Ralph Ewerth; Bernd Freisleben

Computerunterstützte Filmanalyse mit Videana

2007

https://doi.org/10.25969/mediarep/1926

Veröffentlichungsversion / published version
Zeitschriftenartikel / journal article

Empfohlene Zitierung / Suggested Citation:
Ewerth, Ralph; Freisleben, Bernd: Computerunterstützte Filmanalyse mit Videana. In: *Augen-Blick. Marburger Hefte zur Medienwissenschaft*. Heft 39: Technisierung des Blicks (2007), S. 54–66. DOI: https://doi.org/10.25969/mediarep/1926.

Ralph Ewerth und Bernd Freisleben

Computerunterstützte Filmanalyse mit *Videana*

Einleitung

Das Teilprojekt „Methoden und Werkzeuge zur rechnergestützten medienwissenschaftlichen Analyse“ des Kulturwissenschaftlichen Forschungskollegs SFB/FK 615 „Medienumbrüche“ entwickelt zum einen „Mediana“, ein an die Bedürfnisse der medienwissenschaftlichen Projektpartner angepaßtes Datenbanksystem mit dem Ziel, beliebige textuelle und audiovisuelle Datenobjekte zu verwalten und so die medienwissenschaftlichen Arbeitsprozesse zu unterstützen. Zum anderen werden rechnergestützte Verfahren zur medienwissenschaftlichen Analyse digitaler Bild- und Videodaten erforscht und entwickelt, die in die Mediana-Komponente *Videana* integriert sind.

In diesem Beitrag wird gezeigt, wie medienwissenschaftliche Filmanalyse durch rechnergestützte Verfahren quantitativ unterstützt werden kann. Die qualitative Analyse und die Interpretation des in Filmen Gezeigten wird auf absehbare Zeit ausschließlich Menschen vorbehalten sein. Nichtsdestotrotz können Rechner den Menschen bei einigen typischerweise sehr zeitraubenden Prozessen unterstützen bzw. entlasten. Vor allem wird hier die quantitative Analyse folgender Elemente der filmischen Gestaltung betrachtet: Montage der Einstellungen, Kamerabewegung sowie eine Beschreibung des Gezeigten. Zu Letzterem werden speziell Verfahren zum Finden und Erkennen von Gesichtern und Texteinblendungen vorgestellt.

Korte (2001) beschreibt verschiedene Elemente der filmischen Gestaltung, grundlegende Elemente von Einstellungs-/Sequenzprotokollen sowie einige Visualisierungen wie etwa Einstellungsgrafiken, Sequenzdiagramme und Schnittfrequenzdiagramme. Bereits seit einigen Jahren gibt es Softwarepakete, die bei der Erfassung von Kriterien der Filmanalyse in Form eines Einstellungsprotokolls für analog vorliegendes Filmmaterial helfen bzw. zur vereinfachten automatischen Generierung von Visualisierungen (z.B. Szenen- bzw. Sequenzgrafik

oder Schnittfrequenzgrafik) dienen. Zu nennen sind etwa das vom Institut für Medien der Universität Marburg (Giesenfeld) entwickelte System *filmprot* und das an der Hochschule für Bildende Künste Braunschweig entwickelte System *CNfA* (Computergestützte Notation filmischer Abläufe, Korte). Allerdings wurden diese Softwarelösungen u. a. für spezielle Hardware und Videorekorder entwickelt, so daß sie heute zum Teil nicht mehr verfügbar sind. Aktuellere Entwicklungen sind das Programm *Akira* (Kloepfer, Universität Mannheim) sowie das Programm *VideoAS* (Olbrecht/Woelke, Universität Jena), die für das Annotieren von digital vorliegenden Videos konzipiert sind. Allerdings bieten diese keine Möglichkeiten zur automatischen Analyse von Videos.

Abbildung 1 zeigt das Hauptfenster von *Videana*. Links ist ein Fenster zur Wiedergabe des Videos. Auf den beiden Zeitleisten ist die Segmentierung des Videos in Einstellungen sowie das Ergebnis der Gesichtsdetektion visualisiert. Die vertikalen Striche in der Zeitleiste *Cuts* symbolisieren die Schnitte, die Flächen in der Zeitleiste *Faces* die Sequenzen, in denen ein Gesicht frontal gezeigt wurde. Für jedes Ereignis werden zwei Zeitleisten präsentiert: Die jeweils obere Zeitleiste repräsentiert die komplette Dauer des Videos, wohingegen die untere Zeitleiste den oben gestrichelt umrahmten Zeitbereich vergrößert darstellt. Weitere Zeitleisten für die Ereignisse Kamerabewegung und Texteinblendungen kommen hinzu, sofern Analyseergebnisse oder Benutzermarkierungen vorliegen.

