Maximilian Eibl, Martin Gaedke (Hrsg.): INFORMATIK 2017,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2017 1

Aktuelle Entwicklungen in der Automatischen Musikverfolgung

Andreas Arzt[1], Matthias Dorfer[2]

Abstract: Diese Arbeit befasst sich mit aktuellen Entwicklungen in der automatischen Musikverfolgung durch den Computer. Dieser Prozess ist auch unter den Begriffen „Score Following“ oder „Real-time Music Tracking“ bekannt. Es handelt sich dabei um Algorithmen, die einer musikalischen Aufführung „zuhören“, das aufgenommene Audiosignal mit einer (abstrakten) Repräsentation des Notentextes vergleichen und sozusagen in diesem mitlesen. Der Algorithmus kennt also zu jedem Zeitpunkt die Position der Musiker im Notentext. Diese Information erlaubt die Realisierung einer Reihe von Anwendungen, zum Beispiel der automatischen Musikvisualisierung und der automatischen Begleitung. Neben der Vermittlung eines generellen Überblicks, liegt der Schwerpunkt dieser Arbeit auf der Beleuchtung des Aspekts der Flexibilität und der einfacheren Nutzbarkeit dieser Algorithmen. Es wird dargelegt, welche Schritte getätigt wurden (und aktuell getätigt werden) um den Prozess der automatischen Musikverfolgung einfacher zugänglich zu machen. Dies umfasst Arbeiten zur automatischen Identifikation von gespielten Stücken und deren flexible Verfolgung ebenso wie aktuelle Ansätze mithilfe von Deep Learning, die es erlauben Bild und Ton direkt zu verbinden, ohne Umwege über abstrakte und nur unter großem Zeitaufwand zu erstellende Zwischenrepräsentationen.

Keywords: Automatische Musikverfolgung, Score Following, Music Tracking

1 Einleitung

Diese Arbeit beschäftigt sich mit der automatischen Verfolgung von Musik durch Computeralgorithmen – ein Prozess, der auch unter den Begriffen „Score Following“ und „Music Tracking“ bekannt ist. Das Ziel dieses Prozesses ist es, eine Aufführung von (vornehmlich klassischer) Musik mit dem zugrundeliegenden Notentext abzugleichen und sozusagen in diesem live mitzulesen.

Es handelt sich dabei also um Algorithmen, die einer musikalischen Aufführung „zuhören“, das Audiosignal verarbeiten, es mit einer abstrakten Repräsentation des Notentextes vergleichen und diese beiden Zeitreihen (das Audiosignal und den Notentext) miteinander synchronisieren („alignen“). Der Algorithmus berechnet zu jedem Zeitpunkt die Position der Musiker im Notentext. Mithilfe dieser Information lassen sich eine Reihe von Anwendungen realisieren, zum Beispiel automatische Begleitsysteme und die automatische Synchronisation von Medien wie Bild und Text zu der live gespielten Musik für künstlerische Zwecke.

[1] Austrian Research Institute for Artitifical Intelligence, andreas.arzt@ofai.at
[2] Johannes Kepler University Linz, Department of Computational Perception, matthias.dorfer@jku.at

Die ersten Arbeiten zur automatischen Musikverfolgung wurden in den 1980er Jahren veröffentlicht (siehe [Da84, Ve84]). Diese Ansätze waren der Idee eines automatischen Begleitsystems verbunden und basierten noch auf symbolischem Input (sowohl der Notentext als auch die musikalische Aufführung werden als Zeichenketten repräsentiert) und Stringmatchingalgorithmen, um diese beiden Informationen zu synchronisieren. Die Notwendigkeit von symbolischen Daten der musikalischen Aufführung war eine große Einschränkung, da diese Algorithmen somit nur mit speziellen Instrumenten verwendbar waren. Deshalb trat diese Art des Verfolgens von Musik etwas in den Hintergrund (Ausnahmen sind beispielsweise [Na15, Xi16]). Stattdessen liegt der Fokus aktuell hauptsächlich auf Systemen, die direkt auf dem Audiosignal der musikalischen Aufführung das Problem der automatischen Musikverfolgung zu lösen versuchen.

Die meisten aktuellen Verfahren basieren auf diversen Variationen von probabilistischen Modellen, beispielsweise auf Hidden Markov Modellen [Ra10, Ra01, Ra09, OD01, Or03, Co06] und verwandten Modellen [Co10, DP11, LD15], Conditional Random Fields [SYK14, YSK13] und Partikelfiltern [Ko13, MC11, Ot11]. Eine Ausnahme stellen Musikverfolgungssysteme basierend auf Dynamic Time Warping (DTW) dar [AWD08, Ar16].

