Neuronale Netze in der Videoproduktion

Laura Anger
Technische Hochschule Köln
Institut für Medien- und Phototechnik
laura.anger@th-koeln.de

Vera Brockmeyer
Technische Hochschule Köln
Institut für Medien- und Phototechnik
vera-brockmeyer@smail.th-koeln.de

Zusammenfassung

Vera

Schlüsselwörter

Faltungsnetze, Videoproduktion, Stilsynthese, Bewegtbildgenerierung, Deep Writing

Vera

1. Einleitung

Vera

Die Videoproduktion konnte sich im Zuge der Digitalisierung im letzten Jahrzehnt enorm qualitativ verbessern. Jeder der vier Produktionabschnitte (siehe Abbildung 1) konnte verbessert und nachhaltig erleichtert werden. In der Konzeptionsphase konnten die Arbeitsprozesse mit Hilfe des Internets erleichtert werden durch den beschleunigten weltweiten Austausch von Skripten oder den standort-unabhängigen Zugriff auf cloudbasierte Projektmanagement Systeme.

Während der Produktion des Videomaterials unterstützen moderne digitale Kamerasysteme den Kameramann indem sie den Weißabgleich und die Belichtung automatisch berechnen und einstellen. Selbst geringfügige unruhige Bewegungen werden mit Bildstabilisatoren direkt unterdrückt.

Die darauffolgenden Arbeitsprozesse, wie das Schneiden und Editieren des Videomaterials währen der Postproduktionsphase, wurden in den letzten Jahren vereinfacht oder können teilweise durch zuverlässige Algorithmen automatisch durchgeführt werden. Mittlerweile können realistisch virtuelle Bildinhalte von *Computer Generated Imaging* (CGI) Experten mit entsprechender Rechenkapazität in das gedrehte Videomaterial nahtlos rendert werden. Dies ermöglicht es Produktionen fast ausschließlich im Studio zu produzieren und sogar aufwendige Fantasywelten oder aufwendige Stunts mit geringeren Kosten umzusetzen.

Doch gerade qualitativ hochwertige Videoproduktionen erfordern immer noch einen sehr hohen Arbeitsaufwand mit einer große Anzahl an professionellen Mitarbeitern und kostenaufwändigen Materialien. Einen großen Anteil daran hat die Postproduktion in der jede Szene separat editiert und an das Gesamtbild angepasst werden muss. Dieses Gesamtbild muss vorab genau festgelegt werden, denn eine spätere Korrektur erfordert eine vollständige Wiederholung der meisten Arbeitsschritte. Aber auch die Generierung von Bildmaterialien für kurze Schnittszenen oder Webvideos ist sehr zeitaufwändig und kostenintensiv. Ein mehrköpfiges Team mit dem umfangreichen Equipment muss zum Drehort gebracht werden und unter Umständen auch untergebracht werden. Auch äußere Einflüsse, wie zum Beispiel das Wetter, können den Zeitplan verzögern und es entstehen zusätzliche Kosten. Für Studioaufnahmen muss in den meisten Fällen entsprechende Ressourcen angemietet oder auf Dauer gepflegt und in Stand gehalten werden.

In der Zukunft gilt es diesen Arbeits- und Kostenaufwand weiter zu reduzieren indem die einzelnen Arbeitsschritte automatisiert oder teil-automatisiert werden. Eine andere Vision ist es kurze Videoszenen für Webvideos oder Schnittszenen künstlich auf Basis von einzelnen Photographien am Computer

