Device and method for assembling a video clip based on multiple video sequences and text segments

The present invention relates to a device and a method for assembling a video clip based on multiple video sequences and text segments of a text. The device comprises a text module for generating a text feature vector with semantic text features of a text segment. The device further has a transformer module for generating a prediction vector, a selection module for selecting a target vector of a suitable target video sequence from a feature bank, an assignment module for creating a sequence list with target vectors for selected video sequences, and a video clip module for creating video clips from video scenes of a video database based on the sequence list.

Video clips consisting of individual scenes or sequences are used in many areas of daily life, for example in social media, as instructions, or in the form of news reports on television.

The dissemination of news in the form of video clips, so-called video news or news clips, has evolved significantly due to the growth of digital media and widespread internet access. This has led to a shift toward video-based news on platforms such as YouTube and social media. These engaging and accessible formats are preferred over traditional text-based news, particularly by younger people. There is therefore an increasing demand for the rapid creation of video content to capture audience interest in the online news sector.

News broadcasters must therefore produce engaging video clips faster than ever before to ensure their successful positioning in the market. This is partly due to the growing number of news sources and the changing media consumption of the target audience. This development has intensified the need to produce news clips quickly – a requirement that conflicts with traditionally manual and time-consuming video editing processes.

The production of video news comprises complex, time-consuming steps that make it difficult to keep pace with current news. The process begins with writing a news article, which is then converted into a voice-over script for the video. In the production of the news clip, footage is selected and sequenced that matches the sections of the article, whereby the principles of rhythm (i.e., the length of scenes affects the tone of the news clip), continuity (i.e., the content

of successive scenes must be consistent), and montage (i.e., the style of successive scenes follows rules) must be observed to ensure high-quality output.

Apart from individual advances in automating video production, current systems do not meet the efficiency and quality standards required for professional news broadcasts. Extensive manual intervention in the creation of video clips remains necessary, particularly when an assembly is based on various source materials and the generated clips are to be semantically coherent with the editorial text and stylistically consistent with professional editing standards.

There is therefore a great need to improve and streamline the editing and production process of video clips and to propose a system with which the quality of video clips can be maintained and manual interventions can be reduced.

The present object is solved with a device having the features of claim 1 and with a method having the features of claim 15.

In one aspect, the present invention relates to a device comprising a text module for generating a text feature vector with (semantic) text features of a text segment, a transformer module for generating a prediction vector for a video sequence based on the video vector of the preceding video sequence, a selection module for selecting a target vector of a suitable target video sequence (also called target) from a feature bank (also called SourceShots) based on the prediction vector and optionally for generating or selecting a target vector corresponding to the target video sequence, an assignment module for creating a sequence list with target vectors for the selection of video sequences based on the target vectors, and a video clip module for creating a video clip from video sequences of a video database based on the sequence list. The transformer module comprises a cross-modal encoder and a feature decoder and is trained to create the prediction vector based on the video vector and the text feature vector. The video vectors comprise video classification features that comprise or are stylistic means or features.

In a further aspect, the present invention relates to a training device for training a transformer module and for creating video vectors for video sequences based on text segments of a text or news article. The training device comprises a CLIP encoder for creating text feature vectors with text features of a text segment, a CLIP encoder for creating content vectors with video content features for a video sequence, a VGG encoder for creating classification vectors with video classification features for a video sequence, and the transformer module with a cross-modal encoder and a feature decoder. The cross-modal encoder is configured to merge a text feature vector and a content vector as well as a classification vector for a video sequence into

a feature vector and to supply it to the feature decoder, wherein the feature vector preferably has twice the length of the content vector. The feature decoder is configured to create from the feature vector a video vector that has the features of a video sequence present at the input of the training device and preferably has half the length of the feature vector. The training device outputs the video vector at an output and makes it available for further processing. Preferably, the feature vector can comprise further stylistic features such as camera movement or camera angle. This allows additional stylistic parameters to preferably be incorporated into the training.

Further aspects of the invention relate to a corresponding method and a computer program product with program code for performing the steps of the method when the program code is executed on a computer, as well as a storage medium on which a computer program is stored that, when executed on a computer, causes execution of the method described herein.

Preferred embodiments of the invention are described in the dependent claims. It is understood that the features mentioned above and those to be explained below can be used not only in the respectively indicated combination but also in other combinations or alone without departing from the scope of the present invention. In particular, the method and the computer program product can be embodied according to the embodiments described for the device in the dependent claims.

According to the invention, a prediction vector is generated based on a text feature vector comprising semantic text features of a text segment and based on a video vector with features of a video sequence. The video vector of the video sequence is based on a preceding video sequence and comprises its features. The prediction vector is the feature vector generated by the AI unit, the so-called transformer module, which follows the preceding video sequence and comprises its features. The prediction vector thus describes the sought-after, subsequent video sequence.

Based on this prediction vector with features of a video sequence to be selected, which is to follow the previous video sequence and is to match the content of the text features of the text segment relevant to the planned video sequence, the search for a suitable sequence is performed. Based on the prediction vector, a target vector with features of a suitable target video sequence is selected, wherein the target vector is comprised in a feature bank. The feature bank has the video vectors of all available video sequences, i.e., the feature vectors that represent and correspond to the video sequences. From this "pool" of available video sequences, which are represented by their feature vectors, the so-called video vector or target vector, a comparison is made with the prediction vector. The goal is to make a selection of the

most suitable target vector. The target vector that best correlates with the prediction vector is selected. This target vector is now written into a sequence list, preferably appended to an existing or empty sequence list. In this way, after several passes, the sequence list comprises a sequence of target vectors that comprises features of selected video sequences deemed suitable from the "pool" of video sequences represented by the vectors of the feature bank.

