Preprints der Zeitschrift *BIBLIOTHEK – Forschung und Praxis*, 2020, AR 3375 **Pustu-Iren et al.**

AR 3375

Kader Pustu-Iren, Joanna Bars, Markus Mühling, Nikolaus Korfhage, Angelika Hörth, Bernd Freisleben und Ralph Ewerth

Videomining in historischem Material – ein Praxisbericht

Projekt „Visuelle Informationssuche in Video-Archiven“ (VIVA)

Zusammenfassung: Videomining-Algorithmen wie die visuelle Konzeptklassifikation und Personenerkennung sind unerlässlich, um eine feingranulare semantische Suche in großen Videoarchiven wie der historischen Videosammlung der ehemaligen Deutschen Demokratischen Republik (DDR) des Deutschen Rundfunkarchivs (DRA) zu ermöglichen. Wir stellen das Projekt VIVA, unsere Ansätze zur Videoanalyse sowie das VIVA-Softwaretool vor. Letzteres ermöglicht Anwender*innen auf einfache Art, Trainingsdaten zu sammeln, um neue Analysealgorithmen zu trainieren.

Schlüsselwörter: Video Mining Tool, Videoanalyseverfahren, Konzepterkennung, Personensuche, Ähnlichkeitssuche

Video mining on Historical Footage–A Practical Report
Project “Visual Information Search in Video Archives” (VIVA)

Abstract: Video mining algorithms such as concept classification and person recognition enable fine-grained semantic search in large video archives like the historical collection of the former German Democratic Republic (GDR) of the German Broadcasting Archive (DRA). We present the project VIVA, our deep learning approaches, and the VIVA software tool, which allows users to easily acquire data to train analysis algorithms.

Keywords: Video Mining Tool, Historical Video Indexing, Visual Concept Classification, Person Search, Similarity Search

1 Einleitung

Das DRA ist die älteste Gemeinschaftseinrichtung der ARD und archiviert und dokumentiert seit 1952 historisch bedeutende audiovisuelle Medien. Die Schwerpunkte der Bestände sind der Rundfunk vor

1945 sowie das DDR-Fernsehen und der DDR-Hörfunk bis 1991. Am Standort Babelsberg verwaltet das DRA das Programmvermögen des Deutschen Fernsehfunks (DFF) bzw. des Fernsehens der DDR. Die Bestände erstrecken sich von der ersten Sendung 1952 bis zur Einstellung des Sendebetriebs 1991. Das DRA bildet somit die singuläre Anlaufstelle für wissenschaftliche Forschungen in diesem Bereich. Aufgrund seiner Einmaligkeit bietet der Bestand zahlreiche Anknüpfungspunkte für eine Vielzahl von Forschungsfragen, die sowohl die Lebenswelten von DDR-Bürgern, als auch soziokulturelle Funktionen des Fernsehens in der DDR betreffen.

Um die Suche in Videos zu erleichtern, strebt das DRA die Digitalisierung und Indexierung der gesamten Videosammlung an. Wegen der zeitaufwändigen Aufgabe der manuellen Video-Annotation konzentrieren sich menschliche Annotationen auf größere Videosequenzen und Kontexte. Darüber hinaus ist das Auffinden ähnlicher Bilder in großen Multimedia-Archiven manuell nur in begrenztem Maße möglich. Daher bedarf es Methoden der automatisierten Videoanalyse.

Die Bildqualität des dem DRA anvertrauten historischen Materials ist sehr heterogen – so finden sich sowohl Farb-, als auch Schwarz-Weiß-Aufnahmen, Bilder mit geringer Auflösung und teilweise sogar vom Bildschirm abgefilmte Aufnahmen in den Beständen. Dies stellt eine Videomining-Software vor besondere Herausforderungen.

In dem DFG-geförderten Projekt *Visuelle Informationssuche in Video-Archiven (VIVA)* stellen sich das DRA, die Philipps-Universität Marburg und die Technische Informationsbibliothek (TIB) Hannover dieser Herausforderung. In der dreijährigen Projektlaufzeit von 2018 bis 2020 wurden Modelle für 100 zum Teil DDR-spezifische visuelle Konzepte und 100 Persönlichkeiten der ehemaligen DDR trainiert, die im Material des DDR-Fernsehens erkannt werden sollen. Weiterhin wurde eine Ähnlichkeitssuche entwickelt, die eine Trefferliste von Keyframes basierend auf einem einzigen Anfragebild generiert („query by example“). Diese wurde in die Fernseharchivdatenbank FESAD der ARD integriert und befindet sich im Testbetrieb. Ein weiteres Ergebnis des Projekts wird eine Open Source Software sein, die alle Schritte von der Akquise des Trainingsmaterials, über die Annotation einzelner Bilder bis hin zu Training und Verwaltung der Modelle integriert. Ziel ist es, dass Dokumentar*innen den Trainingsprozess mit einer benutzerfreundlichen Oberfläche durchführen und bei Auftreten neuer Anforderungen selbst neue Personen oder Konzepte in das Training der zugrundeliegenden maschinellen Lernmodelle aufnehmen können.

Die folgenden Abschnitte geben einen Überblick zu den im Projekt VIVA durchgeführten Arbeiten, beginnend bei der Sichtung und Auswahl der Trainingsdaten für die visuelle Konzeptklassifikation und Personenerkennung (Abschnitt 2), die entwickelten Algorithmen und Modelle zur Videoanalyse (Abschnitt 3), sowie zum Softwaretool VIVA (Abschnitt 4).

