Vorrichtung und Verfahren zum Zusammenstellen eines Videoclips basierend auf mehreren Videosequenzen und Textsegmenten

Die vorliegende Erfindung betrifft eine Vorrichtung und ein Verfahren zum Zusammenstellen eines Videoclips basierend auf mehreren Videosequenzen und Textsegmenten eines Textes. Die Vorrichtung umfasst ein Text-Modul, um einen Text-Merkmalsvektor mit semantischen Textmerkmalen eines Textsegments zu erzeugen. Die Vorrichtung hat weiter ein Transformer-Modul zum Erzeugen eines Vorhersagevektors, ein Auswahlmodul zum Auswählen eines Ziel-Vektors einer geeigneten Ziel-Videosequenz aus einer Merkmalsbank, ein Zuordnungsmodul zum Erstellen einer Sequenzliste mit Ziel-Vektoren für ausgewählte Videosequenzen und ein Videoclipmodul, um aus Videoszenen einer Videodatenbank Videoclips basierend auf der Sequenzliste zu erstellen.

Videoclips, die aus einzelnen Szenen oder Sequenzen bestehen, werden in vielen Bereichen des täglichen Lebens eingesetzt, beispielsweise in sozialen Medien, als Anleitungen oder in Form von Nachrichtenbeiträgen im Fernsehen.

Die Verbreitung von Nachrichten in Form von Videoclips, sogenannten Videonachrichten oder Nachrichtenclips, hat sich durch das Wachstum der digitalen Medien und den weit verbreiteten Internetzugang erheblich weiterentwickelt. Dies hat zu einer Verlagerung hin zu videobasierten Nachrichten auf Plattformen wie YouTube und sozialen Medien geführt. Diese ansprechenden und zugänglichen Formate werden gegenüber traditionellen textbasierten Nachrichten

insbesondere von jüngeren Leuten bevorzugt. Es besteht deshalb eine steigende Nachfrage nach der schnellen Erstellung von Videoinhalten, um das Interesse des Publikums im Online-Nachrichtenbereich zu wecken.

Nachrichtensender müssen also schneller als je zuvor ansprechende Videoclips produzieren, um ihre erfolgreiche Positionierung auf dem Markt zu gewährleisten. Dies ist zum Teil auf die wachsende Zahl von Nachrichtenquellen und den veränderten Medienkonsum des Zielpublikums zurückzuführen. Diese Entwicklung hat die Notwendigkeit, Nachrichtenclips schnell zu produzieren, verstärkt – eine Anforderung, die im Widerspruch zu den traditionell manuellen und zeitaufwendigen Videobearbeitungsprozessen steht.

Die Produktion von Videonachrichten umfasst komplexe, zeitaufwendige Schritte, die es schwierig machen, mit dem Tempo der aktuellen Nachrichten Schritt zu halten. Der Prozess beginnt mit dem Schreiben eines Nachrichtenartikels, der dann in ein Voice-Over-Skript für das Video umgewandelt wird. Bei der Produktion des Nachrichtenclips werden Aufnahmen ausgewählt und aneinandergereiht, die mit den Abschnitten des Artikels übereinstimmen, wobei die Grundsätze des Rhythmus (d. h. die Länge der Szenen beeinflusst den Ton des Nachrichtenclips), der Kontinuität (d. h. der Inhalt aufeinander folgender Szenen muss konsistent sein) und der Montage (d. h. der Stil aufeinander folgender Szenen richtet sich nach Regeln) beachtet werden müssen, um eine qualitativ hochwertige Ausgabe zu gewährleisten.

Abgesehen von einzelnen Fortschritten bei der Automatisierung der Videoproduktion erfüllen die derzeitigen Systeme nicht die Effizienz- und Qualitätsstandards, die für professionelle Nachrichtensendungen erforderlich sind. Nach wie vor sind umfangreiche manuelle Eingriffe bei der Erstellung von Videoclips notwendig, insbesondere wenn eine Zusammenstellung auf verschiedenen Quellmaterialien beruht und die generierten Clips semantisch kohärent mit dem redaktionellen Text und stilistisch mit den professionellen Bearbeitungsmaßstäben übereinstimmen sollen.

Es besteht daher ein großer Bedarf, den Bearbeitungs- und Herstellungsprozess von Videoclips zu verbessern und zu rationalisieren und ein System vorzuschlagen, mit dem die Qualität der Videoclips beibehalten werden kann und die manuellen Eingriffe reduziert werden können.

Gelöst wird die vorliegende Aufgabe mit einer Vorrichtung mit den Merkmalen des Anspruchs 1 und mit einem Verfahren mit den Merkmalen des Anspruchs 11.

In einem Aspekt betrifft die vorliegende Erfindung eine Vorrichtung, mit einem Text-Modul zum Erzeugen eines Text-Merkmalsvektors mit (semantischen) Textmerkmalen eines Textsegments, einem Transformer-Modul zum Erzeugen eines Vorhersagevektors für eine Videosequenz basierend auf dem Video-Vektor der vorhergehenden Videosequenz, einem Auswahl-Modul zum Auswählen eines Ziel-Vektors einer geeigneten Ziel-Videosequenz (auch Target genannt) aus einer Merkmalsbank (auch SourceShots genannt) basierend auf dem Vorhersagevektor und optional zum Erzeugen bzw. Auswählen eines zu der Ziel-Videosequenz korrespondierenden Ziel-Vektors, einem Zuordnungs-Modul zum Erstellen einer Sequenzliste mit Ziel-Vektoren für die Auswahl von Videosequenzen basierend auf den Ziel-Vektoren und einem Videoclip-Modul zum Erstellen eines Videoclips aus Videosequenzen einer Video-Datenbank basierend auf der Sequenzliste. Das Transformer-Modul umfasst einen Cross-Modal-Encoder und einen Feature-Decoder und ist trainiert, um basierend auf dem Video-Vektor und dem Text-Merkmalsvektor den Vorhersagevektor zu erstellen. Die Video-Vektoren umfassen Video-Klassifizierungsmerkmale, die stilistische Mittel oder Feature umfassen oder sind.

In einem weiteren Aspekt betrifft die vorliegende Erfindung eine Trainingsvorrichtung zum Trainieren eines Transformer-Moduls und zum Erstellen von Video-Vektoren für Videosequenzen basierend auf Textsegmenten eines Textes oder Nachrichtenartikels. Die Trainingsvorrichtung umfasst einen CLIP-Encoder zum Erstellen von Text-Merkmalsvektoren mit Textmerkmalen eines

Textsegments, einen CLIP-Encoder zum Erstellen von Inhalts-Vektoren mit Video-Inhaltsmerkmalen für eine Videosequenz, einen VGG-Encoder zum Erstellen von Klassifikationsvektoren mit Video-Klassifizierungsmerkmalen für eine Videosequenz und das Transformer-Modul mit einem Cross-Modal-Encoder und einem Feature-Decoder. Der Cross-Modal-Encoder ist dazu ausgebildet, um einen Text-Merkmalsvektor und einen Inhalts-Vektor sowie einen Klassifikationsvektor für eine Videosequenz zu einem Featurevektor zusammenzuführen und dem Feature-Decoder zuzuführen, wobei der Featurevektor bevorzugt die doppelte Länge des Inhalts-Vektors hat. Der Feature-Decoder ist dazu ausgebildet, um aus dem Featurevektor einen Video-Vektor zu erstellen, der die Merkmale einer am Eingang der Trainingsvorrichtung anliegenden Videosequenz aufweist und bevorzugt die halbe Länge des Featurevektors hat. Die Trainingsvorrichtung gibt den Video-Vektor an einem Ausgang aus und stellt ihn für eine Weiterverarbeitung zur Verfügung. Bevorzugt kann der Featurevektor weitere stilistische Feature wie z. B. die Kamerabewegung oder den Kamerawinkel umfassen. Damit können bevorzugt zusätzliche stilistische Parameter in das Training einfließen.

Weitere Aspekte der Erfindung betreffen ein entsprechendes Verfahren und ein Computerprogrammprodukt mit Programmcode zum Durchführen der Schritte des Verfahrens, wenn der Programmcode auf einem Computer ausgeführt wird, sowie ein Speichermedium, auf dem ein Computerprogramm gespeichert ist, das, wenn es auf einem Computer ausgeführt wird, eine Ausführung des hierin beschriebenen Verfahrens bewirkt.

Bevorzugte Ausgestaltungen der Erfindung werden in den abhängigen Ansprüchen beschrieben. Es versteht sich, dass die vorstehend genannten und die nachstehend noch zu erläuternden Merkmale nicht nur in der jeweils angegebenen Kombination, sondern auch in anderen Kombinationen oder in Alleinstellung verwendbar sind, ohne den Rahmen der vorliegenden Erfindung zu verlassen. Insbesondere können das Verfahren und das Computerprogrammprodukt entsprechend der für die Vorrichtung in den abhängigen Ansprüchen beschriebenen Ausgestaltungen ausgeführt sein.

Erfindungsgemäß wird basierend auf einem Text-Merkmalsvektor, der semantische Textmerkmale eines Textsegments umfasst, und basierend auf einem Video-Vektor mit Merkmalen einer Videosequenz ein Vorhersagevektor erzeugt. Der Video-Vektor der Videosequenz basiert auf einer vorhergehenden Videosequenz und umfasst deren Merkmale. Der Vorhersagevektor ist der von der KI-Einheit, dem sogenannten Transformer-Modul, erzeugte Merkmalsvektor, der sich an die vorhergehende Videosequenz anschließenden Videosequenz und umfasst deren Merkmale. Der Vorhersagevektor beschreibt also die gesuchte, nachfolgende Videosequenz.

