Praxis

Film ab!

Cognitive Services für die automatisierte Filmanalyse

von Valentin Kuhn

Viele Medienhäuser haben über Jahrzehnte Videoinhalte angesammelt, die zum Teil ohne beschreibende Metadaten archiviert wurden. Die Aufbereitung solcher Archive zur Bereitstellung in VoD-Angeboten ist ein langwieriger und oft manueller Prozess. Die Anbieter stehen vor der Herausforderung, ihre Inhalte durch gute Verschlagwortung für Suchmaschinen und Empfehlungssysteme nutzbar zu machen. In diesem Beitrag untersuchen wir Cognitive Services zur automatisierten Metadaten-Extraktion aus Videomaterial.

Cognitive Systems zur Analyse von Video- und Tonmaterial haben den Zweck, Filme und Videos zu verstehen. Idealerweise könnte man mit diesen Cognitive Systems die Drehbücher des Filmmaterials weitgehend rekonstruieren und damit riesige Video-Archive durchsuchbar machen.

Zum Beispiel haben viele Fernsehsender jahrzehntelang Beiträge gesammelt und archiviert. Zu diesen Beiträgen und Filmen sind nur wenige Informationen vorhanden. Nachträglich Metadaten zu erstellen, wie eine Liste der Schauspieler, ist ein langwieriger und oft auch manueller Prozess. Bei großen Archiven würde diese Arbeit Jahre in Anspruch nehmen. Auch die Art der erfassten Daten ähnelt in ihrer Form oft eher Schlagworten, sodass selten die Abfolge oder Dialoge festgehalten werden.

Video on Demand

Diese Metadaten sind in Archiven und Mediatheken jedoch wichtig, um Inhalte überhaupt auffindbar zu machen: Nutzer wollen in den Beiträgen suchen oder auf Basis der bislang gesehenen Sendungen Empfehlungen ausgesprochen bekommen. Abfragen sollten dabei möglichst natürlich-sprachlich erfolgen: „Wer hat ‚Dumm ist der, der Dummes tut' in welchem Film gesagt?"

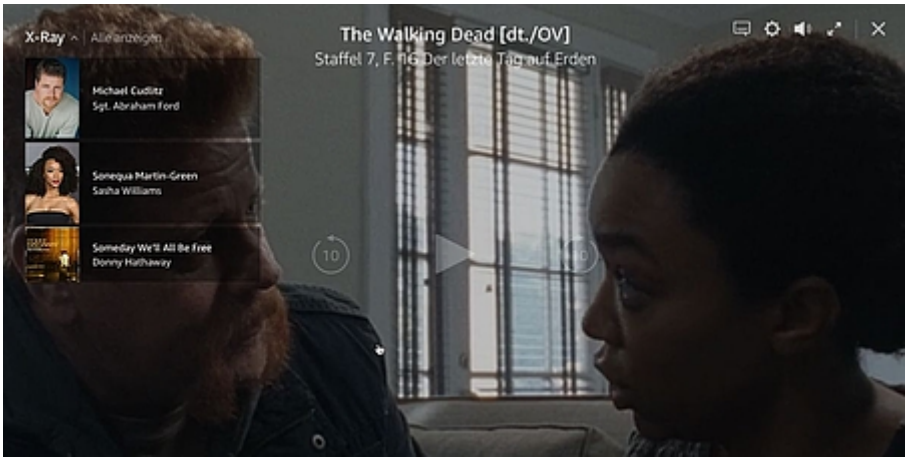

Abb. 1: X-Ray-Overlay von Amazon Video mit erkannten Schauspielern und Hintergrundmusik

Schon heute nutzt Amazon Video die kognitive Filmanalyse, um das X-Ray-Overlay zu ermöglichen. Dieses Overlay listet alle Schauspieler in einer Szene auf und zeigt auch die gespielte Hintergrundmusik, die direkt gekauft werden kann (**s. Abb. 1**). Beim Klicken auf einen der Schauspieler wird eine Liste all seiner Szenen angezeigt, zu denen so auch direkt navigiert werden kann.

