnahmen bleiben unberücksichtigt.

Analyse des Videos (XIV)

Bewegungsaktivität

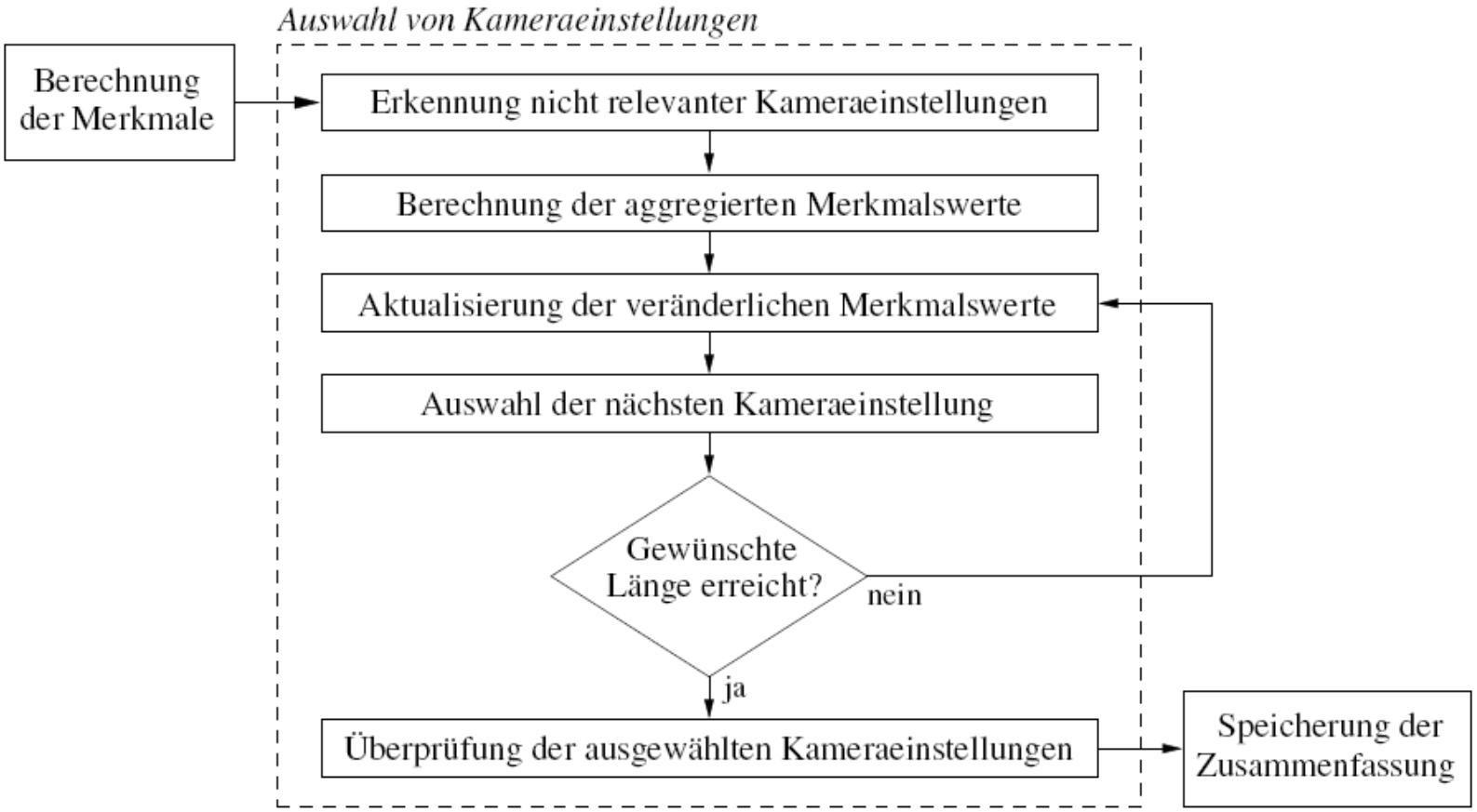
- **Annahme:** Kameraeinstellungen mit starken Bewegungen sind besonders wichtig, da mehrere unterschiedliche Bildinhalte pro Zeitintervall gezeigt werden.
- Eine deutliche Änderung zwischen zwei benachbarten Bildern innerhalb einer Kameraeinstellung kann auf eine schnelle Kamerabewegung, eine Objektbewegung eines großen Objektes oder auf besondere Ereignisse (Lichtänderungen, Feuer, Explosionen) zurückgeführt werden.
- Erstes Verfahren:
  Summe der absoluten Pixeldifferenzen zweier benachbarter Bilder (hohe Werte bei Helligkeitsänderungen, Feuer oder Explosionen)
- Zweites Verfahren:
  Durchschnittliche Länge der Bewegungsvektoren (hohe Werte bei schneller Kamerabewegung).