zu erstellen. Dies erfordert Ansätze die komplexe Zusammenhänge und Erfahrungen wie das menschliche Gehirn vereinen können. Sie sollten im idealen Fall Kreativität umsetzen, Bewegungen und Abläufe voraussagen, bekannte Eigenschaften sinnvoll kombinieren oder erlernte Informationen übertragen und anwenden können. Diese Anforderungen können mit einer Form von künstlichen Intelligenz, den neuronalen Netzen (NN) (siehe Abschnitt 2.3), erfolgreich erfüllt werden, die jenen des menschlichen Gehirn nachempfunden sind. In den letzten Jahren wurden NN stetig weiterentwickelt und es konnten vor allem im Bereich der Medienproduktion bahnbrechende Erfolge erzielt werden. Die vielversprechendsten Erfolge konnten mit einer besonderen Form der NN erzielt werden. Diese Faltungsnetze (CNN) (siehe Abschnitt 2.3) ermöglichen orts- und skalierungs-unabhängige Operationen und somit ideal für mehrdimensionale digitale Signale.

Zu Beginn wurden CNN zur Objektklassifizierung eingesetzt um unter anderem automatisch Metadaten von Bild- oder Videodaten zu generieren und in Datenbanken oder Suchmaschinen einzupflegen. In den letzten Jahren wurden sie auch verstärkt für die Generierung oder Fortsetzung von bekannten Daten oder Signalen eingesetzt. Es konnten klassische Musikstücke sinnvoll beliebig verlängert werden [14] oder bewegte Sequenzen aus einzelnen Bildern generiert werden [15]. Auch in der Postproduktion konnten Bildern einer Stil aufgeprägt werden [7].

In den folgenden Kapiteln wird in den Grundlagen (siehe Kapitel 2) auf den allgemeine Ablauf in der Videoproduktion sowie detailliert die Funktionsweise der NN und CNN beschrieben. Im Anschluss werden in den folgenden Kapiteln verschiedene Entwicklungen von Videoproduktionsmittel vorgestellt, welche verschiedene Formen von NN und im besonderen von CNN nutzen. Drei Ansätze werden detailliert beschrieben und bewertet. Der erste Ansatz [15] beschreibt in Anschnitt 4.2 die Generierung von eine bewegten Bildsequenz aus einem Einzelbild. Die anderen Ansätze [7] [4] erläutert die Übertragung eines Bildstils auf eine andere Videosequenz.

2. Grundlagen

Laura

In Kapitel 2.1 wird zunächst ein allgemeiner Überblick über die verschiedenen Arbeitsschritte einer Videoproduktion. Im drauf folgenden Kapitel werden die Grundlagen von NNs zusammengefasst. Um dann in Kapitel 2.3 vertiefend auf Faltungsnetze einzugehen.

2.1. Videoproduktion

Laura

Mit der Videoproduktion oder auch Filmproduktion wird die Herstellung sowohl von Kino- als auch von Werbe- und Fernsehfilmen zusammengefasst. In Abbildung 1 ist ein Ablaufplan einer typischen Videoproduktion zu sehen. Da es alleine in Deutschland über 850 Produktionsformen gibt [13] (Stand 2014), kann der Ablaufplan nur einen sehr allgemeinen Überblick über die notwendigen Arbeitsschritte bieten.
Der erste Schritt, die Konzeption soll sowohl die Projektentwicklung, als auch die Vorproduktion zusammenfassen. Die sich anschließende Produktionsphase kann grob, wie im Schaubild zu sehen in Bild und Ton unterteilt werden, wobei diese beiden Bereiche nicht immer getrennt betrachtet werden sollten. Die Postproduktion besteht aus vielen verschiedenen Arbeitsschritten, deren Schwerpunkt auf dem Schnitt und der digitalen Bildnachbearbeitung liegt. Der Schritt der Distribution ist hier der Vollständigkeit halber erwähnt und beschreibt die Filmverwertung.

Der Aufbau der folgenden Ausführungen orientiert sich an Abbildung 1 (vgl. Kapitel **??**, 4 und 5).