Auch YouTube-„Content ID" nutzt Filmanalyse-Tools, um Urheberrechtsverletzungen, Nacktheit und Gewalt automatisiert zu erkennen und noch vor der Veröffentlichung herauszufiltern. Darüber hinaus bietet YouTube automatisch generierte Untertitel zu Videos an, die auch per Filmanalyse erstellt werden.

Fachliche Pipeline zur Filmanalyse

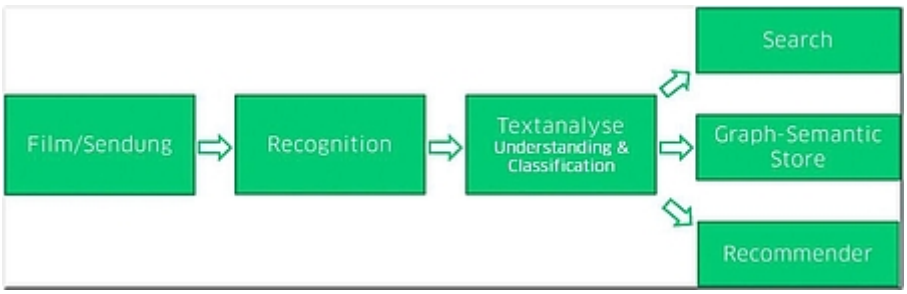

Abb. 2: Pipeline zur Filmanalyse mit Ein- und Ausgaben

Um Nutzern das Durchsuchen von Videoarchiven und das Erstellen von inhaltsbezogenen Empfehlungen zu ermöglichen, haben wir ein Softwaresystem konzipiert, welches Bildmaterial über eine Recognition Engine erfasst (s. Abb. 2). Erprobt wurden dabei die gängigen Cognitive Services auf dem Markt, unter anderem von Microsoft, Google und Amazon, um einen Eindruck von der Leistungsfähigkeit der Cognitive Services zu erhalten. Die auf diesem Weg erstellten Textphrasen werden danach über eine Textanalyse verarbeitet, um dann in Suchmaschinen, Empfehlungssystemen oder (graphenbasierten) Metadaten-Systemen

publiziert zu werden.

Dieser Artikel konzentriert sich auf die „Recognition“-Phase, da diese den ausschlaggebenden Teil der Filmanalyse ausmacht. Dementsprechend betrachten wir im Folgenden hauptsächlich die Analyse von Video-, Bild- und Tonmaterial und lassen die weitere Verarbeitung durch Textanalyse und Publizierungen offen.

Informationen einer Szene

Ein Film besteht aus vielen einzelnen Szenen, die eine Vielzahl an Informationen tragen. Diese Informationen möchten wir zugänglich machen. Unter anderem interessieren wir uns für die beteiligten Personen und deren Eigenschaften wie Emotionen und Alter. Dies bildet die Basis für eine Zuordnung der Schauspieler, wie Amazon Video sie zeigt. Die Audiospur trägt auch viele Informationen, die erfasst werden können. Außerdem sind die vorhandenen Objekte und Gegenstände sowie die Aktivitäten und Interaktionen der Personen von Belang. Nur so können Szenen umfassend beschrieben und Aussagen den richtigen Sprechern zugeordnet werden.