All diesen Systemen ist gemein, dass sie relativ inflexibel sind und die Aufbereitung der für die automatische Verfolgung notwendigen Daten zeitaufwändig ist. Wir werden diese Problematiken in Kapitel 2 detaillierter diskutieren. Daraufhin werden in den Kapiteln 3 und 4 zwei aktuelle Entwicklungen beschrieben, die auf einfachere und flexiblere Verwendung von Musikverfolgungsalgorithmen abzielen. Schließlich werden wir in Kapitel 5 den aktuellen Stand der Forschung nochmals kurz diskutieren und einen Ausblick auf kommende Herausforderungen geben.

2 Präzision vs. Flexibilität in der Automatischen Musikverfolgung

In den letzten Jahren lassen sich zwei Trends im Bereich der Musikverfolgung identifizieren, die auch stark von den jeweiligen Anwendungsgebieten der Algorithmen beeinflusst werden. Einerseits ist dies die Anwendung des *automatischen Begleitens*, und andererseits die *automatische Synchronisation von Visualisierungen*.

Automatische Begleitung hat zum Ziel zur Aufführung eines Solomusikers automatisch Begleitelemente (einzelne Effekten oder die Simulation eines kompletten Orchesters) zu synchronisieren (siehe zum Beispiel [Co10, Ra09]). In diesem Fall ist das Ziel Präzision, beziehungsweise sogar die Vorhersage von Ereignisse, da selbst kleine Abweichungen sich als hörbare Fehler in der Aufführung manifestieren würden.

Um die notwendige Präzision für automatische Begleitung zu erzielen, ist die adäquate Aufbereitung des Notentextes unumgänglich. Dieser muss in einer symbolischen Repräsentation vorliegen, die es erlaubt die erwartete Aufführung des Stückes so exakt wie möglich zu modellieren. Dies umfasst die Definition der Struktur (beispielsweise die genaue Festlegung,

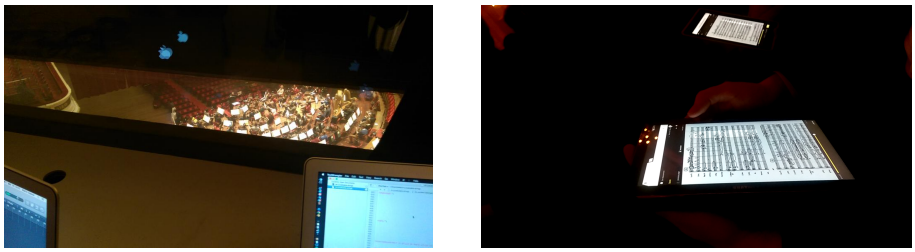

Abb. 1: Impressionen eines Experiments im Rahmen eines regulären Konzertes im Concertgebouw in Amsterdam. Links: Blick vom Kontrollraum auf die Bühne (während einer Orchesterprobe); Rechts: Zur Musik synchronisierter Notentext, angezeigt auf persönlichen Tabletcomputern im Publikum (aus [Ar15]).

welche Wiederholungen gespielt oder ausgelassen werden) ebenso wie die Modellierung von komplexeren Konzepten wie Trillern oder Glissandi, deren spezifische Ausführung nicht explizit durch den Notentext definiert wird (beispielsweise ist die Anzahl der Noten, die ein Triller umfasst, nicht per se festgelegt). In einigen Fällen ist die Erstellung der symbolischen Beschreibung direkt Teil des Kompositionsprozesses – die Komposition wird also bereits in Kombination mit dem spezifischen Musikverfolgungsalgorithmus erarbeitet [CCG14].

Diese Modelle werden speziell auf die geplante Aufführung abgestimmt. Dies geschieht normalerweise indem die Parameter des Modells auf einer Reihe von Probeaufführungen, die möglichst ähnlich der tatsächlichen Aufführung sein sollten, optimiert werden. Ein automatisches Begleitsystem für ein bestimmtes Stück beziehungsweise eine spezielle Aufführung vorzubereiten ist also ein komplexer und zeitaufwändiger Prozess.

Wie bereits erwähnt, ist das zweite große Anwendungsgebiet die *automatische Synchronisation von Visualisierungen* zur Musik. Dies umfasst beispielsweise die Anzeige eines zur Musik synchronisierten Notentextes (inklusive automatischem Umblättern) ebenso wie Medien wie Text, Bild oder Video, die dem Publikum im richtigen Moment präsentiert werden, um einen künstlerischen Mehrwert zu generieren oder einen weiterbildenden Effekt zu erzielen (siehe Abbildung 1, die Impressionen eines Tests einer solchen Anwendung im Concertgebouw in Amsterdam zeigt [Ar15]). Hier ist Präzision zwar noch immer eine wünschenswerte Eigenschaft, aber nicht notwendigerweise im selben Ausmaße. Um zum Beispiel automatisch den Notentext synchronisiert zur Musik anzuzeigen und umzublättern sind normalerweise Abweichungen im Bereich von einigen hundert Millisekunden durchaus akzeptabel.