Once all text feature vectors of the text segments of the entire news text have been processed, the sequence list has all associated video sequences of the video clip as features in the form of target vectors that represent the individual video sequences of the video clip. In a video clip module, the video sequences are selected from a video database based on the target vectors in the feature list and assembled into a video clip. The resulting video clip then comprises the individual video sequences that best represent and reproduce the stored text, for example a news text.

The individual shots of camera footage or archive material from a corresponding database, which are available in the form of video segments, and the corresponding article segments of a text are the basic units for assembling news or video clips. A high-quality news clip with multiple shots or video sequences that is aligned with the underlying texts, particularly news articles, has the features that are deemed desirable. On the one hand, there is a visual representation of the semantics of the news text. On the other hand, appealing transitions between the individual video sequences are realized that comply with the rules for video editing of news clips.

According to the invention, prediction vectors are created in the transformer module, which comprises a cross-modal encoder and a feature decoder, based on a video vector with features of a video sequence and a text feature vector with features of a text segment. Advantageously, the video vector with features of the video sequence and the text feature vector can be merged in the cross-modal encoder so that a feature vector of double length is created. By means of the feature decoder, the suitable features are extracted therefrom to create a feature vector, the prediction vector, whose size corresponds to the size of the video vector or the text feature vector. Preferably, the video vector, the text feature vector, and the prediction vector are each of equal length.

The present invention has the advantage that a video clip can be produced reliably and automatically with high quality even when the video sequences are based on various source materials. The invention enables the generated video clips to be semantically coherent with the editorial text underlying the video clip and, moreover, to be stylistically consistent with and meet professional editing standards. The invention therefore provides an efficient automated

system for editing news clips or video clips with news to support and significantly relieve journalists in their task.

Because the previously known systems focus only on media areas outside of news production, such as feature film production, they cannot be directly applied to the area of news production or documentaries, since significantly different aspects play a decisive role in this area than in films. For journalists and video editors, there are certain stylistic means for producing high-quality news that can only be inadequately represented with previous systems. The present invention therefore focuses on the special stylistic means of news production in order to automate not only content-consistent but also high-quality news production. Within the scope of the invention, it was also recognized that the stylistic means significantly increase and improve the quality of the produced news clips or videos. Their consideration was recognized as a decisive factor for the quality and acceptance of news videos and for increasing information uptake by the viewer.

The present invention thus succeeds in using an embedding of visual language and video editing steps that were learned from an unlabeled video dataset with multiple video sequences. The fully automatic system reduces editing time and significantly improves the quality of the created news clips. The video segments used are aligned with the content of the news article and the respective text segments of the news text, whereby smooth transitions between shots are ensured by adhering to established video editing guidelines. This is achieved by processing the feature-comprising video vector of the preceding video sequence as well as the text features of a text segment intended for the next video sequence in the form of a text feature vector when creating the prediction vector for the next video sequence.

The present invention thus uses stylistic features and features of a video sequence as video features to perform video-text matching. This clearly distinguishes the invention from known systems that only consider content features of a video or a sequence of images. Although it is also known to use an "aesthetic score" for estimating the next video sequence, this only considers aspects such as sharpness or focus, which are technical or qualitative-technical features of a video. However, these features hardly play a role in professional camera footage and are almost always met. They do not incorporate stylistic elements and aspects.

In a preferred embodiment of the invention, the video vector and the prediction vector each comprise text features, video content features, and video classification features. The vectors thus have classification features in addition to content and semantic features. These are suitable for making transitions between individual video sequences smoother and more

pleasant for the viewer. In this way, it is possible to establish the principles of video editing technique in an automated manner.

Preferably, the video classification features are generated by means of a VGG encoder. Such an encoder can, for example, be integrated into the transformer module or the transformer unit. The VGG encoder is based on the concept of transfer learning, i.e., the ability to use existing knowledge developed to solve certain problems to solve new problems. The VGG encoder comprises a neural network of convolutional cells developed at Oxford University. The model achieved an accuracy of over 92% in a competition, which represented the highest value there. The convolution matrices of the neural network use smaller convolution kernels (3x3) and are based on a large database of more than 14 million labeled images divided into over 100 classes.

For example, features of scene composition can be generated and processed with a fine-tuned, specially further-trained "VILA encoder" (e.g., from Google). Motion vectors of the scene can also preferably be generated with a motion encoder. The various video classification features can preferably be merged into a vector.

In a preferred embodiment, the classification features of the video vectors and the prediction vector as well as the target vector are features of the video sequence that can comprise, for example, size features of the shot. The features are members of a list that comprise shot sizes of the video sequence. These can be, for example, information about whether a long shot, a medium shot or medium view, a detail shot, or a portrait shot is present. Further elements of the list of classification features comprise the camera angle used in the video sequence, statements about camera movement in the video sequence, direction of camera movement, movements of persons shown in the video sequence, direction of movement of persons in the sequence, zoom settings, lighting settings, or similar settings used in the video sequence.

