Hochschule der Medien Stuttgart
Aktuelle Themen, AM3
WS2022/2023
Elias Weißenrieder (ew057)
Matrikelnummer: 44980

Künstliche Intelligenz im Foleyschnitt im Bereich der Postproduktion von Schritten

Einleitung

Damit ein Film oder eine Serie die gewünschte Wirkung erzielt ist es heutzutage elementar wichtig, neben der Bildgestaltung und dem Bildschnitt, auch die Tongestaltung im Entstehungsprozess des Films zu berücksichtigen. Deswegen ist die Audiopostproduktion mittlerweile ein wichtiger Bestandteil einer Film- oder Serienproduktion. Sie beschäftigt sich dabei nicht mehr nur mit der Sprachverständlichkeit der Dialoge, sondern besteht häufig aus Nachvertonungen von Geräuschen. Dies ist notwendig, da die Qualität der am Set aufgenommen Tonaufnahmen meist nicht optimal ist.

Die Auslastung von Firmen die solche Dienstleistungen anbieten ist hoch. Grund hierfür ist unteranderem die wachsende Verbreitung von Streaming-Anbietern, wie Netflix, Amazon Prime, Disney+, etc. Mit der Entstehung dieses Marktes kippte das Konsumverhalten der Nutzer. Beispielsweise war es Anfang der 2010er Jahre noch Normalität, dass eine neue Folge einer Serie die Zuschauer frühestens eine Woche nach der letzten Folge erreichte. Mittlerweile sind neue Serien innerhalb von wenigen Wochen komplett verfügbar und jeder kann von überall auf Filme und Serien zugreifen. Dies erhöht natürlich den zeitlichen Druck auf alle am Produktionsprozess beteiligten Gewerke.[1] Außerdem führt dies zu einem Dilemma für die Postproduktionsfirmen, denn diese müssen ein höheres Tempo bei gleichbleibender Qualität anstreben.

In der Branche der Audiopostproduktion existiert der Beruf des Foley-Artist, welcher auch als Geräuschemacher bezeichnet wird. Die Tätigkeit dieses Berufs besteht zu einem großen Teil aus der Vertonung, bzw. Nachvertonung von Schritten. Wenn dieser Produktionsschritt, zumindest teilweise automatisiert werden könnte, wäre dies eine enorme Verringerung des Arbeitsaufwandes.[2] Mir der dadurch freiwerdenden Arbeitskraft könnten dann wiederum aufwändigere Projekte ohne höheren Kostenaufwand umgesetzt werden. Somit würde zusätzlich die generelle Qualität der Postproduktion steigen.[3]

[1] Vgl. Suchant 2022, S.46
[2] Vgl. Suchant 2022, S.19
[3] Vgl. Suchant 2022, S.46

Welche Aufgaben sollte die KI übernehmen können?

Bei der beschriebenen Tätigkeit, der Vertonung von Schrittgeräuschen, könnte ein Programm, welches mit künstlicher Intelligenz ausgestattet ist, hilfreich sein. Dazu müsste es folgende Arbeitsschritte durchlaufen:

Zunächst müsste die Anwendung das Bild- und Originaltonmaterial hinsichtlich der Notwendigkeit von zusätzlichen Schrittgeräuschen analysieren. Denn in manchen Fällen ist es nicht notwendig in der Postproduktion Schrittgeräusche zu ergänzen, bzw. zu ersetzten. In diesen Fällen liefert das Originaltonmaterial vom Dreh eine befriedigende Qualität.

Nachdem dann das Programm Stellen festgestellt hat, die einer Nachvertonung bedürfen, sollte es automatisch die genauen zeitlichen Einsätze des zusätzlichen Tonmaterials erkennen und festlegen.

