Technical Report 69

Automatische Identifikation und Klassifikation von Personen als Key-Visual-Kandidaten

O. Herzog (Antragsteller)
P. Ludes (Antragsteller)
J. Müller
M. Stommel

TZI, Universität Bremen

TZI-Bericht Nr. 69
2013

TZI-Berichte

Herausgeber:
Technologie-Zentrum Informatik und Informationstechnik
Universität Bremen
Am Fallturm 1
28359 Bremen
Telefon: +49 421 218 94090
Fax: +49 421 218 94095
E-Mail: hq@tzi.de
http://www.tzi.de

ISSN 1613-3773

DFG Abschlussbericht

zum Forschungsprojekt (Sachbeihilfe)

Automatische Identifikation und Klassifikation von Personen als Key-Visual-Kandidaten

Förderungszeitraum 01.10.2008 – 05.10.2012

Antragsteller
Prof. Dr. Otthein Herzog
Technologie-Zentrum Informatik und Informationstechnik (TZI)
Fachbereich Mathematik/Informatik
Universität Bremen

Prof. Dr. Dr. (USA) Peter Ludes
School of Humanities and Social Sciences
Jacobs University Bremen

1. Allgemeine Angaben

1.1. DFG-Geschäftszeichen

HE 989/11-1

LU-356/13-1

1.2. Antragsteller

Prof. Dr. Otthein Herzog (1)

Prof. Dr. Dr. (USA) Peter Ludes (2)

1.3. Institut/Lehrstuhl

(1) Technologie-Zentrum Informatik (TZI), Fachbereich Mathematik/Informatik, Universität Bremen

(2) School of Humanities and Social Sciences, Jacobs University Bremen

1.4. Aus DFG-Mitteln bezahlte wissenschaftliche Mitarbeiter mit Angabe des Beschäftigungszeitraums

Björn Gottfried (1): 15.04.2011 - 15.11.2011 (TV-L 13, 100%)

Tobias Kohler (2): 01.10.2009 – 30.09.2011 (TV-L 13, 50%)

Dr. Jan Müller (2): 01.10.2008 – 31.12.2010 (TV-L 13, 50%)

Sarah-Elisa Nees (2): 01.10.2008 – 30.09.2009 (TV-L 13, 50%)

Dr. Martin Stommel (1): 15.10.2008 – 15.04.2011 (TV-L 13, 100%)

Juliana Cunha Costa (2): 06.02.2012 – 05.10.2012 (TV-L 13, 25 %)

1.5. Thema des Projekts

Automatische Identifikation und Klassifikation von Personen als Key-Visual-Kandidaten

1.6. Berichtszeitraum, Förderungszeitraum insgesamt

01.10.2008 – 05.10.2012

1.7. Fachgebiet, Arbeitsrichtung

Informatik, Bildverarbeitung, Kommunikationswissenschaft

1.8. Verwertungsfelder

Interkulturell vergleichende Medieninhaltsanalyse audiovisueller Inhalte

1.9. Liste der wichtigsten Publikationen aus diesem Projekt

a) Arbeiten, die in Publikationsorganen mit einer wissenschaftlichen Qualitätssicherung zum Zeitpunkt der Berichterstellung erschienen oder endgültig angenommen sind:

[ES12] S. Edelkamp, M. Stommel. The Bitvector Machine: A Fast and Robust Machine Learning Algorithm for Non-linear Problems. In P. Flach et al. (eds.). Proc. Europ. Conf. on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD) 2012. Springer: Heidelberg, 2012. S. 175–190.

[Got11] B. Gottfried. Interpreting motion events of pairs of moving objects. GeoInformatica, 15(2):247–271, 2011.

[KL10a] S. Kramer und P. Ludes (Hrsg.) Networks of Culture, Bd. 2 von The World Language of Key Visuals: Computer Sciences, Humanities, Social Sciences, hrsgg. von P. Ludes und O. Herzog. LIT Verlag: Münster. 2010.

[L11] P. Ludes (Hrsg.) Algorithms of Power – Key Invisibles, Bd. 3 von The World Language of Key Visuals: Computer Sciences, Humanities, Social Sciences, hrsgg. von P. Ludes and O. Herzog .LIT Verlag: Münster. 2011.

[Lud11e] P. Ludes. Elemente internationaler Medienwissenschaften: Eine Einführung in innovative Konzepte. VS Verlag, 2011.

[Lud12] P. Ludes. Schlüsselbilder und Schlüssel zu Unsichtbarem: Brasilianische, chinesische, deutsche und US-amerikanische Fernsehsichten, in: Joachim Knape, Anne Ulrich (Hrsg.): Fernsehbilder im Ausnahmezustand. Zur Rhetorik des Televisuellen in Krieg und Krise. Weidler: Berlin, 2012 (neue rhetorik 11). S. 65-96.

[SDH11] M. Stommel, M. Duemcke, O. Herzog. Classification of Semantic Concepts to Support the Analysis of the Inter-Cultural Visual Repertoires of TV News Reviews. German Conf. on Artificial Intelligence (KI), Berlin, Germany, Oct. 4–7, 2011. Springer: Heidelberg, 2011. S. 325-329.

[SH09a] M. Stommel, O. Herzog. Binarising SIFT-Descriptors to Reduce the Curse of Dimensionality in Histogram-Based Object Recognition. In: D. Slezak, S. K. Pal, B.-H. Kang, J. Gu, H. Kurada (eds). Int. Symp. on Signal Processing, Image Processing and Pattern Recognition (SIP), Jeju Island, Korea, 10. –12. December 2009. Springer: Heidelberg, 2009. S. 320–327.

[SH09b] M. Stommel, O. Herzog. SIFT-Based Object Recognition with Fast Alphabet Creation and Reduced Curse of Dimensionality. In D. Bailey (ed.), Int. Conf. on Image and Vision Computing New Zealand (IVCNZ), Wellington, New Zealand, Nov. 23–25, 2009, IEEE, 2009. S. 136-141.

[SH10] M. Stommel und O. Herzog. Learning of Face Components in Coherent and Disturbed Constellations. In A. Bainbridge-Smith and R. Green (eds.), Int. Conf. on Image and Vision Computing New Zealand (IVCNZ), Queenstown, New Zealand, Nov. 8 –9, 2010. IEEE, 2010. S. 1-8.

[SM11] M. Stommel, J. Müller. Automatische, computergestützte Bilderkennung. In T. Petersen, C. Schwender (Hrsg.): Die Entschlüsselung der Bilder: Methoden der Visuellen Kommunikationsforschung. Herbert von Halem: Köln. S. 246–263.

[WSH11] T. Wiedemeyer, M. Stommel, O. Herzog. Wide Range Face Pose Estimation by Modeling the 3D Arrangement of Robustly Detectable Sub-Parts. A. Berciano (Ed.), IAPR International Conference on Computer Analysis of Images and Patterns (CAIP), Seville (Spain), Aug. 29–31, 2011. Springer: Heidelberg, 2011. S. 237–244.

b) Andere Veröffentlichungen

Bücher, Buchkapitel

[Cos13] Juliana Cunha Costa. Criminal violence in Brazilian moving images in 2010 - The Analysis of Visual Narratives' in "Retrospectiva Rede Globo" and "Retrospectiva Rede Record". In Cabecinhas, R., Abadia, L. (eds.). Narratives and social memory: theoretical and methodological approaches. Braga: University of Minho, 2013. http://www.uminho.pt/uploads/eventos/EV_7031/20130328416879876250.pdf, S. 202 -219.

[HHR10] J. Hao, M. Haude und D. M. Reichenbachs. The Olympic Summer Games 2008: A Comparison of Chinese and German Broadcasting on the Opening Day. In S. Kramer und P. Ludes (Hrsg.) Networks of Culture, Bd. 2 von The World Language of Key Visuals: Computer Sciences, Humanities, Social Sciences, hrsgg. von P. Ludes und O. Herzog. Münster: LIT Verlag. 2010. S. 161-186.

[KL10b] S. Kramer und P. Ludes. Introduction. In S. Kramer und P. Ludes (Hrsg.) Networks of Culture, Bd. 2 von The World Language of Key Visuals: Computer Sciences, Humanities, Social Sciences, hrsgg. von P. Ludes und O. Herzog.: LIT Verlag: Münster, 2010. S. 11-24.