Während also ein gewisses Maß an Genauigkeit natürlich weiter erforderlich ist, treten andere Aspekte ebenso in den Vordergrund. Von generellem Interesse ist beispielsweise die Reduktion des Aufwandes, der notwendig ist, um einen Notentext für die automatische Musikverfolgung aufzubereiten. Ein wichtiger Aspekt in vielen Anwendungen ist auch die Flexibilität im Umgang mit der Aufführung an sich. Nützlich kann beispielsweise ein

Algorithmus sein, der „ungeplante“ Abweichungen des Musikers vom Notentext (wie das Auslassen einer Wiederholung) erkennt und dynamisch darauf korrekt und so schnell wie möglich reagiert. In dieser Arbeit werden wir nun das Streben nach Präzision, und damit die Arbeit an automatischen Begleitsystemen, außer Acht lassen und stattdessen zwei aktuelle Forschungsrichtungen zum flexiblen Musikverfolgen genauer betrachten.

Die erste Forschungsrichtung, beschrieben in Kapitel 3, beschäftigt sich mit der Frage nach Flexibilität während einer musikalischen Performance. Wir werden hier einen Ansatz zusammenfassen, der in der Lage ist innerhalb weniger Sekunden Musikstücke zu identifizieren. Damit wird ein Musikverfolgungssystem realisiert, das flexibles Verfolgen auf Basis einer großen Datenbank an Notentexten erlaubt.

Die zweite Forschungsrichtung, präsentiert in Kapitel 4, setzt noch einen Schritt früher an und versucht den Vorverarbeitungsprozess, der klassischerweise notwendig ist um den Notentext für automatische Musikverfolgung aufzubereiten, obsolet zu machen. Dazu werden Verfahren des maschinellen Lernens verwendet, um direkt Korrespondenzen zwischen Bilddaten und Audiosignalen zu lernen. Die präsentierten Ansätze stecken zwar noch in den Kinderschuhen, sind aber erste wichtig Schritte hin zu unserer Vision von einfacher Nutzung von automatischer Musikverfolgung für jedermann: eine Anwendung auf einem (mobilen) Endgerät, die anhand von Fotos von Seiten von Notentexten in der Lage ist, ohne jedweden manuellen Eingriff, eine Aufführung zu verfolgen.

3 Ein flexibles Musikverfolgungssystem basierend auf einer Datenbank symbolischer Notentexte

Der Prozess des automatischen Musikverfolgens ist normalerweise relativ starr definiert. Es wird ein definiertes Stück von Anfang bis Ende verfolgt, wobei keine strukturellen Abweichungen erlaubt sind. Ein flexiblerer Ansatz wird in [Ar16] vorgestellt.

Dieses System basiert auf einer Datenbank von symbolischen Notentexten klassischer Klaviermusik und enthält unter anderem alle Sonaten Mozarts und Beethovens, sowie Chopins Gesamtwerk für Soloklavier (insgesamt mehr als eine Million Noten). Es erlaubt flexible Musikverfolgung auf der gesamten Datenbank von Stücken und kann automatisch das Stück, sowie die exakte Position innerhalb des Stückes erkennen und dann der Aufführung folgen. Ebenso kann es beliebige Sprünge innerhalb eines Stückes und auch zu anderen Stücken detektieren und die korrekte Position wiederfinden. Es kann also jeder Aktion des Musikers / der Musikerin innerhalb dieser Datenbank von Notentexten folgen.

Abbildung 2 skizziert die Funktionsweise dieses Systems. Das Erkennen der Position über alle in der Datenbank vorhandenen Notentexte wird durch eine Kombination aus einem Musiktranskriptionsalgorithmus [BS12] (dieser Algorithmus ist auf Klaviermusik trainiert und schränkt somit die Anwendung des Systems auf ebendiese ein) und einem tempo-invarianten musikalischen Fingerprinting-Algorithmus [ABW12] realisiert. Zuerst

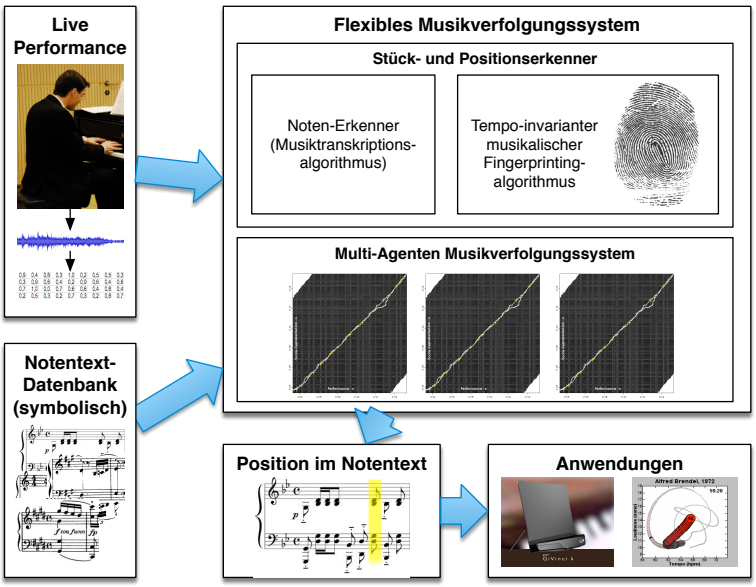

Abb. 2: Schematische Darstellung des flexiblen Musikverfolgungssystems (adaptiert von [Ar16]).