Preferably, the scene composition with respect to the depicted subject is also considered, i.e., whether the person is in the center of the image and clearly recognizable. Image quality such as image blur, camera shake, etc. can also serve as a classification feature. These are important but not to the same extent for the device according to the invention and for the creation of news videos as in systems known in the prior art. In the field of news production, camera images are taken by professional camera operators; therefore, image quality plays a minor role. It can preferably be assumed. Previously, in the prior art, only image quality was considered and not the other stylistic means that are essential for high-quality news production.

A preferred embodiment provides that the video content features of the video vectors, prediction vectors, and/or target vectors were generated by means of a CLIP encoder. A CLIP encoder makes it possible, similar to text-based transformers, to read content from images or video sequences. A CLIP model is a Contrastive Language Image Pre-training model that can predict the most relevant captions or texts for an image without the model being optimized for a specific task. It is based on approximately 400 million pairs of image and text with which it was trained to predict the labels belonging to a specific image. The CLIP encoder can preferably be integrated in the transformer module. However, it can also be configured as a standalone module.

In a preferred embodiment, the text module also comprises a CLIP encoder. The text feature vector is generated by means of this CLIP encoder. It comprises the text features and content of the text segment that is to be assigned to the individual video sequence.

In a preferred embodiment, the transformer module is configured to comprise a plurality of submodules. Each of the submodules is configured to analyze and evaluate a predetermined video classification feature of a video vector. This allows the video classification features to be considered with higher quality when creating the prediction vector.

Each of the submodules in the transformer module preferably comprises a neural network. This neural network is preferably a feed-forward network (FFN). Each individual feed-forward network is specifically trained on a modality or a stylistic feature. In this way, it is possible to consider multiple stylistic features simultaneously, so that when creating a video vector and/or a prediction vector, multiple stylistic features of the video sequences are considered. This improves the quality and the interplay between text and image or image sequence.

In a preferred embodiment, each individual submodule can generate a sub-feature for classification that is considered when creating the prediction vector. The prediction vector then comprises the sum of all sub-features of the submodules.

In a preferred embodiment, the submodule has a neural network or a feed-forward network. Particularly preferably, a plurality of stylistic features or characteristics in the form of video classification features are processed in the feed-forward network.

Also preferred is an embodiment in which each feed-forward network is specialized for a predetermined stylistic means or stylistic feature of a video classification feature. Thus, a specific processing of the individual stylistic features takes place. Preferably, the individual results of the feed-forward networks of the individual submodules are merged and form the

basis for the prediction vector. The prediction vector forms the basis for the video sequence, wherein the target vector or target video vector of a suitable target video sequence is selected from the feature database based on this prediction vector.

In a preferred embodiment, the transformer module is configured to create the prediction vector based on multiple video vectors and multiple text feature vectors. Here, multiple video classification features in the individual video vectors are considered.

In a preferred embodiment, the prediction vector is generated based on the video vectors of a majority of the available video sequences. Particularly preferably, the generation is based on the video vectors of all available video sequences. Simultaneously or alternatively, the generation of the prediction vector can be performed on all available text feature vectors. It is also conceivable that the generation of the prediction vector is based on a majority of the available text feature vectors. Preferably, the prediction vector is created based on video vectors and text feature vectors, particularly preferably based on all available video vectors and text feature vectors. For example, an input vector can be used for this purpose that comprises all segments of the available news text as well as all scenes or video sequences of the available camera footage. Thus, the transformer module receives all available information for its estimation and for generating the prediction vector in order to select the most suitable target vector of a target video sequence based thereon. The transformer module consequently considers all text segments and thus incorporates all text information into its estimation. This has the advantage that the otherwise typically iteratively formed contribution, which only considers the preceding video sequence or scene, is improved since now all available text feature vectors and video vectors can flow into the process of creating the prediction vector and thus into the selection of the target vector. This significantly accelerates the entire process and at the same time improves the quality of the generated news videos.

The generation of the entire target vector thus takes place considering the entire text and all preceding scenes. This provides the transformer module with all available information for its estimation. It considers all text segments and thus incorporates all text information into its estimation.

In this way, the scene sequence for the contribution as a whole can be estimated and the interview content can also be incorporated. Since these (additionally considered) text passages can contain important information, the quality of the estimation improves. Interview passages, as are typical for news broadcasts or videos, can also be considered, which is why special interview tokens (by means of modal type embedding) can be assigned in the feature

vector. The feature vector is the basis for the video vector, on the basis of which the prediction vector is created.

In a preferred embodiment, the creation of the sequence list for the selection of video sequences is performed by appending a target vector to the sequence list. The sequence list is thus extended by further target vectors until all text segments of the text that is to underlie the video clip have been processed. The creation of the sequence list takes place in the assignment module.

In a particular embodiment, the text element of a text comprises a single sentence of the predetermined (overall) text. It is of course also possible for the text element to consist of multiple sentences, preferably up to five sentences. This is advisable when the sentences are related in content.

In a preferred embodiment of the device, the selection module is configured to compare the video vectors stored in the feature database with the prediction vector. Preferably, all video vectors of the feature bank are compared with a prediction vector of the selection module. Particularly preferably, this is done individually. The purpose of the comparison is to find a video vector from the feature bank that corresponds as well as possible with the prediction vector and to transfer this video vector to the sequence list. Preferably, when assessing the fit or matching of the prediction vector with one of the video vectors of the feature bank, the quality of the selection can be assessed. The vector selected from the video vectors is the target vector that is supplied to the sequence list. A quality function can preferably be used to determine the quality for the suitable selection. For example, it is possible to assess the quality based on cosine similarities.