Analyse des Videos (XV)

Gesichter und Objekte

- Große *Gesichter* oder *Objekte* im Bildzentrum einer Kameraeinstellung haben in Dokumentationen häufig eine besondere Bedeutung.
- In Spielfilmen werden häufig die Hauptdarsteller in Nahaufnahme gezeigt.
- In den historischen Dokumentationen werden häufig bekannte Persönlichkeiten gezeigt (Sportler, Wissenschaftler, Politiker).
- Objekte liefern weitere wichtige semantische Informationen über ein Video (Art des Videos, z.B. Sportereignisse).
- Wird ein Objekt besonders häufig im Video erkannt, so sollte es auch in der Zusammenfassung erscheinen.

Auswahl von Kameraeinstellungen (I)



Auswahl von Kameraeinstellungen (II)

Idee

- Berechne *aggregierte Merkmalswerte* berechnet, welche die Informationen auf einen Wert im Intervall [0, 1] abbilden und eine Bewertung von Kameraeinstellungen ermöglichen.
- Der größte Teil der aggregiertenMerkmalswerte wird nur einmal berechnet und ändert sich während der Auswahl der Kameraeinstellungen nicht.
- *Veränderliche Merkmale* müssen nach jeder neu ausgewählten Kameraeinstellung aktualisiert werden.

Auswahl von Kameraeinstellungen (III)

Merkmale	Verfügbare Informationen	Zeitintervall	veränderliches Merkmal
Kamera-bewegung	Art der Kamerabewegung (Zoom, Schwenk), Stärke der Bewegung	Teil einer Kameraeinstellung	nein
Bewegungs-aktivität	Umfang der Bewegungsaktivität	Bild	nein
Gesicht	Größe, Position, Rotationswinkel	Bild	nein
Objekt	Größe, Objektname, Name der Objektklasse, Zuverlässigkeit	Bild	nein
Kontrast	Kontrast eines Bildes	Bild	nein
Gruppen ähnlicher Kamera-einstellungen	Liste mit Kameraeinstellungen	Kameraeinstellung	ja
Szene	Liste mit Kameraeinstellungen	Kameraeinstellung	ja
Zeitliche Verteilung	Entfernung zur nächsten ausgewählten Kameraeinstellung	Kameraeinstellung	ja
Audio	Zeitintervalle der ruhigen Bereiche	Teil des Videos	nein

Auswahl von Kameraeinstellungen (IV)

Kamerabewegung

- Der aggregierte Wert zur Beschreibung der Kamerabewegung C_A wird durch die *Art* der Bewegung C_T , die *Stärke* der Kamerabewegung C_S und deren *Dauer* C_L beeinflusst:

$$C_A = \min\,(C_T + C_S + C_L,\ 1) \quad \text{mit}$$

$$C_S = \min\,(T_S \cdot V_{MV},\ 0{,}5)$$

$$C_L = \min\,(T_L \cdot V_L,\ 0{,}5)$$

Auswahl von Kameraeinstellungen (V)

Kamerabewegung

- Die geringste Bedeutung haben horizontale Schwenks und ausgehende Zoomoperationen ($C_T = 0{,}2$). Selten treten vertikale Schwenks auf, die eine stärkere Gewichtung erhalten ($C_T = 0{,}3$). Die größte Bedeutung haben eingehende Zoomoperationen ($C_T = 0{,}4$), da sie häufig wichtige Objekte im Bildzentrum zeigen.
- Die Stärke der Kamerabewegung C_S wird aus der durchschnittlichen Länge der Bewegungsvektoren des Kameramodells abgeleitet.
- Der Skalierungsfaktor T_L gewichtet die Dauer der erkannten Kamerabewegung.

Auswahl von Kameraeinstellungen (VI)

Kontrast

- In historischen Videos ist die Bildqualität zum Teil so schlecht, dass der Inhalt nur schwer oder gar nicht erkannt werden kann.
- Der Kontrast eines Bildes liefert einen guten Hinweis über die Bildqualität einer Kameraeinstellung.
- Der *aggregierte Kontrast* ist definiert als der durchschnittliche auf das Intervall $[0, 1]$ normierte Kontrast aller Bilder der Kameraeinstellung.