übersetzt der Transkriptionsalgorithmus das Audiosignal der Aufführung in eine Sequenz symbolischer Information (eine Liste von Tonhöhen inklusive deren Beginnzeiten). Mithilfe des Fingerprinting-Algorithmus wird die Datenbank von symbolischen Notentexten nach zum aktuellen Kontext in der Aufführung möglichst ähnlichen Ausschnitten durchsucht. Dieser Prozess läuft dauerhaft im Hintergrund und berechnet Hypothesen, die durch das Multiagenten-Musikverfolgungssystem weiterverarbeitet werden.

Das Multiagenten-Musikverfolgungssystem basiert auf mehreren Instanzen eines Musikverfolgungssystems, das Audio-zu-Audio Alignment mithilfe einer Echtzeitversion des Dynamic Time Warping Algorithmus realisiert [Di05, AWD08, Ar16]. Dieses übernimmt die Positionshypothesen und versucht die musikalische Aufführung anhand dieser Notentextpositionen mitzuverfolgen. Falls es sich bei einer Hypothese um die korrekte Position handelt, wird dies zu geringen Alignmentkosten führen, während der Versuch dies anhand einer inkorrekten Position zu tun zu hohen Kosten führen wird. Zu jedem Zeitpunkt wird die Instanz des Verfolgungsalgorithmus, die die geringsten Alignmentkosten verursacht, als aktiv markiert – sie repräsentiert also die aktuelle Positionshypothese des Systems über alle Stücke in der Datenbank.

Dieses System erlaubt flexibles automatisches Musikverfolgen anhand einer großen Sammlung von (symbolischen) Notentexten klassischer Klaviermusik. Neben Anwendungen für Musiker, wie unterstützende Übungssysteme, bietet es sich an, diese Algorithmen als

Grundgerüst für mobile Anwendungen für Konsumenten klassischer Musik zu nutzen. Beispielsweise ließe sich eine App realisieren, die dem Liebhaber klassischer Klaviermusik vollkommen autonom Zusatzinformationen zur Musik liefert. Diese Anwendung könnte automatisch erkennen welches Stück gespielt wird und den Notentext und künstlerische oder weiterbildende Visualisierungen (etwa die Struktur des Stückes oder Informationen über wichtige musikalische Themen) einblenden. Falls das Stück entsprechend aufbereitet wurde, könnten historische Hintergrundinformation über den Komponisten und das Stück bereitgestellt werden. Ebenso könnte auf weitere berühmte Interpretation des Stücks (und deren Bezugsquellen) hingewiesen werden.

4 Automatische Musikverfolgung direkt anhand des graphischen Notentextes

Ein großes Hindernis, das einer weiter verbreiteten Verwendung von automatischer Musikverfolgung im Weg steht, ist der immense Zeitaufwand der notwendig ist, um die Notentexte für den Verfolgungsprozess aufzubereiten. Der größte Aufwand wird durch die Erstellung beziehungsweise Aufbereitung einer symbolischen Beschreibung des Notentext (zum Beispiel MusicXML oder auch MIDI) verursacht. Diese Beschreibung ist notwendig um den Notentext exakt modellieren zu können.

Es gibt zwar automatische Verfahren via sogenannter Optical Music Recognition (OMR) Algorithmen um aus dem graphischen Notentext diese symbolische Repräsentation zu extrahieren, diese sind aber momentan nur sehr eingeschränkt für komplexere Notentexte geeignet [We15, Ha16, BS15, Re12]. Ein Versuch den Notentext für das zur automatischen Musikverfolgung verwandte Gebiet des Audio Alignment via OMR aufzubereiten, ist in [Ku07] beschrieben.

In diesem Kapitel wird ein neuer Ansatz vorgestellt, der komplett auf jedwede Zwischenrepräsentation verzichtet und direkt auf dem graphischen Notentext arbeitet. Der Ansatz basiert auf aktuellen Forschungsergebnissen aus dem Bereich der künstlichen Intelligenz, genauer gesagt auf künstlichen neuronalen Netzen, welche auch unter dem Schlagwort „Deep Learning“ populär geworden sind. Die angesprochenen Fortschritte im maschinellen Lernen machen es möglich Korrespondenzen zwischen Audiodaten (der musikalischen Aufführung) und Bilddaten (dem Notentext) *direkt* zu lernen. Diese Algorithmen lernen also gleichzeitig einen Notentext zu „lesen“, Musik zu „hören“, und das Gelesene und Gehörte miteinander in Beziehung zu setzen. Die neuronalen Netze erlernen diesen Vorgang rein durch Training anhand von Beispielen.