In a preferred embodiment, the created sequence list, which is created in the assignment module, can comprise interview vectors that correspond to video sequences that comprise or are an interview sequence. In this way, it is possible to consider interview sequences in the video clip to be created. The interview vectors can be feature vectors that also comprise the transcript of the spoken dialogue in the interview sequence.

Corresponding to the device, the method according to the invention can also provide in a preferred embodiment that the generation of the prediction vector for the video sequence is based on multiple video vectors and multiple text feature vectors.

Preferably, according to an embodiment of the method, the generation of the prediction vector is performed based on all available video vectors or on the video vectors of all available video

sequences and/or based on all available text feature vectors. Here, it generally applies that with an increasing number of processed video vectors and text segments, the amount of available information is increased and thus the quality of the subsequent clips is further improved. The same applies to the device.

In a preferred embodiment of the training device, it is provided that the feature vector comprises stylistic means or stylistic features. These stylistic means are the same as those used in generating the prediction vector and selecting a target video sequence.

A preferred embodiment of the training device provides that it is configured such that multiple text feature vectors and/or multiple content vectors and/or multiple classification vectors are used to form the feature vector from which a video vector is created that has the features of an unknown video sequence.

Preferably, the transformer module comprises a neural network, e.g., a deep neural network or a feed-forward network (FFN). Preferably, the transformer module can comprise multiple submodules, each comprising a neural network or an FFN.

Preferably, the training device can be configured to train the submodules of the transformer module and thereby process the stylistic means and stylistic features for which the submodule is to be specifically configured and trained. Preferably, a feature vector (sub-feature vector) can be formed by each submodule.

In a preferred embodiment, the training device can additionally or optionally comprise a VILA encoder (Visual Language Encoder) for creating a vector or feature vector of the scene composition and preferably a score for image quality in order to describe and characterize the video sequence.

A preferred embodiment of the training device can have a motion encoder to create a feature vector of the camera movement for a video sequence. This "motion feature vector" can flow into the feature vector for the video vector, for example by summation or another vector calculation rule.

In a preferred embodiment of the training device, it can have multiple submodules for the stylistic means to be considered, wherein preferably each submodule can be specialized for a stylistic means. In other words, each submodule or each FFN is specifically trained for a modality or a stylistic means. In this way, more stylistic features than just the shot size can be considered.

Instead of processing only, for example, two modalities, namely content and shot size, such as by concatenating CLIP features and VGG features and then linearly mapping them for size reduction, the use of the large amount of information requires special consideration to avoid information loss due to excessive reduction. The entire information summarized in a feature vector is divided into individual sub-feature vectors that can be processed in the multiple submodules without or without practically relevant information loss. After processing in the submodules, the processed sub-feature vectors can be merged back into a feature vector.

The vectors processed in the device, the training device, and the method are feature vectors and comprise features of the text segment or a video sequence. The text vector is a text feature vector with features of a text segment. The prediction vector is a feature vector generated by the device, in particular by the transformer module, with content features as well as classification features for the next video sequence to be selected. The video vector or video feature vector comprises the features of a video sequence, which can comprise both content features and classification features. The target vector is the feature vector found by comparing the prediction vector with video vectors of a feature bank. This is also referred to as target feature vector or matching feature vector. Another designation is target vector.

The invention is described and explained in more detail below with reference to some selected embodiments in connection with the accompanying drawings. They show:

Figure 1 the device according to the invention for assembling a video clip;

Figure 2 the sequence and use of the device when creating a video clip;

Figure 3 a transformer module for generating a video vector;

Figure 4 a training device according to the invention with a transformer module; and

Figure 5 a schematic representation of the method steps of the method for assembling a video clip.

Figure 1 shows a device 10 for assembling a video clip based on multiple video sequences and text segments of a text. The device 10 comprises a transformer module 20 to generate a prediction vector 22 for a video sequence based on a video vector of a preceding video sequence. A text segment 12 is processed in a text module 30, in which a text feature vector 32 is generated from the text features of the text segment 12. This text feature vector 32 represents the semantic content of the text segment 12.

A selection module 40 processes the prediction vector 22, on the basis of which the selection module 40 makes the selection of a target vector 42 of a suitable target video sequence from a feature bank 50. Feature vectors of individual video sequences are stored in the feature bank 50. From this feature bank 50, which can be a database or a corresponding table, the feature vectors that best match the prediction vector 22 generated in the transformer module 20 are selected by means of selection modules 40. The selection can be made based on a quality criterion, for example based on cosine similarities, as exemplarily described in Formula 1.

Formula 1: $$cos(v_g, v_p) = \frac{v_g^T v_p}{\|v_p\|_2 * \|v_p\|_2}$$

Here, $|| x ||_2$ is the L2 norm of a vector x. v_p is the prediction vector 22 and v_g the target vector 42. v^T here means the transpose of the vector v.

In the selection module 40, all target vectors 42 stored in the feature bank 50 are compared with the prediction vector 22. The target vector 42 that has the greatest cosine similarity with the prediction vector 22 is selected and stored in a sequence list 62 in an assignment module 60. The sequence list 62 comprises all target vectors 42 of video sequences that are the basis for the video clip to be created.

In a video clip module 70, a video clip 72 is generated from video sequences 74 based on the sequence list 62. The individual video sequences 74 are the sequences that are described and characterized by the respective target vectors 42. Consequently, a video clip 72 is created in which multiple selected video sequences 74 are strung together. The individual video sequences 74 typically have a length of 30 seconds to 5 minutes, preferably of 10 seconds to 3 minutes, particularly preferably of 30 seconds to 90 seconds.