2 Sichtung und Auswahl von Trainingsmaterial

Der erste Arbeitsschritt seitens des DRA war die Festlegung der Konzept- und Personenlexika. Ziel war es, jeweils 100 Personen und Konzepte im Laufe des Projekts zu trainieren. Da sich das Projekt die Unterstützung und Verbesserung der Recherche in DRA-Beständen auf die Fahnen geschrieben hat, lag dieses Anliegen im Fokus des Interesses. Somit begann im DRA als erstes die Erhebung von Bedarfen aus der Recherche. Fachkolleg*innen aus dem Nutzerservice, der Fernsehrecherche und der Rechteklärung wurden schriftlich und mündlich befragt: Welche Personen oder Konzepte werden in der Recherche besonders häufig nachgefragt, was ist schwierig zu finden, welche Recherchethemen sind jetzt bzw. immer relevant oder könnten demnächst relevant werden? Aus der Auswertung dieser Fragen entstand in Zusammenarbeit mit den Forschungspartnern und ihrem Input aus technischer Sicht eine Liste von 100 Personen und 100 Konzepten, die zur Umsetzung vorgesehen wurden. Die Personenliste wurde aus dem gesamten Sendezeitraum des Deutschen Fernsehfunks und Fernsehens der DDR zusammengestellt und reicht von heute in der breiten Öffentlichkeit weitgehend vergessenen Korrespondenten (z. B. Maxi Haupt) und Komikern der 60er und 70er Jahre bis zu nach dem Ende der DDR bundesweit hochrelevanten Persönlichkeiten – prominentestes Beispiel ist hier sicherlich Angela Merkel. Die visuellen Konzepte decken verschiedenste Kategorien ab, darunter DDR-Alltagsinstitutionen und -infrastruktur (Bäckerei, Fahnenappell, Grillen), Bildeigenschaften (Vogelperspektive, Zeichentrick) oder konkrete Objekte (Fernsehgerät, Hund). Eine der größten Herausforderungen beim Training von Deep-Learning-Verfahren ist die effiziente Bereitstellung eines qualitativ hochwertigen Trainingsmaterials. Auf Seiten des DRA konnte dafür auf bereits bestehende Annotationen mit einer sequenzgenauen Bildbeschreibung zurückgegriffen werden. Trotz des bereits sorgfältig dokumentierten Ausgangsmaterials war die Annotation mit viel Aufwand verbunden, da die bisher für den menschlichen Nutzer ausreichende Annotation (sequenzbasiert, sekundengenau) zu einer framegenauen Annotation erweitert werden musste. Fünf Dokumentar*innen unterstützen das Projekt darin, in einem Annotationstool das Trainingsmaterial zu erstellen und dessen Qualität zu sichern. Im Sinne einer Reduktion von Aufwand wurde das domänenspezifische Bildmaterial des DRA durch eine weitere Materialquelle ergänzt: aus der Kollektion von Google Open Images konnten die Dokumentar*innen zusätzlich tausende passende Bilder auswählen, da es zu den DRA-Konzepten vielfach bereits passende Klassen in Open Images gab. Als hilfreiche Arbeitserleichterung erwies sich das im Projekt bei der Annotation von Personen angewandte Verfahren, die Trainingsbilder automatisch nach Ähnlichkeit der abgebildeten Gesichter zu gruppieren. Da hierdurch nunmehr nicht einzelne Gesichter, sondern ganze Cluster (Gruppen von Bildern) von den annotierenden

Dokumentar*innen bewertet werden mussten, konnte der Annotationsaufwand deutlich verringert werden.

Nach einem ersten Training der Konzepte und Personen wurde das Ergebnis den Experten des DRA noch einmal zur Bewertung vorgelegt. Hierzu wurde auf einer Datenmenge des DRA-Bestands, der nicht zum Trainieren verwendet wurde, ein Test mit entsprechenden Suchanfragen zu Konzepten und Personen durchgeführt. Die hierfür vom System zurückgelieferten ersten 200 Ergebnisse der Trefferliste für jedes Konzept und jede Person wurden dann einer Feedbackannotation unterzogen. So konnte eine erste Einschätzung von der Güte der ersten trainierten Modelle und gleichzeitig weiteres wertvolles Trainingsmaterial zur Verfeinerung der Konzepte gewonnen werden.

Die Ähnlichkeitssuche des VIVA-Projekts wurde bereits testweise in die Fernsehdatenbank FESAD integriert, um erste Erfahrungen und Reaktionen der Nutzer*innen einholen zu können. Parallel zur Weiterentwicklung der Suchtechnologie wird auch weiter an der Darstellung der Ergebnisse gefeilt.

3 Entwickelte Verfahren zur Videoanalyse

Deep-Learning-Ansätze, insbesondere neuronale Netze, werden heutzutage in nahezu allen Bereichen der Computer Vision eingesetzt und übertreffen die menschliche Leistungsfähigkeit in vielen Bereichen wie der Gesichts- oder Konzepterkennung sogar. Diese Deep-Learning-Ansätze zählen zum maschinellen Lernen, das ein Teilgebiet der Künstlichen Intelligenz ist. Die Integration dieser neuen Technologien und Nutzung automatischer Videoannotation in Rundfunk- und Fernsehanstalten ermöglichen, das enorme Potenzial der verfügbaren Multimediadaten zu nutzen. Doch die Automatisierung solcher Verfahren stellt immer noch eine Herausforderung dar, da in der Regel eine Vielzahl von verschiedenen Personen und Konzepten erfasst werden müssen, manchmal über lange Zeiträume und mit einer Vielzahl von Vorkommnissen. Darüber hinaus muss mit der zunehmenden Größe der Archive umgegangen werden. Im Folgenden stellen wir Methoden zur Konzepterkennung, Ähnlichkeitssuche und Gesichtserkennung vor, mit denen wir uns diesen Herausforderungen stellen. In einem ersten Schritt werden die Videosequenzen zeitlich in einzelne Kameraeinstellungen segmentiert und auf dieser Basis Keyframes erzeugt.[1]

3.1 Konzepterkennung

Zur automatischen Erkennung der DRA-spezifischen Konzepte wurde ein Deep-Learning-Ansatz eingesetzt. Es handelt sich um ein künstliches neuronales Netz in Form eines Convolutional Neural

[1] Mühling et al. (2007).