[Lud09b] P. Ludes. Networks of Civilizing Processes Upheavals and Associational Games (Vernetzte Zivilisationsumbrüche und Assoziationsspiele), in: Herbert Willems (Hrsg.) Theatralisierung der Gesellschaft. Band 2: Medientheatralität und Medientheatralisierung, Wiesbaden: VS Verlag für Sozialwissenschaften, 2009. S. 433-447.

[Lud09c] P. Ludes. Rumo a uma "linguagem mundial dos compassos e imagens-chave"? Retrospectivas de fim de ano na TV no Brasil e na Alemanha em 2008. In L. Boccia (Hrsg.) ECUS Cadernos de Pesquisa: Interdisciplinaridade e Cultura. Salvador: Programa Multidisciplinar de Pós-Graduação em Cultura e Sociedade-UFBA, 2009. S. 59-66.

[Lud10] P. Ludes. Key Visual Networks. In S. Kramer und P. Ludes (Hrsg.) Networks of Culture, Bd. 2 von The World Language of Key Visuals: Computer Sciences, Humanities, Social Sciences, hrsgg. von P. Ludes und O. Herzog. LIT Verlag: Münster, 2010. S. 59-74.

[Lud11a] P. Ludes. Foreword: Image Processing, Key Audibles, and Key Invisibles. In P. Ludes (Hrsg.): Algorithms of Power – Key Invisibles, Bd. 3 von The World Language of Key Visuals, hrsgg. von P. Ludes und O. Herzog. LIT Verlag: Münster, 2011. S. 11-12.

[Lud11b] P. Ludes. Towards Bridging the Semantic Gap Between Key Visual Candidates and Algorithms of Power. In P. Ludes (Hrsg.) Algorithms of Power – Key Invisibles, Bd. 3 von The World Language of Key Visuals, hrsgg. von P. Ludes und O. Herzog. LIT Verlag: Münster, 2011. S. 15-48.

[Lud11c] P. Ludes. Multi-Sensory Experiences. In P. Ludes (Hrsg.): Algorithms of Power – Key Invisibles, Bd. 3 von The World Language of Key Visuals: Computer Sciences, Humanities, Social Sciences, hrsgg. von P. Ludes und O. Herzog. LIT Verlag: Münster, 2011. S. 159-174.

[MS11] J. Müller und M. Stommel. Heads of state and common people: perspectives from the computer and social sciences. In P. Ludes (Hrsg.): Algorithms of Power – Key Invisibles, Bd. 3 von The World Language of Key Visuals: Computer Sciences, Humanities, Social Sciences, hrsgg. von P. Ludes und O. Herzog. LIT Verlag: Münster, 2011. S. 49-70.

Technische Berichte:

[KS11] T. Kohler und M. Stommel. Similarity-based clustering of key frames of top athletes from TV annual reviews 1999-2009. Technical Report 62, Center for Computing and Communication Technologies, University of Bremen, Germany, 2011.

[SF11b] M. Stommel und G. Frieder. Evaluation of the Legibility Estimation for 93 Historic Documents. Technical Report 57, Center for Computing and Communication Technologies, University of Bremen, Germany, 2011.

[SF10] M. Stommel und G. Frieder. Automatic Estimation of the Legibility of Binarised Mixed Handwritten and Typed Documents. Technical Report 56, Center for Computing and Communication Technologies, University of Bremen, Germany, 2010.

2. Arbeits- und Ergebnisbericht

2.1. Ausgangslage: Stellen Sie Ausgangsfrage, Zielsetzung und Arbeitshypothesen des Projekts dar.

Zielsetzung: Key Visuals verdichten ähnlich wie Überschriften für Zeitungsartikel die (audio-) visuelle Information in Bildschirmmedien auf ihren Kern. Personen spielen hierbei eine herausragende Rolle. Mit dem Fokus auf die Beispiele „Staatsoberhäupter“, „einfache Leute“, „Spitzensportler“ und „Zuschauer“ als Key-Visual-Kandidaten sollen in diesem Projekt Algorithmen entwickelt werden, die die automatische Detektion und Klassifikation von Personen unterstützen. Die untersuchten Videosequenzen beschränken sich auf Fernsehjahresrückblicke der Jahre 1999-2010 aus Deutschland (ARD) und den USA (ABC, CBS und NBC). Diese ‚Bilder des Jahres‘ enthalten eine hohe Dichte an Key Visuals. Der Untersuchungszeitraum von zwölf Jahren liefert so, gespeist aus den beiden jeweils größten Medienmärkten in Europa und Nordamerika, eine hinreichend homogen formatierte und zugleich visuell variable Datenmenge, die neue Trends erkennen lässt und auch die „Modellfunktion“ von Medienformaten und Präsentationsmustern im US-amerikanischen Fernsehen berücksichtigt.

Dem Projekt liegt die Hypothese zugrunde, daß die zu entwickelnden Methoden zur automatischen Detektion und Klassifikation von Personen es allen Bildwissenschaften ermöglichen werden, umfangreichere visuelle Datenbestände effizienter und besser zu analysieren. Die oft komplexen semantischen Konzepte, die im Fokus von medienwissenschaftlichen Bildanalysen stehen, lassen sich bislang nur eingeschränkt in Algorithmen überführen. Die am Projekt beteiligten Kommunikationswissenschaftler bieten die theoretischen semantischen Vorgaben für die Entwicklung der Algorithmen zur automatischen Erkennung von Key-Visual-Kandidaten und prüfen die jeweils entwickelten Verfahren. Die Verwendung des Konzeptes der Key Visuals als Grundlage von Verfahren zur automatischen Videoanalyse soll zu Synergieeffekten führen, die eine Verringerung des Semantic Gap für die untersuchten Domänen erlauben.

Das Projekt beinhaltet die Modellierung und Erkennung der annotierten Key-Visual-Kandidaten mit den Mitteln der digitalen Bildverarbeitung. Im Antrag werden dazu implizite und explizite Grundannahmen gemacht, auf welchen das Arbeitsprogramm basiert. Zunächst wird vorausgesetzt, dass die genannten relevanten Personengruppen durch Abwandlungen existierender Verfahren über die gesamte Stichprobe detektierbar sind. Darauf aufbauend wird vorausgesetzt, dass biometrische und andere Merkmale existieren und stabil berechenbar sind, und dass sich die genannten Personenkategorien bezüglich dieser Kriterien unterscheiden. Weiterhin wird die Korrektheit und Stabilität gängiger Verfahren und Stichproben vorausgesetzt, so dass auf diesen aufgebaut werden kann.

Nach einer Durchsicht des Datenmaterials und der Literatur wurden die Grundannahmen präzisiert und auf konkrete Verfahren bezogen, welche für die Identifikation von Key-Visual-Kandidaten aussichtsreich erscheinen. Daraus ergaben sich die Hypothesen,

- dass sich Key-Visual-Kategorien durch bestimmte lokale und globale Deskriptoren erkennen lassen, und
- dass visuelle Alphabete durch Clusterung erzeugt werden müssen.

Bezüglich der Erkennung von Personen wurden die Hypothesen aufgestellt,

- dass Gesichtsmerkmale durch die populären und als sehr trennscharf erachteten SIFT-Merkmale erkannt werden können,
- dass sich die Pose eines Gesichts lokal schätzen lässt und
- dass sich auch Seitenansichten von Gesichtern durch lokale Merkmale modellieren und erkennen lassen können.

Dabei wurde auch die numerische Stabilität von SIFT angesichts seiner Hochdimensionalität in Frage gestellt. Bezüglich der Key-Visual-Kategorien Staatsoberhaupt, einfache Leute, Menschenmengen und Sportler wurde überprüft, inwiefern bestimmte qualitativ formulierte formale Darstellungskonventionen auch quantitativ gerechtfertigt sind.

2.2. Beschreibung der durchgeführten Arbeiten

Medien- und Kommunikationswissenschaft:

Zu Beginn (AP MKW-1) wurde, der Kodierung zeitlich vorgelagert, die Recherche nach Videomaterial aus Deutschland (ARD) und USA (ABC; CBS; NBC) und die sukzessive Beschaffung und ggf. Digitalisierung durchgeführt, um über einen entsprechenden Videodatensatz verfügen zu können. Die Tagesrückblicke fanden zunächst keine Beachtung, da sich die Stundenzahl der Jahresrückblicke 1999-2000 als hinreichend aussagekräftig erwiesen. Nach einer ersten Sichtung wurde der auf [Han08] basierende Kodierbogen systematisch in mehreren Feedbackrunden optimiert und an das Videomaterial angepasst.