Since the target vectors 42 of the sequence list 62 comprise the text feature vectors 32 for the respective individual text segments 12, a video clip 72 in the form of a news clip results that comprises multiple video sequences 74 as well as the corresponding text passages for the video sequence 74. In this way, a speaker can speak the corresponding text to the video clip 72. Here, either the underlying text segments 12 can be used or free text oriented solely to the text segments 12 can be spoken. It is possible that the audio signal generated for the text segments 12 is generated with the help of an artificial intelligence or AI unit and assigned to the video sequences 74.

The determined target vector 42 is not only supplied to the assignment module 60, but also to the transformer module 20, where it is processed with the text feature vector 32 of the next text segment 12. In this way, the target vector 42 is considered as video vector 44 of the preceding video sequence 74 in the selection of the next video sequence 74, i.e., the target vector 42. The processing associated therewith is thus continued until all video sequences 74 and text segments 12 have been processed.

Figure 2 shows the cascaded sequence in the generation of a video clip 72 and the assembly of the video clip 72 from multiple video sequences 74 that are based on the selected target vectors 42 of a sequence list 62. Instead of the representation chosen in Figure 1, in which the processing loop is run through multiple times, the individual processing steps are shown here sequentially, wherein the device 10 is shown multiple times in different stages 16, 17, 18, 19.

First, the processing is started with a target vector 42 that represents a best shot vector 14 and is predetermined. The best shot vector 14 represents a predetermined video sequence that serves as the input sequence for the video clip 72. The best shot vector 14 can, for example, be selected and predetermined manually or by a separate AI unit.

In the transformer module 20, first the best shot vector 14 and the text feature vector 32 generated in the text module 30 based on the first text segment 12 are processed. Here, a prediction vector 22 is generated that is processed with the target vectors 42 of the feature bank 50 in the selection module 40. The target vector 42 that best "matches" the prediction vector 22, i.e., has the greatest correspondence or the greatest similarity, is selected.

The selected target vector 42 is inserted into the sequence list 62 present there in the assignment module 60, wherein this target vector 42 is the first entry in the sequence list 62. The target vector of the first stage 16 is supplied to the second stage 17 and processed here with the new text segment 12. The device 10 is thus used again in the second stage 17 for the selection of a suitable target vector 42. This is then appended to the sequence list 62 in the assignment module 60, so that the sequence list 62 grows with each stage and each further video sequence or each further text segment 12.

In a third stage 18, the target vector 42 generated in the second stage 17 is used to generate a new target vector 42. This principle is continued up to an n-th stage 19 until all text segments 12 have been processed. The text segments 12 originate from the text available for a video clip 72, in particular a news text.

At the end of the run through all stages 16 to 19, a video clip has been created that is generated in the video clip module 70 based on the sequence list 62 created in the assignment module 60.

Figure 2 shows that an optional interview bank 52 can be comprised in the feature bank 50. In the underlying news text, which is divided into multiple text segments 12, interview segments 54 can also be comprised that match corresponding interview sequences 55 in the interview bank 52. In a likewise optional interview module 56, the interview sequences 55 of the interview bank 52 are analyzed, for example by a speech-to-text converter, and then compared with the interview segments 54 of an interview segment database. Here, the interview sequences 55 with the greatest similarity and the highest correspondence to the examined interview segments 54 are selected and added in the form of an interview feature vector 58 to the sequence list 62 in the assignment module 60. In the video clip module 70, the corresponding interview sequence 55 of the interview bank 52 is selected and inserted into the video clip 72. This is optional and represents a preferred extension of the video clip 72.

The effective assembly of a video clip 72 or news clip with multiple video segments is thus based on pre-trained representations of the individual video sequences by feature vectors, so-called video vectors. Based on the encoded text segments 12 in the form of text feature vectors 32 and the encoded source material, i.e., the video vectors or target vectors 42 for individual video sequences 74, the workflow of a so-called inference pipeline described in Figure 2 results. In a preferred embodiment, previously marked interview sections in the news article and the corresponding interview recordings or interview sequences 55 can be filtered out, which were classified, for example, by a pre-trained interview classifier. With an index-based text search, the individual interview in-points and interview out-points are inserted into the sequence list 62 with the target vectors 42. At the beginning, a best shot vector 14 is generated that best represents the best shot opening scene or video sequence that is present in the entirety of the video clips 72. This selection is made according to the motto "start with the best scene," i.e., select the best video sequence as the starting point.

The following video sequences are selected with the current text segments 12 and the video vectors of the preceding scene or the target vectors 42. For this purpose, a prediction vector 22 for the next video sequence is generated. The generated prediction vector 22 is then compared with all existing video vectors of the feature bank 50, which each represent the target vectors 42. The target vector 42 that is most similar to the prediction vector is selected. The top candidates, i.e., the target vectors 42 that are most similar to the prediction vector, are inserted into the sequence list 62, on the basis of which the video sequences are selected and assembled into the video clip 72.

Figure 3 shows the transformer module 20 in detail. The transformer module 20 comprises a cross-modal encoder 24 that processes or merges a target vector 42 or a video vector 44 and a text feature vector 32, so that a feature vector 26 is formed. This comprises both the text feature vector 32 and the video vector 44. The feature vector 26 has twice the length of the text feature vector 32 or the video vector 44. In a feature decoder 28 of the transformer module 20, the feature vector 26 is further processed and a prediction vector 22 is generated. This comprises the features for the next video sequence that is to be comprised in the video clip to be generated. The prediction vector 22 has the same length as the target vector 42 or the video vector 44. It is thus preferably half as large as the feature vector 26.