Nachfolgend werden wir zwei unterschiedliche, komplementäre Ansätze vorstellen, die beide dieses Paradigma verfolgen: Automatische Musikverfolgung via Audio-zu-Notentext Matching (siehe Kapitel 4.1) und Automatische Musikverfolgung via Multi-Modalen Joint Embedding Spaces (siehe Kapitel 4.2).

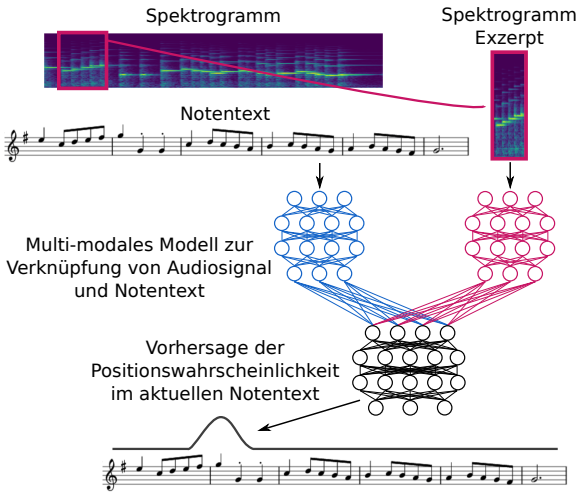

Abb. 3: Schematische Darstellung des automatischen Musikverfolgungsprozesses via Audio-zu-Notentext Matching. Die beiden oberen, getrennten Teile des multi-modalen Netzes verarbeiten die jeweilige Modalität (Notentext und Audio). Der untere Teil des Netzes führt beide Modalitäten zusammen und sagt die Positionswahrscheinlichkeit im aktuellen Notentext vorher.

Beide Arbeiten sind noch frühe Prototypen und wurden bisher hauptsächlich mit aus MIDI Dateien synthetisierten Audiodaten (via Klaviersoundfonts) trainiert und getestet. Nichtsdestotrotz handelt es sich um vielversprechende Ansätze, die in einigen Jahren automatische Musikverfolgung für jedermann zugänglich machen könnten. Anstelle einer komplizierten Datenaufbereitung würde es ausreichen direkt auf einem mobilen Endgerät den Notentext zu fotografieren. Ohne weitere manuelle Eingriffe könnte danach das Musikverfolgen gestartet werden.

4.1 Automatische Musikverfolgung via Audio-zu-Notentext Matching

Die Idee dieses Ansatzes ist gleichzeitig den aktuellen Kontext im Audiosignal und ein Bild des Notentextes zu verarbeiten und in Beziehung zueinander zu setzen. In [DAW16] wurden erste Versuche auf sehr reduzierter Klaviermusik (monophoner Notentext und monophones Audiosignal) beschrieben. Ein Schema dieses Ansatzes ist in Abbildung 3 zu sehen. Da eine genaue Beschreibung des Algorithmus den Rahmen dieser Arbeit sprengen würde, skizzieren wir nachfolgend seine generelle Funktionsweise.

Das zentrale Element dieses Ansatzes sind zwei parallele Convolutional Neural Networks, je eines für die Audioeingabedaten beziehungsweise den Notentext. Diese beiden Netze lernen abstrakte interne Repräsentationen der jeweiligen Modalität die im Anschluss durch einen gemeinsam Layer miteinander verbunden werden. Das Ziel dieses multimodalen Netzwerks

ist es, eine Wahrscheinlichkeitsverteilung abzuschätzen, die den aktuellen Audiokontext in Relation zum Notentext setzt – also wiedergibt, wie wahrscheinlich der gegebene Ausschnitt an Audiodaten den jeweiligen Positionen im Notentext entspricht. Gegeben diese Verteilung kann die aktuelle Position durch finden des Maximums dieser Verteilung berechnet werden.

4.2 Automatische Musikverfolgung via Multi-Modalen Joint Embedding Spaces

Der zweite auf multi-modalen neuronalen Netzen basierende Ansatz ist enger mit der flexiblen Musikverfolgung, wie in Kapitel 2 vorgestellt, verwandt. Hier wird das Problem des Findens von Korrespondenzen zwischen Audiodaten und Bilddaten vom Blickpunkt einer Retrievalaufgabe zwischen den beiden Modalitäten betrachtet [DAW17].