Wo anfänglich von einer Einteilung der kleinsten zu kodierenden Einheiten anhand der automatisierten Shotgrenzen-Erkennung ausgegangen worden war, ist man im weiteren Verlauf zu einer manuell definierbaren segmentbasierten sekundengenauen Kodierung übergegangen. Der Prozess der Kodierung erstreckte sich über einen Großteil der Projektlaufzeit und stellte die Grundlage für die weiteren Arbeitspakete dar.

Nach einer ersten Sichtung des Videomaterials und einer vorläufigen Annotierung konnte eine Reihe von Videosequenzen und Standbildern zur Bestimmung des Trainingsmaterials (AP MKW-2) herangezogen werden. Weiterhin wurde die Methode gewählt, aus den zu untersuchenden Videosequenzen einzelne Standbilder (sog. Keyframes) zu extrahieren, die ein Training der entwickelten Machine-Learning-Algorithmen ermöglichen sollten. Die automatisierte, sekundengenaue Frame-Extraktion wurde sukzessive von einer manuellen, pixelgenauen Markierung von Gesichtern durch Kodierer begleitet, so dass eine hinreichende Trainingsmenge für die im Rahmen des Projekts verwendete Supervised-Learning-Software definiert werden konnte.

Mittels quantitativer Analysen und statistischer Auswertungen wurde in einem weiteren Arbeitsschritt (AP MKW-3) nach stabilen und regelmäßig auftretenden Mustern gesucht, die schließlich in eine Klassifikation der Key Visuals nach Art und Häufigkeit mündete. Insbesondere wurde hier auf Grundlage der dargestellten Typifizierungen „einfache Leute“ und „Staatsoberhäupter“ die Methode der 4D-Verteilung der im Videostream erkannten Gesichter gewählt (s. Abb. 1), die deutliche Trends zutage fördern konnte und eine ansatzweise Klassifikation der selektierten Ausschnitte zulässt.

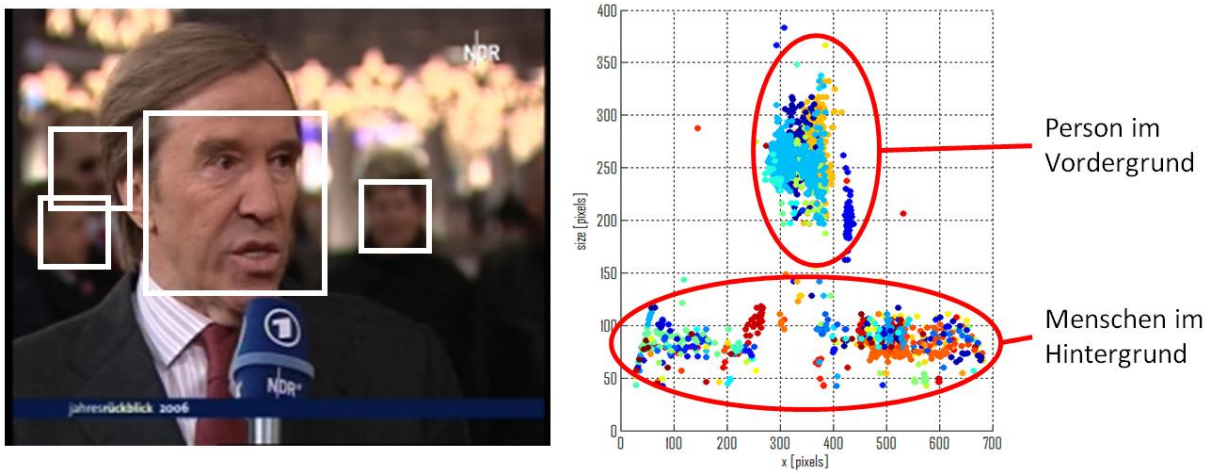

Abbildung 1: Links: Gesichtsdetektionen eines Frames aus einem längeren Shot. Rechts: Projektion der 4D-Verteilung der Gesichtsdetektionen des Shots auf Größe und horizontale Koordinate. Die Darstellung gibt deutlichen Aufschluß über den Aufbau der Szene.

In einem letzten Arbeitsschritt konnte eine Evaluation der identifizierten Key Visuals angegangen werden. Die Hypothesengenerierung erfolgte sowohl induktiv als auch deduktiv. So konnte eine Reihe von Hypothesen geprüft und bestätigt bzw. falsifiziert werden, z.B.

- dass Staatsoberhäupter im Vergleich zu einfachen Leuten als aktiver agierende Darstellungstypen fungieren,
- dass Staatsoberhäupter als Stellvertreter für Mächtige im Sinne einer Bedeutungsgröße mehr Bildschirmfläche einnehmen,
- oder dass in US-amerikanischen Jahresrückblicken im Vergleich zu deutschen das Staatsoberhaupt häufiger in militärischen Zusammenhängen gezeigt wird.

In Bezug auf die Darstellung „einfacher Leute“ konnte herausgearbeitet werden,

- dass ein länderübergreifender Trend vorherrscht, diese Akteure häufig als Opfer von Naturkatastrophen, Unfällen und von Verbrechen zu präsentieren.

Die darüber hinausgehende manuelle Kodierung und Auswertung der Darstellungen von Spitzensportlern und Publikum erlaubte eine systematische Bearbeitung einzelner ausgewählter Forschungshypothesen [KS11]: So konnten mit Hilfe einer ähnlichkeitsbasierten Clusterung von Keyframes auf der Grundlage multidimensionaler Skalierung einzelne im Antrag formulierte Hypothesen bestätigt bzw. falsifiziert werden.

Bildverarbeitung

Es traten keine Ereignisse auf, die gravierende Abweichungen vom ursprünglichen Arbeitsplan notwendig erscheinen ließen. Die durchgeführten Literaturdurchsichten und Experimente lassen jedoch rückblickend eine differenziertere Sicht auf einige Grundannahmen des Antrags zu.

Zunächst (AP INF-1) wurde die Eignung der in der Literatur beschriebenen Verfahren für das Videomaterial überprüft. Dieses ist bezüglich der meisten für die Bildverarbeitung relevanten Randbedingungen (z.B. Beleuchtung, Pose, Perspektive, Hintergrund) kaum eingeschränkt. Viele Verfahren, welche durchgängig bestimmte dynamische, perspektivische oder texturbezogene Eigenschaften voraussetzen, sind daher nicht einsetzbar (z.B. Verfahren aus der Videoüberwachung oder konturbasierte Verfahren). Der im Antrag genannte Gesichtsdetektor von Lienhart und Maydt (2002) beispielsweise erreichte auf der Stichprobe nur eine Precision von etwa 50% bei einem Recall von 70%-90%, je nachdem ob auch schlecht aufgelöste Gesichter erkannt werden sollten. Es wurde daher entschieden, einen kompositionellen Ge

Abbildung 2: Beispielframes ausgewählter Key Visuals.

sichtsdetektor zu entwickeln, der für eine größere Variation an Posen einsetzbar ist und teilweise Verdeckungen toleriert. Die Hypothese, dass Personen über die gesamte Stichprobe hinweg mit einem Verfahren detektierbar sind, wurde daher auf die Detektion von Gesichtern eingeschränkt.

Zur Charakterisierung von Personen und ihres Kontexts (AP INF-2) wurde die Gesichtsdetektion in einem 4D-Koordinatensystem als Merkmal ausgewählt. Kleidungsbezogene Merkmale wurden aufgrund ihrer vielfältigen Erscheinung verworfen. Stattdessen wurden Kontextmerkmale hinzugenommen, nämlich lokale Texturen, Texturgradienten, lokale und globale Farbhistogramme und die Kameraperspektive. Zudem wurde der gesprochene Text als Merkmal ausgewählt. Hier ist jedoch zu beachten, dass entsprechende Software noch nicht vollständig zuverlässig und die manuelle Annotation aufwendig ist. Darüber hinaus korrespondiert der Text oft nicht mit dem Bild. Der Sprecher tritt auch oft selbst nicht als Akteur in Erscheinung. Der Antrag verweist auf weitere schwierige und ungelöste Probleme, deren Behandlung jedoch in jeweils spezialisierten Projekten angemessener erscheint.