Networks (CNN), das anhand eines Trainings mit zigtausenden Beispielbildern die Fähigkeit „erlernt", verschiedene visuelle Konzepte zu erkennen und zu unterscheiden. Deep CNNs sind flexibel, da die verwendeten und für die Konzepte charakteristischen Merkmale nicht mehr von Hand definiert, sondern automatisch gelernt werden. Solche „tiefen" neuronalen Netze haben eine potenziell hohe Anzahl von Schichten von Neuronen und werden aktuell verstärkt für komplexe Problemstellungen in der Bildverarbeitung eingesetzt, wie z.B. für die Bilderkennung.[2] Um die Genauigkeit der Modelle weiter zu verbessern, basieren die neuesten Netzwerkarchitekturen auf der Neural Architecture Search.[3] Hierbei handelt es sich um eine Alternative zur zeitaufwendigen manuellen Netzwerkarchitekturoptimierung. Es wird ein Meta-Lernverfahren genutzt, um nicht nur die Gewichte, sondern die Netzwerkarchitektur selbst zu lernen. Eine für die Bilderkennung optimierte Architektur ist das NASnet,[4] die auch in unserem Fall für die Erkennung der DRA-Konzepte eingesetzt wurde.

Im Bereich der Bilderkennung beschränken sich die Ansätze meist auf „Single-Label" Probleme, bei denen jedem Bild genau ein Label zugeordnet wird. In der Realität, so auch in unserem Fall, handelt es sich um Multi-Label Daten, bei denen in einem Bild mehrere Konzepte auftreten können. Zu diesem Zweck wurde die Netzwerkarchitektur angepasst und eine spezielle Fehlerfunktion entwickelt, die im Falle von Multi-Label Daten zu einer für das Training günstigeren Gewichtung von positiven und negativen Trainingsbeispielen sorgt. Neben der Netzwerkarchitektur hängt die Erkennungsgenauigkeit eines neuronalen Netzes entscheidend von der Qualität des Trainingsmaterials ab. Damit das neuronale Netz für jedes Konzept lernt, wie es sich von anderen Konzepten unterscheidet, muss es möglichst viele Beispielbilder „gesehen" haben. Je nach Komplexität eines Konzepts mussten mindestens 100 Beispielbilder gefunden werden. Eine große Bandbreite annotierter Bilder, die unterschiedliche Jahres- und Tageszeiten, Jahrzehnte und Perspektiven der einzelnen Konzepte widerspiegelt, sichert das Erkennen in unbekanntem Bildmaterial.

Um den manuellen Aufwand der Trainingsdatenakquise zu minimieren, wurde auf eine iterative Verbesserung der Modelle gesetzt (siehe Abb. 1). Ausgehend von einer initialen Menge von Trainingsdaten, die sowohl aus DRA-spezifischen als auch aus bereits annotierten öffentlich verfügbaren Trainingsbeispielen besteht, wurde ein Basismodell trainiert und evaluiert. Dieses Modell wurde auf die unbekannten Testdaten angewendet, um diejenigen Keyframes zu finden, die die gesuchten Konzepte enthalten. Diese Ergebnisse werden den Benutzer*innen als Antwort auf

[2] Mühling et al. (2017).
[3] Tan et al. (2019).
[4] Zoph et al. (2017).

eine Suchanfrage in Form einer Trefferliste zur Feedbackgenerierung präsentiert. Die mittels Benutzerfeedback positiv bzw. negativ gekennzeichneten Beispiele werden dem Trainingsmaterial hinzugefügt und führen somit zu einer iterativen Verbesserung. Der aktuelle Trainingsdatensatz umfasst ca. 140000 Bilder. Auf einem Testdatensatz mit 9231 Keyframes wurde mit dem zuvor beschriebenen Ansatz eine sehr gute durchschnittliche Average Precision von 80,0% erreicht. Die Average Precision ist ein Gütemaß, das die Qualität einer Trefferliste beurteilt und bei der Berechnung des (hier wiederum verwendeten) Durchschnitts der Präzision der Liste (über alle Ränge mit korrektem Ergebnis) die weit oben platzierten Treffer stärker gewichtet. Die Abb. 1 zeigt die Top-15-Trefferliste für das Konzept „Bäckerei“, die in ungefähr 20000 Stunden historischem Videomaterial des DRA gefunden wurden.