Während des Trainings werden dem neuronalen Netz wieder Paare von kurzen Audiodaten und Ausschnitten von Bildern von Notentexten präsentiert. Allerdings wird in diesem Ansatz versucht beide Modalitäten in einem gemeinsamen Einbettungsraum („Embedding Space“) zu repräsentieren. Dazu werden während des Trainings für beide Modalitäten Projektionen (in unserem Fall nichtlineare Projektionen mittels neuronaler Netze) mit folgender Eigenschaft gelernt: Die Projektionen (Einbettungskoordinaten) zusammengehöriger Audio- und Bilddaten sollen kleine Distanzen zueinander aufweisen, wohingegen die Projektionen nicht korrespondierender Paare möglichst weit auseinander liegen sollen. Abbildung 4 skizziert die Architektur dieses Retrievalnetzwerks wobei $f(\mathbf{a}, \Theta_f)$ und $g(\mathbf{i}, \Theta_g)$ die Projektionen von Audio $\mathbf{a}$ und Bildausschnitt $\mathbf{i}$ in den Einbettungsraum bezeichnen. Θ_f und Θ_g sind die Parameter der jeweiligen Netze. Um die oben beschriebenen Lagebeziehungen zwischen den Projektionen von korrespondierenden und nicht korrespondierenden Paaren zu erreichen, werden die Netzwerke dahingehenden optimiert, einen auf Kosinusdistanz basierenden *paarweisen Rankingfehler* zu minimieren (für eine detaillierte Beschreibung dieses Optimierungsvorgangs verweisen wir auf [DAW17]).

Nach dem Training können beide Projektionen unabhängig voneinander verwendet werden. Zum Beispiel kann zuerst eine Referenzdatenbank bestehend aus Bildern von Notentexten mithilfe der gelernten Projektion f für Notentextbilder in diesen Einbettungsraum projiziert werden. Danach kann eine Audioabfrage mithilfe der Projektion g für die Audiodaten in denselben Raum transferiert werden. Auf Basis der oben beschriebenen Distanz (bzw. Lage) Eigenschaften kann der Raum nun mittels Kosinusdistanz nach Notentextausschnitten, die der Audioabfrage ähnlich sind, durchsucht werden. Abbildung 5 skizziert diesen Vorgang.

Dieses Szenario ist in der Literatur auch als Cross-Modality Retrieval bekannt. Um damit automatische Musikverfolgung zu realisieren, wird zuerst der Notentext (oder, wie in der flexiblen Musikverfolgen in Kapitel 2 eine Sammlung von Notentexten) in den Suchraum eingebettet. Während der Aufführung wird der aktuelle Audiokontext laufend als Abfrage verwendet. Das tatsächliche Musikverfolgen findet dann durch die zeitliche Analyse der Abfrageergebnisse analog zu traditionellen Ansätzen (etwa mithilfe von DTW) statt.

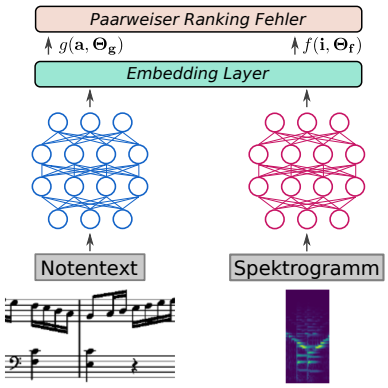

Abb. 4: Architektur des Retrievalnetzwerks. Das Netzwerk wird mit dem Ziel die Ähnlichkeit von korrespondierenden Paaren von Audio- und Notentextdaten zu maximieren trainiert. Dies wird durch das Minimieren eines paarweisen Rankingfehlers erreicht.

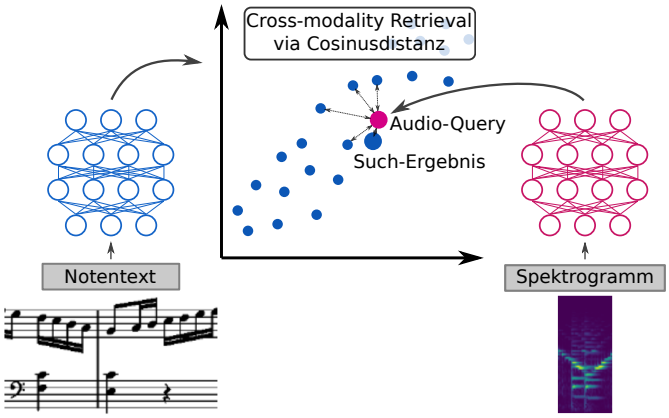

Abb. 5: Der Retrievalprozess. Blaue Punkte repräsentieren eingebettete Notentextinstanzen. Der rote Punkt symbolisiert die Projektion der Audioabfrage in ebendiesen Raum. Der große blaue Punkt ist die Notentextinstanz, die der eingebetteten Audioabfrage am nächsten ist.