Für Kontextinformationen wurde eine algorithmische Umsetzung in Anlehnung an den populären Bag-of-Features-Ansatz gewählt, wohingegen für die Erkennung von Gesichtern ein geometrisch selektiverer Konstellationenansatz gewählt wurde (AP INF-3). Für die Gesichtserkennung wurden SIFT-Merkmale als Basis gewählt, deren mögliche relative 3D-Positionen mittels Voting-Verfahren überprüft wurden. Die Gesichtserkennung wurde anhand der Feret-Stichprobe getestet, welche dazu in erheblichem Umfang erweitert und sogar korrigiert wurde. Zur Analyse von Kontextinformationen wurden mehrere Merkmalsdetektoren und -deskriptoren kombiniert und getestet. Hierbei wurde entdeckt, dass der populäre SIFT-Operator dem Fluch der Dimension unterliegt, was für die Abstraktion lokaler Merkmale nachteilig ist. Dieses Problem konnte durch eine Merkmalsbinarisierung gelöst werden. In der Literatur wird in diesem Zusammenhang auch fast obligatorisch eine Clusterung von SIFT-Merkmalen durchgeführt, welche gelegentlich jedoch als sehr aufwendig beschrieben wird. Dieser Problemfall wurde rekonstruiert und als ernsthaftes Hindernis bewertet. Als Lösung wurde eine Synthesemethode für künstliche Code-Books entwickelt, welche bei geringer Laufzeit einen hohen Grad an Flexibilität bietet. Dadurch konnten hohe Klassifikationsraten in der Szenenanalyse erreicht werden. Dagegen zeigte sich, dass Kamerabewegungen auf dem verwendeten Videomaterial schwer zu detektieren sind. Eine entsprechende Software ergab unter 14 möglichen Bewegungen jedoch für die Erkennung statischer Szenen zufriedenstellende Resultate.

Die modellierten Merkmale wurden genutzt, um aus einer manuellen Annotation abgeleitete Key-Visual-Kandidaten (s. Abb. 2) automatisch zu klassifizieren (AP INF-4). Dazu wurden die modellierten Merkmale mittels Support Vector Machines und des Pyramid Match Kernels kombiniert. Für die visuell definierten Kategorien Menschenmenge, Studio/Nicht-Studio, Fußball/Eishockey und Computergraphik/Natürlich wurden hohe Klassifikationsraten erzielt. Mit Hilfe des SURF-Operators konnten die Klassen "Sports People", "Symbols Splitscreen",

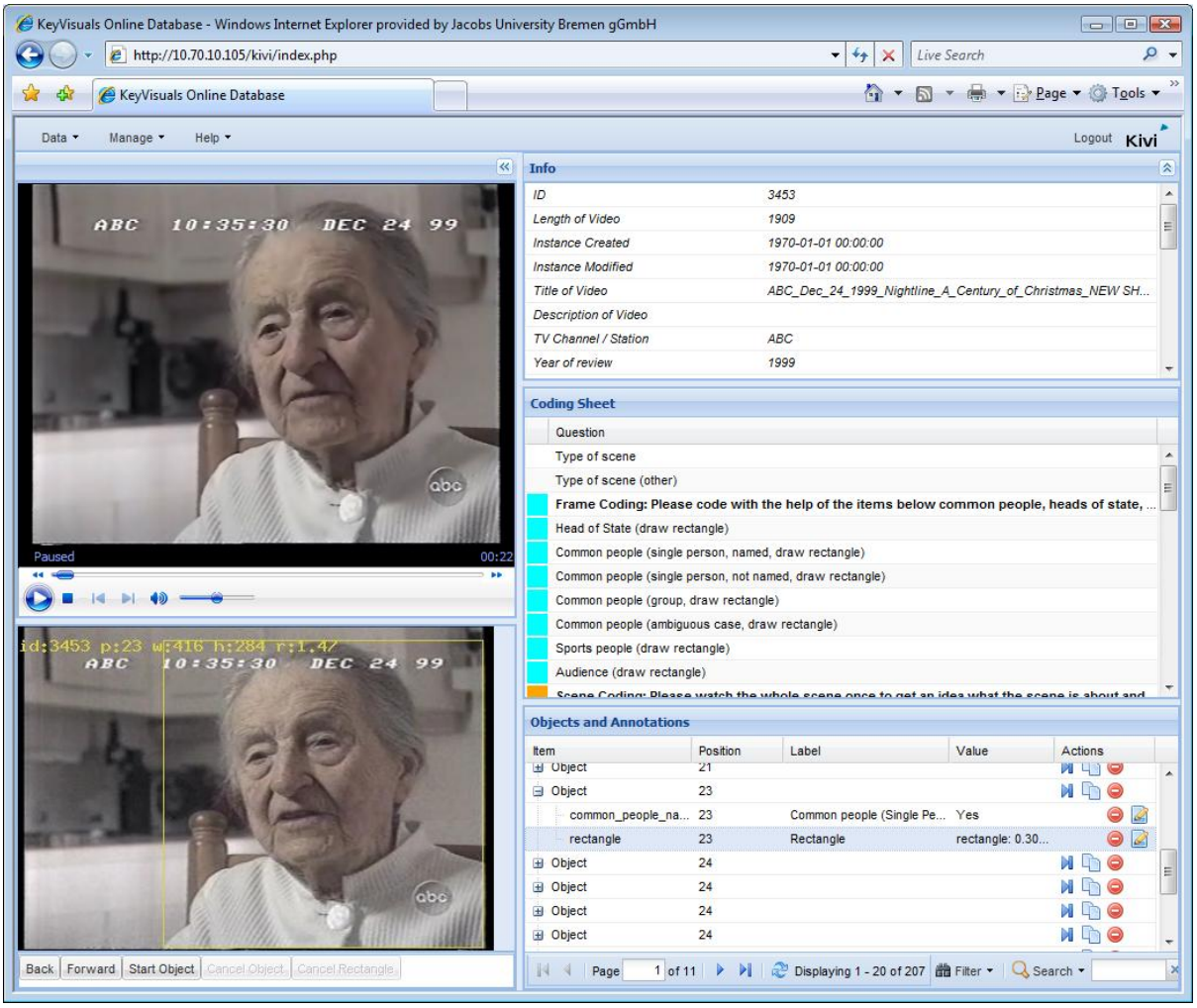

Abbildung 3: Ansicht des im Rahmen des Projekts weiterentwickelten Annotationswerkzeugs 'Kivi'. Zu sehen sind der Videoplayer (oben links), der Kodierbogen (Mitte rechts), die Einzelansicht eines Frames zur Markierung von Bildausschnitten (unten links) und eine Übersicht der kodierten Elemente (unten rechts).

"Head of State", "Military and Security" und "Natural Disasters" unterschieden werden mit Erkennungsraten erheblich über der Chance-Performance. Auf der Basis der Gesichtsdetektionen konnten weitere verfeinerte Kategorien klassifiziert werden, darunter ebenfalls Menschenmengen, frontal und zentral dargestellte Personen, Begleitpersonen sowie die Perspektive der Aufnahme. Eine medienwissenschaftlich formulierte Hypothese zur Darstellung von Staatsoberhäuptern gegenüber einfachen Personen konnte aufgrund der Messwerte differenziert werden.

2.3. Darstellung der erzielten Ergebnisse

Rund 24 Stunden Jahresrückblickssendungen aus Deutschland und den USA aus den Jahren 1999 bis 2010 konnten auf der Grundlage eines von Hanitzsch [Han08] verwendeten Kodierbogens mit Hilfe der Kodiersoftware ‚Kivi' (s. Abb. 3) hinreichend annotiert werden. Darüber hinaus wurden erste Annotationen von brasilianischen (Rede Globo, Rede Record) sowie chinesischen Jahresrückblickssendungen (CCTV 1, 4, 9) basierend auf diesem Schema entwickelt. Die von der Software unterstützte Funktion des Datenexports erlaubt die Extraktion quantitativer und qualitativer Daten zur statistischen Aufbereitung der videosegmentbasierten Annotationen. Übereinstimmungstests und jeweils einzeln projektweise durchge-

führte Kodierverfahren mit mehreren Kodierern aus unterschiedlichen Herkunftsländern machten die

Messung der Intercoder-Reliability mit Hilfe von Krippendorff's Alpha zu einem zentralen Merkmal in Bezug auf den Grad übereinstimmender Annotationen.