Die Analyse von Multimediadaten ist im Übrigen nicht nur für medienwissenschaftliche Zwecke interessant. Mit zunehmenden Rechner- und Datenkapazitäten heutiger Rechner und der inzwischen umfangreicheren Ausstattung mobiler Geräte (MP3-Player, Mobiltelefone etc.) haben multimediale Dateien eine immense Verbreitung gefunden. So ist seit geraumer Zeit die effiziente Suche nach Informationen in Videos bzw. generell in multimedialen Datenbeständen ein von vielen Forschern weltweit mit hoher Aufmerksamkeit verfolgtes Forschungsgebiet.

Rechts neben dem Wiedergabefenster sind die einzelnen Einstellungen durch jeweils drei kleine Bilder symbolisiert (Anfang, Mitte, und Ende der Einstellung). Durch einfaches Anklicken mit der Maus kann zu der jeweiligen Position im Video gesprungen werden, durch Doppelklick wird das Video an dieser Stelle gestartet.

Das Video- und Filmanalysewerkzeug Videana

Wie bereits einleitend erwähnt, wird der größte Vorteil der Rechnerunterstützung in der Automatisierung formaler, zeitintensiver Analyseschritte gesehen. Zu nennen ist etwa die zeitliche Segmentierung eines Videos in Kameraeinstellungen, hierbei die Identifikation der Montageart (harter Schnitt, Aufblende, Abblende, Überblende etc.), das Finden und Erkennen von eingeblendetem Text, das Erkennen von Kamera- und Objektbewegung, das Erkennen der verwendeten Einstellungsgröße, Informationen über die Präsenz der Akteure, die Art der auditiven Signale etc. In *Videana* sind gegenwärtig Funktionen zur Schnittdetektion, zum Finden von eingeblendetem Text, zur Bestimmung der Kamerabewegung und zur Detektion und Wiedererkennung von frontal erscheinenden Gesichtern realisiert. Die graphische Benutzerschnittstelle von *Videana* ermöglicht, Videos anzuzeigen, abzuspielen und bildgenau an eine bestimmte Position des Videos zu gelangen. Weiterhin sind Algorithmen zur Schnittdetektion, Textdetektion und -segmentierung (notwendig für eine anschließende Erkennung mit einer Optical Character Recognition (OCR) Software), Gesichtsdetektion und der Bestimmung der Kamerabewegung integriert worden, die mittels eines „Plugin-Konzepts“ auf einfache Art und Weise ergänzt, entfernt oder ausgetauscht werden können. Automatisch erzeugte Analyseergebnisse können jederzeit von Benutzerseite wieder manuell korrigiert werden.

Sobald eine zeitliche Segmentierung in Einstellungen erfolgt ist, wird für das erste, mittlere und letzte Einzelbild jeder Einstellung (optional für Szenen)

je ein „Icon“ erstellt und angezeigt (siehe Abbildung 1). Diese Ansicht kann auch auf Szenenbasis gezeigt werden, zum jetzigen Zeitpunkt müssen Einstellungen allerdings noch manuell von Benutzerseite zu einer Szene zusammengefaßt werden.

Es können Diagramme bezüglich der Schnittfrequenz und der Helligkeitsdynamik innerhalb eines Videos generiert werden. In Abbildung 2 ist ein solches Schnittfrequenzdiagramm für einen 30-minütigen Videoausschnitt zu sehen. Die Ergebnisse der verschiedenen Detektoren werden auf einer jeweils eigenen Zeitleiste visualisiert. Einzelne Kameraeineinstellungen, Ereignisse wie graduelle Übergänge, und Text- oder Gesichtsobjekte können mit beliebigen Kommentaren und mit Schlüsselwörtern annotiert werden. Sowohl die extrahierten Multimedia-Metadaten, als auch die während des medienwissenschaftlichen Arbeitsprozesses manuell ergänzten Kommentare werden dem MPEG-7 Standard (Martinez, 2002) konform gespeichert. Der MPEG-7 Standard formalisiert die Darstellung solcher Daten und ermöglicht den Datenaustausch zwischen unterschiedlichen Anwendungen.