Basierend auf diesem Vorhersagevektor mit Merkmalen einer auszuwählenden Videosequenz, die sich an die vorherige Videosequenz anschließen soll und mit dem Inhalt der für die geplante Videosequenz relevanten Textmerkmale des Textsegments übereinstimmen soll, wird die Suche nach einer geeigneten Sequenz durchgeführt. Basierend auf dem Vorhersagevektor wird ein Ziel-Vektor mit Merkmalen einer geeigneten Zielvideosequenz ausgewählt, wobei der Ziel-Vektor in einer Merkmalsbank umfasst ist. Die Merkmalsbank weist die Video-Vektoren aller zur Verfügung stehenden Videosequenzen auf, also die Merkmalsvektoren, die die Videosequenzen repräsentieren und mit ihnen korrespondieren. Aus diesem „Fundus“ an zur Verfügung stehenden Videosequenzen, die durch ihre Merkmalsvektoren, den sogenannten Video-Vektor oder Ziel-Vektor, repräsentiert werden, wird mit dem Vorhersagevektor verglichen. Ziel ist es, eine Auswahl des am geeignetsten Ziel-Vektors zu treffen. Dabei wird der Ziel-Vektor ausgewählt, der mit dem Vorhersagevektor am besten korreliert. Dieser Ziel-Vektor wird nun in eine Sequenzliste geschrieben, bevorzugt an eine bestehende oder leere Sequenzliste angehängt. Auf diese Weise umfasst die Sequenzliste nach mehreren Durchgängen eine Aneinanderreihung von Ziel-Vektoren, die Merkmale von ausgewählten und als geeignet eingestuften Videosequenzen aus dem „Fundus“ von Videosequenzen umfasst, die durch die Vektoren der Merkmalsbank repräsentiert werden.

Sobald alle Text-Merkmalsvektoren der Textsegmente des gesamten Nachrichtentextes abgearbeitet sind, weist die Sequenzliste alle zugehörigen Videosequenzen des Videoclips als Merkmale in Form von Ziel-Vektoren auf, die die einzelnen Videosequenzen des Videoclips repräsentieren. In einem Videoclip-Modul werden auf Basis der Ziel-Vektoren in der Merkmalsliste die Videosequenzen aus einer Videodatenbank ausgewählt und zu einem Videoclip zusammengefügt. Der dabei entstehende Videoclip umfasst dann die einzelnen Videosequenzen, die den hinterlegten Text, beispielsweise einen Nachrichtentext, am besten repräsentieren und wiedergeben.

Die einzelnen Aufnahmen des Kameramaterials oder auch von Archivmaterial einer entsprechenden Datenbank, die in Form von Videosegmenten zur Verfügung stehen, und die entsprechenden Artikelsegmente eines Textes sind die Grundeinheiten für die Zusammenstellung von Nachrichten- oder Videoclips. Ein qualitativ hochwertiger Nachrichtenclip mit mehreren Aufnahmen bzw. Videosequenzen, der auf den zugrundeliegenden Texten, insbesondere Nachrichtenartikeln, abgestimmt ist, weist dabei die Merkmale auf, die als wünschenswert eingestuft werden. Zum einen erfolgt eine visuelle Darstellung der Semantik des Nachrichtentextes. Zum anderen sind ansprechende Übergänge zwischen den einzelnen Videosequenzen realisiert, die den Regeln für den Videoschnitt von Nachrichtenclips entsprechen.

Erfindungsgemäß werden in dem Transformer-Modul, das einen Cross-Modal-Encoder und einen Feature-Decoder umfasst, Vorhersagevektoren erstellt, die auf einem Video-Vektor mit Merkmalen einer Videosequenz und einem Text-Merkmalsvektor mit Merkmalen eines Textsegments basieren. In vorteilhafter Weise können der Video-Vektor mit Merkmalen der Videosequenz und der Text-Merkmalsvektor in dem Cross-Modal-Encoder zusammengefügt werden, sodass ein Merkmalsvektor der doppelten Länge entsteht. Mittels des Feature-Decoders werden hieraus die geeigneten Merkmale extrahiert, um einen Merkmalsvektor, den Vorhersagevektor, zu erstellen, dessen Größe der Größe des Video-Vektors bzw. des Text-Merkmalsvektors entspricht. Bevorzugt sind der

Video-Vektor, der Text-Merkmalsvektor und der Vorhersagevektor jeweils gleich lang.

Die vorliegende Erfindung hat den Vorteil, dass ein Videoclip auch dann zuverlässig und automatisiert mit einer hohen Qualität hergestellt werden kann, wenn die Videosequenzen auf verschiedenen Quellmaterialien beruhen. Die Erfindung ermöglicht es, dass die generierten Videoclips semantisch kohärent mit dem redaktionellen Text sind, der dem Videoclip zugrunde liegt, und darüber hinaus stilistisch mit professionellen Bearbeitungsmaßstäben übereinstimmen und diese erfüllen. Die Erfindung bietet daher ein effizientes automatisiertes System für die Bearbeitung von Nachrichtenclips oder Videoclips mit Nachrichten, um Journalisten bei ihrer Aufgabe zu unterstützen und deutlich zu entlasten.

Weil sich die bisher bekannten Systeme nur auf Medienbereiche außerhalb der Nachrichtenproduktion fokussieren, wie etwa Spielfilm-Produktion, lassen sie sich nicht direkt auf den Bereich von Nachrichtenproduktion oder Dokumentationen anwenden, da in diesen Bereich deutlich andere Aspekte eine entscheidende Rolle spielen als bei Filmen. Für Journalisten und Videoeditoren gibt es bestimmte stilistische Mittel, um qualitativ hochwertige Nachrichten zu produzieren, die sich mit den bisherigen Systemen nur unzureichend abbilden lassen. Die vorliegende Erfindung konzentriert sich daher auf die besonderen stilistischen Mittel der Nachrichtenproduktion, um nicht nur eine inhaltlich konsistente, sondern auch eine qualitativ hochwertige Nachrichtenproduktion zu automatisieren. Im Rahmen der Erfindung wurde auch erkannt, dass die stilistischen Mittel die Qualität der produzierten Nachrichtenclips oder -videos deutlich erhöhen und verbessern. Ihre Berücksichtigung wurde als entscheidender Faktor für die Qualität und Akzeptanz der Nachrichtenvideos und für die Erhöhung der Informationsaufnahme durch den Betrachter erkannt.

Der vorliegenden Erfindung gelingt es also, eine Einbettung von Bildsprache und Videobearbeitungsschritten zu verwenden, die aus einem nicht beschrifte-

ten Videodatensatz mit mehreren Videosequenzen gelernt wurden. Das vollautomatische System reduziert die Bearbeitungszeit und verbessert die Qualität der erstellten Nachrichtenclips deutlich. Die verwendeten Videosegmente sind am Inhalt des Nachrichtenartikels und den jeweiligen Textsegmenten des Nachrichtentextes ausgerichtet, wobei fließende Übergänge zwischen den Aufnahmen gewährleistet werden, indem die etablierten Richtlinien für den Videoschnitt eingehalten werden. Dies gelingt dadurch, dass bei der Erstellung des Vorhersagevektors für die nächste Videosequenz der Merkmale umfassende Video-Vektor der vorhergehenden Videosequenz sowie die für die nächste Videosequenz bestimmten Textmerkmale eines Textsegments in Form eines Text-Merkmalsvektors verarbeitet werden.

Die vorliegende Erfindung verwendet also stilistische Merkmale und Features einer Videosequenz als Video-Feature, um ein Video-Text-Matching durchzuführen. Damit unterscheidet sich die Erfindung deutlich von den bekannten Systemen, die lediglich inhaltliche Merkmale eines Videos oder einer Sequenz von Bildern berücksichtigen. Zwar ist auch bekannt, zur Schätzung der nächsten Videosequenz einen „aesthetic score“ zu verwenden. Dieser berücksichtigt aber nur Aspekte wie Schärfe oder Fokus, was technische bzw. qualitativ-technische Merkmale eines Videos sind. Diese Merkmale spielen jedoch bei professionellem Kameramaterial kaum eine Rolle und sind fast immer erfüllt. Sie beziehen eben keine stilistischen Elemente und Aspekte mit ein.

In einer bevorzugten Ausgestaltung der Erfindung umfassen der Video-Vektor und der Vorhersagevektor jeweils Textmerkmale, Video-Inhaltsmerkmale und Video-Klassifikationsmerkmale. Die Vektoren haben also neben inhaltlichen und semantischen Merkmalen auch Klassifizierungsmerkmale. Diese sind geeignet, um Übergänge zwischen einzelnen Videosequenzen fließender und für den Betrachter angenehmer zu machen. Auf diese Weise ist es möglich, die Grundsätze der Videoschnitttechnik automatisiert zu etablieren.

Vorzugsweise werden die Videoklassifikationsmerkmale mittels eines VGG-Encoders erzeugt. Ein derartiger Encoder kann beispielsweise in das Transformer-Modul oder die Transformer-Einheit integriert sein. Der VGG-Encoder beruht auf dem Konzept des Transfer-Learnings, also der Fähigkeit, vorhandenes Wissen, das zur Lösung bestimmter Probleme entwickelt wurde, zur Lösung neuer Probleme zu nutzen. Der VGG-Encoder umfasst ein neuronales Netzwerk aus Faltungszellen, die an der Universität Oxford entwickelt wurden. Das Modell erreichte in einem Wettbewerb eine Genauigkeit von über 92 %, was dort den höchsten Wert darstellte. Die Faltungsmatrizen des neuronalen Netzwerks verwenden kleinere Faltungskerne (3x3) und basieren auf einer großen Datenbank von mehr als 14 Millionen gelabelten Bildern, die in über 100 Klassen eingeteilt sind.

Beispielsweise können mit einem fine-getunten, speziell weitertrainierten "VILA-Encoder" (z.B. von Google) Merkmale der Szenenkomposition erzeugt und verarbeitet werden. Ebenfalls können bevorzugt Bewegungsvektoren der Szene mit einem Bewegungs-Encoder erzeugt werden. Die verschiedenen Video-Klassifizierungsmerkmale können bevorzugt zu einem Vektor zusammengefügt werden.