5 Zusammenfassung und Ausblick

Diese Arbeit beschäftigt sich mit flexiblen Algorithmen zur automatischen Musikverfolgung. Zum einen wurde eine Methode diskutiert, die flexibles Verfolgen basierend auf einer Datenbank von Notentexten erlaubt. Diese Methode ist bereits robust genug um komplexe Klaviermusik zu verfolgen und könnte als Grundgerüst für (mobile) Anwendungen verwendet werden. Zum anderen wurden in dieser Arbeit Ansätze diskutiert, die es erlauben automatische Musikverfolgung flexibler einzusetzen, indem die aufwändige explizite Erstellung einer symbolischen Repräsentation vermieden wird. Hierzu wurden zwei miteinander

verwandte Methoden beschrieben, die direkte Korrespondenzen zwischen Bilddaten des Notentextes und Audiodaten einer Aufführung automatisch lernen. Dies bedeutet, dass die Qualität des gelernten Modells stark von der Anzahl und Qualität der verfügbaren Trainingsdaten abhängt. Die Erweiterung dieses Trainingsdatensatzes hat also hohe Priorität um noch generellere Modelle zur Verbindung von Audio- und Bilddaten zu trainieren. Das Hauptziel dieser Vergrößerung und Diversifizierung des Trainingsdatensatzes ist es diesen Ansatz auch mit Audiodaten von tatsächlichen Aufführungen klassischer Klaviermusik (beziehungsweise zu einem späteren Zeitpunkt beliebiger klassischer Musik) verwenden zu können – bisher wurden diese Modelle wie erwähnt mit aus MIDI generierten Audiodaten getestet. Dies ist nicht trivial, da die neuronalen Netze dazu lernen müssen mit Herausforderungen wie asynchronen Notenonsets innerhalb eines Akkords, die Verwendung der Pedale des Klavieres und variierende Dynamik entsprechend zu modellieren. Die bisherigen Ergebnisse sowie der potentielle Nutzen eines generellen notentextbasierten automatischen Musikverfolgungssystems deuten für uns jedoch darauf hin, dass es vielversprechend ist diese Forschungsrichtung verstärkt zu verfolgen.

Acknowledgements

Diese Arbeit wurde von den österreichischen Bundesministerien BMVIT und BMWFW und dem Bundesland Oberösterreich (COMET Center SCCH) sowie von der Europäischen Union (European Research Council, ERC Grant Agreement 670035, project CON ESPRESSIONE) unterstützt. Die Tesla K40, die für diese Arbeit verwendet wurde, wurde von der NVIDIA Corporation bereitgestellt.

Literaturverzeichnis

[ABW12] Arzt, Andreas; Böck, Sebastian; Widmer, Gerhard: Fast Identification of Piece and Score Position via Symbolic Fingerprinting. In: Proceedings of the International Society for Music Information Retrieval Conference (ISMIR). Porto, Portugal, S. 433–438, 2012.

[Ar15] Arzt, Andreas; Frostel, Harald; Gadermaier, Thassilo; Gasser, Martin; Grachten, Maarten; Widmer, Gerhard: Artificial Intelligence in the Concertgebouw. In: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI). Buenos Aires, Argentina, S. 2424–2430, 2015.

[Ar16] Arzt, Andreas: Flexible and Robust Music Tracking. Dissertation, Johannes Kepler University, Linz, Austria, 2016.

[AWD08] Arzt, Andreas; Widmer, Gerhard; Dixon, Simon: Automatic Page Turning for Musicians via Real-Time Machine Listening. In: Proceedings of the European Conference on Artificial Intelligence (ECAI). Patras, Greece, S. 241–245, 2008.

[BS12] Böck, Sebastian; Schedl, Markus: Polyphonic piano note transcription with recurrent neural networks. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). Kyoto, Japan, S. 121–124, 2012.

[BS15] Byrd, Donald; Simonsen, Jakob Grue: Towards a standard testbed for optical music recognition: Definitions, metrics, and page images. Journal of New Music Research, 44(3):169–195, 2015.

[CCG14] Coffy, Thomas; Cont, Arshia; Giavitto, Jean-Louis: AscoGraph: A User Interface for Sequencing and Score Following for Interactive Music. In: Proceedings of the International Computer Music Conference (ICMC). Athens, Greece, 2014.

[Co06] Cont, Arshia: Realtime Audio to Score Alignment for Polyphonic Music Instruments, using Sparse Non-Negative Constraints and Hierarchical HMMS. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). Toulouse, France, S. 245–248, 2006.

[Co10] Cont, Arshia: A Coupled Duration-Focused Architecture for Real-Time Music-to-Score Alignment. IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(6):974–987, 2010.

[Da84] Dannenberg, Roger B.: An On-Line Algorithm for Real-Time Accompaniment. In: Proceedings of the International Computer Music Conference (ICMC). Paris, France, S. 193–198, 1984.

[DAW16] Dorfer, Matthias; Arzt, Andreas; Widmer, Gerhard: Towards score following in sheet music images. In: Proceedings of the International Society for Music Information Retrieval Conference (ISMIR). S. 789–795, 2016.

[DAW17] Dorfer, Matthias; Arzt, Andreas; Widmer, Gerhard: Learning Audio-Sheet Music Correspondences for Score Identification and Offline Alignment. In: Proceedings of the International Society for Music Information Retrieval Conference (ISMIR). 2017.

[Di05] Dixon, Simon: An On-Line Time Warping Algorithm for Tracking Musical Performances. In: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI). Edinburgh, Scotland, S. 1727–1728, 2005.