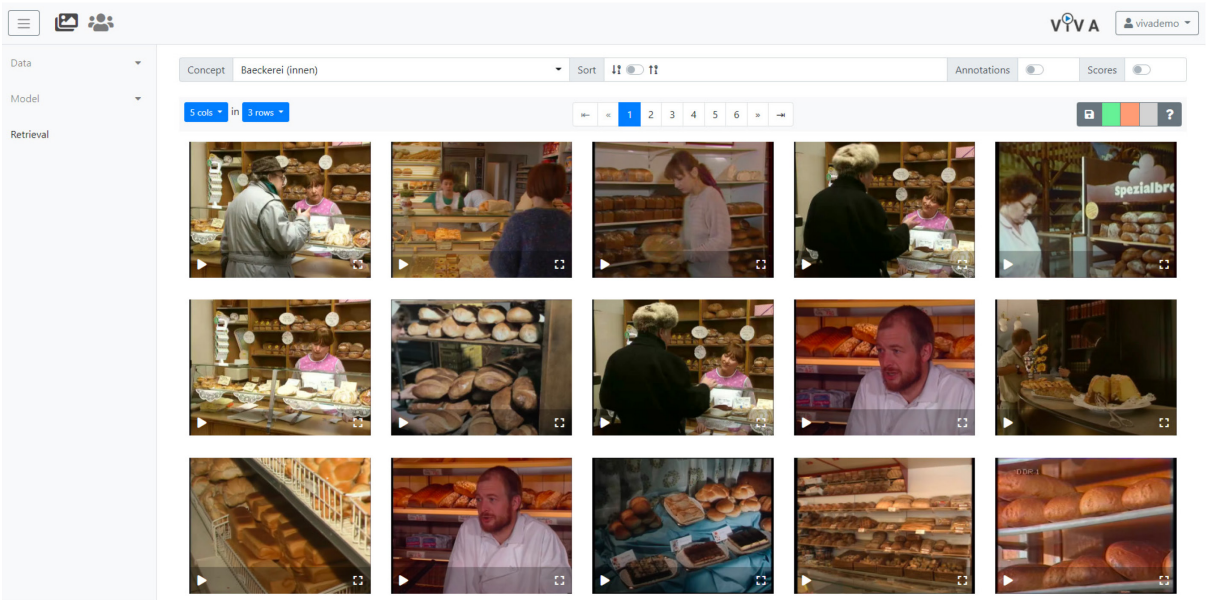


Abb. 1: Trefferliste für das visuelle Konzept „Bäckerei“

3.2 Ähnlichkeitssuche

Die bildbasierte Ähnlichkeitssuche erlaubt es, Anfragen an das System in Form von Beispielbildern zu stellen. Hierbei wählen Benutzer*innen ein Anfragebild aus, welches die Suchintention möglichst gut widerspiegelt. Dieses Anfragebild kann entweder aus externen Quellen stammen (Upload eines Bildes oder Eingabe einer Webadresse) oder auch ein Bild aus dem bereits erschlossenen Bestand sein. Als Ergebnis wird eine sortierte Trefferliste zurückgeliefert, die visuell ähnliche Bilder enthält. Im VIVA-Tool besteht die Möglichkeit, die Ähnlichkeitssuche im Rahmen der Datenakquise zu verwenden. Hier wird zunächst eine Bildersuche im Web eingesetzt (Web Crawling), um Trainingsbilder zu finden. Die so erhaltenen Trainingsbilder werden beispielsweise genutzt, um ähnliche Bilder im Videobestand des DRA zu suchen und entsprechend zu annotieren. So ermöglicht

die Ähnlichkeitssuche, rasch eine große Menge an Trainingsmaterial für (neue) Konzepte im gesamten Bestand zu erschließen, wobei der manuelle Annotationsaufwand in vielen Fällen stark reduziert wird.

Das entwickelte Verfahren zur Ähnlichkeitssuche basiert auf Merkmalsrepräsentationen, die mithilfe von tiefen neuronalen Netzen in Form von CNNs gelernt werden.[5] Der Vorteil dieser Methode besteht in einer Ähnlichkeitsfunktion, die dem menschlichen Ähnlichkeitsempfinden besonders nahe kommt. Um eine effiziente Suche in großen Bild- und Videobeständen zu ermöglichen wurde eine Netzwerkarchitektur entworfen, die Bilder auf Binärcodes abbildet. Hierzu wurde die NASNet Architektur des neuronalen Netzes um geeignete Schichten für die Kodierung der Merkmale und Fehlerfunktion beim Trainieren (Encoding und Loss Layer) erweitert. Das Modell zur Ähnlichkeitssuche wurde sowohl auf öffentlich verfügbaren Trainingsdaten (Places[6] und ImageNet[7]) als auch auf DRA-spezifischem Trainingsmaterial trainiert, um eine gute Generalisierbarkeit zu erreichen. In der Anwendungsphase erlaubt die kompakte Repräsentation der gelernten Binärcodes einen schnellen Abgleich (Matching) von Bildern. Zur weiteren Optimierung der Laufzeit wurde ein zweistufiger Ansatz mit verschieden großen Binärcodes verfolgt. Hierbei wird in einem ersten Schritt eine sehr effiziente Grobsuche basierend auf 64-Bit Binärcodes unter der Verwendung von Indexstrukturen durchgeführt. Im zweiten Schritt werden 256-Bit Binärcodes genutzt, um die potentielle Trefferliste aus der Grobsuche nochmal zu verfeinern. Dieser Ansatz wurde für die Integration in FESAD in Elastic Search[8] realisiert. Die Laufzeit einer Anfrage in Elastic Search auf einem Bestand von ca. 10 Mio. Bildern beträgt rund 256 Millisekunden auf einem Intel-Core-Prozessor i7-4771 mit 3,5 GHz Taktfrequenz. Die Abb. 2 zeigt die Trefferliste für ein Anfragebild aus dem Archivbestand zum Thema „Demonstration“.