In einer bevorzugten Ausführungsform sind die Klassifizierungsmerkmale der Video-Vektoren und des Vorhersagevektors sowie des Ziel-Vektors Merkmale der Videosequenz, die beispielsweise Größenmerkmale der Aufnahme umfassen können. Die Merkmale sind Mitglieder einer Liste, die Einstellungsgrößen der Videosequenz umfassen. Dies können beispielsweise Informationen darüber sein, ob eine Totalansicht, eine Normalbild- oder Medium-Ansicht, eine Detail-Ansicht oder eine Porträt-Ansicht vorliegt. Weitere Elemente der Liste der Klassifizierungsmerkmale umfassen den verwendeten Kamerawinkel der Videosequenz, Aussagen zur Bewegung der Kamera in der Videosequenz, Bewegungsrichtung der Kamera, Bewegungen von Personen, die in der Videosequenz gezeigt sind, Bewegungsrichtung von Personen in der Sequenz, Zoomeinstellungen, Beleuchtungseinstellungen oder ähnliche Einstellungen, die bei der Videosequenz verwendet wurden.

Bevorzugt werden auch die Szenenkomposition bezüglich des abgebildeten Subjekts berücksichtigt, d.h. ist die Person in der Bildmitte und gut erkennbar. Ebenso kann als Klassifikationsmerkmal auch die Bildqualität wie Bildunschärfe, Kamerawackler, etc. dienen. Diese sind zwar wichtig, aber für die erfindungsgemäße Vorrichtung und für die Erstellung von Nachrichtenvideos nicht in dem Maße wie in den im Stand der Technik bekannten Systemen. Im Bereich der Nachrichtenproduktion werden die Kamerabilder von professionellen Kameraleuten aufgenommen; daher spielt die Bildqualität eine geringe Rolle. Sie kann bevorzugt vorausgesetzt werden. Bisher wurden im Stand der Technik nur die Bildqualität betrachtet und nicht die anderen stilistischen Mittel, die essenziell für eine qualitativ hochwertigen Nachrichtenproduktion sind.

Eine bevorzugte Ausführungsform sieht vor, dass die Video-Inhaltsmerkmale der Video-Vektoren, Vorhersagevektoren und/oder Ziel-Vektoren mittels eines CLIP-Encoders erzeugt wurden. Ein CLIP-Encoder ermöglicht es, ähnlich wie bei textbasierten Transformern Inhalte aus Bildern oder Videosequenzen herauszulesen. Ein CLIP-Modell ist ein Contrastive Language Image Pre-training-Modell, das die relevantesten Bildunterschriften oder Texte für ein Bild vorhersagen kann, ohne dass das Modell für eine bestimmte Aufgabe optimiert ist. Es basiert dabei auf rd. 400 Millionen Paaren aus Bild und Text, mit denen es für die Vorhersage der zu einem bestimmten Bild gehörenden Beschriftungen trainiert wurde. Der CLIP-Encoder kann bevorzugt in dem Transformer-Modul integriert sein. Er kann jedoch auch als eigenständiges Modul ausgebildet sein.

In einer bevorzugten Ausführungsform umfasst das Text-Modul ebenfalls einen CLIP-Encoder. Der Text-Merkmalsvektor wird mittels dieses CLIP-Encoders erzeugt. Er umfasst die Textmerkmale und Inhalte des Textsegments, das der einzelnen Videosequenz zugeordnet werden soll.

In einer bevorzugten Ausführungsform ist das Transformer-Modul so ausgebildet, dass es eine Mehrzahl von Submodulen umfasst. Jedes der Submodule ist dazu ausgebildet, ein vorgegebenes Video-Klassifizierungsmerkmal eines Vi-

deovektors zu analysieren und auszuwerten. Damit lassen sich die Videoklassifikationsmerkmale beim Erstellen des Vorhersagevektors qualitativ hochwertiger berücksichtigen.

Jedes der Submodule in dem Transformer-Modul umfasst bevorzugt ein neuronales Netzwerk. Dieses neuronale Netzwerk ist bevorzugt ein Feed-Forward-Network, (FFN). Jedes einzelne Feed-Forward-Network wird speziell auf eine Modalität oder ein stilistisches Merkmal trainiert. Auf diese Weise ist es möglich, mehrere stilistische Merkmale gleichzeitig zu berücksichtigen, sodass bei der Erstellung eines Videovektors und/oder eines Vorhersagevektors mehrere stilistische Merkmale der Videosequenzen berücksichtigt werden. Hierdurch werden die Qualität und das Zusammenspiel zwischen Text und Bild bzw. Bildfolge, verbessert.

In einer bevorzugten Ausführungsform kann jedes einzelne Submodul ein Submerkmal zur Klassifizierung erzeugen, das beim Erstellen des Vorhersagevektors berücksichtigt wird. Der Vorhersagevektor umfasst dann die Summe aller Submerkmale der Submodule.

In einer bevorzugten Ausführungsform weist das Submodul ein neuronales Netzwerk oder ein Feed-Forward-Network auf. Besonders bevorzugt wird in dem Feed-Forward-Network eine Mehrzahl von stilistischen Features oder Merkmalen in Form von Video-Klassifikationsmerkmalen verarbeitet.

Ebenso bevorzugt ist eine Ausführungsform, in der jedes Feed-Forward-Network spezialisiert ist auf ein vorgegebenes stilistisches Mittel oder stilistisches Merkmal eines Video-Klassifizierungsmerkmals. Es findet somit eine spezifische Abarbeitung der einzelnen stilistischen Merkmale statt. Vorzugsweise werden die einzelnen Ergebnisse der Feed-Forward-Networks der einzelnen Submodule zusammengeführt und bilden die Basis für den Vorhersagevektor. Der Vorhersagevektor bildet die Grundlage für die Videosequenz, wobei der Ziel-Vektor oder Ziel-Videovektor einer geeigneten Ziel-Videosequenz aus der Merkmalsdatenbank basierend auf diesem Vorhersagevektor ausgewählt wird.

In einer bevorzugten Ausführungsform ist das Transformer-Modul dazu ausgebildet, um den Vorhersagevektor basierend auf mehreren Videovektoren und mehreren Text-Merkmalsvektoren zu erstellen. Hierbei werden mehrere Video-Klassifikationsmerkmale in den einzelnen Videovektoren berücksichtigt.

In einer bevorzugten Ausführungsform wird der Vorhersagevektor basierend auf den Videovektoren einer Mehrheit der zur Verfügung stehenden Videosequenzen erzeugt. Besonders bevorzugt erfolgt das Erzeugen auf den Videovektoren aller zur Verfügung stehenden Videosequenzen. Gleichzeitig oder alternativ kann die Erzeugung des Vorhersagevektors auf allen zur Verfügung stehenden Text-Merkmalsvektoren durchgeführt werden. Ebenso ist denkbar, dass die Erzeugung des Vorhersagevektors auf einer Mehrheit der zur Verfügung stehenden Text-Merkmalsvektoren basiert. Bevorzugt wird der Vorhersagevektor basierend auf Videovektoren und Text-Merkmalsvektoren erstellt, besonders bevorzugt basierend auf allen zur Verfügung stehenden Videovektoren und Text-Merkmalsvektoren. Beispielsweise kann hierzu ein Inputvektor verwendet werden, der alle Segmente des zur Verfügung stehenden Nachrichtentextes umfasst, sowie alle Szenen oder Videosequenzen des zur Verfügung stehenden Kameramaterials. Somit erhält das Transformer-Modul alle zur Verfügung stehenden Informationen für seine Schätzung und für die Erzeugung des Vorhersagevektors, um basierend hierauf den geeignetsten Zielvektor einer Ziel-Videosequenz auszuwählen. Das Transformer-Modul betrachtet folglich alle Textsegmente und bezieht damit alle Textinformationen in seine Schätzung ein. Dies hat den Vorteil, dass der sonst typischerweise iterativ gebildete Beitrag, der jeweils nur die vorhergehende Videosequenz oder Szene berücksichtigt, verbessert wird, da nun alle zur Verfügung stehenden Text-Merkmalsvektoren und Videovektoren in den Prozess des Erstellens des Vorhersagevektors und somit in die Auswahl des Ziel-Vektors einfließen können. Damit wird das gesamte Verfahren deutlich beschleunigt und gleichzeitig die Qualität für die der erzeugten Nachrichtenvideos verbessert.

Die Generierung des gesamten Zielvektors erfolgt also unter Berücksichtigung des gesamten Textes und aller vorhergehenden Szenen. Damit stehen dem

Transformer-Modul alle verfügbaren Informationen für seine Schätzung zur Verfügung. Es berücksichtigt alle Textsegmente und bezieht somit alle Textinformationen in seine Schätzung ein.

Auf diese Weise können die Szenenabfolge für den Beitrag als Ganzes geschätzt und auch die Interviewinhalte mit einbezogen werden. Da diese (zusätzlich berücksichtigten) Textpassagen wichtige Informationen enthalten können, verbessert sich die Qualität der Schätzung. Auch können Interview-Passagen, wie sie für Nachrichtensendungen oder -videos typisch sind, berücksichtigt werden, weshalb spezielle Interview-Token (mittels Modal Type Embedding) im Featurevektor vergeben werden können. Der Featurevektor ist Grundlage für den Video-Vektor, auf dessen Basis der Vorhersagevektor erstellt wird.

In einer bevorzugten Ausführungsform erfolgt das Erstellen der Sequenzliste für die Auswahl von Videosequenzen durch Anfügen eines Ziel-Vektors an die Sequenzliste. Die Sequenzliste wird also durch weitere Ziel-Vektoren erweitert, bis alle Textsegmente des Textes, der dem Videoclip unterlegt sein soll, abgearbeitet sind. Das Erstellen der Sequenzliste erfolgt in dem Zuordnungsmodul.