Danach sollten, passend zu der jeweiligen Szene, vorproduzierte Tonsamples aus einer vorhandenen Datenbank ausgewählt werden. Bei dieser Auswahl spielen folgende Faktoren der Schrittgeräusche eine wichtige Rolle: Untergrund, Raum in Verbindung mit benutztem Schuhwerk und Kraft, sowie die körperliche Verfassung der gehenden Person.[4] Die Analyse dieser Faktoren könnte beispielsweise über das vorhandene Bildmaterial, oder den Originalton der Szene erfolgen. Die Umsetzung dieses Arbeitsschritts könnte mittels eines MIDI-Datenstroms erfolgen.[5] MIDI steht in diesem Fall für Musical Instrument Digital Interface und ist ein Übertragungsprotokoll aus der Musikbranche. Es übersendet Steuerdaten mit denen zum Beispiel Samplemaschinen angesteuert werden können.[6] Die Analyse des Bild- und Originaltonmaterials liefert dann den MIDI-Datenstrom. In diesem sind die Werte für die beschriebenen Faktoren, die für die Auswahl der Tonsamples aus der Datenbank wichtig sind, enthalten. Dies hätte weiterhin den Vorteil, dass eine große Kompatibilität für Weiterentwicklungen in Form von externen Plugins möglich wäre. Denn diese könnten sehr einfach zusätzlich durch den MIDI-Datenstrom angesteuert werden.

Im Anschluss an diesen Arbeitsschritt der Analyse sollte das KI gestützte Programm die ausgewählten Tonsamples an die festgelegten Stellen einfügen. Bei diesem Schritt ist allerdings zu beachten, dass das Programm nicht stets die gleichen Tonsamples nacheinander verwenden kann. Dies würde dazu führen, dass der Ton der Szene statisch und unrealistisch wirkt. Das Programm muss also zusätzlich über eine gewisse Zufalls-Algorithmik verfügen, welche aus einem Pool von „Tonschnipseln“ auswählt und gleichzeitig verhindert, dass identische Tonsamples nacheinander abgespielt werden.[7]

Als letzten Schritt sollte das Programm den Audiopegel der eingefügten Schritttonsamples an den Dialogpegel anpassen. Dies ist notwendig, da bei falschem Verhältnis der beiden Pegel, die Sprachverständlichkeit leidet und der Ton der Szene unrealistisch wirkt.

[4] Vgl. Suchant 2022, S.30
[5] Vgl. Suchant 2022, S.29
[6] Vgl. Dickreiter 1997, S.93
[7] Vgl. Suchant 2022, S.30

Für welche Anwendungsbereiche eignet sich KI in der Postproduktion?

Im Folgenden werden die möglichen Anwendungsbereiche des im vorherigen Kapitel beschriebenen KI-Foleyschnittprogramm diskutiert. Die unterschiedlichen Formate, im fiktiven Bereich, für die in Deutschland hauptsächlich Audiopostproduktion betrieben wird, sind der Kinofilm, der Fernsehfilm und die Vorabendserie.

Dabei besitzt der deutsche Kinofilm in der Regel meist ein Produktionsbudget von einer bis zehn Millionen Euro. Der Fernsehfilm stellt das Hauptgeschäft von deutschen Produktionsfirmen dar, sein Budget beträgt im Durchschnitt ca. 2,5 Millionen Euro.[8] Die Vorabendserie besitzt, beispielsweise bei Produktionen für den Sender ZDF, nur noch 405 Tausend Euro pro 45 Minuten Folge.[9] An dieser Differenz ist zu erahnen, dass die Verhältnisse des Budgets der drei Formate sich auch auf die Postproduktion ableiten lassen. Das führt bei den Produktionsfirmen dazu, dass nicht für alle Produktionen der gleiche Arbeitsaufwand investiert werden kann.

Um die Anwendungsmöglichkeiten des KI gestützten Foleyschnittprogramms in den verschiedenen Formaten zu überprüfen, lassen sich zwei durchschnittliche Kinofilme, Fernsehfilme und Vorabendserien anführen, die hinsichtlich der Kriterien Kameraperspektive, Kameraverhalten und Schnitttechnik analysiert werden.