[5] Mühling et al. (2019), Korfhage et al. (2020).
[6] Zhou et al. (2017).
[7] Deng et al. (2009).
[8] https://www.elastic.co/.

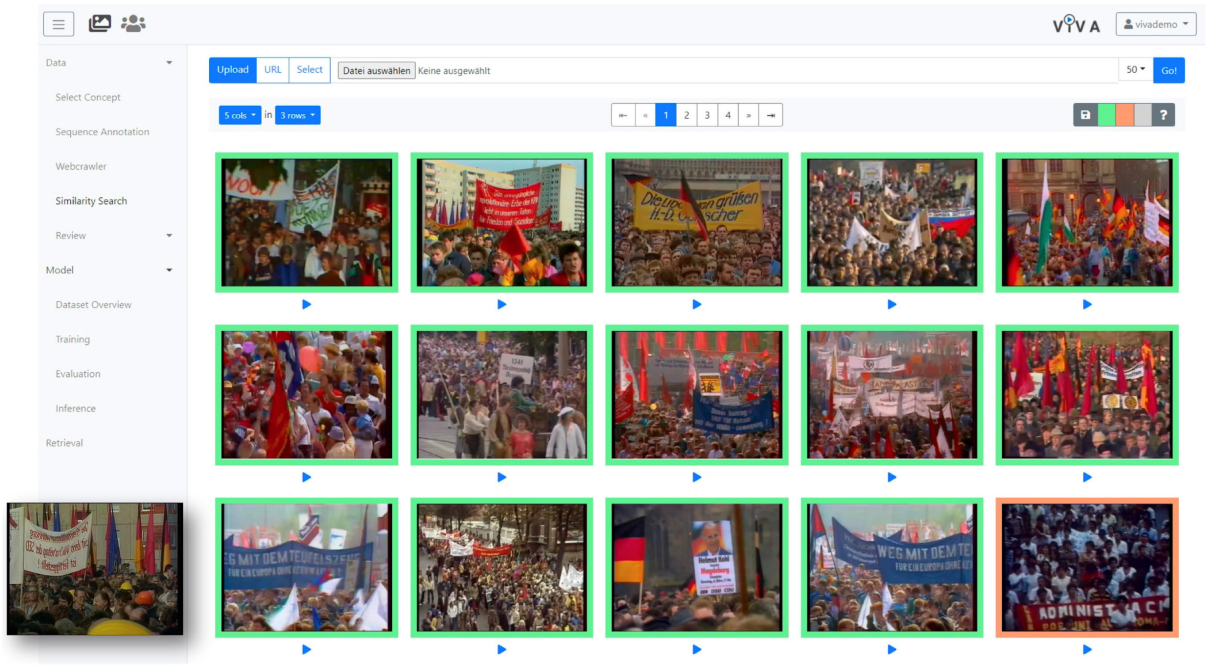


Abb. 2: Ergebnisse der Ähnlichkeitssuche für das angezeigte Anfragebild (links unten) einer Demonstration

3.3 Personenerkennung

Auf Grundlage des vom DRA definierten Personenlexikons wurde ein Gesichtserkennungsansatz entwickelt, der mehrere Schritte umfasst. So besteht die Folge Schritte zur Gesichtsverarbeitung aus den folgenden Komponenten: Gesichtsdetektion, Gesichtsausrichtung (Face Alignment), Merkmalsextraktion und Gesichtserkennung. Um Gesichter im historischen Bildmaterial zunächst zu detektieren, wird der RetinaFace-Detektor[9] verwendet. Eine anschließende frontale Ausrichtung der detektierten Gesichter mittels des dlib Shape Predictors[10] sorgt dafür, dass detektierte Gesichter in ihrer Pose normalisiert werden.

Um den Annotationsaufwand zu verringern und unerwünschte (Hintergrund-)Personen in Bildern zu eliminieren, wurden detektierte Gesichter im Trainingsmaterial mittels eines agglomerativen Clustering-Verfahrens gruppiert. Für jede zu trainierende Person wurde der Annotationsaufwand auf jeweils 100 Cluster beschränkt. Resultierende Trainingsbilder der DDR-Persönlichkeiten werden durch Merkmalsvektoren repräsentiert, die auf neuronalen Netzen beruhen. Für die Merkmalsextraktion wird FaceNet[11] verwendet. Das hier verwendete Modell[12] wurde auf einem einschlägigen Datensatz (VGGFace2[13]) mit über 9000 Identitäten und 3,3 Mio. Bildern trainiert und

[9] Yang et al. (2016).
[10] Kazemi et al. (2014).
[11] Schroff et al. (2015).
[12] https://github.com/davidsandberg/facenet.
[13] Cao et al. (2018).

erreicht auf dem etablierten LFW-Benchmark (LFW: Labeled Faces in the Wild) eine sehr hohe Genauigkeit von 99,65%. Für die Identifikation der DDR-Persönlichkeiten wurde ein Klassifikator mit Support Vector Machines (SVM) trainiert, der neue Gesichter mittels Merkmalsvektoren diesen Personen zuordnet und Wahrscheinlichkeiten schätzt. Auf einem Testdatensatz, bestehend aus insgesamt 5762 annotierten Keyframes und Webbildern wurde mit dem zuvor beschriebenen Ansatz für die DDR-Persönlichkeiten eine Mean Average Precision von 85,0% erreicht.