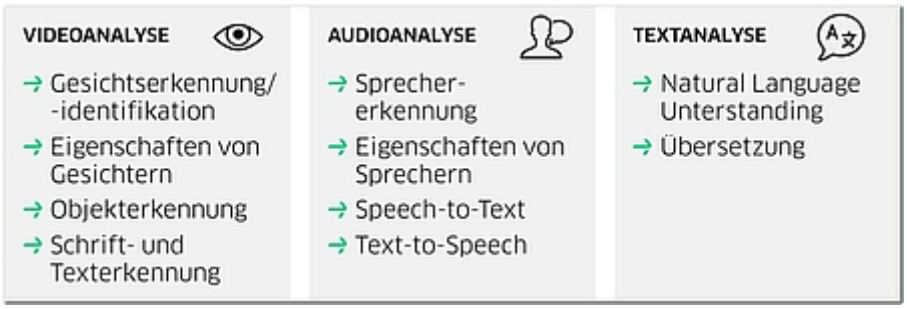

Abb. 3: Cognitive Services in der Filmanalyse

Da es bisher sehr wenige Anbieter mit integrierten Filmanalyse-Tools gibt, bietet sich eine Aufteilung in „Video-“ und „Audioanalyse“ gemäß **Abbildung 3** an. Für die Videoanalyse können nun klassische Bildanalyse-Werkzeuge auf Einzelbilder angewandt werden, und die Audiospur kann mit bestehenden Speech-to-Text-Verfahren transkribiert werden. Anschließend wird der transkribierte Text mit Textanalyse-Werkzeugen auf seinen Inhalt untersucht.

Videoanalyse

Hierzu gehören Gesichtserkennung und -identifikation sowie das Erkennen von Eigenschaften erkannter Gesichter, beispielsweise Emotionen und Alter. Weitere Cognitive Services der Videoanalyse sind Objekterkennung und Schrift- beziehungsweise Texterkennung. Alle diese Dienste extrahieren Informationen aus optischen Daten wie Bildern und Videos, betrachten aber nicht die Tonspur.

Audioanalyse

Hierzu kommen andere Cognitive Services zum Einsatz, die auf menschliche Sprache spezialisiert sind. Die Sprechererkennung ordnet einzelne Aussagen unterschiedlichen Sprechern zu und kann dabei möglicherweise auch Eigenschaften des Sprechers wie Geschlecht und Alter erkennen. Weiterhin wird das Gesprochene mittels Speech-to-Text in maschinenlesbaren und indizierbaren Text umgewandelt. Das Gegenstück, Text-to-Speech, gehört zwar auch dieser Kategorie an, ist für die Filmanalyse aber weniger bedeutend.

Textanalyse

Zur Extraktion von Informationen aus den per Speech-to-Text gewonnenen Texten wird Natural Language Understanding eingesetzt, das Kontext und Zusammenhänge erkennen kann. Da diese Services in der Regel nur eine Sprache gleichzeitig unterstützen, müssen gemischtsprachliche Texte zuerst übersetzt werden.

Analyse eines Films

Abb. 4: Extraktion der Tonspur und Einzelbilder eines Videos mit FFmpeg

In der Analyse halten wir uns an die zuvor beschriebenen Kategorien Audio-, Video- und Textanalyse. Dazu extrahieren wir die Tonspur und einige markante Einzelbilder aus dem Video. Hierfür kann FFmpeg [FFmpeg] verwendet werden (**s. Abb. 4**). So können wir für jede Kategorie verschiedene Cognitive Services nutzen und die Stärken verschiedener Dienste ausnutzen.

Im Folgenden stellen wir einige Dienste vor, die wir exemplarisch zur Analyse einer Sendung des ZDF-Nachrichtenmagazins heute-journal verwendet haben. Dabei konzentrieren wir uns auf die großen Cloud-Services von Google, Amazon und Microsoft und greifen für spezialisierte Anwendungen auf kleinere Anbieter zurück.

Audioanalyse: Zuerst analysieren wir, wer etwas sagt, wann er es sagt und was er sagt. Bei der Analyse des „Wer“ unterscheidet man dabei zwischen drei Qualitätsstufen:

Einige Dienste führen keine Sprechererkennung durch

Speaker Diarization bezeichnet die Zuordnung von Textphrasen zu anonymen Sprechern, beispielsweise „m1“ für den ersten erkannten männlichen Sprecher in einer Audiodatei

Speaker Recognition identifiziert darüber hinaus die einzelnen Sprecher anhand von Sprachsamples

Ein bekannter Service auf dem Gebiet der Spracherkennung ist das Google Cloud Speech-to-Text API. Diese Programmierschnittstelle erkennt relativ zuverlässig den gesprochenen Text, was jedoch auch an der sehr klaren Aussprache der Nachrichtensprecherin Marietta Slomka liegt. Darüber hinaus bietet das API kaum weitere Informationen an und beherrscht beispielsweise keine Sprechererkennung.