Der erste Film soll dabei “Who Am I – Kein System ist sicher“ aus dem Jahr 2014 von Baran bo Odar sein. Direkt zu Beginn des Films wird deutlich, dass dieser von viel Dynamik lebt. So wechseln die Perspektiven permanent zwischen „Halbnah“, „Nah“ und „Groß“. Nur wenig der Handlung wird in der „Halbtotalen“ oder der „Totalen“ erzählt. Auch die Kameraführung ist sehr dynamisch und verfügt selten über konventionelle Kamerafahrten. Außerdem ist das Schnittverhalten des Films ebenfalls sehr schnell und nahezu unübersichtlich. Diese Bedingungen sorgen dafür, dass Bewegungen der Personen schwer zu interpretieren sind.[10]

Der zweite Film, „Die Hochzeit“ aus dem Jahr 2020 von Till Schweiger hingegen zeigt ein anderes Verhalten. Es finden viele Einstellungen in der „Totalen“ oder der „Halbtotalen“ statt. Im Unterschied zu „Who Am I – Kein System ist sicher“ arbeitet der Film wenig in nahen Kameraeinstellungen. Des Weiteren wirken die Kamerafahrten weniger dynamisch, was dazu führt, dass der Film eher ruhiger und klarer wirkt. Alle Bewegungen der Figuren in den Szenen sind gut zu interpretieren und für den menschlichen Betrachter gut nachzuvollziehen. Trotzdem sind auch in diesem Beispiel viele Schnitte vorhanden.[11]

Durch die komplexen, oder nicht immer genau interpretierbaren Bewegungen der Personen, sowie die schnellen Schnitte sind Kinofilme eher ungeeignet für die Einführung eines KI gestützten Foleyschnitts.[12]

Im deutschen Fernsehfilm herrschen jedoch andere Voraussetzungen. Dies

[8] ZDF
[9] ZDF
[10] Filme: Netflix
[11] Vgl. Suchant 2022, S.20
[12] Vgl. Suchant 2022, S.22

ist am Beispiel der ersten Folge des Dreiteilers „Ku'damm 59“ des ZDF erkennbar. Die Kameraperspektive verfolgt lange, zusammenhängende Einstellungen in der „Halbtotalen“ und „Halbnahen“. Dabei bleibt die Bildführung stets stabil und wenig dynamisch. Die Bewegungen der einzelnen Figuren sind meist in einer einzelnen Kameraeinstellung abgebildet. Schnelle Schnitte sind nur bei Dialogen, bei denen die sprechenden Personen keine Bewegungen machen vorhanden. Der Film wirkt sehr ruhig und ist durch eine gute Übersichtlichkeit geprägt. Alle Schritte der Figuren sind also klar zu erkennen und können gut verortet werden.[13]

Ein ähnliches Bild ergibt sich bei Betrachtung der Folge „Tyrannenmord“ der ARD Vorabendserie „Tatort“. Auch hier wirkt das Bild klar strukturiert und es sind wenige aufregende Schnitttechniken zu erkennen. Die Kameraführung ist stabil und wenig dynamisch. Die Szenen wirken sehr aufgeräumt, was dazu führt, dass man als Zuschauer eindeutig die Bewegungsmuster jeder Figur erkennen kann. Somit fällt auch die Interpretation der Schritte leicht.[14]

Anders als bei Kinofilmen eignen sich deutsche Fernsehfilmproduktionen deutlich besser für den Einsatz der Foleyschnitt-KI. Dies liegt daran, dass die Bewegungen gut zu interpretieren und wenige schnelle Schnitte vorhanden sind.[15]

Als Beispiel für eine deutsche Vorabendserie wurde Folge 284 der ARD-Serie „In aller Freundschaft – Die Jungen Ärzte“ gewählt. Direkt zu Beginn der Folge fällt auf, dass es viele Szenen gibt, in denen die Figuren gehen. Ähnlich wie in den Fernsehfilmen sind meist die Kameraeinstellungen „Totale“ und „Halbtotale“ gewählt. Nahaufnahmen werden nur selten geschnitten. Der Fokus liegt auf der Handlung, weniger auf der künstlerischen Umsetzung der Kameraführung, weswegen diese auch weniger aufgeregt ist. Der Zuschauer kann der Szene klar folgen und ihm wird wenig Interpretation abverlangt. Der Ablauf der Szenen folgt einer klaren Struktur.[16]