Figure 4 shows the training device 80 according to the invention with the transformer module 20. The training device 80 is used to generate video vectors 44 from multiple video sequences 74 and multiple text segments 12 that are suitable for describing the video sequences 74 as well as possible and for allowing a selection of video sequences 74 based on the features contained in the video vectors 44.

The text segments 12 are processed in the text module 30. They are parts of an overall text, for example a news text or news article. Preferably, each text segment 12 comprises a main clause of the news article. The news text is usually divided into individual text segments 12 in order to be able to comply with the general video editing rule "one shot per main clause." The text module 30 preferably comprises a CLIP encoder 82 that is used to convert the content part of the text into a feature vector, here the text feature vector 32.

The video sequences 74 are categorized with a (different) CLIP encoder 82, wherein the content features of the video are extracted. Based thereon, a content vector 84 with video content features for the respective video sequence 74 is generated. In addition, the video sequences 74 are processed by means of a VGG encoder 86 to generate a classification vector 88. The classification vector 88 comprises video classification features that can comprise, for example, shot sizes of the video sequence 74, corresponding views comprised in the video sequence 74, camera angles, movements within the video or movements of the camera as well as directions of movement of persons or objects within the video or the video sequence 74 and/or the camera, zoom settings, or similar settings. The classification vector 88 thus represents the stylistic part of the video sequence 74.

Content vector 84 and classification vector 88 are merged into the video vector 44, which comprises both the semantic essence of a video sequence 74 as well as its stylistic features.

Feature vector 26 and text feature vector 32 are further processed in the cross-modal encoder 24, so that the feature vector 26 is formed. The further processing of the feature vector 26 takes place in the feature decoder 28, so that a prediction vector 22 is generated.

Thus, using text-image pairs or text-video sequence pairs, it is possible to train the transformer module 20 or the training device 80 such that exactly the shot features can be predicted that correlate with certain text segments 12 of news articles. This is done based on the available video segments. The use of the cross-modal encoder 24 and the feature decoder 28 facilitate the assignment process.

The video sequences 74 used both for the training device 80 and for the device according to the invention according to Figures 1 to 3 can consist of both current news clips based on a news article and archive material or archive video sequences. Each video sequence 74 preferably has a corresponding news article intended for the voice-over. In addition, a list with manual editorial cuts often exists.

The video vector 44 or prediction vector 22 based on the feature vector 26, which is available at the output of the training device 80, is supplied to an evaluation module 90 to assess the quality of the generated prediction vector 22 or video vector 44. For the generation of video clips based on video segments and the text-based assembly of the video clips, finding video segments that are content-relevant and of appropriate size is important. For this purpose, the transformer module 20 of the training device 80 is trained for fine-tuning the feature vector 26 using the text features from the text segment 12 and the features of the preceding video sequence 74. The training pairs thus comprise a feature vector for the preceding shot, a feature vector of the target shot, and a text of the text segment. The CLIP encoders 82 and VGG encoders 86 used generate content features and size features for the video, while the text features are also derived with a CLIP encoder 82. The encoded sequences generated here are processed by a stack of transformer encoding layers, wherein each layer has a consistent architecture.

The video vectors 44 generated in the transformer module 20 are then evaluated by means of a quality criterion in the evaluation module 90, wherein preferably in a similarity module 92 the cosine similarities between the generated video vector 44 and video vectors 44 of a database that correspond to the text segment are assessed. Here, according to Formula 2, each video vector 44 of a video sequence 74 is compared with other video sequences 74 and their feature vectors as negative samples. According to Formula 2, contrastive learning is performed, wherein v_p reflects the prediction vector 22 and v_j represents video vectors 44 as comparison samples. τ is a temperature parameter.

$$L_C = -log \frac{\exp(\frac{\cos(v_g, v_p)}{\tau})}{\exp(\frac{\cos(v_g, v_p)}{\tau}) + \sum \exp(\frac{\cos(v_j, v_p)}{\tau})}$$ Formula 2

In an entropy module 94 of the evaluation module 90, the cross-entropy losses are used as a training objective for quality assessment. Starting from the multimodal transformer module 20 and the generated prediction vectors 22, three 1x3 CONV layers are used in the entropy module 94, followed by a ReLU activation. This entropy module 94 is trained to classify the stylistic shot size of the scenes.

Finally, sigmoidal activations are added to improve the prediction vectors 22. The cross-entropy losses can be calculated with

$$L_S = -\lambda(v * \log\ v_p + (1 - v) * \log\ (1 - v_p))$$ Formula 3

with v_p = prediction vector 22; v = target vector 42 or video vector 44.

For this purpose, the combination of the individual cross-entropy losses over all video sequences 74 in the training set is set as the overall training objective. These entropy losses can be described by

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

, wherein L = total loss, L_s = binary cross-entropy loss (from Formula 3) of the prediction vector 22 and L_c = contrastive learning (from Formula 2). $\alpha$and β are hyperparameters and control the robustness of the total loss; N is the number of individual training sets in the overall training set.

The training set comprises the video sequences 74 and the text segments 12 that are present at the input of the training device 80 and the features generated therefrom at the input of the transformer module 20.