Aufgrund der unterschiedlichen Mediensysteme und Senderformate war in der Folge eine breite Varianz der untersuchten Jahresrückblickssendungen festzustellen, die auf eine qualitative und quantitative Ausdifferenzierung des Formats der Fernsehjahresrückblicke über verschiedene Medienkulturen hinweg bei gleichzeitiger Stabilität des Auftretens bestimmter nationaler, transkultureller und globaler Schlüsselbilder, hindeuten.

Unsere detaillierte Untersuchung von Fernsehjahresrückblicken 1999 bis 2010 ergibt folgendes Bild: Die Länder und Regionen in Fernsehjahresrückblicken in CBS (1999 bis 2010) und CCTV 4/CCTV 9 (2008 und 2009) sowie Jahrhundert-Rückblicken in den USA zeigen eindeutige Schwerpunkte und blinde Flecken [Lud11e]. Fernsehrückblicke zeigen demnach historische Machtdarstellungsunterschiede [MS11a] ebenso wie die begrenzte Wahrnehmung internationaler Beziehungspartner. So war Deutschland für das 20. Jahrhundert im Fokus der US-Rückblicke [Lud12] – im vergangenen Jahrzehnt hingegen der Irak. Die berücksichtigten chinesischen Fernsehjahresrückblicke zeigen, wie sehr die USA als Gegenüber wahrgenommen werden. Betrachten wir die verschiedenen Akteurstypen, ergeben sich eindeutige Unterschiede zwischen dem Genre der Jahrhundert-Rückblicke, in dem Augenzeugen dominieren, CBS-Fernsehjahresrückblicken 1999-2010, mit den Staatsoberhäuptern auf Rang 1 und den Journalisten/Fotografen auf Rang 1 bei CCTV 4 und 9. Beispiele für die Darstellung von Kriminalität in brasilianischen Jahresrückblicken bietet [Cos13].

Zur automatisierten Videoanalyse wurden Code-Book-basierte Verfahren sowohl mit [SH10, WSH11] als auch ohne zusätzliche Geometrieinformation [SH09a,SH09b,Sto10] entwickelt, um Gesichter und Szenen zu erkennen. In diesem Zusammenhang wurde zunächst gezeigt [SH09a], dass die Objekterkennung mit den populären SIFT-Deskriptoren dem sogenannten Fluch der Dimensionalität [BGRS99] unterliegt, welcher den Vergleich von Mustern beeinträchtigt. Als Lösung wurde eine Merkmalsbinarisierung in Kombination mit dem Hamming-Abstand vorgestellt. Für den Spezialfall von SIFT-Deskriptoren konnte die extrem zeitaufwendige Clusterung zur Erstellung des Merkmalsalphabets durch eine von den Eingabedaten entkoppelte Merkmalssynthese ersetzt werden. Die Methode ist insbesondere im Training extrem schnell und liefert konkurrenzfähige Klassifikationsraten in der Bilderkennung [SH09b, Sto10].

In der Gesichtserkennung eignet sich die Methode zur Detektion von Gesichtsmerkmalen (s. Abb. 4) mit sehr hoher Genauigkeit (98%). Dabei ist die Methode sogar selektiv genug, um von einzelnen Gesichtsmerkmalen wie einem Auge, der Nase oder dem Mund die Pose des Gesichts mit einer Genauigkeit von etwa 60% zu schätzen [SH10]. Ausgehend von diesen Erkenntnissen wurde ein Verfahren zur gleichzeitigen Detektion und Schätzung der Pose von Gesichtern entwickelt [WSH11]. Die Pose wird dabei modellbasiert aufgrund der relativen Lage der detektierten Gesichtsmerkmale erkannt. Hierzu wurde die bekannte Feret-Stichprobe [PRD96] um die Positionen von 15 Gesichtsmerkmalen in etwa 11.000 Portraitbildern erweitert. Dabei wurden auch einige Fehler der Feret-Stichprobe erkannt und korrigiert. Das Ergebnis ist eine sehr genaue Schätzung der Pose über 180 Grad bezüglich des Gierwinkels und 30 Grad bezüglich der Nick- und Rollwinkel.

Es wurde ferner gezeigt, dass die Parametrisierung der Merkmalsbinarisierung ein eigenständiger Prozessschritt ist. Daher kann auf die erneute Schätzung der Parameter zur Laufzeit verzichtet werden, was für die Geschwindigkeit und Vorhersagbarkeit günstig ist. Verzichtet man auf die dichte Merkmalsabtastung, wie sie in der Bildklassifikation oft eingesetzt wird, ergibt sich so eine im Vergleich zum originalen SIFT-Deskriptor höhere Geschwindigkeit und höhere Zuverlässigkeit mit vielfältiger Anwendbarkeit. Der praktische Einsatz des Verfahrens auf einem mobilen Roboter führte beispielsweise zu einer um bis zu 10% niedrigeren Fehlerrate in der Stereoanalyse und einer um 16% höheren Geschwindigkeit beim

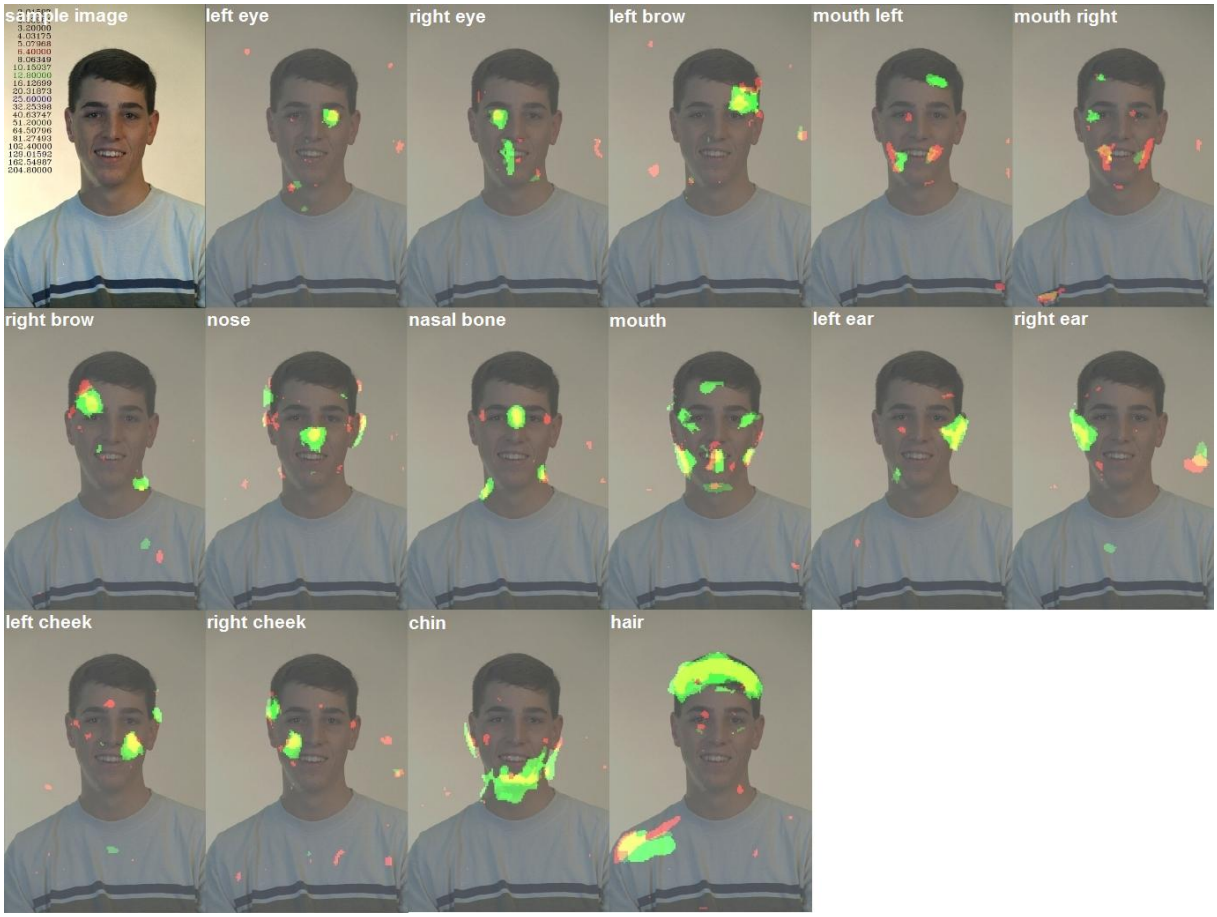

Abbildung 4: Detektion von Gesichtsmerkmalen.