Auswahl von Kameraeinstellungen (VII)

Ähnlichkeit von Kameraeinstellungen

- Kameraeinstellungen mit visueller Ähnlichkeit werden gemeinsamen Gruppen zugeordnet. Um einen möglichst umfangreichen Überblick in der Zusammenfassung zu geben, sollten Kameraeinstellungen aus unterschiedlichen Gruppen ausgewählt werden.
- Die Bewertung C_i einer Gruppe i hängt von dessen Länge ab, d. h. von der Summe der Längen aller Kameraeinstellungen, die dieser Gruppe zugeordnet sind:

$$C_i = \frac{1}{\max_j\{D_j\}} \cdot \frac{D_i}{1 + S_i^2}, \qquad j = 1 \ldots N.$$

Auswahl von Kameraeinstellungen (VIII)

Ähnlichkeit von Kameraeinstellungen

- D_i definiert die Länge der Gruppe i, S_i gibt die Anzahl der bereits ausgewählten Kameraeinstellungen dieser Gruppe an.
- Die größte Gruppe innerhalb des Videos definiert den Gewichtungsfaktor zur Normierung von C_i auf das Intervall $[0,1]$.
- Mit der Auswahl einer Kameraeinstellung aus der Gruppe i erhöht sich S_i um eins, so dass für den weiteren Auswahlprozess der aggregierte Wert dieser Gruppe sinkt und bevorzugt Kameraeinstellungen aus anderen großen Gruppen ausgewählt werden.

Auswahl von Kameraeinstellungen (IX)

Szenen

- Benachbarte Kameraeinstellungen einer Szene sollen in der Zusammenfassung enthalten sein.
- Eine einzelne Kameraeinstellung liefert häufig nicht ausreichend Informationen, um den Inhalt der Szene zu verstehen.
- Andererseits wiederholen sich bei mehr als zwei ausgewählten Kameraeinstellungen einer Szene die Inhalte, und der Zugewinn an Informationen nimmt deutlich ab.

Auswahl von Kameraeinstellungen (X)

Heuristik zur Bewertung der Szenen

- Initialisiert den Wert für jede Kameraeinstellung zunächst mit 0,5.
- Reduziere den Wert, falls zwei oder mehr Kameraeinstellungen einer Szene für die Zusammenfassung ausgewählt wurden.
- Bei genau einer ausgewählten Kameraeinstellung erhalten die Werte der angrenzenden Kameraeinstellungen derselben Szene den Maximalwert von eins und die Werte der anderen Kameraeinstellungen dieser Szene werden auf null reduziert.
- Diese Heuristik begünstigt die Auswahl von genau zwei benachbarten Kameraeinstellungen.

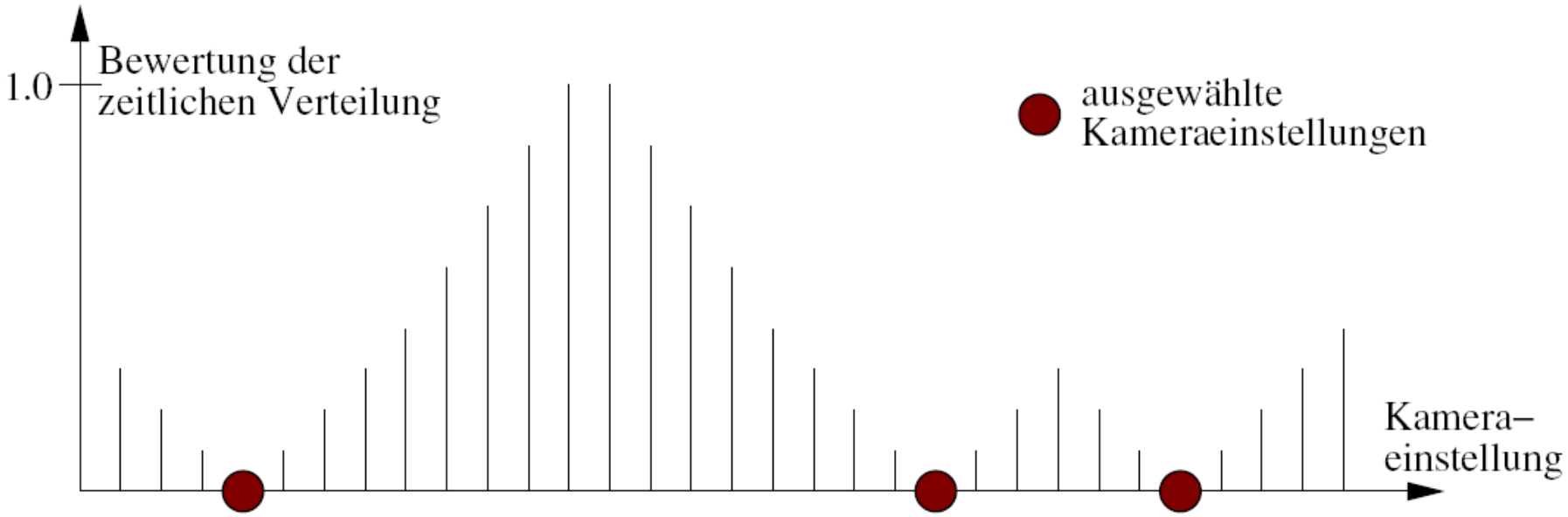
Auswahl von Kameraeinstellungen (XI)

