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

Abstract

A device (10) for assembling a video clip (72) based on 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) based on 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) 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); and 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). The transformer module (20) comprises a cross-modal encoder (24) and a feature decoder (28) and is configured to create the prediction vector (22) based on a video vector (44) and a text feature vector (32).

(Figure 1)