Ein leicht unterschiedliches Bild ergibt sich bei Betrachtung der Folge „Halle forever“ der ZDF-Vorabendserie „Blutige Anfänger“. Die allgemeine Bildsprache ist hier kurzweiliger gehalten. Jedoch sind in den Kameraperspektiven Parallelen zu der ARD-Serie zu erkennen. Bildschnitte zwischen unterschiedlichen Kameraperspektiven bei gleicher Handlung sind in „Blutige Anfänger“ häufiger vorhanden. Jedoch zieht sich die Art der Schnitte durch die gesamte Folge. So ist beispielsweise eine Szene zu nennen, in der eine Person eine Treppe hochgeht. In der ersten Einstellung ist die Person von unten und in der zweiten Einstellung aus der Draufsicht abgebildet. Die Kameraführung bleibt dabei jedoch wie üblich stets unaufgeregt und stabil. Dies führt dazu, dass sich längere zusammenhängende Einstellungen doch gut beurteilen lassen.[17]

Dies lässt das Fazit für Vorabendserien zu, dass bei dieser Art von Filmen die Anwendung der Foleyschnitt-KI

[13] ZDF-Mediathek
[14] ARD-Mediathek
[15] Vgl. Suchant 2022, S.22
[16] ARD-Mediathek
[17] Vgl. Suchant 2022, S.21

möglich wäre. Da wie bereits erwähnt in solchen Vorabendserien viele Szenen zu sehen sind, in denen Figuren gehen, würde der Einsatz der KI dem Foley-Artist viel Arbeit abnehmen.[18]

Welche Methoden zur Erkennung des zeitlichen Einsatzes der Schritte könnten gewählt werden?

Für die prinzipielle Erkennung der Stellen, an denen die nachträglich vertonten Schnitte genau platziert werden sollen, eignen sich zwei Methoden. Zum einen die Objekt-Verfolgung anhand des Bildmaterials und zum anderen die Analyse des Originaltonmaterials.[19]

Die Objekt-Verfolgung, auch Object-Tracking genannt, wird heutzutage sehr ausgiebig genutzt. So wird sie zum Beispiel zur Überwachung der Verkehrslage im Straßenverkehr oder in Filmproduktionen verwendet. Mittlerweile reicht die Rechenleistung moderner Heimcomputer aus, sodass nahezu jeder diese Technik auf seiner Hardware nutzen kann. Bei der Objekt-Verfolgung wird die Verschiebung eines Objekts in Abhängigkeit von der Zeit erfasst. Dabei können einzelne oder auch mehrere Objekte verfolgt werden.[20] Es gibt unterschiedliche Herangehensweisen Objektverfolgung zu realisieren.

So werden bei bereichsbasierter Objekt-Verfolgung verschiedene Variationen des Hintergrundes genutzt, um ein dynamisches Hintergrundgeschehen zu erzeugen. Dieses wird dann vom aktuellen Frame subtrahiert und übrig bleibt das zu verfolgende Objekt.

Bei aktiver konturbasierter Objekt-Verfolgung werden die Umrisse des Objekts als Bewegungskonturen dargestellt. Diese werden dynamisch über die Zeit aktualisiert.[21]

Des Weiteren existiert noch die merkmalsbasierte Objekt-Verfolgung. Diese orientiert sich an Merkmalen höherer Ordnung.

Abschließend lässt sich noch das modellbasierte Tracking anführen, welches aus Vorwissen projizierte Objektmodelle an die Bilddaten anpasst.[22]

Generell basieren diese Verfahren der Objekt-Verfolgung auf einer Variante des „Deep Learnings“. Das Programm lernt aus dem vorherigen und dem nachfolgenden Bild, wie es auf die Bewegung eines Fixpunktes schließen kann.[23]

Diese Tracking Funktionen könnte sich die Foleyschnitt-KI zunutze machen.[24] Wenn die KI die Kameraperspektive kennt, in der die Einstellung gedreht wurde, kann sie die Gehbewegung der Figur mittels Objektverfolgung erkennen. Bei seitlicher Kameraperspektive wäre dies beispielsweise ein wellenförmiger Graph. Die Tiefpunkte dieses Graphen sind die Auftritte der Person und somit die Momente in denen man die Schritte der Figur hören sollte.