Eine Alternative zur Google Cloud bietet das kleinere Speechmatics, das mit Speaker Diarization aufwartet. Diese gibt zur ersten Phrase Marietta Slomkas „f1“ aus, was die erste weibliche Sprecherin in der Tonspur bezeichnet, identifiziert diese Sprecherin jedoch nicht.

Anbieter	Sprachen	Zeit-Indizierung	SpeakerDiarization/ Recognition	Training
Microsoft Bing Speech	15	✗	✗/✗ (Recognition aus 10 Personen)	Akustikmodell, Wörterbuch
Google Speech-to-Text	über 120	✓	✗/✗	Wörterbuch
SpeechMatics	75	✓	✓/✗	✗
Amazon Transcribe	2 (kein Deutsch)	✓	✓/✗	Wörterbuch
IBM Watson Speech to Text	10	✓	✓/✗	Akustikmodell, Wörterbuch

Tabelle 1: Speech-to-Text-Werkzeuge verschiedener Anbieter

Leider gibt es nur wenige Dienste zur Speaker Recognition auf dem Markt. Die Microsoft Speaker Recognition aus der Microsoft Bing Speech Suite beispielsweise ist nur auf die Erkennung von zehn vordefinierten Sprechern ausgelegt. Govivace aus dem Forensikbereich soll auch mehrere Millionen Stimmen unterstützen, lässt sich aber nicht ausprobieren, ohne dass man zuvor Verschwiegenheitserklärungen unterzeichnet hat. **Tabelle 1** veranschaulicht die Funktionen verschiedener Speech-to-Text-Dienste.

Videoanalyse

Die optische Analyse dient der Erkennung von Personen und von im Bild dargestelltem Text. Face Recognition erkennt Personen und deren Position im Bild ebenso wie das Alter und Emotionen in Gesichtern. Mit diesen Informationen lassen sich Personen identifizieren (Face Identity). Darüber hinaus können Objekte und Zusammenhänge von Personen und Objekten im Bildmaterial erkannt werden, um den Kontext besser zu verstehen. Zusätzlich kann eine Szenenerkennung (Shot Detection) feststellen, wann eine Szene beginnt und endet. Dies ermöglicht eine Zuordnung von Handlungen und Interaktionen der Personen zu einer Szene.

```
{
  "entity": {
    "description": "television presenter", ...
  },
  "categoryEntities": [
    {
      "description": "person", ...
    }
  ],
  "segments": [
    {
      "segment": {
        "startTimeOffset": "274.420s",
        "endTimeOffset": "284.120s"
      },
      "confidence": 0.6814277
    }
  ]
}
```

Listing 1: JSON-Auszug aus der Google-Cloud-Video-Intelligence-Analyse

Ein Anbieter von Videoanalyse-Werkzeugen ist Google mit dem Video Intelligence API. Die darin enthaltene Shot Detection unterscheidet einzelne Szenen. Die Label Detection erzeugt Labels und gibt den Zeitindex deren Auftretens zurück. **Listing 1** zeigt beispielhaft das erkannte Label „television presenter“, das zum ersten Mal nach 274 Sekunden auftritt und 10 Sekunden sichtbar ist.

Genauere Beschreibungen der Bilder erzeugt eine Bildanalyse wie das Google Cloud Vision API, das auf die Frames der Shot Detection angewendet werden kann. So werden aussagekräftige Labels zur Beschreibung der jeweiligen Frames generiert und dargestellte Texte transkribiert, was das Video Intelligence API nicht leistet. Pro Label wird eine Konfidenz zurückgegeben, mit welcher das System die Richtigkeit der Zuordnung des Labels zu diesem Bild schätzt.