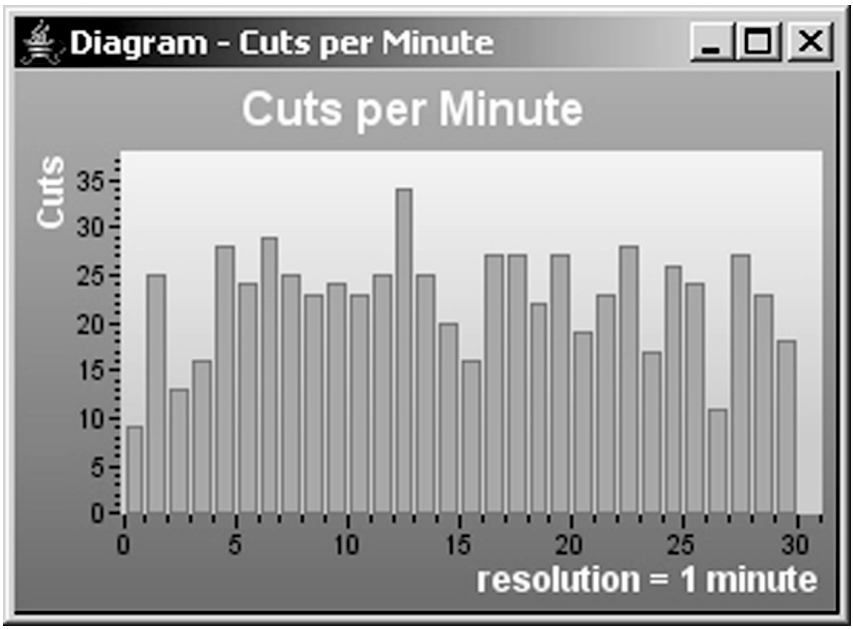

Abbildung 2 Ausgabe eines Schnittfrequenzdiagramms in Videana für einen 30-minütigen Ausschnitt aus einem Spielfilm.

Erkennung von Schnitten und graduellen Einstellungswechseln

Eine der wichtigsten Aufgaben der digitalen Videoanalyse ist die Unterteilung einer Videosequenz in ihre grundlegenden Einheiten, die Einstellungen („Shots"). Unter einer Einstellung wird im Allgemeinen eine audiovisuelle Sequenz verstanden, die eine kontinuierliche Aufzeichnung ohne Unterbrechung der Aufnahme darstellt. Die Übergänge (oder Transitionen) zwischen Einstellungen können abrupt oder graduell sein; abrupte Übergänge werden auch als (harte) Schnitte bezeichnet, graduelle Übergänge resultieren aus dem Einsatz chromatischer oder räumlicher Editiereffekte, wie z.B. Ein- bzw. Ausblendungen („Fade in/out"), Überblendungen („Dissolve") oder Verschiebungen („Wipe").

Seit Anfang der 1990er Jahre wurde eine Vielzahl von Segmentierungsverfahren (bekanntere Ansätze stammen z. B. von Yeo/Liu (1995), Hanjalic (2002) oder Bescos (2004)) vorgeschlagen, insbesondere auch für die Schnitterkennung. Zur Erkennung gradueller Übergänge gibt es sowohl allgemeine Ansätze als auch auf bestimmte Effekte (wie etwa Überblendung (Hanjalic, 2002), Ein- und Ausblende (Truong et al., 2000)) spezialisierte Detektoren. Viele der Ansätze zur Schnitterkennung basierten auf dem Vergleich von zwei aufeinander folgenden Einzelbildern („Frames"). Jüngere Ansätze (Tahaghoghi et al. (2005), Yuan et al. (2005)) vergleichen alle Bilder innerhalb eines kurzen Zeitfensters miteinander, um so zu robusteren Ergebnissen zu kommen. In der seit 2001 jährlich durchgeführten Vergleichsstudie TRECVID konnten im Jahr 2005 solche Ansätze die besten Erkennungsraten erzielen: so wurden ca. 95% der Schnitte gefunden (Erkennungsrate, „Recall"), und ebenfalls 95% aller von diesen Detektoren gemeldeten Schnitte waren auch tatsächlich solche (Präzision des Ergebnisses, „Precision"). Der von den Autoren entwickelte Ansatz (Ewerth/Freisleben, 2004) gehörte bei dieser Studie, an der 21 Institute aus aller Welt teilnahmen, zu den 5 Ansätzen, die sowohl eine Erkennungsrate als auch eine Präzision von mindestens 90% erreichen konnten. Die Erkennung gradueller Übergänge hat noch nicht diese Güte erreicht. Hier liegen die Erkennungsrate und die Präzision der besten Ansätze (Amir et al. 2005; Yuan et al., 2005) bei ca. 80%.

Finden und Erkennen von eingeblendetem Text

Eingeblendeter Text gibt oftmals wichtige Hinweise über das im Bild Gezeigte. So sind z. B. in Nachrichtensendungen die Texthinweise eng mit dem aktuellen Nachrichtenbeitrag verknüpft, in frühen Filmen (ohne Sprache) wiederum wurde das Gezeigte mit Zwischentiteln ergänzt. Begrifflich sind die Algorithmen

zu unterscheiden darin, ob es sich um einen Ansatz zur Textdetektion, Textlokalisierung, Textverfolgung (in Videos), Textsegmentierung (auch Textextraktion genannt) oder Texterkennung handelt (Jung et al., 2004). Ein Textdetektor liefert als Ergebnis Informationen darüber, ob und ggf. wo sich in einem Bild oder in einer Kameraeinstellung Text befindet. Die Textsegmentierung hingegen verarbeitet lokalisierten Text dergestalt, daß der Hintergrund des Textes entfernt wird, so daß schließlich ein Ergebnisbild entsteht, das schwarzen Text auf einem weißen Hintergrund zeigt. Dies ist notwendig, um mit einer OCR Software, die ein Bild mit dem gezeigten Text in eine maschinenlesbare Form transformiert, ein optimales Erkennungsergebnis erreichen zu können. Etwaige Hintergrundinformationen beeinträchtigen in der Regel das Erkennungsresultat. In Abbildung 3 werden beispielhaft Ergebnisse von Textlokalisierung, Textsegmentierung und -erkennung gezeigt.