2.2. Neuronale Netze

Laura

NNs finden unter anderem Anwendung bei der Steuerung von Robotern, Börsenkursanalysen, Medizin oder Fahrzeugsteuerung. In der Bildverarbeitung werden NNs vor allem zur Klassifizierung genutzt.
Sie sind vom menschlichen Gehirn inspiriert, welches

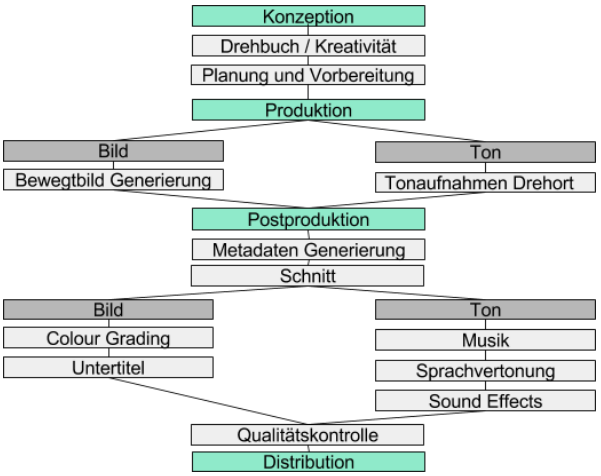

Abbildung 1. Grundlegende Arbeitsschritte einer Videoproduktion.

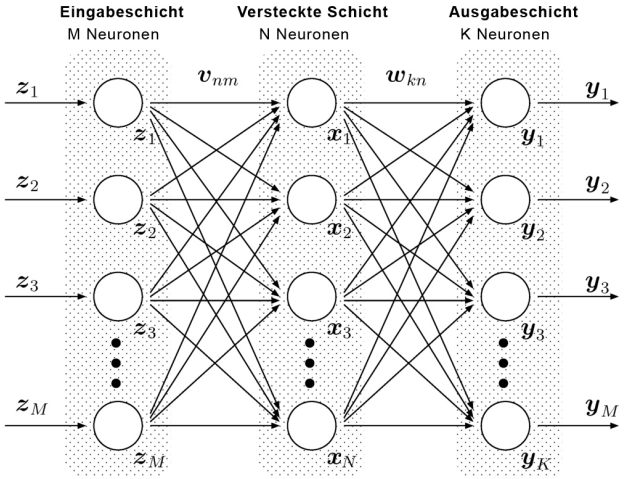

Abbildung 2. Prinzipieller Aufbau MLP nach [6].

laut [5] ein nicht-lineares, komplexes und hoch paralleles System zur Verarbeitung von Informationen darstellt. Ähnlich wie dieses bestehen künstliche NNs aus einer Menge an simulierten Neuronen, die untereinander verbunden sind und in Schichten organisiert sind. Es gibt verschiedene Arten der Vernetzung, die, wie in [5] und [12] beschrieben, in rück- und vorwärts gekoppelte Modelle unterteilt werden können.

Am häufigsten kommen sogenannte *Multilayer Perceptrons* (MLP) [1][8][12] zum Einsatz. Wie der Name vermuten lässt, werden hierbei die Neuronen in Schichten angeordnet. Ein solcher Aufbau ist beispielhaft in Abbildung 2 zu sehen. Dieses MLP besteht aus eine Eingabe- und Ausgabeschicht mit M bzw K Neuronen und einer versteckten Schicht mit N Neuronen. Es handelt sich um ein vorwärtsgekoppeltes Modell, bei welchem jedes Neuron einer Schicht mit jedem Neuron der darauffolgenden Schichte verbunden ist. Dies nennt man volle Verbindung. Die Eingangsschicht dient zum Verteilen der Eingangswerte z_m mit $m = 1, ..., M$. Die Ausgabe eines jeden Neurons in der versteckten Schicht, dargestellt durch x_n, lässt sich durch Formel 1 berechnen. Hierbei steht v_{nm} für die jeweilige Gewichtung der Verbindungen zwischen den Neuronen der Eingabe- und der versteckten Schicht und f für die Aktivierungfunktion [12][5] des jeweiligen Neurons.

$$x_n = f\,(\sum_{m=1}^{M} v_{nm} z_m) \tag{1}$$

Die Ausgangswerte y_k, mit $k = 1, ..., K$, lassen sich äquivalent unter Hereinnahme der Werte x_n und der Gewichte w_{kn}, sowie einer Aktivierungsfunktion g berechnen, und gelten als Vertrauenswerte. Sie müssen gemäß der Aufgabenstellung interpretiert werden.