In besonderer Ausführungsform umfasst das Textelement eines Textes einen einzelnen Satz des vorgegebenen (Gesamt-)Textes. Möglich ist selbstverständlich auch, dass das Textelement aus mehreren Sätzen, bevorzugt bis zu fünf Sätzen, bestehen kann. Dies bietet sich an, wenn die Sätze inhaltlich zusammenhängend sind.

In einer bevorzugten Ausführungsform der Vorrichtung ist das Auswahlmodul dazu ausgebildet, die in der Merkmalsdatenbank hinterlegten Video-Vektoren mit dem Vorhersagevektor zu vergleichen. Bevorzugt werden alle Video-Vektoren der Merkmalsbank mit einem Vorhersagevektor des Auswahlmoduls verglichen. Besonders bevorzugt erfolgt dies einzeln. Zweck des Vergleichs ist es, einen möglichst gut mit dem Vorhersagevektor korrespondierenden Video-Vektor aus der Merkmalsbank zu finden und diesen Video-Vektor in die Sequenzliste zu übernehmen. Bevorzugt kann bei der Beurteilung der Passung oder des

Matchings des Vorhersagevektors mit einem der Video-Vektoren der Merkmalsbank die Güte der Auswahl beurteilt werden. Der aus den Video-Vektoren ausgewählte Vektor ist der Ziel-Vektor, der der Sequenzliste zugeführt wird. Zur Bestimmung der Güte für die geeignete Auswahl kann vorzugsweise eine Gütefunktion verwendet werden. Beispielsweise ist es möglich, die Güte anhand von Kosinus-Ähnlichkeiten zu beurteilen.

In einer bevorzugten Ausführungsform kann die erstellte Sequenzliste, die in dem Zuordnungsmodul erstellt wird, Interview-Vektoren umfassen, die mit Videosequenzen korrespondieren, die eine Interview-Sequenz umfassen oder sind. Auf diese Weise ist es möglich, Interview-Sequenzen in dem zu erstellenden Videoclip zu berücksichtigen. Die Interview-Vektoren können dabei Merkmalsvektoren sein, die auch den Transkript des gesprochenen Dialogs in der Interview-Sequenz umfassen.

Entsprechend der Vorrichtung kann auch das erfindungsgemäße Verfahren in einer bevorzugten Ausführungsform vorsehen, dass das Erzeugen des Vorhersagevektors für die Videosequenz basierend auf mehreren Videovektoren und mehreren Text-Merkmalsvektoren erfolgt.

Bevorzugt wird gemäß einer Ausführungsform des Verfahrens die Erzeugung des Vorhersagevektors basierend auf allen zur Verfügung stehenden Videovektoren bzw. auf den Videovektoren aller zur Verfügung stehenden Videosequenzen und/oder basierend auf allen zur Verfügung stehenden Text-Merkmalsvektoren durchgeführt. Hierbei gilt generell, dass mit steigender Anzahl von verarbeiteten Videovektoren und Textsegmenten die Anzahl der zur Verfügung stehenden Informationen erhöht wird und sich damit die Qualität der nachfolgenden Clips weiter verbessert. Gleiches gilt für die Vorrichtung.

In einer bevorzugten Ausführungsform der Trainingsvorrichtung ist vorgesehen, dass der Featurevektor stilistische Mittel oder stilistische Features umfasst. Diese stilistischen Mittel sind die gleichen, die bei dem Erzeugen des Vorhersagevektors und der Auswahl einer Ziel-Videosequenz verwendet werden.

Eine bevorzugte Ausführungsform der Trainingsvorrichtung sieht vor, dass sie derart ausgebildet ist, dass mehrere Text-Merkmalsvektoren und/oder mehrere Inhaltsvektoren und/oder mehrere Klassifikationsvektoren verwendet werden, um den Featurevektor zu bilden, aus dem ein Videovektor erstellt wird, der die Merkmale einer unbekannten Videosequenz aufweist.

Bevorzugt umfasst das Transformer-Modul ein Neuronales Netzwerk, z.B. ein Deep Neural Network oder ein Feed-Forward-Network (FFN). Bevorzugt kann das Transformer-Modul mehrere Submodule umfassen, die jeweils ein Neuronales Netzwerk oder ein FFN umfassen.

Vorzugsweise kann die Trainingsvorrichtung dazu ausgebildet sein, um die Submodule des Transformer-Moduls zu trainieren und hierbei die stilistischen Mittel und stilistischen Merkmale zu verarbeiten, für die das Submodul spezifisch ausgebildet und trainiert werden soll. Bevorzugt kann von jedem Submodul ein Featurevektor (Sub-Freature-Vektor) gebildet werden.

In einer bevorzugten Ausgestaltung kann die Trainingsvorrichtung zusätzlich oder optional einen VILA-Encoder (Visual Language Encoder) zum Erstellen eines Vektors oder Merkmalsvektors der Szenenkomposition und bevorzugt eines Scores für die Bildqualität umfassen, um die Videosequenz zu beschreiben und charakterisieren.

Eine bevorzugte Ausführungsform der Trainingsvorrichtung kann einen Bewegungs-Encoder aufweisen, um einen Merkmalsvektor der Kamerabewegung für eine Videosequenz zu erstellen. Dieser „Bewegungs-Merkmalsvektor“ kann in den Featurevektor für den Video-Vektor einfließen, beispielsweise durch Summation oder eine andere Vektor-Rechenvorschrift.

In einer bevorzugten Ausführungsform der Trainingsvorrichtung kann sie mehrere Submodule für die zu berücksichtigenden stilistischen Mittel aufweisen, wobei bevorzugt jedes Submodul auf ein stilistisches Mittel spezialisiert sein kann.

Mit anderen Worten wird jedes Submodul oder jedes FFN speziell für eine Modalität oder ein stilistisches Mittel trainiert. Auf diese Weise können mehr stilistische Features als nur die Einstellungsgröße berücksichtigen werden.

Statt nur beispielsweise zwei Modalitäten, nämlich Inhalt und Einstellungsgröße, zu verarbeiten, etwa in dem CLIP-Feature und VGG-Feature aneinandergehängt und dann für eine Reduktion der Größe linear abgebildet werden, erfordert die Verwendung der vielen Informationen eine besondere Berücksichtigung, um einen Informationsverlust durch zu große Reduktion zu vermeiden. Die gesamten Informationen, die in einem Featurevektor zusammengefasst sind, werden in einzelne Sub-Featurevektoren aufgeteilt, die in den mehreren Submodulen ohne oder ohne in der Praxis relevanten Informationsverlust verarbeitet werden können. Nach der Bearbeitung in den Submodulen können die verarbeitenden Sub-Featurevektoren wieder zu einem Featurevektor zusammengeführt werden.

Die in der Vorrichtung, der Trainingsvorrichtung und dem Verfahren verarbeiteten Vektoren sind Merkmalsvektoren und umfassen Merkmale des Textsegments bzw. einer Videosequenz. Der Textvektor ist dabei ein Text-Merkmalsvektor mit Merkmalen eines Textsegments. Der Vorhersagevektor ist ein von der Vorrichtung, insbesondere von dem Transformer-Modul erzeugter Merkmalsvektor mit inhaltlichen Merkmalen wie auch mit Klassifikationsmerkmalen für die nächste auszuwählende Videosequenz. Der Video-Vektor oder Video-Featurevektor umfasst die Merkmale einer Videosequenz, die sowohl inhaltliche Merkmale als auch Klassifikationsmerkmale einschließen können. Der Ziel-Vektor ist der durch den Vergleich des Vorhersagevektors mit Video-Vektoren einer Merkmalsbank gefundene Merkmalsvektor. Dieser wird auch als Ziel-Featurevektor oder Matching-Featurevektor bezeichnet. Eine andere Bezeichnung ist Target-Vektor.

Die Erfindung wird nachfolgend anhand einiger ausgewählter Ausführungsbeispiele im Zusammenhang mit den beiliegenden Zeichnungen näher beschrieben und erläutert. Es zeigen:

Figur 1 die erfindungsgemäße Vorrichtung zum Zusammenstellen eines Videoclips;

Figur 2 den Ablauf und Einsatz der Vorrichtung beim Erstellen eines Videoclips;

Figur 3 ein Transformer-Modul zum Erzeugen eines Video-Vektors;

Figur 4 eine erfindungsgemäße Trainingsvorrichtung mit einem Transformer-Modul; und

Figur 5 eine schematische Darstellung der Verfahrensschritte des Verfahrens zum Zusammenstellen eines Videoclips.

Figur 1 zeigt eine Vorrichtung 10 zum Zusammenstellen eines Videoclips basierend auf mehreren Videosequenzen und Textsegmenten eines Textes. Die Vorrichtung 10 umfasst ein Transformer-Modul 20, um einen Vorhersagevektor 22 für eine Videosequenz basierend auf einem Video-Vektor einer vorhergehenden Videosequenz zu erzeugen. Ein Textsegment 12 wird in einem Text-Modul 30 verarbeitet, in dem aus den Textmerkmalen des Textsegments 12 ein Text-Merkmalsvektor 32 erzeugt wird. Dieser Text-Merkmalsvektor 32 gibt den semantischen Inhalt des Textsegments 12 wieder.

Ein Auswahlmodul 40 verarbeitet den Vorhersagevektor 22, auf dessen Basis das Auswahlmodul 40 die Auswahl eines Ziel-Vektors 42 einer geeigneten Ziel-Videosequenz aus einer Merkmalsbank 50 trifft. In der Merkmalsbank 50 sind Merkmalsvektoren von einzelnen Videosequenzen hinterlegt. Aus dieser Merkmalsbank 50, die eine Datenbank oder eine entsprechende Tabelle sein kann, werden mittels Auswahlmodulen 40 die Merkmalsvektoren ausgewählt, die am besten zu dem Vorhersagevektor 22 passen, der in dem Transformer-Modul 20 erzeugt wurde. Die Auswahl kann anhand eines Gütekriteriums erfolgen, beispielsweise anhand von Kosinus-Ähnlichkeiten, wie sie exemplarisch in Formel 1 beschrieben sind.