Abbildung 3 In den Bildern links sind Textlokalisierungsergebnisse zu sehen, in der Mitte das Ergebnis der Textsegmentierung, im Zuge derer der Bildhintergrund entfernt und der Text schwarz markiert wurde, und schließlich rechts die Ergebnisse der Erkennungssoftware (OCR).

Forschung im Bereich der automatischen Text- bzw. Zeichenerkennung (OCR) in Dokumenten (in der Regel handelt es sich hier um „gescannte" Textseiten) wird schon seit Jahrzehnten betrieben. Textdetektion, Textsegmentierung und schließlich Texterkennung in Bildern und Videos sind inzwischen auch seit mehr als 10 Jahren Gegenstand der Forschung; eine Vielzahl von Methoden ist hierzu entstanden. Jung et al. (2004) geben einen Überblick über die Forschungsarbeiten auf diesem Gebiet. In der Arbeitsgruppe der Autoren wurden sowohl Ansätze zur Textdetektion (Gllavata/Ewerth/Freisleben, 2004a) und Textsegmentierung (Gllavata/Ewerth/Stefi/Freisleben, 2004; Gllavata/

Freisleben, 2005) als auch für das Verfolgen von bewegten Texteinblendungen über mehrere Einzelbilder hinweg (Gllavata/Ewerth/Freisleben, 2004b) entwickelt. Auf einer Testmenge von Bildern konnte mittels eines vorgeschlagenen Ansatzes zur Segmentierung von Text die Worterkennungsrate (Zeichenerkennungsrate) von 62% auf 79% (bzw. von 76% auf 91%) verbessert werden (Gllavata/Freisleben, 2005). Momentan wird an der Integration einer OCR Software in *Videana* gearbeitet: Nach erfolgter Integration wird das System in der Lage sein, entsprechende Kameraeinstellungen automatisch mit den gefundenen und erkannten Wörtern zu annotieren.

Finden und Erkennen von Gesichtern

Im Bereich der Gesichtsverarbeitung in Bildern und Videos ist zu unterscheiden zwischen Gesichtsdetektion und –erkennung. Ein umfassender Überblick zur Gesichtsdetektion wird von Yang et al. (2002) gegeben, ein umfassender Übersichtsartikel zur Gesichtserkennung wurde von Zhao et al. (2003) veröffentlicht. Der Begriff Gesichtsdetektion wird analog zu Textdetektion verwendet: Ein Gesichtsdetektor gibt an, ob sich in einem Bild oder in einer Einstellung ein Gesicht befindet, in der Regel ist dies auch eng mit der Gesichtslokalisierung verbunden. In *Videana* wurde ein Ansatz von Viola und Jones (2004) zur Detektion frontal erscheinender Gesichter integriert, der in der *Intel Open Source Computer Vision Library* (*OpenCV*) verfügbar ist. Für diesen werden auf relevanten Standard-Testmengen (130 Bilder mit insgesamt 507 Gesichtern) Detektionsraten von 92.1% berichtet, bei einer Anzahl von 50 falschen Treffern.

Im Bereich der Gesichtserkennung wird zwischen den folgenden Anwendungsszenarien unterschieden:

Identifikation
: Es wird die Identität einer dem System präsentierten Aufnahme eines Gesichts in einer Datenbank gesucht, bzw. wird das Gesicht als unbekannt klassifiziert.

Verifikation
: In diesem Fall wird überprüft, ob die dem System präsentierte Aufnahme der angegebenen Identität entspricht.

Im Rahmen des *Face Recognition Vendor Test* (*FRVT*) aus dem Jahr 2005 hat sich gezeigt (Phillips et al., 2005), daß Gesichtserkennung unter bestimmten Bedingungen Erfolgsraten von über 90% hat. So existieren seit einigen Jahren auch kommerzielle Lösungen mit Gesichtserkennungstechnologie. Zusam-

menfassend kommt die Studie zu folgenden Schlußfolgerungen:

- Die besten Systeme erreichten bei Innenaufnahmen eine Identifikationsrate von 90%, bei einer Rate falscher Alarme von 1%.
- Die besseren Gesichtserkennungssysteme waren nicht sensitiv bezüglich normaler Beleuchtungsänderungen bei Innenaufnahmen.
- Dreidimensionale Modelle (zum Morphen einer Pose in eine frontale Position) verbesserten die Erkennungsrate.
- Die Erkennung von Personen in Bildern, die außen aufgenommen wurden, funktioniert noch nicht zufriedenstellend (Identifikationsrate von 50% bei 1% Fehlerrate).
- Es gab keinen Unterschied in der Erkennungsleistung in Abhängigkeit von der Aufnahmequelle (Videosequenzen vs. Einzelbilder).
- Jüngere Personen sind schwieriger zu erkennen als ältere Personen.
- Männer wurden besser erkannt als Frauen.
- Die Identifikationsleistung verringert sich linear im Verhältnis zum Logarithmus der Datenbankgröße (Anzahl der Personen).