Zeitliche Verteilung der Kameraeinstellungen

- Innerhalb einer Zusammenfassung soll der gesamte Inhalt und nicht nur einzelne Teile des Videos gezeigt werden.
- Eine gute zeitliche Verteilung ist besonders für Dokumentationen und Nachrichtensendungen wichtig, für die ein Überblick über das Video gegeben werden soll.
- Bei Spielfilmen muss diese Heuristik eingeschränkt werden, da in einer Vorschau beispielsweise das spannende Ende des Filmes nicht aufgedeckt werden soll.
- Ungeeignet ist die Heuristik zur Bewertung der zeitlichen Verteilung für Zusammenfassungen von Sportveranstaltungen, da besondere Aktionen und Ereignisse relevant sind, die nicht gleichmäßig über die gesamte Länge des Videos verteilt sind.
- Der aggregierte Wert wird aus dem Abstand der Kameraeinstellung zu der am nächsten gelegenen ausgewählten Kameraeinstellung abgeleitet und auf das Intervall $[0, 1]$ normiert.

Auswahl von Kameraeinstellungen (XII)

Zeitliche Verteilung der Kameraeinstellungen



Erzeugung einer Zusammenfassung (I)

Auswahl von Kameraeinstellungen

- Der *gewichtete Wert einer Kameraeinstellung* R_i wird definiert als:

$$R_i = \sum_j \alpha_j \cdot F_{i,j}.$$

- Der aggregierte Wert $F_{i,j}$ eines Merkmals j der Kameraeinstellung i wird mit den Faktoren j gewichtet, die individuelle Präferenzen eines Benutzers widerspiegeln.
- Die aggregierten Merkmalswerte und der gewichtete Wert werden zunächst für alle Kameraeinstellungen berechnet.
- Die Kameraeinstellung mit dem maximalen Wert für R_i wird für die Zusammenfassung ausgewählt.
- Falls die Zusammenfassung noch nicht die gewünschte Länge erreicht hat, werden die dynamischen Merkmalswerte aktualisiert, und eine weitere Kameraeinstellung wird ausgewählt.

Erzeugung einer Zusammenfassung (II)

Gesichter	0,38
Szenen	0,50
Bewegte Objekte	0,00
Kontrast	0,91
Bewegungsaktivität	0,20
Kamerabewegung	0,00
Zeitliche Verteilung	0,55
Gruppen ähnlicher Kameraeinstellungen	0,84
Summe	**3,38**

Gesichter	0,00
Szenen	0,50
Bewegte Objekte	0,00
Kontrast	0,94
Bewegungsaktivität	0,91
Kamerabewegung	0,00
Zeitliche Verteilung	1,00
Gruppen ähnlicher Kameraeinstellungen	0,69
Summe	**4,04**

Gesichter	0.00
Szenen	0.50
Bewegte Objekte	0.00
Kontrast	0.32
Bewegungsaktivität	0.09
Kamerabewegung	0.00
Zeitliche Verteilung	0.48
Gruppen ähnlicher Kameraeinstellungen	0.53
Summe	**1,92**