2.3. Faltungsnetze

Laura

Im Folgenden wird genauer auf Faltungsnetze eingegangen, da diese die Grundlage, für die meisten der in den folgenden Kapiteln vorgestellten Ansätze, bilden. Vereinfacht ausgedrückt besteht ein Faltungsnetz aus einer Vernetzung von Faltungsoperationen mit unterschiedliche Filtermasken. Faltungsnetze kommen, bedingt durch ihre Architektur, oft zum Einsatz, wenn große Datenmengen von einem NN verarbeitet werden sollen. Ein schematischer Aufbau ist in Abbildung 3 zu sehen.

Jedes Pixel eines Feldes, das auf der Abbildung zu sehen ist, wird durch ein Neuron repräsentiert. Die Felder sind in Schichten organisiert. Die Eingangsschicht fungiert, vergleichbar wie bei den MLPs aus Kapitel 2.2, als Verteiler der Information an die

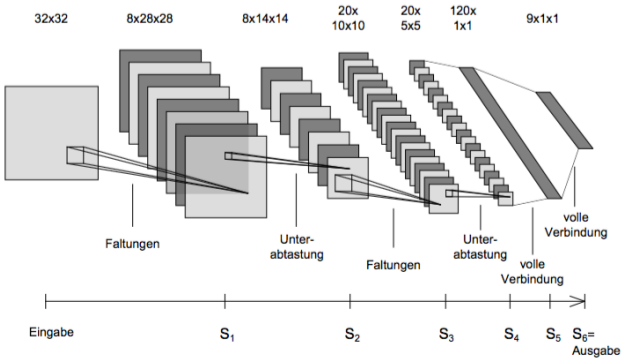

Abbildung 3. Prinzipieller Aufbau Faltungsnetz nach [10].

Neuronen der nächsten Schicht S_1. Die Besonderheit eines Faltungsnetzes sind die sich abwechselnd durchgeführte Faltung und anschließende Unterabtastung. Zwischen den Schichten S_4 und S_6 ähnelt das Modell einem MLP, da die Neuronen schichtweise voll verbunden sind.

Im Allgemeinen wird für eine Faltung eine Filtermaske h, also eine endlicher zweidimensionaler Koeffizientensatz, wie in Formel 2 zu sehen, verwendet. Hierbei stehen x und y jeweils für die horizontale bzw. die senkrechte Bildkoordinate. Die Anzahl der Koeffizienten a_{xy}, wird in der Horizontalen mit N_{hx} und im Vertikalen mit N_{hy} bezeichnet.

$$h(x,y) = \begin{cases} 0 \; für \; x < -\lfloor \frac{N_{hx}-1}{2} \rfloor \; \vee \; y < -\lfloor \frac{N_{hy}-1}{2} \rfloor \\ 0 \; für \; x > \lfloor \frac{N_{hx}-1}{2} \rfloor \quad \vee \; y > \lfloor \frac{N_{hy}-1}{2} \rfloor \\ a_{xy} \end{cases} \tag{2}$$

Formel 3 beschreibt die Faltung eines Eingangssignals s mit einer Filtermaske h, wobei I das Ausgangssignal in Abhängigkeit von x und y beschreibt.

$$I(x,y) = (s * h)(x,y) = \sum_{m_x=-\infty}^{\infty} \sum_{m_y=-\infty}^{\infty} s(m_x, m_y) \cdot h(x - m_x, y - m_y) \tag{3}$$

Im Fall eines Faltungsnetzes wird die Faltung, die wie in Abbildung 3 zu sehen, beispielsweise zwischen der Eingangsschicht und S_1 vollzogen wird, wird durch die Verbindung zwischen den Neuronen zweier Felder modelliert. Dabei entsprechen die Gewichte der Neuronen genau den Filterkoeffizienten a_{xy}. Für ein jedes Feld sind diese Koeffizienten konstant, was bedeutet, dass alle Neuronen eines Feldes mit nur einem Gewicht auskommen. Dieses Prinzip nennt man geteilte Gewichte.