Aus diesen Ergebnissen ist bereits ersichtlich, daß es schwierig ist, Videos mit dem Auftreten von Personen zu indexieren, insbesondere dann, wenn Außenaufnahmen in einem Video vorkommen. Das Problem der Außenaufnahmen sowie das Ausnutzen der Vielzahl von Einzelbildern in einer Videosequenz sind sicher zukünftig Gegenstand der Forschung.

Auch wenn ein Gesichtsidentifikationssystem für medienwissenschaftliche Zwecke interessant sein kann (“In welchen Kameraeinstellungen war eine gegebene Person zu sehen?“), wurde im Rahmen des Projekts zunächst ein allgemeines System entwickelt (Ewerth/Mühling/Freisleben, 2006), das für ein beliebiges Video einen Index über das Auftreten verschiedener Personen erstellt. Voraussetzung ist lediglich eine Segmentierung des Videos in Kameraeinstellungen, optional auch eine Szenensegmentierung. Das Ergebnis ist eine Menge von Personen, für die jeweils eine Liste mit den Nummern der Einstellungen ausgegeben wird, in denen eine Person X (aus der Menge aller Personen) zu sehen war. Das System ist prinzipiell sowohl für das Finden und Erkennen von frontal als auch von im Profil gezeigten Gesichtern konzipiert. Allerdings ist der Detektor der verwendeten *OpenCV Library* (*OpenCV*) für Profilgesichter noch nicht ausgereift genug, so daß momentan nur frontal gezeigte Gesichter verarbeitet werden. Die recht präzise räumliche Detektion von frontal gezeigten Gesichtern, insbesondere der Augenpositionen, ist hingegen ein guter Anhaltspunkt, um Rotation von Gesichtern in der Bildebene, wie sie durch einen zur Seite geneigten Kopf entstehen, zu korrigieren (Abbildung 4). Nach einer ersten Gruppierungsphase werden die Gruppen der Personen, die in

mehr als einer vorher definierten Mindestanzahl von Einstellungen zu sehen waren, einer weiteren Analyse unterzogen. Ziel dieser Analyse ist es, die Merkmale eines Gesichts in einer Gruppe zu bestimmen, die es am besten von den Gesichtern der anderen Gruppen unterscheiden. Schließlich wird aufgrund dieser für jede Gruppe separat selektierten Merkmale ein erneuter Klassifikationsprozeß durchgeführt. Erste Ergebnisse für das Erkennen von frontal gezeigten Gesichtern sind sehr viel versprechend, insbesondere konnten die Ergebnisse durch die Korrektur der Rotation geneigter Gesichter und dem anschließenden Lernen der charakteristischen Gesichtsmerkmale signifikant verbessert werden: Für einen Ausschnitt aus einer Fernsehdiskussionsrunde wurden für 5 der 6 gezeigten Personen hinreichend große Gruppen (Cluster) erzeugt, so daß sie als ein Repräsentant einer Person zum Lernen der Gesichtsmerkmale genutzt wurden. Die Erkennungsrate betrug im besten Fall 84% bei einer Präzision der Personengruppierungen von 94% (d. h. nur 6% der einer Gruppe zugeordneten Personen entsprachen nicht der Hauptperson der Gruppe), das Basissystem erreichte bei gleicher Präzision lediglich eine Erkennungsrate von 71%.

Abbildung 4, obere Reihe: Beispiele von geneigten Köpfen, die unter anderem zu einer in der Bildebene rotierten Gesichtsdarstellung führen. In der unteren Reihe sind die Gesichter zu sehen, nachdem sie von dem Gesichtserkennungssystem in *Videana* anhand der Augenpositionen zurückgedreht wurden. Dies ist wichtig für den späteren Vergleich zwischen zwei Gesichtern.