Der beschriebene Ansatz wurde verwendet, um Keyframes des DRA-Videobestands nach dem Vorkommen der Persönlichkeiten zu durchsuchen und diese mit entsprechenden Annotationen und Wahrscheinlichkeiten zu versehen. Der daraus resultierende Index von Personenvorkommen kann schließlich bei Suchanfragen verwendet werden. Abb. 3 zeigt die Top-15-Trefferliste für die Person Maybrit Illner, die in dem historischen Videomaterial des DRA gefunden wurden.

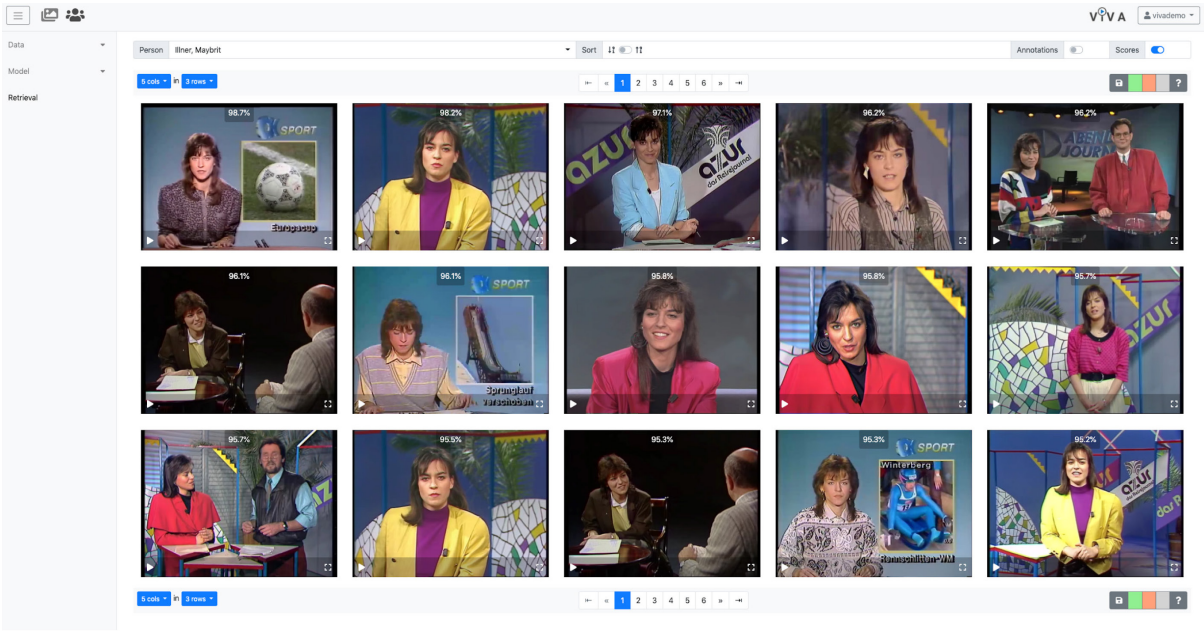


Abb. 3: Trefferliste der Personenerkennung für die Suchanfrage „Maybrit Illner“

3.4 Weitere Bildanalyseverfahren für Videoarchive

Im Zuge des Projekts sind auch weitere Verfahren und Forschungsarbeiten zu Bild- und Videoanalyse in Video-Archiven entstanden. So untersucht eine gemeinsame Arbeit[14] der Projektpartner, ob der Schwierigkeitsgrad eines Konzepts Rückschlüsse auf die benötigte Trainingsmenge und die daraus resultierende Performanz des Konzepts zulässt. Hierfür wurde die Inter-Coder-Übereinstimmung zwischen Laien und Experten bei der Annotation von Bildern für DDR-spezifische Persönlichkeiten

[14] Pustu-Iren et al. (2019).

und Konzepte ermittelt. Unter anderem hat die Studie ergeben, dass Archivmitarbeitende die Personen der DDR-Geschichte konsistenter annotieren und besser identifizieren konnten als die Nicht-Expert*innen. Außerdem wurde am Beispiel von Konzepten gezeigt, dass Erkennungsraten aus der Konsistenz, d.h. der Qualität der Annotationen der Trainingsdaten, sowie der Anzahl der Trainingsbilder geschätzt werden können. Die Arbeit könnte zukünftig derart ausgebaut werden, dass die benötigte Trainingsmenge je nach „Schwierigkeit" der zu lernenden Person oder des Konzepts geschätzt wird, um den Annotationsaufwand zu optimieren und somit möglichst gering halten zu können.

Des Weiteren wurden Methoden zur (weltweiten) Schätzung des Aufnahmeorts von Bildern entwickelt,[15] die auf neuronalen Netzen basieren. Diese beziehen Kontextinformationen eines Bilds ein, um das Bild mit dem für die erkannte Szene (Stadt, Natur, Innenraum) trainierten Netz in eine der vordefinierten Zellen auf der Erdoberfläche einzuordnen. Die Methoden zur Geolokalisierung werden derzeit erweitert, um eine Schätzung des Aufnahmeorts in historischem DDR-Material zu ermöglichen.

Außerdem wurde ein System zur Personenerkennung im Bildinhalt von im Internetarchiv gespeicherten Webnachrichten entwickelt.[16] So ermöglicht das System, Beziehungen von Personen der Öffentlichkeit genauer zu erschließen. Interessante Personen einer Domäne wie z.B. Politik werden mithilfe von Bildern aus dem Web automatisch identifiziert und anschließend im Bildmaterial des Internetarchivs mittels eines effizienten Merkmalvergleichs erkannt. Der Ansatz wird anhand von zwei Beispieldomänen, Politik und Unterhaltung, demonstriert und umfasst die Visualisierung von Ko-Okkurrenzen zwischen Personen.