Im Faltungsnetz wird nach jeder Faltung eine Unterabtastung durchgeführt um zu gewährleisten, dass die Dimension der Eingangsdaten schrittweise an die Dimension des Ausgangsvektors angepasst wird. Hierzu wird meist eine bilineare Unterabtastung um den Faktor 2 vorgenommen.

Wie zu Beginn des Kapitels erwähnt, haben Faltungsnetze gegenüber den MLPs den Vorteil, dass sie nahezu beliebig hochskaliert werden können und somit gut geeignet für große Datenmengen sind. Dies liegt vor allem daran, dass die Neuronen nur lokal verbunden sind und sich somit das Prinzip der geteilten Gewichte zu Nutze gemacht werden kann. Ein weiterer Vorteil von Faltungsnetzen, der vor allem in der Bildverarbeitung genutzt wird, ist das sie translationsinvariant sind.

3. Konzeption

Vera

Vor der eigentlichen Produktion von Videomaterial muss das Projekt zunächst konzipiert und detailliert geplant werden. Dies bezieht sich vor allem auf die kreativen Prozesse des Drehbuchschreibens und die darauffolgende gesamte Projektplanung und -vorbereitung. Naturgemäß sind NN weniger sinnvoll für die Planung und das Management von Projekten. Doch in den letzten Jahren wurde damit begonnen mit Hilfe von NN kreative Prozesse umzusetzen und zu erforschen ob sie Kreativität entwickeln können. In diesem Sinne ist Kreativität auch mit sinnvollen und selbstständigem Denken sowie Entscheiden gleichzusetzen. Beides ist im Entstehungsprozess von Drehbüchern unumgänglich.

Es gibt bereits erste Versuche mit Hilfe von NN automatisch sinnvolle Texte und Dialoge zu erstellen, die auf bekannten Texten und Storylines basieren. Dieses Verfahren wird in der Literatur auch *Deep Writing* ge-

nannt.Es können auch Romane, Dialoge oder Songtexte automatisch mit einem LSTM Recurrent NN erstellt werden [2].

3.1. Aktueller Stand

Vera

Ein solches LSTM Recurrent NN wurde mit allen bekannten Episoden der Serie *Silicon Valley* trainiert [3]. Nach dem erfolgreichen Training sind Wörter die häufig zusammenhängen in einem mathematischem Model gruppiert [2]. Zu Beginn des Schreibprozesses wird ein beliebiges Wort genutzt um den ersten Satz zu initialisieren. Das NN sucht im Anschluss das Wort, welches am häufigsten nach dem Startwort in den Trainingsdaten genannt wurde. Mit diesem Wortpaar wird nach dem selben Prinzip das dritte Wort des Textes ermittelt. Dieser Prozess wird solange wiederholt bis der generierte Text die gewünschte Länge erreicht hat.

Das Ergebnis sind Sätze, welche eine korrekte Grammatik aufweisen und häufig inhaltlich einen Sinn ergeben [3]. Trotzdem ergeben die generierten Sätze beziehungsweise Dialoge sind nicht kohärent und folgen keiner ersichtlichen Storyline.

Der Film *Sunspring* [9] wurde nach einem ähnlichen Prinzip erstellt und im Anschluss von einem professionellen Filmteam realisiert. Der größte Unterschied zu [3] ist, dass das NN nicht nur Wörter unterscheidet. Stattdessen wird zunächst alles in Buchstaben zerlegt und dann in neue Wörter und Sätze zusammengesetzt. Das verwendete NN nannte sich selber *Benjamin* und wurde mit einer großen Anzahl an Drehbüchern von Science Fiction Filmen aus den 80er und 90er Jahren trainiert.