Tracking von Personen [SLHK11]. Die Methode wurde von uns in abstrakter Form als "Bit-vector Machine" in die Literatur eingeführt [ES12]. Bezüglich der Erkennung und Klassifikati

on beweglicher Objektverbünde wurde die Abbildbarkeit auf eine qualitative Darstellung [Got11] entdeckt, welche die relativen Objektbewegungen erschöpfend beschreibt.

Um die Komplexität der Struktur visueller Daten zu ergründen, wurden Objektdarstellungen auf der Basis von formalen Sprachmodellen untersucht. Üblicherweise werden hierzu in der Literatur Abwandlungen kontextfreier Sprachen beschrieben, welche geometrische Abhängigkeiten definieren, jedoch keinen Laufzeitvorteil bieten. Aus diesem Grund wurde die Eignung von Church-Rosser-Sprachen zur Objekterkennung untersucht [MS11b]. Diese sind aufgrund ihrer nur linearen Komplexität attraktiv. Es wurde daher eine zweidimensionale Church-Rosser-Bildsprache entwickelt, welche im Gegensatz zu vielen anderen Ansätzen die Lokalität des Bildes erhält. Es konnte theoretisch gezeigt werden, dass zu dieser Bildsprache ein gleichbedeutender deterministischer, schrumpfender, zweidimensionaler Restart-Automat existiert, welcher die Erkennung von Worten der Sprache in linearer Zeit bezogen auf die Fläche leistet.

Die genannten Methoden wurden eingesetzt, um die in Abschnitt 3.2 genannten Key-Visual-Kandidaten zu klassifizieren [LHMS10, MKS10, L11, SDH11, SM11]. Die Ergebnisse gingen in das von Petersen und Schwender herausgegebene Methodenbuch zur visuellen Kommunikationsforschung ein [SM11], wo sie den Stand der Forschung mit definieren.

Aufgrund der Vielfalt des Videomaterials und der oft symbolischen Nebenbedeutungen vieler Key-Visual-Kategorien scheint es immer weniger einzelne Operatoren zu geben, welche für ganze semantische Konzepte und die ganze Stichprobe gültig sind. Es ist daher nötig, viele, oft auch sehr schwer zu detektierende Merkmale zu kombinieren, um jeweils Untergruppen

einer komplexeren Klasse zu erkennen. Die Gültigkeit der Einsatzbedingungen der Verfahren lässt sich nicht mehr vorher absehen, sondern muss durch eigene Klassifikatoren oder die Verfahren selbst überprüft werden. Daher wurde ein neuer Ansatz entwickelt, bei dem die Operatorausgabe durch den Algorithmus selbst bewertet wird, um die Gültigkeit der Einsatzbedingungen zu prüfen und gute Verfahrensparameter zu finden. Die Methode wurde mit vielversprechendem Ergebnis auf einem Document-Enhancement-Filter getestet [SF11].

[BGRS99] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft. When is ”Nearest Neighbor” Meaningful. In Int. Conf. on Database Theory, 1999. S. 217–235,

[Got11] B. Gottfried. Interpreting motion events of pairs of moving objects. GeoInformatica, 15(2):247–271, 2011.

[Han08] T. Hanitzsch. Codebook for content analysis – Foreign TV news project. 2008. Abgerufen am 20.11.2008 und 14.12.2012 von www.thomas-hanitzsch.de/docs/Lehre/Codebook.pdf

[LHMS10] P. Ludes, O. Herzog, J. Müller, und M. Stommel. Automated Identification and Classification of State Heads and Common People as Key Visual Candidates. VisComX Intro-Workshop, Jacobs University Bremen, September 2010.

[Lud12] P. Ludes. Schlüsselbilder und Schlüssel zu Unsichtbarem: Brasilianische, chinesische, deutsche und US-amerikanische Fernsehsichten, in: Joachim Knape / Anne Ulrich (Hrsg.): Fernsehbilder im Ausnahmezustand. Zur Rhetorik des Televisuellen in Krieg und Krise. Weidler: Berlin, 2012 (neue rhetorik 11). S. 65-96.

[MKS10] J. Müller, T. Kohler, M. Stommel. Computer-based content analysis of television footage from four countries: Coding technique and intercoder reliability testing. Vortrag auf der Doing Global Media Studies Conference in Bremen, October 11–12, 2010, pre-conference to the 3rd European Communication Conference.

[MS11a] J. Müller, M. Stommel. Heads of state and common people: perspectives from the computer and social sciences. In P. Ludes (Hrsg.): Algorithms of Power – Key Invisibles, Bd. 3 von The World Language of Key Visuals: Computer Sciences, Humanities, Social Sciences, hrsgg. von P. Ludes und O. Herzog. LIT Verlag: Münster, 2011, S. 49–70.

[MS11b] H. Messerschmidt, M. Stommel. Church-Rosser Picture Languages and Their Applications in Picture Recognition. Journal of Automata, Languages and Combinatorics (JALC), Otto-von-Guericke-Universität Magdeburg, vol. 16, no. 2–4, 2011, S.165–194,.

[PRD96] Phillips, P. J., Rauss, P. J., Der, S. Z.: FERET (Face Recognition Technology) Recognition Algorithm Development and Test Results. October 1996. Army Research Lab Technical Report 995.

[SDH11] Martin Stommel, Martina Duemcke, Otthein Herzog). Classification of semantic concepts to support the analysis of the inter-cultural visual repertoires of TV news reviews. In Joscha Bach and Stefan Edelkamp (eds.). KI'11: Proceedings of the 34th Annual German Conference on Advances in Artificial Intelligence, Berlin, Germany, October 4-7, 2011. Springer: Heidelberg, 2011. S. 325-329.

[SH09a] M. Stommel, O. Herzog. Binarising SIFT-Descriptors to Reduce the Curse of Dimensionality in Histogram-Based Object Recognition. In: D. Slezak, S. K. Pal, B.-H. Kang, J. Gu, H. Kurada (Eds), Int. Symp. on Signal Processing, Image Processing and Pattern Recognition (SIP), Jeju Island, Korea, 10. –12. Dec. 2009. Springer: Heidelberg, 2009. S. 320–327.

[SH09b] M. Stommel, O. Herzog. SIFT-Based Object Recognition With Fast Alphabet Creation and Reduced Curse of Dimensionality. In D. Bailey (ed.), Int. Conf. on Image and Vision Computing New Zealand (IVCNZ), Wellington, New Zealand, Nov. 23–25, 2009, IEEE, 2009, S. 136–141.

[SH10] M. Stommel, O. Herzog (2010). Learning of Face Components in Coherent and Disturbed Constellations. In Andrew Bainbridge-Smith, Richard Green (eds.). Proc. 25th International Conf. on Image and Vision Computing New Zealand (IVCNZ) 2010, Queenstown, 8 - 9 November 2010. S. 1-8.

[SLHK11] M. Stommel, M. Langer, O. Herzog, K.-D. Kuhnert. A Fast, Robust and Low Bit-Rate Representation for SIFT and SURF Features. IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR), Kyoto, Nov. 1–5, 2011, IEEE. S. 278–283.

[SF11] M. Stommel, G. Frieder. Automatic Estimation of the Legibility of Binarised Historic Documents for Unsupervised Parameter Tuning. 11th IAPR International Conference on Document Analysis and Recognition (ICDAR), Beijing, China, Sept. 18–21, 2011, S. 104–108.

[Sto10] M. Stommel. Binarising SIFT-Descriptors to Reduce the Curse of Dimensionality in Histogram-Based Object Recognition. Intern. Journal of Signal Processing, Image Processing and Pattern Recognition (IJSIP), Volume 3, 2010:25–36.

[WSH11] Thiemo Wiedemeyer, Martin Stommel, Otthein Herzog. Wide Range Face Pose Estimation by Modelling the 3D Arrangement of Robustly Detectable Sub-Parts. In Pedro Real, Daniel Diaz-Pernil,Helena Molina-Abril, Ainhoa Berciano, and Walter G. Kropatsch (eds.). Proc. CAIP 2011 - 14th International Conference on Computer Analysis of Images and Pattern. Sevilla, Spain, August 29-31, 2011. Springer: Heidelberg, 2011. S. 237-244.