4 Software-Tool VIVA

Die zuvor beschriebenen Algorithmen zur Konzepterkennung, Ähnlichkeitssuche, Gesichtsdetektion sowie -erkennung sind in einem Videoanalyse- und Such-Tool integriert. Die im Rahmen des Projekts entstandene Software ermöglicht es Fachanwender*innen bestehende Videoanalyseverfahren anzupassen, um Modelle für neue Personen oder Konzepte zu lernen. Hierfür bietet die Client/Server-basierte Webanwendung eine intuitive Benutzeroberfläche (siehe Abb. 1 bis Abb. 3). So bietet die grafische Nutzeroberfläche einerseits Bereiche für das Daten- und Trainingsmanagement, um neue, inhaltsbasierte Videoanalysen durchzuführen und andererseits einen Bereich, um Suchergebnisse zu visualisieren.

[15] Müller-Budack et al. (2018a).
[16] Müller-Budack et al. (2018b).

Das Datenmanagement hilft im Einzelnen dabei, semi-automatisch neue Trainingsbilder zu akquirieren. Hierfür muss zunächst eine neue Klasse (Konzept oder Person) definiert oder eine in der Tabellenübersicht bestehende ausgewählt werden. Dabei wird angezeigt, wie viele positiv, negativ oder neutral annotierte Bilder für bestehende Klassen bereits verfügbar sind. Das Tool ermöglicht es, Bilder aus dem Web zu „crawlen“, eigenes Bildmaterial hochzuladen oder weitere Bilder aus dem Archivbestand mittels der Ähnlichkeitssuche vorzuschlagen. Potentielle Trainingsbilder der Klasse können jeweils über eine benutzerdefinierbare Gitteransicht effizient annotiert werden.

Funktionalitäten im Trainingsmanagement sind an die Bedürfnisse der Fachanwender angepasst. So kann ein neuer Trainingsprozess nach der Annotationsphase durch einen einzigen Knopfdruck angestoßen werden. In einem vorherigen Schritt gibt eine Datensatzübersicht in Form eines Balkendiagramms Aufschluss darüber, ob genügend Trainingsbeispiele pro Klasse vorhanden sind und insbesondere ein balancierter Trainingsdatensatz vorliegt. Während des Trainings werden vereinfachte Informationen zum Trainingsfortschritt angezeigt. Die Güte des Modells wird automatisch ausgewertet und dem Benutzer ebenfalls in Form eines Balkendiagramms für die einzelnen Klassen präsentiert. Gegebenenfalls können auch Testergebnisse des vorherigen Modells betrachtet werden. Wenn die Ergebnisse des trainierten Modells zufriedenstellend sind, kann in einer weiteren Komponente des Bereichs die neue Indexierung des Archivmaterials angestoßen und über angezeigte Fortschrittsinformationen überwacht werden.

Die Ergebnisse zu einer gewählten Person bzw. zu einem Konzept werden den Nutzer*innen in Form einer Rangliste von Videosequenzen präsentiert, in der die einzelnen Videoaufnahmen durch das Schlüsselbild mit der höchsten Wahrscheinlichkeit dargestellt werden. Für jedes Schlüsselbild in der Gitteransicht kann die entsprechende Videosequenz abgespielt werden.

Da die Personen- und Konzepterkennung zwar ähnliche, aber nicht identische Schritte für die automatische Materialindexierung nach sich ziehen, kann ein Benutzer mittels der Schaltflächen für Konzepte und Personen, die vereinfacht Piktogramme gekennzeichnet sind, zwischen der Konzept- und Personenanwendung wechseln. So ist zum Beispiel ein bedeutender Unterschied, dass bei Personen feingranularer Gesichter in Bildern annotiert werden müssen und daher bei der Aggregation von Beispielbildern im Hintergrund der Anwendung eine Gesichtsdetektion durchgeführt werden muss, bevor die Nutzer*innen Bilder annotieren können. Weitere Unterschiede, z. B. im Training, sind in Abschnitt 3.1 und 3.3 beschrieben.

5 Fazit

Nach dem Projekt ist vor der nächsten spannenden Phase. Nach der Integration der Metadaten in die Fernsehdatenbank des Deutschen Rundfunkarchivs wird eine weitere Erprobungsphase folgen: Wie helfen die automatisiert erzeugten Metadaten in der alltäglichen Recherche durch Dokumentare und Wissenschaftler? Sind weitere Schritte der Datenaggregation oder -visualisierung nötig, um die Ergebnisse optimal nutzen zu können? Wie werden im Archivbereich zukünftig manuell und automatisch erzeugte Metadaten miteinander koexistieren? Wir freuen uns auf die nächsten spannenden Erfahrungen mit all diesen Themen – umso mehr, wenn das im Projekt entwickelte VIVA-Annotationstool auch andere Archive dabei unterstützt, einen weiteren Schritt auf diesem Weg zu gehen.

Danksagung

Das Projekt VIVA wird unter dem Titel „Entwicklung eines Softwaresystems zur Szenen- und Personenerkennung für die automatische Erschließung von wissenschaftlich genutzten Videoarchiven“ von der Deutschen Forschungsgemeinschaft (DFG) gefördert (Projektnummer: 388420599).

Literaturverzeichnis

Cao, Qiong; Shen, Li; Xie, Weidi; Parkhi, Omkar; Zisserman, Andrew (2018): VGGFace2: A Dataset for Recognising Faces across Pose and Age. In: *Proceedings of the International Conference on Automatic Face & Gesture Recognition (FG)*, Xian, China, 67–74.