Heraus kam ein Drehbuch, dass ähnlich wie [3] zwar grammatikalisch korrekte Sätze erschuf, die aber häufig inhaltlich keinen Sinn ergaben. Ein anderes Problem war der Umgang mit Namen, da diese sprachlich anders behandelt werden. Aus diesem Grund mussten alle Charaktere nur mit einzelnen Buchstaben benannt werden. Dies hatte zur Folge, dass das NN zwei Charaktere mit der selben Bezeichnung betitelt wurden und nachträglich umbenannt wurden.

Aus den mangelnden Zusammenhängen lässt sich ableiten, dass derzeit NN keine kreativen Prozesse simulieren können. Selbst mit einer sehr großen Anzahl von Trainingsdaten konnten keine kohärenten Dialoge generiert werden und keine konstante Storyline verfolgt werden. Weiterführend, sind die NN nicht in der Lage neue Charaktere oder Geschichten zu erfinden, sonder kombiniert lediglich bekannte wörtliche Zusammenhänge neu. Somit gibt es zur Zeit keinen brauchbaren Ansatz, welcher den Aufwand des Drehbuchschreibens minimieren könnte.

4. Produktion

Laura

In diesem Kapitel werden zunächst Ansätze vorgestellt, die auf Grundlage von NNs Arbeitsschritte bei der Produktion von Videos übernehmen bzw. vereinfachen könnten (vgl. Kapitel 4.1). Dazu ist an zu merken, dass diese Ansätze meist in einem anderen Kontext entwickelt wurden und ggf. eine Anpassung an die Standards einer Produktion stattfinden müsste.

In Kapitel 4.2 wird ein Ansatz zur automatisierten Generierung von Szenendynamiken in Hinsicht auf Funktionsweise und Arbeitserleichterung für die Videoproduktion analysiert.

4.1. Aktueller Stand

Laura

4.1.1. Ton

- Algo Rythm:

4.2. Szenendynamik nach Vondrick

Laura

5. Postproduktion

Vera

In diesem Kapitel wird die Verwendung von NN in der Videopostproduktion beschrieben. Gerade in diesem Produktionsabschnitt finden NN eine breites Anwendungsfeld, da gerade die CNN im Bereich der

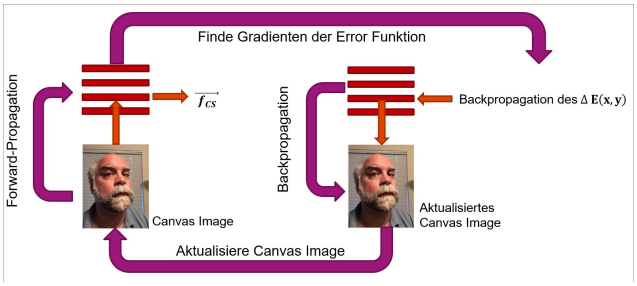

Abbildung 4. Ablaufdiagramm des Deep Dream Verfahrens.

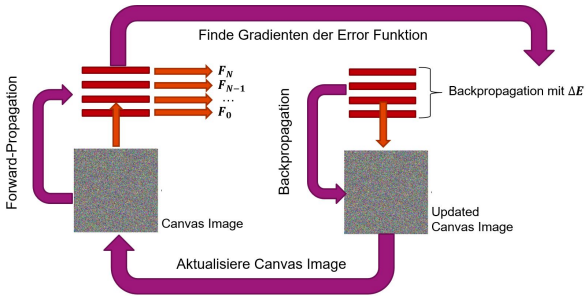

Abbildung 5. Ablaufdiagramm des Deep Style Verfahrens.