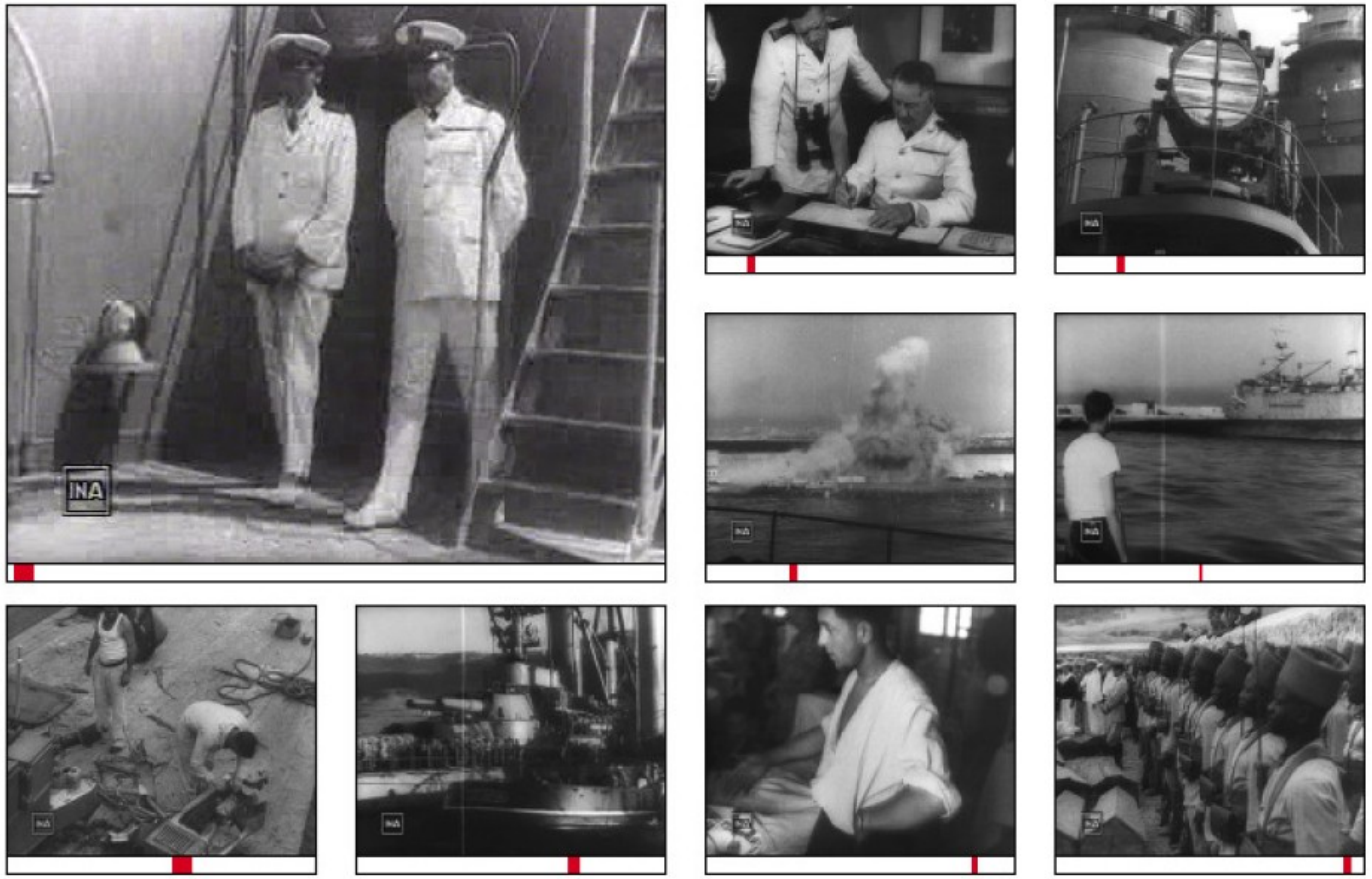
Erzeugung einer Zusammenfassung (III)

Überprüfung der ausgewählten Kameraeinstellungen

- Direkt aufeinander folgende Kamerabewegungen erzeugen einen unprofessionellen Eindruck des Videos, so dass Kameraeinstellungen mit deutlichen Kameraoperationen an Aufnahmen mit statischer Kamera angrenzen sollten.
- Zum besseren Verständnis der Handlung sollten mindestens zwei Kameraeinstellungen einer Szene ausgewählt werden.
- Die durchschnittliche Bewegungsaktivität sollte in der Zusammenfassung nicht wesentlich höher als im Originalvideo sein.
- Die Länge der Zusammenfassung sollte ungefähr der durch den Benutzer spezifizierten Länge entsprechen.
- Die Audiospur sollte nur in ruhigen Bereichen geschnitten werden.