Formel 1:

$$cos(v_g, v_p) = \frac{v_g^T v_p}{\|v_p\|_2 * \|v_p\|_2}$$

Dabei ist || x ||$_2$ die L2-Norm eines Vektors x. v_p ist der Vorhersagevektor 22 und v_g der Ziel-Vektor 42. v^T bedeutet hierbei die Transponierte des Vektors v.

In dem Auswahlmodul 40 werden alle in der Merkmalsbank 50 hinterlegten Ziel-Vektoren 42 mit dem Vorhersagevektor 22 verglichen. Der Ziel-Vektor 42, der die größte Kosinus-Ähnlichkeit mit dem Vorhersagevektor 22 aufweist, wird ausgewählt und in einem Zuordnungsmodul 60 in einer Sequenzliste 62 hinterlegt. Die Sequenzliste 62 umfasst alle Ziel-Vektoren 42 von Videosequenzen, die Basis für den zu erstellenden Videoclip sind.

In einem Videoclip-Modul 70 wird ein Videoclip 72 aus Videosequenzen 74 basierend auf der Sequenzliste 62 erzeugt. Die einzelnen Videosequenzen 74 sind die Sequenzen, die durch die jeweiligen Ziel-Vektoren 42 beschrieben und charakterisiert werden. Es entsteht folglich ein Videoclip 72, in dem mehrere ausgewählte Videosequenzen 74 aneinandergereiht sind. Die einzelnen Videosequenzen 74 haben typischerweise eine Länge von 30 Sekunden bis 5 Minuten, vorzugsweise von 10 Sekunden bis 3 Minuten, besonders bevorzugt von 30 Sekunden bis 90 Sekunden.

Da in den Ziel-Vektoren 42 der Sequenzliste 62 die Text-Merkmalsvektoren 32 für die jeweiligen, einzelnen Textsegmente 12 umfasst sind, ergibt sich ein Videoclip 72 in Form eines Nachrichtenclips, der mehrere Videosequenzen 74 sowie die entsprechenden Textpassagen für die Videosequenz 74 umfasst. Auf diese Weise kann ein Sprecher den entsprechenden Text zu dem Videoclip 72 sprechen. Hierbei können entweder die zugrundeliegenden Textsegmente 12 verwendet werden oder einzig an den Textsegmenten 12 orientierter freier Text gesprochen werden. Möglich ist es, dass das zu den Textsegmenten 12 erzeugte Audiosignal mit Hilfe einer künstlichen Intelligenz oder KI-Einheit generiert und den Videosequenzen 74 zugeordnet wird.

Der ermittelte Ziel-Vektor 42 wird nicht nur dem Zuordnungsmodul 60 zugeführt, sondern auch dem Transformer-Modul 20, wo er mit dem Text-Merkmalsvektor 32 des nächsten Textsegments 12 verarbeitet wird. Auf diese Weise wird der Ziel-Vektor 42 als Video-Vektor 44 der vorhergehenden Videosequenz 74 mit bei der Auswahl der nächsten Videosequenz 74, also des Ziel-Vektors 42, berücksichtigt. Die hiermit einhergehende Verarbeitung wird also fortgesetzt, bis alle Videosequenzen 74 und Textsegmente 12 abgearbeitet sind.

Figur 2 zeigt den kaskalierten Ablauf bei der Erzeugung eines Videoclips 72 und das Zusammenstellen des Videoclips 72 aus mehreren Videosequenzen 74, die auf den ausgewählten Ziel-Vektoren 42 einer Sequenzliste 62 basieren. Anstelle der in Figur 1 gewählten Darstellung, bei dem die Verarbeitungsschleife mehrfach durchlaufen wird, sind die einzelnen Verarbeitungsschritte hier sequenziell dargestellt, wobei die Vorrichtung 10 jeweils mehrfach in unterschiedlichen Stufen 16, 17, 18, 19 gezeigt ist.

Zunächst wird die Verarbeitung mit einem Ziel-Vektor 42 gestartet, der einen Best-Shot-Vektor 14 darstellt und vorgegeben wird. Der Best-Shot-Vektor 14 repräsentiert eine vorbestimmte Videosequenz, die als Eingangssequenz für den Videoclip 72 dient. Der Best-Shot-Vektor 14 kann beispielsweise manuell oder von einer separaten KI-Einheit ausgewählt und vorgegeben sein.

In dem Transformer-Modul 20 wird zunächst der Best-Shot-Vektor 14 und der in dem Text-Modul 30 erzeugte Text-Merkmalsvektor 32 auf Basis des ersten Textsegments 12 verarbeitet. Hier wird ein Vorhersagevektor 22 erzeugt, der mit den Ziel-Vektoren 42 der Merkmalsbank 50 in dem Auswahlmodul 40 verarbeitet wird. Dabei wird der Ziel-Vektor 42 ausgewählt, der am besten mit dem Vorhersagevektor 22 “matched“, also die größte Übereinstimmung bzw. die größte Ähnlichkeit aufweist.

Der ausgewählte Ziel-Vektor 42 wird in dem Zuordnungsmodul 60 in die dort vorhandene Sequenzliste 62 eingefügt, wobei dieser Ziel-Vektor 42 der erste Eintrag in der Sequenzliste 62 ist. Der Ziel-Vektor der ersten Stufe 16 wird der

zweiten Stufe 17 zugeführt und hier mit dem neuen Textsegment 12 verarbeitet. Die Vorrichtung 10 wird also in der zweiten Stufe 17 erneut für die Auswahl eines geeigneten Ziel-Vektors 42 verwendet. Dieser wird dann in dem Zuordnungsmodul 60 an die Sequenzliste 62 angefügt, sodass die Sequenzliste 62 mit jeder Stufe und jeder weiteren Videosequenz bzw. jedem weiteren Textsegment 12 wächst.

In einer dritten Stufe 18 wird der in der zweiten Stufe 17 erzeugte Ziel-Vektor 42 verwendet, um einen neuen Ziel-Vektor 42 zu erzeugen. Dieses Prinzip wird bis zu einer n-ten Stufe 19 weitergeführt, bis alle Textsegmente 12 abgearbeitet sind. Die Textsegmente 12 entstammen dem für einen Videoclip 72 zur Verfügung stehenden Text, insbesondere einem Nachrichtentext.

Am Ende des Durchlaufs aller Stufen 16 bis 19 ist ein Videoclip entstanden, der in dem Videoclip-Modul 70 basierend auf der dem Zuordnungsmodul 60 erzeugten Sequenzliste 62 generiert wird.

Figur 2 zeigt, dass in der Merkmalsbank 50 eine optionale Interviewbank 52 umfasst sein kann. In dem zugrundeliegenden Nachrichtentext, der in mehrere Textsegmente 12 unterteilt ist, können auch Interviewsegmente 54 umfasst sein, die zu entsprechenden Interviewsequenzen 55 in der Interviewbank 52 passen. In einem ebenfalls optionalen Interview-Modul 56 werden die Interviewsequenzen 55 der Interviewbank 52 analysiert, beispielsweise durch einen Sprache-zu-Text-Umwandler (Speech to Text), und dann mit den Interviewsegmenten 54 einer Interviewsegmentdatenbank verglichen. Hierbei werden die Interviewsequenzen 55 mit der größten Ähnlichkeit und der höchsten Übereinstimmung zu den untersuchten Interviewsegmenten 54 ausgewählt und in Form eines Interview-Merkmalsvektors 58 in die Sequenzliste 62 in dem Zuordnungsmodul 60 hinzugefügt. In dem Videoclip-Modul 70 wird die entsprechende Interviewsequenz 55 der Interviewbank 52 ausgewählt und in den Videoclip 72 eingefügt. Dies ist optional und stellt eine bevorzugte Erweiterung des Videoclips 72 dar.

Die effektive Zusammenstellung eines Videoclips 72 oder Nachrichtenclips mit mehreren Videosegmenten basiert also auf vortrainierten Repräsentationen der einzelnen Videosequenzen durch Merkmalsvektoren, sogenannten Video-Vektoren. Basierend auf den codierten Textsegmenten 12 in Form von Text-Merkmalsvektoren 32 und dem codierten Quellmaterial, also den Video-Vektoren bzw. Ziel- Vektoren 42 für einzelne Videosequenzen 74, ergibt sich der in Figur 2 beschriebene Arbeitsablauf einer sogenannten Inferenzpipeline. In einer bevorzugten Ausführungsform können dabei zuvor markierte Interviewabschnitte im Nachrichtenartikel und die entsprechenden Interviewaufnahmen bzw. Interviewsequenzen 55 herausgefiltert werden, die beispielsweise durch einen vortrainierten Interview-Klassifikator klassifiziert wurden. Mit einer indexbasierten Textsuche werden die einzelnen Interview-in-points und Interview-out-points in die Sequenzliste 62 mit den Ziel-Vektoren 42 eingefügt. Zu Beginn wird ein Best-Shot-Vektor 14 erzeugt, der am besten die Best-Shot-Anfangsszene oder Videosequenz repräsentiert, die in der Gesamtheit der Videoclips 72 vorhanden ist. Diese Auswahl erfolgt nach dem Motto „beginne mit der besten Szene“, wähle also die beste Videosequenz als Startpunkt aus.

Die folgenden Videosequenzen werden mit den aktuellen Textsegmenten 12 und den Video-Vektoren der vorhergehenden Szene bzw. den Ziel-Vektoren 42 ausgewählt. Hierzu wird ein Vorhersagevektor 22 für die nächste Videosequenz erzeugt. Der erzeugte Vorhersagevektor 22 wird dann mit allen vorhandenen Video-Vektoren der Merkmalsbank 50, die jeweils die Ziel-Vektoren 42 darstellen, verglichen. Dabei wird der Ziel-Vektor 42 ausgewählt, der dem Vorhersagevektor am ähnlichsten ist. Die Topkandidaten, also die Ziel-Vektoren 42, die dem Vorhersagevektor am Ähnlichsten sind, werden in die Sequenzliste 62 eingefügt, auf deren Basis die Videosequenzen ausgewählt und zu dem Videoclip 72 zusammengefügt werden.