Bildverarbeitung in den letzten Jahre sehr erfolgreich eingesetzt werden konnten. Es stellte sich auch heraus, dass CNN auch für *Text-to-Speech* Verfahren erfolgreich eingesetzt werden können.

Im nächsten Abschnitt werden zunächst die interessantesten auf NN-basierenden Bildverabeitungs- und Text-to-Speech Ansätze in der Postproduktion vorgestellt und bewertet. Auf zwei Verfahren zur Stilsynthese wird in den darauffolgenden Anschnitten näher eingegangen inklusive einer ausführliche Analyse über die Verwendung der Verfahren in der professionellen Videoproduktion. Zuletzt wird ein Ausblick auf Einsatzmöglichkeiten gegeben.

5.1. Aktueller Stand

Vera

Metadaten generierung, Text-to-speech, Stilsynthese

Ein bekannter Ansatz ist *Clarifai* [11], welcher eine Bibliothek mit konfigurierten CNN anbietet um optimierte Datenbanken anzulegen und zu verwalten.

5.2. Faltungsnetze zur Stilsynthese

5.3. Deep Dream

5.3.1. Verfahren

5.3.2. Fazit

5.4. Deep Style

5.4.1. Verfahren

5.4.2. Fazit

5.5. Ausblick

6. Fazit

Vera

Literatur

[1] H. Braun. *Neuronale Netze Optimierung durch Lernen und Evolution*. Springer, 1997.

[2] M. Deutsch. How to write with artificial intelligence. https://medium.com/deep-writing/how-to-write-with-artificial-intelligence-45747ed073c. Abgerufen: 19.08.2016.

[3] M. Deutsch. Silicon valley: A new episode written by ai. https://medium.com/deep-writing/silicon-valley-a-new-episode-written-by-ai-a8f832645bc2. Abgerufen: 19.08.2016.

[4] L. A. Gatys, A. S. Ecker, and M. Bethge. A neural algorithm of artistic style. *CoRR*, abs/1508.06576, 2015.

[5] S. Haykin. *Neural Networks A Comprehensive Foundation*. Macmillan College Publishing Company, 1994.

[6] T. Isokawa, H. Nishimura, and N. Matsui. Quaternionic multilayer perceptron with local analyticity. *Information*, 3(4):756, 2012.

[7] A. Mordvintsev. Inceptionism: Going deeper into neural networks. https://research.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html. Abgerufen: 16.08.2016.

[8] D. Nauck, F. Klawonn, and R. Kruse. *Neuronale Netze und Fuzzy Systeme*. Vieweg, 1994.

[9] A. Newitz. Movie written by algorithm turns out to be hilarious and intense. https://arstechnica.com/gaming/2016/06/an-ai-wrote-this-movie-and-its-strangely-moving/. Abgerufen: 19.08.2016.

[10] M. Osadchy, Y. LeCun, and M. Miller. Synergistic face detection and pose estimation with energy-based models. *Journal of Machine Learning Research*, 8:1197–1215, 2007.

[11] T. Simonite. A startup's neural network can understand video. https://www.technologyreview.com/s/534631/a-startups-neural-network-can-understand-video/. Abgerufen: 16.08.2016.

[12] J. Stanley and E. Bate. *Neuronale Netze Computersimulation biologischer Intelligenz*. Systhema Verlag GmbH, 1991.

[13] statista. Anzahl der aktiven film- und fernsehproduktionsfirmen in deutschland in den jahren 1998 bis 2014. https://de.statista.com/statistik/daten/studie/243238/umfrage/anzahl-der-film-und-fernsehproduktionsfirmen-in-deutschland/. Abgerufen: 10.08.2016.

[14] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu. Wavenet: A generative model for raw audio. *CoRR*, abs/1609.03499, 2016.

[15] C. Vondrick, H. Pirsiavash, and A. Torralba. Generating videos with scene dynamics. *CoRR*, abs/1609.02612, 2016.