Allerdings ist dies bei rotierenden, bzw. bewegenden Kamerapositionen nicht trivial lösbar. Das Programm müsste

[18] Vgl. Suchant 2022, S.22
[19] Vgl. Suchant 2022, S.24
[20] Vgl. Singh 2020, S.77
[21] Vgl. Singh 2020, S.77
[22] Vgl. Al Naijar 2014, S.120
[23] Vgl. Held 2016
[24] Vgl. Suchant 2022, S.25

die Dimensionen des virtuellen Koordinatensystems kennen, um die Tiefpunkte zu berechnen. Beispielsweise gibt es bei einer Einstellung, bei der sich die Person direkt auf die Kamera zu bewegt nur eine Koordinatenachse.

Eine andere Methode, um die Einsätze der Trittgeräusche zu erkennen ist die Analyse des Originaltonmaterials. Schrittgeräusche können meist sehr gut ausgemacht werden. Es gibt zwei Arten das vorhandene Audiosignal auf Schrittsounds zu analysieren:

Mittels Analyse des Amplitudenverlaufs des Settonsignals können Schritte, welche nichts anderes als kurzzeitige plötzliche Geräusche sind und deswegen relativ gut zu identifizieren sind, ausgemacht werden.[25]

Des Weiteren kann durch eine Frequenzanalyse im Originalton Schritte ausgemacht werden. Grund dafür ist der relativ markante klangliche Aufbau von Schritten.[26] Diese Variante hat allerdings den Nachteil, dass nur so lange alle Schritte im Originalton die gleiche Frequenzverteilung aufweisen, eine eindeutige Identifizierung erfolgen kann.[27]

Allerdings ist auch die Methode der Analyse des Originaltons nicht komplett fehlersicher. So können Schrittgeräusche beispielsweise bei lauten Umgebungsgeräuschen, bzw. Naturgeräuschen nicht mehr valide erkannt werden.[28] Außerdem ist bei der Anwendung dieser Methode keine figurenbezogene Zuweisung möglich, wenn mehrere Personen gehen.[29]

Für die zukünftige Fehlersichere Nutzung der Foleyschnitt-KI ist es daher sicherlich empfehlenswert eine Kombination aus beiden Methoden anzuwenden, um die Einsätze der Schritte zuverlässig zu erkennen.

[25] Vgl. Suchant 2022, S.26
[26] Vgl. Flückiger 2002, S.517
[27] Vgl. Suchant 2022, S.26
[28] Vgl. Suchant 2022, S.27
[29] Vgl. Suchant 2022, S.27

Quellen

- Suchant, Michel (2022): Entwurf einer Anwendung zur Vereinfachung des Foleyschnitts. Zugl. Mittweida, Fachhochsch., Bachelorarbeit, 2022.
- Dickreiter, Michael (1997): Handbuch der Tonstudiotechnik. 6., verbesserte Auf-lag. 2 Bände. München: K. G. Saur
- ZDF (2021): Programmprofile und -kosten. https://www.zdf.de/zdfunterneh-men/transparenz-programmkosten-100.html geprüft am 19.05.22
- Netflix: www.Netflix.com
- ZDF-Mediathek
- ARD-Mediathek
- Singh, Rajiv; Nigam, Swati; Singh, Amit Kumur; Elhoseny, Mohamed (2020): In-telligent Wavelet Based Techniques for Advanced Multimedia Applications. Cham: Springer
- Al Najjar, Mayssaa; Shantous, Milad; Bayoumi, Magdy (2014): Viedeo Surveil-lance for Sensor Platforms. Algorithms and Architectures. Cham: Springer
- Held, David; Thrun, Sebastian; Savarese, Silvio (2016): Learning to Track at 100 FPS with Deep Regression Networks. Online verfügbar bei Springer: https://link.springer.com/chapter/10.1007/978-3-319-46448-0_45 geprüft am 19.05.22
- Flückiger, Barbara (2002): Sound Design. Die virtuelle Klangwelt eines Films. 2. Aufl. Marburg: Schüren