Figur 3 zeigt das Transformer-Modul 20 im Detail. Das Transformer-Modul 20 umfasst einen Cross-Modal-Encoder 24, das einen Ziel-Vektor 42 bzw. einen Video-Vektor 44 und einen Text-Merkmalsvektor 32 verarbeitet bzw. zusammengefügt, sodass ein Featurevektor 26 gebildet wird. Dieser umfasst sowohl

den Text-Merkmalsvektor 32 als auch den Video-Vektor 44. Der Featurevektor 26 hat die doppelte Länge des Text-Merkmalsvektors 32 bzw. des Video-Vektors 44. In einem Feature-Decoder 28 des Transformer-Moduls 20 wird der Featurevektor 26 weiterverarbeitet und ein Vorhersagevektor 22 erzeugt. Dieser umfasst die Merkmale für die nächste Videosequenz, die in dem Videoclip, der zu erzeugen ist, umfasst sein soll. Der Vorhersagevektor 22 hat die gleiche Länge wie der Ziel-Vektor 42 bzw. der Video-Vektor 44. Er ist damit bevorzugt halb so groß wie der Featurevektor 26.

Figur 4 zeigt die erfindungsgemäße Trainingsvorrichtung 80 mit dem Transformer-Modul 20. Die Trainingsvorrichtung 80 wird dazu eingesetzt, um aus mehreren Videosequenzen 74 und mehreren Textsegmenten 12 Video-Vektoren 44 zu erzeugen, die geeignet sind, um die Videosequenzen 74 bestmöglich zu beschreiben und eine Auswahl von Videosequenzen 74 auf Grundlage der in den Video-Vektoren 44 enthaltenen Merkmalen zu erlauben.

Die Textsegmente 12 werden in dem Text-Modul 30 verarbeitet. Sie sind Teile eines Gesamttextes, beispielsweise eines Nachrichtentextes oder Nachrichtenartikels. Vorzugsweise umfasst jedes Textsegment 12 einen Hauptsatz des Nachrichtenartikels. Der Nachrichtentext wird üblicherweise derart in einzelne Textsegmente 12 unterteilt, um der generellen Videobearbeitungsregel „eine Aufnahme pro Hauptsatz“ entsprechen zu können. Das Text-Modul 30 umfasst bevorzugt einen CLIP-Encoder 82, der verwendet wird, um den inhaltlichen Teil des Textes in einen Merkmalsvektor, hier den Text-Merkmalsvektor 32, umzuwandeln.

Die Videosequenzen 74 werden mit einem (anderen) CLIP-Encoder 82 kategorisiert, wobei die inhaltlichen Merkmale des Videos extrahiert werden. Hierauf basierend wird ein Inhalts-Vektor 84 mit Videoinhaltsmerkmalen für die jeweilige Videosequenz 74 erzeugt. Zudem werden die Videosequenzen 74 mittels eines VGG-Encoders 86 bearbeitet, um einen Klassifikationsvektor 88 zu erzeugen. Der Klassifikationsvektor 88 umfasst Videoklassifikationsmerkmale, die bei-

spielsweise Einstellungsgrößen der Videosequenz 74, entsprechende Ansichten, die in der Videosequenz 74 umfasst sind, Kamerawinkel, Bewegungen innerhalb des Videos oder Bewegungen der Kamera sowie Bewegungsrichtungen von Personen oder Gegenständen innerhalb des Videos bzw. der Videosequenz 74 und/oder der Kamera, Zoomeinstellungen oder ähnliche Einstellungen umfassen kann. Der Klassifikationsvektor 88 repräsentiert somit den stilistischen Teil der Videosequenz 74.

Inhalts-Vektor 84 und Klassifikationsvektor 88 werden zu dem Video-Vektor 44 zusammengefügt, der sowohl die semantische Essenz einer Videosequenz 74 wie auch deren stilistische Merkmale umfasst.

Featurevektor 26 und Text-Merkmalsvektor 32 werden in dem Cross-Modal-Encoder 24 weiterverarbeitet, so dass der Featurevektor 26 gebildet wird. Die Weiterverarbeitung des Featurevektors 26 erfolgt in dem Feature-Decoder 28, sodass ein Vorhersagevektor 22 erzeugt wird.

Somit gelingt es also unter der Verwendung von Text-Bild-Paaren bzw. Text-Videosequenz-Paaren das Transformer-Modul 20 bzw. die Trainingsvorrichtung 80 so zu trainieren, dass genau die Aufnahmemerkmale vorhergesagt werden können, die mit bestimmten Textsegmenten 12 von Nachrichtenartikeln korrelieren. Dies erfolgt auf Grundlage der zur Verfügung stehenden Videosegmente. Die Verwendung des Cross-Modal-Encoders 24 und des Feature-Decoders 28 erleichtern den Prozess der Zuordnung.

Die verwendeten Videosequenzen 74 sowohl für die Trainingsvorrichtung 80 als auch für die erfindungsgemäße Vorrichtung nach den Figuren 1 bis 3 können sowohl aus aktuellen Nachrichtenclips auf Grundlage eines Nachrichtenartikels als auch auf Archivmaterial bzw. Archivvideosequenzen bestehen. Dabei hat jede Videosequenz 74 bevorzugt einen entsprechenden Nachrichtenartikel, der für die Vertonung vorgesehen ist. Zudem existiert häufig eine Liste mit manuellen redaktionellen Schnitten.

Der auf dem Featurevektor 26 beruhende Video-Vektor 44 bzw. Vorhersagevektor 22, der im Ausgang der Trainingsvorrichtung 80 zur Verfügung steht, wird einem Evaluierungsmodul 90 zugeführt, um die Güte des erzeugten Vorhersagevektors 22 bzw. Video-Vektors 44 zu beurteilen. Für die Erzeugung von Videoclips basierend auf Videosegmenten und die textbasierte Zusammenstellung der Videoclips ist das Auffinden von Videosegmenten, die inhaltsrelevant und von angemessener Größe sind, wichtig. Dazu wird das Transformer-Modul 20 der Trainingsvorrichtung 80 zur Feinabstimmung des Featurevektors 26 unter Verwendung der Textmerkmale aus dem Textsegment 12 und der Merkmale der vorhergehenden Videosequenz 74 trainiert. Die Trainingspaare umfassen also einen Merkmalsvektor für die vorangegangene Aufnahme, einen Merkmalsvektor der Zielaufnahme und einen Text des Textsegments. Die verwendeten CLIP-Encoder 82 und VGG-Encoder 86 erzeugen Inhaltsmerkmale und Größenmerkmale für das Video, während die Textmerkmale ebenfalls mit einem CLIP-Encoder 82 abgeleitet werden. Die hierbei erzeugten codierten Sequenzen werden durch einen Stapel von Transformer-Codierungsschichten verarbeitet, wobei jede Schicht eine konsistente Architektur aufweist.

Die in dem Transformer-Modul 20 erzeugten Video-Vektoren 44 werden dann mittels eines Gütekriteriums in dem Evaluierungsmodul 90 bewertet, wobei bevorzugt in einem Ähnlichkeitsmodul 92 die Kosinus-Ähnlichkeiten zwischen dem erzeugten Video-Vektor 44 und Video-Vektoren 44 einer Datenbank, die dem Textsegment entsprechen, beurteilt werden. Hierbei wird gemäß Formel 2 jeder Video-Vektor 44 einer Videosequenz 74 mit anderen Videosequenzen 74 und deren Merkmalsvektoren als Negativproben verglichen. Dabei wird gemäß Formel 2 ein kontrastives Lernen durchgeführt, wobei v_p den Vorhersagevektor 22 widerspiegelt und v_j Video-Vektoren 44 als Vergleichsprobe darstellt. τ ist ein Temperaturparameter.

$$\mathcal{L}_c = -log \frac{\exp\left(cos(v_g, v_p)/\tau\right))}{\exp\left(cos(v_g, v_p)/\tau\right)) + \sum \exp\left(cos(v_j, v_p)/\tau\right))} \qquad \text{Formel 2}$$

In einem Entropiemodul 94 des Evaluierungsmoduls 90 werden die Kreuzentropieverluste als Trainingsziel zur Gütebewertung herangezogen. Ausgehend von dem multimodalen Transformer-Modul 20 und den erzeugten Vorhersagevektoren 22 werden in dem Entropiemodul 94 drei 1x3 CONV-Schichten verwendet, gefolgt von einer ReLU-Aktivierung. Dieses Entropiemodul 94 wird trainiert, um die stilistische Einstellungsgröße der Szenen zu klassifizieren.

Schließlich werden sigmoidale Aktivierungen hinzugefügt, um die Vorhersagevektoren 22 zu verbessern. Die Kreuzentropieverluste lassen sich mit

$$L_S = -\lambda\,(v * \log v_p + (1-v) * \log(1-v_p)) \qquad \text{Formel 3}$$

mit v_p = Vorhersagevektor 22; v = Ziel-Vektor 42 bzw. Video-Vektor 44

berechnen. Zu diesem Zweck wird als gesamtes Trainingsziel die Kombination der einzelnen Kreuzentropieverluste über alle Videosequenzen 74 im Trainingssatz gesetzt. Diese Entropieverluste lassen sich durch

$$L = \frac{1}{N}\sum_{i=0}^{N}(\alpha * L_C + \beta * L_S) \qquad \text{Formel 4}$$

beschreiben, wobei L = Gesamtverlust, L_s = binäre Kreuzentropieverlust (aus Formel 3) des Vorhersagevektors 22 und L_c = kontrastives Lernen (aus Formel 2). α und β sind Hyperparameter und kontrollieren die Robustheit des Gesamtverlustes; N ist die Anzahl der einzelnen Trainingssätze in dem Gesamt-Trainingssatz.