Figure 5 shows the basic sequence of the method for assembling a video clip based on multiple video sequences and text segments of a text, in particular a news text.

In a first step S10, a text feature vector with semantic text features of the text segment is generated. In a further step S12, a prediction vector for a video sequence is generated based

on a video vector of a preceding video sequence and based on the text feature vector of the current text segment.

A step S14 relates to selecting a target vector of a target video sequence from a feature bank based on the prediction vector. In this step, the current text segment is thus linked with the preceding video sequence in order to generate therefrom a feature vector for the next video sequence.

Based on this prediction vector, in step S14 the target vector for the next, i.e., the subsequent video sequence, also called target video sequence, is generated.

A step S16 relates to creating a sequence list with target vectors for the selection of video sequences based on the target vectors. This is preferably done by appending a new target vector to the sequence list or to the target vectors already present in the sequence list.

In a step S18, video sequences are selected from a video database based on the sequence list in order to generate a video clip. The video database can be a database that can comprise multiple video sequences from current films or videos, but also archive material, which preferably can likewise consist of shorter video sequences or short video clips. Alternatively, the video database can also be configured on an external server or on multiple storage media, for example by individual video clips that can be stored on the storage media or at various locations.

In a preferred embodiment, steps S10 to S16 are run through multiple times, namely again for each new text segment, in order to generate in this way a sequence list with multiple target vectors.

In a preferred embodiment, an interview feature vector can be generated in a step S20. This is inserted into the sequence list. The interview feature vector corresponds to an interview sequence or to a video sequence comprising an interview sequence.

Stylistic features within the meaning of this invention are, for example, features that comprise the overall content correspondence of text or text segments and scenes or video sequences or images. For this purpose, the contents of both the text or the text sequences as well as the scenes, images, or video sequences can be represented in the form of characteristic quantities or coded digits, for example in the form of numerical values or in the form of colors or color adjustments.

A correspondence (of coded features) can be determined when both the coding of the text and of the scenes and video sequences are the same or largely the same, or when the feature vectors resulting from the respective text segments and video scenes have a high correspondence of individual elements of the vector.

Further stylistic means within the meaning of the invention relate to the sequence of shot sizes of a video scene or video sequence. These are, for example, the long shot, medium shot, portrait shot, or detail shot. The stylistic means in this regard provides information about how the corresponding stylistic sizes or features are changed. It has proven advantageous when a video sequence shows a long shot, the following video sequence shows a medium view or portrait view, and the subsequent video sequence comprises a detail shot.

A further stylistic feature for video classification or the video classification feature comprises camera movement. Here, it can be considered that the camera movement is relatively calm and that the same camera movement does not occur multiple times in succession in different sequences. For example, within the scope of the invention, it has proven disadvantageous when the camera performs a panning movement in the same direction multiple times in succession. This creates the impression for the viewer that they are rotating around their own axis, which is not wanted and not desired, particularly in news broadcasts and news videos.

A further stylistic feature or stylistic characteristic as a video classification feature is scene quality. This comprises features such as image sharpness, exposure, or camera shake or unpredictable camera movements.

The further stylistic features to be preferably considered also comprise scene compositions. These features express whether the described subject or object that is comprised in the corresponding text segments is also found in the video sequence. In particular, it can be considered and expressed here where the object is arranged or whether the object or subject is in the center of the image and centrally located or at an edge position within the image.

The consideration of stylistic features in the video classification features that are considered when creating the prediction vector leads to the successive video sequences and the corresponding text segments matching each other and fitting together well in the sense that, particularly for news videos and information videos, a calm and informative image and scenario is created that, on the one hand, matches and corresponds to the stored text and the individual text segments and, on the other hand, creates an atmosphere for the viewer in which they can best absorb the information content generated by the images and the text.

The stylistic means comprise the content correspondence of text and scenes, the sequence of shot sizes (long shot, medium view, detail view), the sequence of camera movements (standstill, pan, or zoom), the scene composition (position of the subject), and the image quality.

The invention has been comprehensively described and explained with reference to the drawings and the description. The description and explanation are to be understood as examples and not as limiting. The invention is not limited to the disclosed embodiments. Other embodiments or variations will become apparent to those skilled in the art when using the present invention as well as upon careful analysis of the drawings, the disclosure, and the following patent claims.

In the patent claims, the words "comprise" and "with" do not exclude the presence of further elements or steps. The indefinite article "a" or "an" does not exclude the presence of a plurality. A single element or a single unit can perform the functions of several of the units mentioned in the patent claims. An element, a unit, a device, and a system can be implemented partially or completely in hardware and/or in software. The mere mention of some measures in several different dependent patent claims is not to be understood as meaning that a combination of these measures cannot also be advantageously used. A computer program can be stored/distributed on a non-volatile data carrier, for example on an optical storage medium or on a solid-state drive (SSD). A computer program can be distributed together with hardware and/or as part of hardware, for example via the Internet or via wired or wireless communication systems. Reference signs in the patent claims are not to be understood as limiting.

Patent Claims

1. A device for assembling a video clip (72) based on multiple video sequences (74) and text segments (12) of a text, comprising:

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);

a selection module (40) for selecting a target vector (42) of a suitable target video sequence from a feature bank (50) based on the prediction vector (22);

an assignment module (60) for creating a sequence list (62) with target vectors (42) for the selection of video sequences (74) based on the target vectors (42);

a video clip module (70) for creating a video clip (72) from video sequences (74) of a video database based on the sequence list (62);

wherein

the transformer module (20) comprises a cross-modal encoder (24) and a feature decoder (28); and

the transformer module (20) is configured to create the prediction vector (22) based on a video vector (44) of the preceding video sequence (74) and on a text feature vector (32);

the video vectors (44) comprise video classification features that comprise stylistic features.