[DP11] Duan, Zhiyao; Pardo, Brian: A state space model for on-line polyphonic audio-score alignment. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). Prague, Czech Republic, S. 197–200, 2011.

[Ha16] Hajič jr, Jan; Novotnỳ, Jirı; Pecina, Pavel; Pokornỳ, Jaroslav: Further Steps towards a Standard Testbed for Optical Music Recognition. In: Proceedings of the International Society for Music Information Retrieval Conference (ISMIR). New York, USA, S. 157–163, 2016.

[Ko13] Korzeniowski, Filip; Krebs, Florian; Arzt, Andreas; Widmer, Gerhard: Tracking Rests and Tempo Changes: Improved Score Following with Particle Filters. In: Proceedings of the International Computer Music Conference (ICMC). Perth, Australia, 2013.

[Ku07] Kurth, Frank; Müller, Meinard; Fremerey, Christian; Chang, Yoon-ha; Clausen, Michael: Automated Synchronization of Scanned Sheet Music with Audio Recordings. In: Proceedings of the International Society for Music Information Retrieval Conference (ISMIR). Wien, Österreich, S. 261–266, 2007.

[LD15] Li, Bochen; Duan, Zhiyao: Score Following for Piano Performances with Sustain-Pedal Effects. In: Proceedings of the International Society for Music Information Retrieval Conference (ISMIR). Málaga, Spain, S. 469–475, 2015.

[MC11] Montecchio, Nicola; Cont, Arshia: A unified approach to real time audio-to-score and audio-to-audio alignment using sequential Montecarlo inference techniques. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). Prague, Czech Republic, S. 193–196, 2011.

[Na15] Nakamura, Eita; Cuvillier, Philippe; Cont, Arshia; Ono, Nobutaka; Sagayama, Shigeki: Autoregressive Hidden Semi-Markov Model of Symbolic Music Performance for Score Following. In: Proceedings of the International Society for Music Information Retrieval Conference. Málaga, Spain, S. 392–398, 2015.

[OD01] Orio, Nicola; Déchelle, François: Score Following Using Spectral Analysis and Hidden Markov Models. In: Proceedings of the International Computer Music Conference (ICMC). Havana, Cuba, 2001.

[Or03] Orio, Nicola; Lemouton, Serge; Schwarz, Diemo; Schnell, Norbert: Score Following: State of the Art and New Developments. In: Proceedings of the International Conference on New Interfaces for Musical Expression (NIME). Montréal, Canada, S. 36–41, 2003.

[Ot11] Otsuka, Takuma; Nakadai, Kazuhiro; Takahashi, Toru; Ogata, Tetsuya; Okuno, Hiroshi G.: Real-Time Audio-to-Score Alignment using Particle Filter for Co-player Music Robots. EURASIP Journal on Advances in Signal Processing, 2011(2011:384651), 2011.

[Ra01] Raphael, Christopher: A Bayesian Network for Real-Time Musical Accompaniment. In: Advances in Neural Information Processing Systems (NIPS). Vancouver, Canada, S. 1433–1439, 2001.

[Ra09] Raphael, Christopher: Current Directions with Music Plus One. In: Proceedings of the Sound and Music Computing Conference (SMC). Porto, Portugal, S. 71–76, 2009.

[Ra10] Raphael, Christopher: Music Plus One and Machine Learning. In: Proceedings of the International Conference on Machine Learning (ICML). Haifa, Israel, S. 21–28, 2010.

[Re12] Rebelo, Ana; Fujinaga, Ichiro; Paszkiewicz, Filipe; Marcal, Andre R. S.; Guedes, Carlos; Cardoso, Jaime S.: Optical music recognition: state-of-the-art and open issues. International Journal of Multimedia Information Retrieval, 1(3):173–190, 2012.

[SYK14] Sako, Shinji; Yamamoto, Ryuichi; Kitamura, Tadashi: Ryry: A Real-Time Score-Following Automatic Accompaniment Playback System Capable of Real Performances with Errors, Repeats and Jumps. In: Proceedings of the International Conference on Active Media Technology (AMT). Warsaw, Poland, S. 134–145, 2014.

[Ve84] Vercoe, Barry: The Synthetic Performer in the Context of Live Performance. In: Proceedings of the International Computer Music Conference (ICMC). Paris, France, S. 199–200, 1984.

[We15] Wen, Cuihong; Rebelo, Ana; Zhang, Jing; Cardoso, Jaime: A new optical music recognition system based on combined neural network. Pattern Recognition Letters, 58:1–7, 2015.

[Xi16] Xia, Guangyu: Expressive Collaborative Music Performance via Machine Learning. Dissertation, Carnegie Mellon University, Pittsburgh, USA, 2016.

[YSK13] Yamamoto, Ryuichi; Sako, Shinji; Kitamura, Tadashi: Robust on-line algorithm for real-time audio-to-score alignment based on a delayed decision and anticipation framework. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). Vancouver, Canada, S. 191–195, 2013.