Erkennung von Kamerabewegungen

In der Filmgestaltung ist der Einsatz der Kamera ein wesentliches Mittel des ästhetischen Ausdrucks. Formate zur Videokompression wie MPEG-1 oder MPEG-2 unterstützen eine Bewegungsschätzung auf Pixelblockbasis für aufeinander folgende Videobilder, um die große zeitliche Redundanz in Videos für die Kompression der Daten auszunutzen. Die Rechenzeiten für die Extraktion solcher Bewegungsvektoren sind im Vergleich zu der Dekodierung eines

Vollbildes und der Berechnung eines optischen Flußfeldes (Berechnung der Bewegung für jedes Pixel) sehr gering. Allerdings ist ein großer Teil dieser Vektoren häufig „verrauscht“ und nicht optimal im Sinne einer Bewegungsbeschreibung. Aufbauend auf diesen Beobachtungen wurde ein eigener Ansatz (Ewerth/Schwalb/Tessmann/Freisleben, 2004) entwickelt, der MPEG-Bewegungsvektoren zur Berechnung der Kameraparameter verwendet. Die „unzuverlässigen“ Bewegungsvektoren eines Vektorfeldes werden zunächst in einem Vorverarbeitungsschritt mit einer effektiven Methode entfernt („Outlier Removal“). Mit den verbleibenden Bewegungsvektoren werden mit Hilfe des Nelder-Meade Minimierungsalgorithmus die Parameter eines 3D-Kameramodells geschätzt. Das verwendete Modell hat den Vorteil, daß es prinzipiell die Unterscheidung zwischen Translation und Rotation der Kamera (in der entsprechenden Richtung) zuläßt. Experimente mit aufwendig hergestellten, synthetischen Videosequenzen konnten zeigen, daß das Entfernen der unzuverlässigen Bewegungsvektoren zu deutlich besseren Ergebnissen führt. So konnte für Zoom-In und Zoom-Out eine Erkennungsrate und eine Präzision von 99% (98% und 94% ohne „Outlier Removal“) erreicht werden, die Ergebnisse für die Rotation um die z-Achse verbesserten sich von 86% auf 95% (Erkennungsrate) und von 75% auf 89% (Präzision). Mit diesem System haben die Autoren auch an der TRECVID Evaluation 2005 teilgenommen, wobei an dem sogenannten „low-level-feature task“ zur Kamerabewegung insgesamt 12 Institute teilgenommen haben. Für diese Evaluation waren insgesamt 140 Nachrichtenvideos mit einer jeweiligen Dauer von 30 bis 60 Minuten zu analysieren. Die eingereichten Ergebnisse sollten all die Kameraeinstellungen enthalten, die horizontale, vertikale Kamerabewegung oder einen Zoom (in/out) beinhalteten. Aus diesen 140 Videos wurden seitens der Veranstalter letztendlich ca. 2000 Kameraeinstellungen zur Auswertung ausgewählt, die eine Bewegung oder Zoom eindeutig bzw. eindeutig nicht erkennen ließen. Neben guten Ergebnissen bei der Erkennung horizontaler Bewegung (76% Erkennungsrate, 92% Präzision) konnte das System der Autoren das zweitbeste Ergebnis bzgl. vertikaler Bewegung (72% Erkennungsrate, 96% Präzision) und das beste Ergebnis bei der Erkennung von Kamerazooms (89% Erkennungsrate, 93% Präzision) erreichen.

Anwendungsbeispiel: Analyse von Computerspielen

Das Projekt B9 „Mediennarrationen und Medienspiele“ (R. Leschke) des SFB *Medienumbrüche* untersucht die Hybridformen von Spiel und Erzählung, die sich in Computerspielen und Spielfilmen seit den 1990er Jahren verstärkt beo-

bachten lassen. Diese sollen formalästhetisch und funktionslogisch analysiert und in einer Typologie zusammengefaßt werden. Neben der Unterstützung dieser Forschungsaktivitäten mit Hilfe der Basisfunktionalität von *Videana* wird gegenwärtig ein System entwickelt, das die zugrundeliegenden Merkmale narrativer und spielerischer Sequenzen in Computerspielen und Spielfilmen automatisch erlernen soll. Hierzu können neben den Kategorien „Spiel" und „Narration" weitere spezielle Kategorien wie z. B. Gewalt oder Suche (Exploration) definiert werden, darüber hinaus ist mit diesem System aber durchaus die Definition beliebiger Kategorien möglich.

Zum Lernen und Klassifizieren werden dem System die Ergebnisse der vorgenannten Ansätze zur Schnitterkennung, Textdetektion, Gesichtsdetektion und Kamerabewegung sowie Informationen über die Farbverteilung in jedem Einzelbild präsentiert. Erste Experimente zeitigten interessante Ergebnisse: So wurde zunächst eine aufgenommene interaktive Spielsequenz des Spiels „Max Payne" genutzt, um die Eigenschaften auf Ebene der automatisch extrahierbaren Merkmale von narrativen Sequenzen und von interaktiven spielerischen Sequenzen (mit den Unterkategorien Suche/Exploration und Gewalt) zu lernen. Das erlernte Modell wurde dann angewendet auf einen 30-minütigen Ausschnitt des Films „Matrix Reloaded", der einige Sequenzen enthält, die an Computerspiele erinnern. Interessanterweise werden aber fast alle Einzelbilder vom trainierten System als narrativ (d.h. nicht interaktiv) markiert (wie es ja auch formal korrekt ist, schließlich handelt es sich ja um einen Spielfilm), bis auf ein paar wenige Ausnahmen: So wurden einige Einzelbilder einer sehr langen Kampfsequenz sowie einer weiteren „Actionsequenz" als nicht narrativ klassifiziert. In einem nächsten Schritt kann nun analysiert werden, welche Merkmale diese Sequenzen von den anderen auf der Ebene der Signalverarbeitung und des maschinellen Lernens unterscheiden. Es obliegt den Medienwissenschaftlern, diese Ergebnisse zu interpretieren und ggf. weiter zu verfolgen oder zu verwerfen.