2.4. Ausblick auf zukünftige Arbeiten

Die hohe Bandbreite an Jahresrückblicksformaten in den beiden Mediensystemen Deutschlands und der USA sowie der jeweils untersuchten TV-Sender deutet darauf hin, dass sowohl Machart als auch Länge und Verortung der Jahresrückblickssendungen im Sendeschema der jeweiligen TV-Sender je unterschiedliche Funktionen erfüllen. Während die US-amerikanischen Rückblicke im Untersuchungszeitraum eher episodisch als Teil der halbstündigen Abendnachrichten fungieren, lässt sich in Deutschland beobachten, dass Jahresrückblickssendungen als abendfüllende Nachrichtensendungen, teilweise im Talk-Show- und Unterhaltungsformat, präsentiert werden. Für die chinesischen und brasilianischen Jahresrückblicke lässt sich von ähnlichen Divergenzen ausgehen, auf die jedoch spezifisch einzugehen wäre, was zukünftiger Forschungsanstrengungen bedarf, die vor dem Hintergrund der Kombination mit den vorliegenden Ergebnissen überaus vielsprechend erscheinen, wie insbesondere die Nutzung des webbasierten Software-Tools ‚Kivi', das inzwischen unter der Affero General Public License (AGPL) vorliegt.

Die maschinelle Bildverarbeitung hat entscheidende Fortschritte für die automatisierte Identifikation und Klassifikation von dynamischen Bildtypen erreicht. Sie konstruiert mit anwendungsspezifischen Modellen „bottom-up" die Bedeutung von Bildern für möglichst breite Anwendungsfelder. Dem gegenüber konzentriert sich die Medien- und Kommunikationswissenschaft bisher fast ausschließlich auf „top-down"-Interpretationen, die nur qualitativ definiert sind. Die maschinelle Bildverarbeitung algorithmisiert bisher vor allem die automatisierte Identifikation und Klassifikation von „low-level"-Merkmalen (wie Pixel oder unbenannte Objekte) von Standbildern und dynamischen Bildtypen. Sie versucht damit, Bedeutung allein aus der „bottom-up"-Richtung zu erfassen. Weitgehend unerforscht ist dagegen in der Bildverarbeitung die Behandlung semantischer Konzepte auf einer sehr hohen Abstraktionsebene, wie sie in der Medien- und Kommunikationswissenschaft sowie der experimentellen Wahrnehmungspsychologie untersucht werden. Es wird ein Antrag mit dem Titel „Identifikation und Klassifikation audiovisueller Narrative in Bildschirmmedien: Brasilien, China, Deutschland und die USA" erarbeitet. In Zusammenarbeit mit Forschern aus den Bereichen Bildverarbeitung, Kommunikationswissenschaft, Linguistik, der Universität Bremen und der Jacobs University Bremen soll medienwissenschaftliches Kontextwissen operationalisiert werden, insbesondere Modelle audiovisueller Narrative. Die so algorithmisierte Identifikation

und Klassifikation systematisiert quantitativ Typen dynamischer Bildmedienmuster für historisch neuartige und umfangreiche Bilddaten.
Um höhere semantische Konzepte modellieren zu können, sind leistungsfähigere Bildverarbeitungsmethoden zur Objekterkennung nötig. Aufgrund ihrer Robustheit gegenüber Beleuchtungseinflüssen, Verformungen und Fehldetektionen werden hierzu derzeit bevorzugt kompositionelle Modelle eingesetzt, die Objekte durch die geometrische Struktur ihrer Einzelteile beschreiben. Die Erscheinung der Teile wird durch lokale Deskriptoren beschrieben. Da über die statistischen Prinzipien, welche die lokale Erscheinung an die Objektstruktur binden, bisher wenig bekannt ist, setzt das Training der Modelle meistens entweder an der Erscheinung oder der Struktur an. Die prinzipielle Schwierigkeit, den einen Aspekt zu modellieren, ohne den jeweils anderen zu kennen, schränkt viele Ansätze stark ein, beispielsweise bezüglich der Anzahl modellierbarer Objektansichten. Aus diesem Grund wird ein Projektvorschlag erarbeitet, der auf die Analyse der erscheinungsbezogenen und strukturellen Bildeigenschaften in gegenseitiger Abhängigkeit abzielt. Dabei soll die in dem vorliegend abgeschlossenen Projekt entwickelte robuste und bit-effiziente Darstellung der Erscheinung bezüglich einer effizienten, dichten Merkmalsabtastung weiterentwickelt werden. Die Struktur der Merkmale soll auf die Übereinstimmung mit der Klasse der Church-Rosser-Sprachen untersucht werden, welche die Perspektive auf eine lineare Laufzeit bieten. In den resultierenden Modellen sollen invariante Abhängigkeiten zwischen erscheinungsbasierten und strukturellen Modelleigenschaften identifiziert und auf ihre Eignung als Merkmal in der Objekterkennung untersucht werden. Der Projektvorschlag wird als DFG-Einzelantrag eingereicht werden. Der in diesem Projekt entwickelte neuartige Ansatz, die Einsatzgrenzen eines Filters durch den Filter selbst zu ermitteln und dabei geeignete Parametrisierungen selbst zu wählen, bietet ebenfalls vielversprechende Perspektiven bei weiterer Erforschung. Der Vorteil des Verfahrens liegt darin, dass eine manuelle Optimierung eines Filters auf spezielles Bildmaterial weitgehend automatisiert wird. Dadurch werden große Datenmengen für die digitale Bearbeitung und Analyse zugänglich.

Ausgehend von den im Forschungsprojekt erarbeiteten Methoden und Ergebnissen ohne zeitliche Ausdehnung wird beantragt werden, in einem Nachfolgeprojekt die wichtigsten dynamischen Ereignistypen in Informationsprogrammen in Bildschirmmedien zu analysieren, um die Unterschiede der Narrative zwischen ausgewählten westlichen und nicht-westlichen Bildschirmmedien aus Deutschland und den USA im Vergleich zu Brasilien und China herauszuarbeiten und ihre zeitliche Dynamik zu bestimmen. Dazu werden die Ereignistypen bezüglich der Handlungen der dargestellten Personen mit den Mitteln der digitalen Bildverarbeitung modelliert und automatisch im Videomaterial detektiert. Dies wird eine detaillierte Erfassung semantisch bestimmter Merkmale und damit eine Quantifizierung der Narrative in dem untersuchten Videomaterial erlauben.

2.5. Interdisziplinäre Weiterentwicklung

Die in den Arbeitspaketen der Medien- und Kommunikationswissenschaft entwickelten Ansätze zur Identifikation und Annotation von Personen auf der Grundlage von Keyframes dienen der Definition einer Trainingsmenge für einen Algorithmus überwachten Lernens. Im Zusammenhang mit der in den informatikbezogenen Arbeitspaketen entwickelten Merkmalsdarstellung sind diese von allgemeinem Interesse in allen Bereichen der Bildverarbeitung, die mit der Abstraktion oder dem Vergleich lokaler Merkmale zu tun haben (z.B. Objekterkennung, Mustervergleich, Stereoanalyse). Neben der Anwendung in den Medienwissenschaften wurde das Verfahren insbesondere in der Robotik bereits erfolgreich getestet. Eine entsprechende Publikation ist unter Begutachtung [SLHK]. Die Methode zur automatischen Schätzung der Anwendbarkeit eines Verfahrens wurde in der Dokumentbildverarbeitung zur Filterung historischer Schriften eingesetzt und auf einer entsprechenden Konferenz veröffentlicht [SF11a].

2.6. Verwertungspotenzial

Die im Rahmen des Projekts erzeugten Annotationen des Videomaterials von Jahresrückblickssendungen aus Deutschland und den USA liefern eine breite Datenbasis für die weitere Untersuchung ähnlicher Formate aus anderen Mediensystemen und -kulturen, mit deren Hilfe Unterschiede und Gemeinsamkeiten über nationale und kulturelle Grenzen hinweg aufgezeigt werden können.