Basierend darauf, können nur die wahrscheinlichsten Labels für das Bild beziehungsweise für die zugehörige Szene übernommen werden. Durch die Kombination von Video Intelligence API und Cloud Vision API können nun Szenen zeitlich getrennt und unabhängig voneinander beschrieben werden, was das Auffinden einer ganz bestimmten Szene ermöglicht.

```
"faces": [
  {
    ...
    "confidence": 0.9968,
    "name": "Marietta Slomka",
    "description": "Marietta Slomka ist eine deutsche Journalistin...",
    "title": "Journalistin",
    "appearances": [
      {
        "startTime": "0:00:09.28",
        "endTime": "0:00:52.66",
        "startSeconds": 9.3,
        "endSeconds": 52.7
      },...
    ],
    "seenDuration": 149.1,
    "seenDurationRatio": 0.2499
  },...
]
```

Listing 2: JSON-Auszug der Microsoft-Video-Indexer-Analyse

Ein weiterer Anbieter für Videoanalyse ist Microsofts Video Indexer. Dieser erkennt eigenständig Gesichter und kann deren

Identität über Bing herausfinden, wie in **Listing 2** zu sehen ist.

Genauere Ergebnisse wie Emotionen, Alter und Geschlecht lassen sich jedoch besser mit dem Microsoft Face API zur Gesichtserkennung erzeugen. Auch die Labels, die Microsoft Video Indexer erzeugt, können mit Microsoft Computer Vision zur Objekterkennung genauer erfasst werden. Hierfür liefert der Video Indexer die Key-Frames des Videos zurück, auf welchen eine Bildanalyse ausgeführt werden kann. Im Gegensatz zum Google Video Intelligence API bezieht Microsoft Video Indexer auch die Tonspur mit ein und transkribiert diese. Zusätzlich lassen sich vor- und selbstdefinierte Markennamen gesondert erkennen. In unserem heute-journal-Beispielvideo wurde deshalb der Name des Co-Moderators Heinz Wolf fälschlicherweise als vordefinierte Marke „Heinz“ (Ketchup) erkannt.

Neben Google und Microsoft bietet auch Amazon mit seiner AWS Rekognition Video eine Videoanalyse-Suite an, die vor allem eine sehr zuverlässige „Celebrity Recognition“, basierend auf den Daten der Internet Movie Database IMDb [IMDb], durchführt. Diese Erkennung steckt hinter dem eingangs gezeigten „X-Ray“, das Informationen zu Schauspielern während ihrer gespielten Szenen anzeigt.

Anbieter	Szene	OCR	Face Recognition	Face Identity	Labels	Training
Microsoft Video Indexer	✓	✓	✓	✓	✓	Linguistik, Marken, Personen
Microsoft Computer Vision	✗	✓	✓	✓	✓	Labels über Bilder
Google Cloud Video Int.	✓	✓ (nur Englisch)	✗	✗	✓	✗
Google Cloud Vision	✗	✓	✓	✗	✓	✗
Amazon Rekognition Video	✓	✓	✓	✓ (IMDb)	✓	Personen über Bilder
Amazon Rekognition Image	✗	✓	✓	✓ (IMDb)	✓	Personen über Bilder

Tabelle 2: Video- und Bildanalyse-Werkzeuge verschiedener Anbieter

Besonders bei der Celebrity Recognition stellt sich die Frage, wie gut sich diese Systeme auf eigene Bedürfnisse anpassen lassen. Soll beispielsweise ein Video mit weniger gut bekannten Personen indiziert werden, die keinen Eintrag in der IMDb haben, so kann Amazon Rekognition mit eigenen Bildern nachtrainiert werden. Dies ist auch bei den entsprechenden Microsoft-Services möglich. Die Gemeinsamkeiten und Unterschiede verschiedener Video- und Bildanalyse-Services stellt **Tabelle 2** dar.