Der Trainingssatz umfasst die Videosequenzen 74 und die Textsegmente 12, die am Eingang der Trainingsvorrichtung 80 und die daraus erzeugten Merkmale am Eingang des Transformer-Moduls 20 anliegen.

Figur 5 zeigt den prinzipiellen Ablauf des Verfahrens zum Zusammenstellen eines Videoclips basierend auf mehreren Videosequenzen und Textsegmenten eines Textes, insbesondere eines Nachrichtentextes.

In einem ersten Schritt S10 wird ein Text-Merkmalsvektor mit semantischen Textmerkmalen des Textsegments erzeugt. In einem weiteren Schritt S12 erfolgt ein Erzeugen eines Vorhersagevektors für eine Videosequenz basierend auf einem Video-Vektor einer vorhergehenden Videosequenz und basierend auf dem Text-Merkmalsvektor des aktuellen Textsegments.

Ein Schritt S14 betrifft das Auswählen eines Ziel-Vektors einer Zielvideosequenz aus einer Merkmalsbank basierend auf dem Vorhersagevektor. In diesem Schritt wird also das aktuelle Textsegment mit der vorhergehenden Videosequenz verknüpft, um daraus einen Merkmalsvektor für die nächste Videosequenz zu erzeugen.

Basierend auf diesem Vorhersagevektor wird in dem Schritt S14 der Ziel-Vektor für die nächste, also die nachfolgende Videosequenz, auch Zielvideosequenz (Target Video Sequenz) genannt, erzeugt.

Ein Schritt S16 betrifft das Erstellen einer Sequenzliste mit Ziel-Vektoren für die Auswahl von auf den Ziel-Vektoren basierenden Videosequenzen. Dies erfolgt bevorzugt durch Anfügen eines neuen Ziel-Vektors an die Sequenzliste bzw. die in der Sequenzliste bereits vorhandenen Ziel-Vektoren.

In einem Schritt S18 werden basierend auf der Sequenzliste Videosequenzen aus einer Videodatenbank ausgewählt, um einen Videoclip zu erzeugen. Die Videodatenbank kann eine Datenbank sein, die mehrere Videosequenzen von aktuellen Filmen oder Videos, aber auch Archivmaterial umfassen kann, was bevorzugt ebenfalls aus kürzeren Videosequenzen oder kurzen Videoclips bestehen kann. Alternativ kann die Videodatenbank auch auf einem externen Server oder auf mehreren Speichermedien ausgebildet sein, beispielsweise durch einzelne Videoclips, die auf den Speichermedien oder an verschiedenen Orten gespeichert sein können.

In einer bevorzugten Ausführungsform werden die Schritte S10 bis S16 mehrfach durchlaufen, nämlich für jedes neue Textsegment erneut, um auf diese Weise eine Sequenzliste mit mehreren Ziel-Vektoren zu erzeugen.

In einer bevorzugten Ausführungsform kann in einem Schritt S20 ein Interview-Featurevektor oder Interview-Merkmalsvektor erzeugt werden. Dieser wird in die Sequenzliste eingefügt. Der Interview-Merkmalsvektor korrespondiert mit einer Interviewsequenz oder mit einer eine Interviewsequenz umfassenden Videosequenz.

Stilistische Merkmale im Sinne dieser Erfindung sind beispielsweise Merkmale, die die gesamtheitliche inhaltliche Übereinstimmung von Text bzw. Textsegmenten und Szenen bzw. Videosequenzen oder Bildern umfassen. Dazu können die Inhalte sowohl des Textes oder der Textsequenzen wie auch der Szenen, Bilder oder Videosequenzen in Form von charakteristischen Größen oder codierten Ziffern vorgenommen werden, beispielsweise in Form von Zahlenwerten oder in Form von Farben oder Farbabstimmungen.

Eine Übereinstimmung (von codierten Features) kann dann festgestellt werden, wenn sowohl die Codierung des Textes als auch der Szenen und Videosequenzen gleich oder größtenteils gleich sind oder die sich aus den jeweiligen Textsegmenten und Videoszenen ergebenden Merkmalsvektoren eine große Übereinstimmung an einzelnen Elementen des Vektors aufweisen.

Weitere stilistische Mittel im Sinne der Erfindung betreffen die Abfolge von Einstellungsgrößen einer Videoszene oder Videosequenz. Dies sind beispielsweise die Totale, Medium-Darstellung, Porträtdarstellung oder Detaildarstellung. Das stilistische Mittel in diesem Bezug gibt eine Auskunft darüber, wie die entsprechenden stilistischen Größen oder Merkmale geändert werden. Es hat sich als vorteilhaft erwiesen, wenn eine Videosequenz eine Totale zeigt, die darauf folgende Videofrequenz eine Medium-Ansicht oder Porträtansicht und die anschließende Videosequenz eine Detaildarstellung umfasst.

Ein weiteres stilistisches Feature für die Videoklassifizierung oder das Video-Klassifizierungsmerkmal umfasst die Kamerabewegung. Hierbei kann berücksichtigt werden, dass die Kamerabewegung relativ ruhig ist und nicht die gleiche Bewegung der Kamera mehrfach hintereinander in verschiedenen Sequenzen

auftritt. Beispielsweise hat es sich im Rahmen der Erfindung als nachteilig erwiesen, wenn die Kamera mehrmals hintereinander eine Schwenkbewegung in die gleiche Richtung durchführt. Dies ruft bei dem Betrachter den Eindruck hervor, dass er sich um die eigene Achse dreht, was insbesondere bei Nachrichtensendungen und Nachrichtenvideos nicht gewollt und nicht gewünscht ist.

Ein weiteres stilistisches Feature oder stilistisches Merkmal als Video-Klassifizierungsmerkmal ist die Szenenqualität. Hierzu gehören Merkmale wie Bildschärfe, Belichtung oder Kamerawackler oder nicht vorhersehbare Kamerabewegungen.

Zu den weiteren bevorzugt zu berücksichtigenden stilistischen Merkmalen gehören auch Szenenkompositionen. Diese Merkmale bringen zum Ausdruck, ob das beschriebene Subjekt oder Objekt, das in den entsprechenden Textsegmenten umfasst ist, sich auch in der Videosequenz wiederfindet. Insbesondere kann hierbei berücksichtigt und zum Ausdruck gebracht werden, wo das Objekt angeordnet ist oder ob das Objekt oder Subjekt sich in der Bildmitte und zentral befindet oder an einer Randposition innerhalb des Bildes.

Die Berücksichtigung von stilistischen Merkmalen oder Features in den Video-Klassifizierungsmerkmalen, die bei der Erstellung des Vorhersagevektors berücksichtigt werden, führt dazu, dass die aufeinanderfolgenden Videosequenzen und die entsprechenden Textsegmente miteinander matchen und in dem Sinne gut zusammenpassen, dass insbesondere für Nachrichtenvideos und Informationsvideos ein ruhiges und informatives Bild und Szenario entsteht, das zum einen zu dem hinterlegten Text und den einzelnen Textsegmenten passt und korrespondiert und zum anderen bei dem Betrachter eine Atmosphäre hervorruft, in der er den durch die Bilder und den Text generierten Informationsgehalt bestmöglich aufnehmen kann.

Die stilistischen Mittel umfassen die inhaltliche Übereinstimmung von Text und Szenen, die Abfolge der Einstellungsgrößen (Totale, Medium-Ansicht, Detailansicht), die Abfolge der Kamerabewegungen (Stillstand, Schwenk oder Zoom), die Szenenkomposition (Position des Subjekts) und die Bildqualität.

Die Erfindung wurde anhand der Zeichnungen und der Beschreibung umfassend beschrieben und erklärt. Die Beschreibung und Erklärung sind als Beispiel und nicht einschränkend zu verstehen. Die Erfindung ist nicht auf die offenbarten Ausführungsformen beschränkt. Andere Ausführungsformen oder Variationen ergeben sich für den Fachmann bei der Verwendung der vorliegenden Erfindung sowie bei einer genauen Analyse der Zeichnungen, der Offenbarung und der nachfolgenden Patentansprüche.

In den Patentansprüchen schließen die Wörter „umfassen“ und „mit“ nicht das Vorhandensein weiterer Elemente oder Schritte aus. Der undefinierte Artikel „ein“ oder „eine“ schließt nicht das Vorhandensein einer Mehrzahl aus. Ein einzelnes Element oder eine einzelne Einheit kann die Funktionen mehrerer der in den Patentansprüchen genannten Einheiten ausführen. Ein Element, eine Einheit, eine Vorrichtung und ein System können teilweise oder vollständig in Hard- und/oder in Software umgesetzt sein. Die bloße Nennung einiger Maßnahmen in mehreren verschiedenen abhängigen Patentansprüchen ist nicht dahingehend zu verstehen, dass eine Kombination dieser Maßnahmen nicht ebenfalls vorteilhaft verwendet werden kann. Ein Computerprogramm kann auf einem nichtflüchtigen Datenträger gespeichert/vertrieben werden, beispielsweise auf einem optischen Speicher oder auf einem Halbleiterlaufwerk (SSD). Ein Computerprogramm kann zusammen mit Hardware und/oder als Teil einer Hardware vertrieben werden, beispielsweise mittels des Internets oder mittels drahtgebundener oder drahtloser Kommunikationssysteme. Bezugszeichen in den Patentansprüchen sind nicht einschränkend zu verstehen.