2. The device according to claim 1, characterized in that the video vector (44) and the prediction vector (42) comprise text features, video content features, and video classification features.

3. The device according to the preceding claim, characterized in that the video classification features are generated by means of a VGG encoder (86), wherein the VGG encoder (86) is preferably integrated in the transformer module (20).

4. The device according to claim 2 or 3, characterized in that the video classification features comprise features of the video sequence (74), wherein the features are elements of a list relating to shot sizes of the video sequence (74) such as long shot, detail view, portrait view, or medium view, camera angle used in the video sequence (74), movement of the camera in the video sequence (74), direction of movement of the camera, movement of persons in the video sequence (74), direction of movement of persons in the video sequence (74), zoom settings, and similar settings of the video sequence (74).

5. The device according to claim 2, characterized in that the video content features are generated by means of a CLIP encoder (82), wherein the CLIP encoder (82) is preferably integrated in the transformer module (20).

6. The device according to one of the preceding claims, characterized in that the transformer module (20) comprises a plurality of submodules, wherein each submodule has a neural network and is configured to analyze and evaluate a predetermined video classification feature of a video vector (44) in order to consider the video classification feature when creating the prediction vector (22), wherein the submodule is preferably a feed-forward network and wherein preferably the video classification feature comprises a stylistic feature.

7. The device according to one of the preceding claims, characterized in that the transformer module (20) is configured to create the prediction vector (22) based on multiple video vectors (44) and multiple text feature vectors (32), preferably based on the video vectors (44) of all available video sequences (74) and/or on all available text feature vectors (32).

8. The device according to one of the preceding claims, characterized in that the text module (30) comprises a CLIP encoder (82), wherein the text feature vector (32) is generated by means of the CLIP encoder (82).

9. The device according to one of the preceding claims, characterized in that the creation of the sequence list (62) for the selection of video sequences (74) is performed by appending a target vector (42) to the sequence list (62).

10. The device according to one of the preceding claims, characterized in that the text segment (12) comprises a sentence of the predetermined text.

11. The device according to one of the preceding claims, characterized in that the selection module (40) is configured to compare the video vectors (44) stored in the feature bank (50) with the prediction vector (22), preferably individually, and particularly preferably the quality of the selection of the target vector (42) is assessed, very preferably by means of a quality function.

12. The device according to one of the preceding claims, characterized in that the sequence list (62) comprises interview feature vectors (58) that correspond to video sequences (74) that comprise an interview sequence (55).

13. A training device for training a transformer module (20) and for creating video vectors (44) for video sequences (74) based on text segments (12) of a text or news article, comprising:

a CLIP encoder (82) for creating text feature vectors (32) with text features of a text segment (12);

a CLIP encoder (82) for creating content vectors (84) with video content features for a video sequence (74);

a VGG encoder (86) for creating classification vectors (88) with video classification features for a video sequence (74);

the transformer module (20) with a cross-modal encoder (24) and a feature decoder (28),

wherein

the cross-modal encoder (24) is configured to merge a text feature vector (32) and a content vector (84) as well as a classification vector (88) for a video sequence (74) into a feature vector (26) and to supply it to the feature decoder (28), wherein the feature vector (26) preferably has a greater length than the content vector (84), very preferably twice the length;

the feature vector (26) comprises stylistic means or features;

the feature decoder (28) is configured to create from the feature vector (26) a video vector (44) that has the features of a video sequence (74) present at the input of the training device (80) and preferably has half the length of the feature vector (26);

wherein the training device (80) outputs the video vector (44) at an output and makes it available for further processing.

14. The training device according to the preceding claim, characterized in that the training device is configured to form the feature vector (26) from multiple text feature vectors (32) and/or multiple content vectors (84) and/or multiple classification vectors (88).

15. A method for assembling a video clip (72) based on multiple video sequences (74) and text segments (12) of a text, comprising the following steps:

a) generating a text feature vector (32) with text features of a text segment (12);

b) generating a prediction vector (22) for a video sequence (74) based on a video vector (44) of a preceding video sequence (74) and based on the text feature vector (32) of the text segment (12);

c) selecting a target vector (42) of a target video sequence from a feature bank (50) based on the prediction vector (22);

d) creating a sequence list (62) with target vectors (42) for the selection of video sequences (74) based on the target vectors (42) by appending a target vector (42) to the sequence list (62);

e) creating a video clip (72) from video sequences (74) of a video database based on the sequence list (62).

16. The method according to the preceding claim, characterized in that steps a) to d) are executed multiple times to generate the sequence list (62).

17. The method according to claim 15 or 16, characterized in that an interview feature vector (58) is appended to the sequence list (62), which corresponds to a video sequence (74) comprising an interview sequence (55).

18. The method according to one of claims 15 to 17, characterized in that the generation of the prediction vector (22) for the video sequence (74) is performed based on multiple video vectors (44) and multiple text feature vectors (32), preferably based on the video vectors (44) of all available video sequences (74) and/or on all available text feature vectors (32).

19. A computer program product comprising instructions that, when the program is executed by a computer, cause the computer to perform the steps of the method according to claim 15.