Die Erkennung von Schnittgrenzen ist für die Medienwissenschaft von Bedeutung, da sie darstellerische und potentiell auch inhaltliche Einheiten in Videos trennen. Die Analyse von Videoschnitten lässt Rückschlüsse auf die Videostruktur zu und erlaubt anhand statistischer Kriterien die Abgrenzung von Material aus verschiedenen Domänen. Eine automatische Schnitterkennung ist auch als Vorverarbeitungsschritt für nachfolgende Bildverarbeitungsalgorithmen relevant, um die zeitliche Vermischung von Merkmalen über darstellerische Brüche hinweg zu verhindern.

Die erzielte Erkennung von Gesichtern nicht nur aus Frontal-, sondern insbesondere auch aus Seitansichten ist von unbestrittenem Interesse für Anwendungen in der Videoüberwachung, dem Video-Retrieval, der Robotik, der Mensch-Maschine-Interaktion, der Unterhaltungsindustrie, mobilen Anwendungen und sozialen Netzwerken. Die entwickelten Methoden der Videoinhaltsanalyse dienen der Verwaltung und Analyse von Multimediadaten. Sie fördern damit sowohl die inhaltliche Weiterverarbeitung als auch Erforschung von Videomaterial durch Redakteure beziehungsweise Medien- und Kommunikationswissenschaftler.

2.7. Beteiligte Wissenschaftler

Antragsteller:

Prof. Dr. Otthein Herzog:

- Aufbau einer funktionierenden Kooperation zwischen Bildverarbeitung und Kommunikationswissenschaft
- Projektleitung der Bildverarbeitung
- Mitarbeit bei der Ausarbeitung der Konzepte für die Lösung der Forschungsaufgaben, insbesondere für die Aufarbeitung der höheren semantischen Funktionalitäten.
- Mitarbeit an Publikationen: [SDH11], [SH09a], [SH09b], [SH10], [SM11], [WSH11].

Prof. Dr. Dr. (USA) Peter Ludes:

- Aufbau einer funktionierenden Kooperation zwischen Kommunikationswissenschaft und Bildverarbeitung,
- Projektleitung der Kommunikationswissenschaft; Vorstellung der Ergebnisse z.B. in Cross Unit Theme Session, Chair und Keynote der International Communication Association 2009,
- auf Sitzungen der European Sociological Association 2009, 2011 und 2012
- Integration ausgewählter Ergebnisse in „Elemente internationaler Medienwissenschaften“ 2011 [Lud11] und „Visuelle Rhetorik“ [Lud12].

Projektmitarbeiter (in alphabetischer Reihenfolge):

Juliana Costa: Aufbereitung von Forschungsergebnissen, Video-Inventarverwaltung und Katalogisierung, Schulung und Anleitung von Kodierern.

PD Dr.-Ing. habil. Björn Gottfried: Untersuchung zur Modellierbarkeit relativer Objektbewegungen [Got11].

Tobias Kohler: Aufbereitung von Forschungsergebnissen, Beschaffung von deutschen und US-amerikanischen Video-Samples, Video-Inventarverwaltung und Katalogisierung, Organisation eines Workshops, in dessen Folge [L11] veröffentlicht wurde, Schulung und Anleitung von Kodierern.

Dr. Hartmut Messerschmidt: Mitarbeit auf dem Gebiet formale Sprachen/Code books, spezialisiert auf Church-Rosser Picture Languages und ihre Anwendung in der Bildanalyse [SM11b].

Dr. Jan Müller: Entwicklung des digitalen Videoinhaltsanalyse-Tools ‚Kivi' inkl. Benutzerdokumentation, Veröffentlichung wissenschaftlich-methodischer Beiträge in [MS10a] und [SM11].

Sarah-Elisa Nees: Systematische Datenbankrecherche und Zusammenstellung des Video-Samples der deutschen und US-amerikanischen Fernsehjahresrückblicke, Sample-Einteilung, Anpassung des Kodierbogens, Anleitung von Kodierern.

Dr.-Ing. Martin Stommel: Binäre Merkmalsdarstellung, Gesichtserkennung, Szenenklassifikation, Veröffentlichung wissenschaftlich-methodischer Beiträge in [MS11a], [MS11b], [SM11], [SDH11], [SH09a], [SH09b], [SH10], [SLHK11], [SF11], [Sto10], [WSH11].

2.8. Qualifikation des wissenschaftlichen Nachwuchses

An der Universität Bremen erwarb Frau Martina Duemcke den Bachelor-Grad und Herr Thiemo Wiedemeyer das Diplom mit Abschlussarbeiten im Rahmen des Projekts. Beide Abschlussarbeiten wurden auf internationalen Konferenzen publiziert. Frau Duemcke setzte ihre Master-Studien an der ETH Zürich fort, Herr Wiedemeyer ist gegenwärtig Doktorand an der Universität Bremen.
Frau Sarah Nees erhielt im Anschluss an ihre wiss. Mitarbeit ein Doktorandenstipendium in der Bremen International Graduate School of Social Sciences. Herr Jan Müller erhielt im Anschluss an seine wiss. Mitarbeit eine langfristige wiss. Mitarbeiterstelle an der Universität Lüneburg. Herr Tobias Kohler ist, mit anfänglicher Unterstützung durch ein Förderprogramm der Stadt Bremen, in der Softwarebranche tätig. Frau Costa ist Doktorandin in Intercultural Humanities an der Jacobs University Bremen.

3. Zusammenfassung

Obwohl das Format der Fernsehjahresrückblicke sowohl qualitativ als auch quantitativ über verschiedene Medienkulturen hinweg ausdifferenziert ist, konnte die Stabilität des Auftretens bestimmter nationaler, trans-kultureller und globaler Schlüsselbilder gezeigt werden. An den thematischen Schwerpunkten der Jahresrückblicke in den USA und Deutschland im vergangenen Jahrzehnt zeigen sich historische Machtdarstellungsunterschiede sowie die begrenzte Wahrnehmung internationaler Beziehungspartner. So war Deutschland für das 20. Jahrhundert im Fokus der US-Rückblicke – im vergangenen Jahrzehnt hingegen der Irak. Die berücksichtigten chinesischen Fernsehjahresrückblicke zeigen, wie sehr die USA als Gegenüber wahrgenommen werden. Differenzierte Analysen nach Ländern, Jahren, Genre und Akteurstypen wurden durchgeführt; denn erst eine mit Text- und Statistikauswertungen gleich berechtigte Konzentration auf visuelle Daten kann sonst unterbelichtete Perspektiven eines neuen Strukturwandels der Öffentlichkeit erhellen und hierdurch lassen sich nicht in Worte übersetzbare Semantiken als Schlüssel zu sonst unsichtbar bleibenden sozialen Figurationen und Prozessen erschließen. Die unterschiedlichen Darstellungshäufigkeiten von Staatsoberhäuptern zu einfachen Leuten z, B. in den von uns untersuchten Fernsehjahresrückblikken aus Brasilien, China, Deutschland und den USA, hier nur für 2008 bis 2010, sind signifikant, für Deutschland 1:4,6, die USA 1:2,2, Brasilien 1:2,3 und China 1:20,0. Damit sind die brasilianischen, deutschen und US-amerikanischen Machtpräsentationen wesentlich ähnlicher untereinander als mit China.

Um die Analysen durchführen zu können, wurden Werkzeuge zur manuellen und automatischen Videoannotation erstellt. Die Software erlaubt den Datenexport zu Statistikwerkzeugen für weitere Analysen. Automatisierte Übereinstimmungstests anhand der Messung von Krippendorffs Alpha zwischen verschiedenen Annotationen lassen auf die Verlässlichkeit des Datenmaterials schließen. Die Annotation wurde durch automatische Verfahren zur Erkennung von Schnittgrenzen, Gesichtern und verschiedenen Arten von Schlüsselbildern unterstützt. Hierzu wurde ein Gesichtsdetektor entwickelt, der nicht nur frontale, sondern auch stark seitliche Ansichten erkennt. Eine effiziente Merkmalsdarstellung wurde entwickelt, welche einen schnellen und numerisch stabilen Vergleich visueller Muster erlaubt, welcher nicht durch die Hochdimensionalität der Vektordarstellungen beeinträchtigt ist. Es wurden erfolgreiche Anwendungen des Verfahrens in der Objekterkennung, Gesichtserkennung, Stereoanalyse und dem Objekttracking demonstriert und damit die Grundlagen für die automatisierte Erkennung eines Satzes von Key Visuals geschaffen.