Ergebnisse (I)

Beispiel einer statischen Zusammenfassung



Ergebnisse (II)

Beispiel einer statischen Zusammenfassung

Ergebnisse (III)

Beispiel einer Zusammenfassung in Form einer Kollage

Ergebnisse (IV)

Beispiel einer dynamischen Zusammenfassung

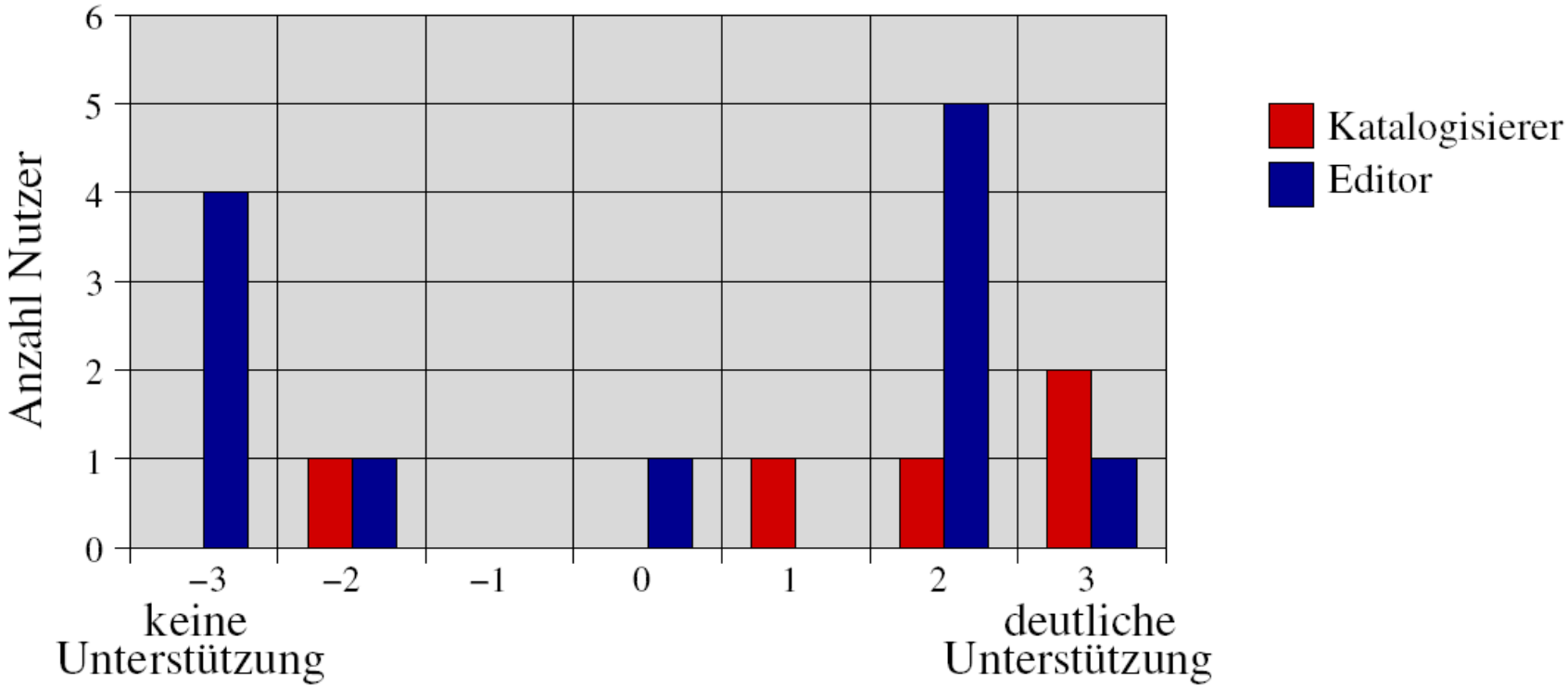
Video-Zusammenfassung

nicht relevante Kameraeinstellungen
(sehr geringe Merkmalswerte)

Ergebnisse (V)

Ergebnisse von Benutzerbefragungen

”Erwarten Sie, dass automatisch erzeugte Zusammenfassungen Ihre Arbeit unterstützen werden?”



Ergebnisse (VI)

Es ist nicht möglich die optimale Zusammenfassung eines Videos zu erzeugen. Zwei Personen würden unterschiedliche Kameraeinstellungen für Zusammenfassungen auswählen. Eine automatisch erzeugte Zusammenfassung wird eine dritte Auswahl treffen, die auch nicht notwendigerweise optimal ist.

Fragen ?