Einschätzung der Videoanalyse

Auffällig ist, dass keiner der Services zuverlässig Aktivitäten in den Videos erkennt. Es werden hauptsächlich Personen und Objekte erfasst und Aktivitäten beinahe komplett ignoriert. So erkennt Google Video Intelligence beispielsweise 114 Personen und Objekte wie „businessperson“ und „building“, aber nur 6 Aktivitäten wie „sitting“ und „driving“.

Außerdem sind Videoanalyseverfahren leider nicht eigenständig nutzbar, sondern spielen ihre Mächtigkeit nur im Zusammenspiel mit einer Analyse einzelner Frames aus. Hier könnte auch beliebig gemischt werden, wenn sich ein anderer Dienst besser eignet.

Textanalyse

Die Spracherkennung ist bereits auf einem Niveau, welches menschlichen Fähigkeiten nahekommt, wenn nicht sogar teilweise überlegen ist. Die Intention des gesprochenen Wortes jedoch richtig zu verstehen, ist eine Herausforderung der Textanalyse (**s. Abb. 2**). Wer beispielsweise bei einer Hotline an einen Sprachassistenten gerät, verzweifelt schon mal an einem „Das habe ich leider nicht verstanden“. Limitiert man das Sprachvokabular jedoch auf spezifische Bereiche mit einem reduzierten Wortschatz oder auf weniger komplexe Kontexte wie das Smart Home, so lassen sich bereits erstaunliche Ergebnisse erzielen. Systeme hierfür sind beispielsweise das Google Cloud Natural Language API, Microsoft LUIS und für On-premise-Nutzung Snips NLU.

Fazit

Bei der Bild- und Videoanalyse steht man erst am Anfang. Viele für den Menschen einfach zu erkennende Situationen überfordern noch die Algorithmen der vorhandenen Services. So wird beispielsweise eine Szene des heute-journals, die die Folgen eines Erdbebens in Japan zeigt, von Google Video Intelligence nicht in den Video-Labels erwähnt. Auch keine verwandten Labels wie „damage“ oder „destruction“ wurden für diesen Teil des Videos erzeugt. Während Personen im Video schon recht gut erkannt und Schlagwörter zu Szenen erzeugt werden, ist es für die Systeme noch sehr schwer, Tätigkeiten zu erkennen.

Trotzdem können mit den erzeugten Metadaten schon viele Informationen bereitgestellt werden, die eine Indizierung von Filmen

in Video-on-Demand-Angeboten zulassen. Auch die eingangs erwähnte Suche nach einem Zitat wie „Dumm ist der, der Dummes tut“ wird durch Text-to-Speech ermöglicht. Nur die Zuordnung des Zitats gestaltet sich durch mangelhafte Angebote zur Sprechererkennung noch als schwierig.

Basis

Der Artikel basiert auf Accso-internen Erkenntnissen zur Filmanalyse, die schon im Rahmen der JAX 2018 von Thomas Jäger präsentiert wurden. Eine Aufzeichnung des Vortrags ist auf YouTube verfügbar [YT].

Links

[AWS] Machine Learning in AWS

[FFmpeg]FFmpeg multimedia framework

[Goo]Cloud Machine Learning-Dienste, Google

[IMDb]IMDb

[Mic]Cognitive Services, Microsoft Azure

[YT] Th. Jäger, Cognitive Services – Kognitive Services in der Filmanalyse, Vortrag auf der JAX 2018, 30.5.2018

Valentin Kuhn

arbeitet als Junior Software Engineer bei Accso und studiert im Master Informatik an der Technischen Universität Darmstadt. Er ist spezialisiert auf mobile Anwendungen im Android-Umfeld und beschäftigte sich zuletzt mit intuitiven Interaktionsmethoden wie Sprachsteuerung.

E-Mail: valentin.kuhn(at)accso.de

Bildnachweise:

Accso

Kontakt & Anfrage