(12) NACH DEM VERTRAG ÜBER DIE INTERNATIONALE ZUSAMMENARBEIT AUF DEM GEBIET DES PATENTWESENS (PCT) VERÖFFENTLICHTE INTERNATIONALE ANMELDUNG

(19) **Weltorganisation für geistiges Eigentum**
Internationales Büro

(43) **Internationales Veröffentlichungsdatum**
21. August 2025 (21.08.2025)

WIPO | PCT

(10) **Internationale Veröffentlichungsnummer**
WO 2025/172294 A1

(51) **Internationale Patentklassifikation:**
H04N 21/8549 (2011.01) *G06N 3/08* (2023.01)
H04N 21/234 (2011.01) *G06V 10/82* (2022.01)

(21) **Internationales Aktenzeichen:** PCT/EP2025/053593

(22) **Internationales Anmeldedatum:**
11. Februar 2025 (11.02.2025)

(25) **Einreichungssprache:** Deutsch

(26) **Veröffentlichungssprache:** Deutsch

(30) **Angaben zur Priorität:**
10 2024 103 869.1
12. Februar 2024 (12.02.2024) DE
10 2024 114 279.0
22. Mai 2024 (22.05.2024) DE

(71) **Anmelder: HOCHSCHULE RHEINMAIN** [DE/DE]; Kurt-Schumacher-Ring 18, 65197 Wiesbaden (DE).

(72) **Erfinder: NARROSCHKE, Matthias**; Unter den Eichen 5, 65195 Wiesbaden (DE). **QUANDT, Dennis**; Unter den Eichen 5, 65195 Wiesbaden (DE). **RUPPEL, Wolfgang**; Unter den Eichen 5, 65195 Wiesbaden (DE). **ALTMEYER, Philipp**; Unter den Eichen 5, 65195 Wiesbaden (DE).

(74) **Anwalt: DURM PATENTANWÄLTE**; Moltkestr. 45, 76133 Karlsruhe (DE).

(81) **Bestimmungsstaaten** *(soweit nicht anders angegeben, für jede verfügbare nationale Schutzrechtsart)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU,

WO 2025/172294 A1

(54) Title: APPARATUS AND METHOD FOR COMPILING A VIDEO CLIP ON THE BASIS OF MULTIPLE VIDEO SEQUENCES AND TEXT SEGMENTS

(54) Bezeichnung: VORRICHTUNG UND VERFAHREN ZUM ZUSAMMENSTELLEN EINES VIDEOCLIPS BASIEREND AUF MEHREREN VIDEOSEQUENZEN UND TEXTSEGMENTEN

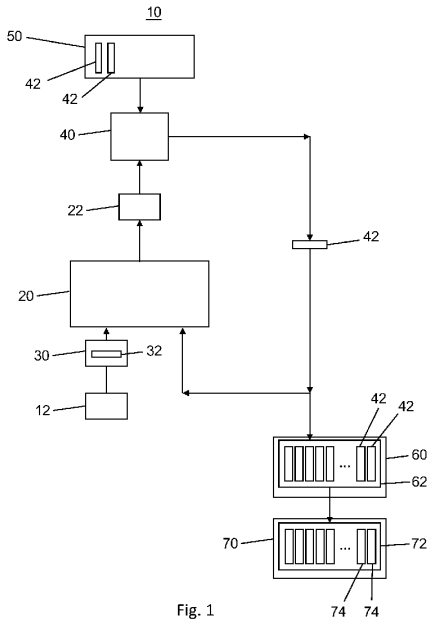


Fig. 1

(57) Abstract: An apparatus (10) for compiling a video clip (72) on the basis of multiple video sequences (74) and text segments (12) of a text comprises a text module (30) for generating a text feature vector (32) with text features of a text segment (12); a transformer module (20) for generating a prediction vector (22) for a video sequence (74) on the basis of the video vector (44) of the preceding video sequence (74); a selection module (40) for selecting a target vector (42) of a suitable target video sequence from a feature bank (50) on the basis of the prediction vector (22); an assignment module (60) for creating a sequence list (62) with target vectors (42) for selecting video sequences (74) on the basis of the target vectors (42); and a video clip module (70) for creating a video clip (72) from video sequences (74) of a video database on the basis of the sequence list (62). The transformer module (20) comprises a cross-modal encoder (24) and a feature decoder (28) and is designed to create the prediction vector (22) on the basis of a video vector (44) and a text feature vector (32).

(57) Zusammenfassung: Eine Vorrichtung (10) zum Zusammenstellen eines Videoclips (72) basierend auf mehreren Videosequenzen (74) und Textsegmenten (12) eines Textes umfasst 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) basierend auf dem Video-Vektor (44) der vorhergehenden 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); und ein Videoclip-Modul (70) zum Erstellen eines Videoclips (72) aus

[Fortsetzung auf der nächsten Seite]

LY, MA, MD, MG, MK, MN, MU, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UY, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) **Bestimmungsstaaten** *(soweit nicht anders angegeben, für jede verfügbare regionale Schutzrechtsart)*: ARIPO (BW, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, SD, SL, ST, SZ, TZ, UG, ZM, ZW), eurasisches (AM, AZ, BY, KG, KZ, RU, TJ, TM), europäisches (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, ME, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Veröffentlicht:

— *mit internationalem Recherchenbericht (Artikel 21 Absatz 3)*

— *in Schwarz-Weiss; die internationale Anmeldung enthielt in ihrer eingereichten Fassung Farbe oder Graustufen und kann von PATENTSCOPE heruntergeladen werden.*

Videosequenzen (74) einer Video-Datenbank basierend auf der Sequenzliste (62). Das Transformer-Modul (20) umfasst einen Cross-Modal-Encoder (24) und einen Feature-Decoder (28) und ist dazu ausgebildet, um basierend auf einem Video-Vektor (44) und einem Text-Merkmalsvektor (32) den Vorhersagevektor (22) zu erstellen.