Deng, Jia; Dong, Wei; Socher, Richard; Li, Li-Jia; Li, Kai; Li, Fei-Fei (2009): ImageNet: A large-scale hierarchical image database. In: *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 248–55.

Deng, Jiankang; Guo, Jia; Zhou, Yuxiang; Yu, Jinke; Kotsia, Irene; Zafeiriou, Stefanos (2020): RetinaFace: Single-stage Dense Face Localisation in the Wild. In: *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 5202–11.

Kazemi, Vahid; Sullivan, Josephine (2014): One Millisecond Face Alignment with an Ensemble of Regression Trees. In: *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 1867–74.

Korfhage, Nikolaus; Mühling, Markus; Freisleben, Bernd (2020): Intentional Image Similarity Search. In: *IAPR Workshop on Artificial Neural Networks in Pattern Recognition*, 23–35.

Mühling, M.; Ewerth, R.; Stadelmann, T.; Zöfel, C.; Shi, B.; Freisleben, B. (2007): University of Marburg at TRECVID 2007: Shot Boundary Detection and High Level Feature Extraction. In: *Online Proceedings of TREC Video Retrieval Evaluation Workshop 2007*. Verfügbar unter https://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.7.org.html.

Mühling, Markus; Korfhage, Nikolaus; Müller, Eric; Otto, Christian; Springstein, Matthias; Langelage, Thomas; Veith, Uli; Ewerth, Ralph; Freisleben, Bernd (2017): Deep learning for content-based video retrieval in film and television production. *Multimedia Tools and Applications*, 76 (21), 22169–94.

Mühling, Markus; Meister, Manja; Korfhage, Nikolaus; Wehling, Jörg; Hörth, Angelika; Ewerth, Ralph; Freisleben, Bernd (2019): Content-based video retrieval in historical collections of the German broadcasting archive. *International Journal on Digital Libraries*, 20 (2), 167–83.

Müller-Budack, Eric; Pustu-Iren, Kader; Diering, Sebastian; Ewerth, Ralph (2018b): Finding Person Relations in Image Data of News Collections in the Internet Archive. In: *Proceedings of International Conference on Theory and Practice of Digital Libraries (TPDL)*, Porto: Portugal, 229–40.

Müller-Budack, Eric; Pustu-Iren, Kader; Ewerth, Ralph (2018a): Geolocation Estimation of Photos using a Hierarchical Model and Scene Classification. In: *Proceedings of European Conference on Computer Vision (ECCV)*, 575–92.

Pustu-Iren, Kader; Mühling, Markus; Korfhage, Nikolaus; Bars, Joanna; Bernhöft, Sabrina; Hörth, Angelika; Freisleben, Bernd; Ewerth, Ralph (2019): Investigating Correlations of Inter-coder Agreement and Machine Annotation Performance for Historical Video Data. In: *Proceedings of the International Conference on Theory and Practice of Digital Libraries (TPDL)*, Oslo: Norwegen, 107–14.

Schroff, Florian; Kalenichenko, Dmitry; Philbin, James (2015): FaceNet: A unified embedding for face recognition and clustering. In: *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, 815–23.

Tan, Mingxing; Le, Quoc V. (2019): Efficientnet: Rethinking model scaling for convolutional neural networks. In: *Proceedings of the International Conference on Machine Learning (ICML)*, 6105–14.

Zhou, Bolei; Lapedriza, Agata; Khosla, Aditya; Oliva, Aude; Torralba, Antonio (2017): Places: A 10 million image database for scene recognition. In: *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 40 (6), 1452–64.

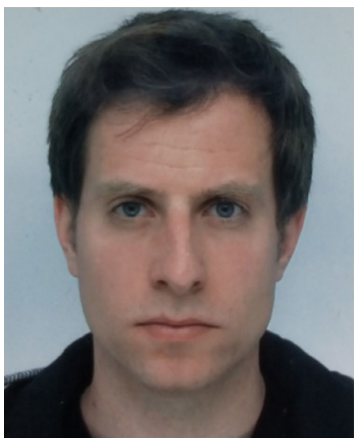
Zoph, Barret; Le, Quoc V. (2017): Neural architecture search with reinforcement learning. In: *arXiv preprint arXiv:1611.01578*.

Kader Pustu-Iren
Technische Informationsbibliothek (TIB)
Welfengarten 1B
D-30167 Hannover

kader.pustu@tib.eu

Joanna Bars
Deutsches Rundfunkarchiv (DRA)
Marlene-Dietrich-Allee 20
D-14482 Potsdam
joanna.bars@dra.de

Dr. Markus Mühling
Philipps-Universität Marburg
Hans-Meerwein-Str. 6
D-35032 Marburg
muehling@informatik.uni-marburg.de



Nikolaus Korfhage
Philipps-Universität Marburg
Hans-Meerwein-Str. 6
D-35032 Marburg
korfhage@informatik.uni-marburg.de

Angelika Hörth
Deutsches Rundfunkarchiv (DRA)
Marlene-Dietrich-Allee 20
D-14482 Potsdam
dra-babelsberg@dra.de

Prof. Bernd Freisleben
Philipps-Universität Marburg
Hans-Meerwein-Str. 6
D-35032 Marburg
freisleb@informatik.uni-marburg.de

Prof. Ralph Ewerth
Technische Informationsbibliothek (TIB)
Leibniz-Informationszentrum für Technik und Naturwissenschaften
Welfengarten 1B
D-30167 Hannover
ralph.ewerth@tib.eu