Zusammenfassung und Fazit

In diesem Beitrag wurde das in der Arbeitsgruppe der Autoren entwickelte Softwaresystem *Videana* vorgestellt, das der Unterstützung der medienwissenschaftlichen Filmanalyse dient. Neben einer Vorstellung der wesentlichen Komponenten und der Realisierung der graphischen Benutzerschnittstelle von *Videana* wurden experimentelle Ergebnisse für die einzelnen Ansätze zur Schnitterkennung, Textdetektion und -erkennung, Gesichtsdetektion und -er-

kennung und zur Bestimmung von Kamerabewegungen präsentiert. Schließlich wurde noch eine Anwendung dieser Software für das Projekt „Mediennarrationen und Medienspiele“ vorgestellt, im Rahmen derer automatisch extrahierte Merkmale genutzt wurden, um statistische Modelle über die Videocharakteristika von interaktiven Computerspielsequenzen und narrativen Sequenzen in Spielfilm maschinell zu „lernen“ und diese zum Finden derselben in anderen Videoaufnahmen anzuwenden.

Abschließend sei nochmals hervorgehoben, daß *Videana* MedienwissenschaftlerInnen keine qualitativen Analysen oder Interpretationen anbieten kann, da die „Intelligenz“ heutiger Hardware/Softwaresysteme hierzu nicht ausreicht. Vielmehr bietet *Videana* medienwissenschaftlichen NutzerInnen Möglichkeiten zur Bearbeitung von Video-Metainformationen sowie eine Fülle von Werkzeugen an, die zeitaufwändige Arbeiten automatisieren. Die zugrunde liegenden Algorithmen entsprechen dem aktuellen Stand der einschlägigen Forschung im Bereich des *Video Indexing und Retrieval*. Nach Kenntnis der Autoren stellt *Videana* in der hier vorgestellten Form eine für MedienwissenschaftlerInnen bislang einzigartige Software dar.

Literatur

Amir, A./Iyengar, G./Argillander, J./Campbell, M./Haubold, A./Ebadollahi, S./Kang, F./Naphade, M. R./Natsev, A./Smith, J. R./Teši•, J./Volkmer, T.: IBM Research TRECVID-2005 Video Retrieval System. In: TRECVID Online Proceedings,

Auf: http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html, Abgerufen am 20. 04. 2006.

Bescos, J.: Real Time Shot Change Detection Over Online MPEG-2 Video. In: *IEEE Transactions on Circuits and Systems for Video Technology* 1 (2004). No. 4. S. 475-484.

Ewerth, Ralph/Mühling, Markus/Freisleben, Bernd: Self-Supervised Learning of Face Appearances in Videos. In: Kettenberger, P. (Hrsg.): Proceedings of the 8th IEEE International Symposium on Multimedia. San Diego, CA, 2006, S. 78-85.

Ewerth, Ralph/Freisleben, Bernd: Video Cut Detection without Thresholds. In: Bartkowiak, M./Domanski, M./Grajek, T./Stasinski, R./Swierczynski, R./Rosinski, T. (Hrsg.): Proceedings of the 11th International Workshop on Systems, Signals and Image Processing. Poznan, Polen, 2004. S. 227-230.

Ewerth, Ralph/Freisleben, Bernd: Improving Cut Detection Algorithmus for MPEG Videos by GOP-oriented Frame Difference Normalization. In: Kittler, J./Petrou, M./Nixon, M. S. (Hrsg.): Proceedings of 17th International Conference on Pattern Recognition. Vol. 2. Cambridge (UK) 2004. S. 807-810.

Ewerth, Ralph/Schwalb, Martin/Tessmann, Paul/Freisleben, Bernd: Estimation of Arbitrary Camera Motion in MPEG Videos. In: Kittler, J./Petrou, M./Nixon, M. S. (Hrsg.): Proceedings of 17th International Conference on Pattern Recognition. Vol. 1. Cambridge (UK) 2004. S. 512-515.

Giesenfeld, G./Sanke, P.: Ein komfortabler Schreibstift für spezielle Aufgaben: Vorstellung des Filmprotokollierungssystems 'Filmprot' (Vers. 1.01). In: Korte, H./Faulstich, W.: Filmanalyse interdisziplinär. Göttingen 1991. S. 135-146.