Patentansprüche

1. Vorrichtung zum Zusammenstellen eines Videoclips (72) basierend auf mehreren Videosequenzen (74) und Textsegmenten (12) eines Textes; umfassend

ein Text-Modul (30) zum Erzeugen eines Text-Merkmalsvektors (32) mit Textmerkmalen eines Textsegments (12);

ein Transformer-Modul (20) zum Erzeugen eines Vorhersagevektors (22) für eine Videosequenz (74);

ein Auswahl-Modul (40) zum Auswählen eines Ziel-Vektors (42) einer geeigneten Ziel-Videosequenz aus einer Merkmalsbank (50) basierend auf dem Vorhersagevektor (22);

ein Zuordnungs-Modul (60) zum Erstellen einer Sequenzliste (62) mit Ziel-Vektoren (42) für die Auswahl von Videosequenzen (74) basierend auf den Ziel-Vektoren (42);

ein Videoclip-Modul (70) zum Erstellen eines Videoclips (72) aus Videosequenzen (74) einer Video-Datenbank basierend auf der Sequenzliste (62);

wobei

das Transformer-Modul (20) einen Cross-Modal-Encoder (24) und einen Feature-Decoder (28) umfasst; und

das Transformer-Modul (20) dazu ausgebildet ist, um basierend auf einem Video-Vektor (44) der vorhergehenden Videosequenz (74) und auf einem Text-Merkmalsvektor (32) den Vorhersagevektor (22) zu erstellen;

die Video-Vektoren (44) Video-Klassifizierungsmerkmale umfassen, die

stilistische Feature umfassen.

2. Vorrichtung nach Anspruch 1, dadurch gekennzeichnet, dass der Video-Vektor (44) und der Vorhersagevektor (42) Textmerkmale, Video-Inhaltsmerkmale und Video-Klassifizierungsmerkmale umfassen.

3. Vorrichtung nach dem vorhergehenden Anspruch, dadurch gekennzeichnet, dass die Video-Klassifizierungsmerkmale mittels eines VGG-Encoders (86) erzeugt sind, wobei der VGG-Encoder (86) bevorzugt in dem Transformer-Modul (20) integriert ist.

4. Vorrichtung nach Anspruch 2 oder 3, dadurch gekennzeichnet, dass die Video-Klassifizierungsmerkmale Merkmale der Videosequenz (74) umfassen, wobei die Merkmale Elemente einer Liste betreffend Einstellungsgrößen der Videosequenz (74) wie Totale, Detailansicht, Portraitansicht oder Mediumansicht, verwendeter Kamerawinkel der Videosequenz (74), Bewegung der Kamera in der Videosequenz (74), Bewegungsrichtung der Kamera, Bewegung von Personen in der Videosequenz (74), Bewegungsrichtung von Personen in der Videosequenz (74), Zoomeinstellungen und ähnliche Einstellungen der Videosequenz (74) sind.

5. Vorrichtung nach Anspruch 2, dadurch gekennzeichnet, dass die Video-Inhaltsmerkmale mittels eines CLIP-Encoders (82) erzeugt sind, wobei der CLIP-Encoder (82) bevorzugt in dem Transformer-Modul (20) integriert ist.

6. Vorrichtung nach einem der vorhergehenden Ansprüche, dadurch gekennzeichnet, dass das Transformer-Modul (20) eine Mehrzahl von Submodulen umfasst, wobei jedes Submodul ein neuronales Netzwerk aufweist und dazu ausgebildet ist, ein vorgegebenes Video-Klassifizierungsmerkmal eines Video-Vektors (44) zu analysieren und auszuwerten, um das Video-Klassifizierungsmerkmal bei dem Erstellen des Vorhersagevektors (22) zu berücksichtigen, wobei das Submodul bevorzugt ein

Feed-Forward-Network ist und wobei bevorzugt das Video-Klassifizierungsmerkmal ein stilistisches Feature umfasst.

7. Vorrichtung nach einem der vorhergehenden Ansprüche, dadurch gekennzeichnet, dass das Transformer-Modul (20) dazu ausgebildet ist, um den Vorhersagevektor (22) basierend auf mehreren Video-Vektoren (44) und mehreren Text-Merkmalsvektoren (32) zu erstellen, bevorzugt basierend auf den Video-Vektoren (44) aller zur Verfügung stehenden Videosequenzen (74) und/oder auf allen zur Verfügung stehenden Text-Merkmalsvektoren (32) zu erstellen.

8. Vorrichtung nach einem der vorhergehenden Ansprüche, dadurch gekennzeichnet, dass das Text-Modul (30) einen CLIP-Encoder (82) umfasst, wobei der Text-Merkmalsvektor (32) mittels des CLIP-Encoders (82) erzeugt wird.

9. Vorrichtung nach einem der vorhergehenden Ansprüche, dadurch gekennzeichnet, dass das Erstellen der Sequenzliste (62) für die Auswahl von Videosequenzen (74) durch Anfügen eines Ziel-Vektors (42) an die Sequenzliste (62) erfolgt.

10. Vorrichtung nach einem der vorhergehenden Ansprüche, dadurch gekennzeichnet, dass das Textsegment (12) einen Satz des vorgegebenen Textes umfasst.

11. Vorrichtung nach einem der vorhergehenden Ansprüche, dadurch gekennzeichnet, dass das Auswahl-Modul (40) dazu ausgebildet ist, um die in der Merkmalsbank (50) hinterlegten Video-Vektoren (44) mit dem Vorhersagevektor (22) zu vergleichen, bevorzugt jeweils einzeln, und besonders bevorzugt die Güte der Auswahl des Ziel-Vektors (42) beurteilt wird, sehr bevorzugt mittels einer Gütefunktion.

12. Vorrichtung nach einem der vorhergehenden Ansprüche, dadurch gekennzeichnet, dass in der Sequenzliste (62) Interview-Merkmalsvektoren

(58) umfasst sind, die mit Videosequenzen (74) korrespondieren, die eine Interviewsequenz (55) umfassen.

13. Trainingsvorrichtung zum Trainieren eines Transformer-Moduls (20) und zum Erstellen von Video-Vektoren (44) für Videosequenzen (74) basierend auf Textsegmenten (12) eines Textes oder Nachrichtenartikels; umfassend

einen CLIP-Encoder (82) zum Erstellen von Text-Merkmalsvektoren (32) mit Textmerkmalen eines Textsegments (12);

einen CLIP-Encoder (82) zum Erstellen von Inhalts-Vektoren (84) mit Video-Inhaltsmerkmalen für eine Videosequenz (74);

einen VGG-Encoder (86) zum Erstellen von Klassifikationsvektoren (88) mit Video-Klassifizierungsmerkmalen für eine Videosequenz (74);

das Transformer-Modul (20) mit einem Cross-Modal-Encoder (24) und einem Feature-Decoder (28),

wobei

der Cross-Modal-Encoder (24) dazu ausgebildet ist, um einen Text-Merkmalsvektor (32) und einen Inhalts-Vektor (84) sowie einen Klassifikationsvektor (88) für eine Videosequenz (74) zu einem Featurevektor (26) zusammenzuführen und dem Feature-Decoder (28) zuzuführen, wobei der Featurevektor (26) bevorzugt eine größere Länge als der Inhalts-Vektor (84) hat, sehr bevorzugt die doppelte Länge;

der Featurevektor (26) stilistische Merkmale oder Feature umfasst;

der Feature-Decoder (28) dazu ausgebildet ist, um aus dem Featurevektor (26) einen Video-Vektor (44) zu erstellen, der die Merkmale einer am Eingang der Trainingsvorrichtung (80) anliegenden Videosequenz (74) aufweist und bevorzugt die halbe Länge des Featurevektors (26) hat;

wobei die Trainingsvorrichtung (80) den Video-Vektor (44) an einem Ausgang ausgibt und für eine Weiterverarbeitung zur Verfügung stellt.

14. Trainingsvorrichtung nach dem vorhergehenden Anspruch, dadurch gekennzeichnet, dass die Trainingsvorrichtung dazu ausgebildet ist, um aus mehreren Text-Merkmalsvektoren (32) und/oder mehreren Inhalts-Vektoren (84) und/oder mehreren Klassifikationsvektoren (88) den Featurevektor (26) zu bilden.

15. Verfahren zum Zusammenstellen eines Videoclips (72) basierend auf mehreren Videosequenzen (74) und Textsegmenten (12) eines Textes; umfassend die folgenden Schritte:

a) Erzeugen eines Text-Merkmalsvektors (32) mit Textmerkmalen eines Textsegments (12);

b) Erzeugen eines Vorhersagevektors (22) für eine Videosequenz (74) basierend auf einem Video-Vektor (44) einer vorhergehenden Videosequenz (74) und basierend auf dem Text-Merkmalsvektor (32) des Textsegments (12);

c) Auswählen eines Ziel-Vektors (42) einer Ziel-Videosequenz aus einer Merkmalsbank (50) basierend auf dem Vorhersagevektor (22);

d) Erstellen einer Sequenzliste (62) mit Ziel-Vektoren (42) für die Auswahl von auf den Ziel-Vektoren (42) basierenden Videosequenzen (74) durch Anfügen eines Ziel-Vektors (42) an die Sequenzliste (62);

e) Erstellen eines Videoclips (72) aus Videosequenzen (74) einer Video-Datenbank basierend auf der Sequenzliste (62).

16. Verfahren nach dem vorhergehenden Anspruch, dadurch gekennzeichnet, dass die Schritte a) bis d) mehrfach ausgeführt werden, um die Sequenzliste (62) zu erzeugen.

17. Verfahren nach Anspruch 15 oder 16, dadurch gekennzeichnet, dass ein Interview-Merkmalsvektor (58) in die Sequenzliste (62) angefügt wird, der mit einer eine Interviewsequenz (55) umfassenden Videosequenz (74) korrespondiert.

18. Verfahren nach einem der Ansprüche 15 bis 17, dadurch gekennzeichnet, dass das Erzeugen des Vorhersagevektors (22) für die Videosequenz (74) basierend auf mehreren Video-Vektoren (44) und mehreren Text-Merkmalsvektoren (32) erfolgt, bevorzugt basierend auf den Video-Vektoren (44) aller zur Verfügung stehenden Videosequenzen (74) und/oder auf allen zur Verfügung stehenden Text-Merkmalsvektoren (32).

19. Computerprogrammprodukt umfassend Befehle, die bei der Ausführung des Programms durch einen Computer diesen veranlassen, die Schritte des Verfahrens nach Anspruch 11 auszuführen.