Gllavata, Julinda/Ewerth, Ralph/Freisleben, Bernd: Text Detection in Images Based on Unsupervised Classification of High-Frequency Wavelet Coefficients. In: Kittler, J./Petrou, M./Nixon,

M. S. (Hrsg.): Proceedings of 17th International Conference on Pattern Recognition. Vol. 1. Cambridge (UK) 2004. S. 425-428.

Gllavata, Julinda/Ewerth, Ralph/Freisleben, Bernd: Tracking Text in MPEG Videos. In: Schulzrinne, H./Dimitrova, N./Sasse, A./Moon, S. B./Lienhart, R. (Hrsg.): Proceedings of ACM Multimedia. New York 2004. S. 240-243.

Gllavata, Julinda/Ewerth, Ralph/Freisleben, Bernd: A Text Detection, Localization and Segmentation System for OCR in Images. In: Werner, B. (Hrsg.): Proceedings of the 6th IEEE Int. Symposium on Multimedia Software Engineering. Miami 2004. S. 310-317.

Gllavata, Julinda/Ewerth, Ralph/Stefi, Teuta/Freisleben, Bernd: Unsupervised Text Segmentation Using Color and Wavelet Features. In: Enser, P./Kompatsiaris, Y./O'Connor, N. E./Smeaton, A. F./Smeulders, A. W. M. (Hrsg.): Lecture Notes on Computer Science: Proceedings of the 3rd International Conference on Image and Video Retrieval. Dublin 2004. S. 216-224.

Hanjalic, A.: Shot Boundary Detection: Unraveled and Resolved? In: *IEEE Transactions on Circuits and Systems for Video Technology* 12 (2002). No. 2. S. 90-105.

OpenCV: Intel's Open Source Computer Vision Library.

Auf: http://www.intel.com/technology/computing/opencv/ Abgerufen am 20. 04. 2006.

Jung, Keechul/Kim, Kwang In/Jain, Anil K.: Text Information Extraction in Images and Video: A Survey. In: *Pattern Recognition* 37 (2004). Elsevier, Großbritannien, S. 977 – 997.

Korte, Helmut: Projektbericht CNfA – Computergestützte Notation filmischer Abläufe – Erweiterte und aktualisierte Fassung. In: IMF-Schriften, Heft 1, Braunschweig, 1992.

Korte, Helmut: Handbuch CNfA, Prototyp 3, Computergestützte Notation filmischer Abläufe. Braunschweig 1994.

Korte, Helmut: Einführung in die Systematische Filmanalyse. Erich Schmidt Verlag, Berlin 2001.

Martinez, J. M.: MPEG–7 Overview. Technical Report N4980, ISO/IEC JTC1/SC29/WG11. Klagenfurt 2002.

Phillips, P. J./Grother, P./Micheals, R. J./Blackburn, D. M./Tabassi, E./Bone, J. M.: FRVT 2002: Overview and Summary. Auf: http://www.frvt.org/FRVT2002/documents.htm, Abgerufen am 03. 05. 2005.

Tahaghoghi, S. M. M./Thom, J. A./Williams, H. E./Volkmer, T.: Video Cut Detection Using Frame Windows. In: *Proc. of the Twenty-Eighth Australasian Computer Science Conf.* 38 (2005). S. 193-199.

TREC Video Retrieval Evaluation, Auf: http://www-nlpir.nist.gov/projects/trecvid/, Abgerufen am 20. 04. 2006.

Truong, B. T./Dorai, C./Venkatesh, S.: New Enhancements to Cut, Fade, and Dissolve Detection Processes in Video Segmentation. In: Paknikar, S./Kankanhalli, M./Ramakrishnan, K. R./Srinivasan, S. H./Ngoh, L. H.: Proceedings of the 8th ACM International Conference on Multimedia. Marina del Rey 2000.S. 219 – 227.

Viola, P./Jones, M.: Robust Real-Time Face Detection. In: *International Journal of Computer Vision*, 57 (2004). Kluwer Academic Publishers, Niederlande. No. 2. S. 137-154.

Yang, M.-H./Kriegman, D. J./Ahuja, N.: Detecting Faces in Images: A Survey. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24 (2002). No. 1. S. 34-58.

Yeo, B./Liu, B.: Rapid Scene Analysis on Compressed Video. In: *IEEE Transactions on Circuits and Systems for Video Technology* 5 (1995). No. 6. S. 533-544.

Yuan, J./Xiao, L./Wang, D./Ding, D./Zuo, Y./Tong, Z./Liu, X./Xu, S./Zheng, W./Li, X./Si, Z. /Li, J./Lin, F./Zhang, B.: Tsinghua University at TRECVID 2005. In: Online Proceedings of TRECVID Conference Series 2005

Auf: http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html Abgerufen am 20.04. 2006.

Zhao, W./Chellappa, R./Phillips, P. J./Rosenfeld, A.: Face Recognition: A Literature Survey. In: *ACM Computing Surveys* 35 (2003). Issue 4. S. 399-458.