IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 22 December 2022, accepted 15 March 2023, date of publication 20 March 2023, date of current version 24 March 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3259733

RESEARCH ARTICLE

Simultaneous Video Retrieval and Alignment

WON JO[1], GEUNTAEK LIM[2], YUJIN HWANG[2], GWANGJIN LEE[2], JOONSOO KIM[3], JOUNGIL YUN[3], JIYOUNG JUNG[4], AND YUKYUNG CHOI[1,2], (Member, IEEE)

[1]Department of Artificial Intelligence, Sejong University, Gwangjin-gu, Seoul 05006, Republic of Korea
[2]Department of Intelligent Mechatronics Engineering, Sejong University, Gwangjin-gu, Seoul 05006, Republic of Korea
[3]Electronics and Telecommunications Research Institute (ETRI), Yuseong-gu, Daejeon 34129, Republic of Korea
[4]Department of Artificial Intelligence, University of Seoul, Dongdaemun-gu, Seoul 02504, Republic of Korea

Corresponding author: Yukyung Choi (ykchoi@sejong.ac.kr)

This work was supported by the Institute for Information and Communications Technology Planning and Evaluation (IITP) grant funded by the Korea Government (MSIT) under Grant 2020-0-00011 (Video Coding for Machine, 50%), Grant 2021-0-02067 (Next Generation AI for Multi-Purpose Video Search, 30%), and Grant 2021-0-00755 (Dark Data Analysis Technology for Data Scale and Accuracy Improvement, 20%).

ABSTRACT With the growth of the video streaming industry, video retrieval and video alignment are facing high levels of demand. Several studies have demonstrated the feasibility of these methods for various problems related to video retrieval and alignment independently, but testing in a unified framework has never been done. However, in real-world applications, it is also simultaneously necessary not only to find which video pairs are similar (video retrieval), but also to align the positions of the pairs that are related (video alignment). In this paper, we present a novel task: simultaneous video retrieval and alignment. As a solution to this task, a Simultaneous video Retrieval and Alignment framework, abbreviated as SRA, is proposed, which is a two-stage approach consisting of a foreground proposal stage and a downstream stage to efficiently process untrimmed videos. Furthermore, two criteria are suggested to support the new task: a metric mAP@J assessing how highly related videos are ranked and how well relevant positions are assigned in those videos, and a dataset FIVR+A that includes video-level relationships and hierarchical segment-level annotations. Finally, we conduct multi-pronged analyses to assess how our approach handles the new task in various experiments.

INDEX TERMS Computer vision, content based retrieval, information retrieval.

I. INTRODUCTION

Information retrieval can be understood as the task of finding the best-matched material within a large collection. The history of information retrieval started with document searches using keywords and evolved rapidly into searching images using a query image, then finally to retrieval of a best-matched video using a query video. Currently, due to the development of the internet infrastructure, the video streaming industry is in the spotlight. YouTube, one of the largest video streaming platforms, has more than two billion users per month who spend more than one billion hours per day watching videos on the platform [1]. Along with the rapid growth of the video streaming market, more diverse functions are required for the task of video retrieval. For example, a user may simply want to find a new video that is most similar to the one he just watched. Another user may want to find the movie that contains a short video clip that she saw in an attractive trailer (video retrieval). Yet another user may want to find and watch from the exact location at which a short video clip starts within the entire movie (video alignment).

There have been many studies on the tasks of video retrieval and alignment relevant to these needs. The video retrieval, known as Content-Based Video Retrieval (CBVR) [2], aims to find the most relevant video in a database given a video query. CBVR started with Near-Duplicate Video Retrieval (NDVR), which searches for visually identical videos, especially on a dataset [3]. For instance, a pattern-based approach [4] and a graph-based approach [5] detect copied videos from keyframe-level representations. Following that, Fine-grained Incident Video Retrieval (FIVR) and Event Video Retrieval (EVR), which search for videos of

The associate editor coordinating the review of this manuscript and approving it for publication was Adnan Abid.

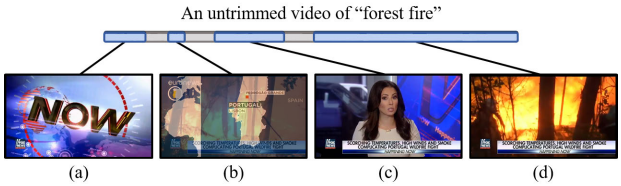


FIGURE 1. Composition example of an untrimmed video related to a forest fire: (a) A text-effect scene for the news opening; (b) The process of a scene transition; (c) A scene in which the announcer provides an explanation; and (d) A scene showing the forest fire.

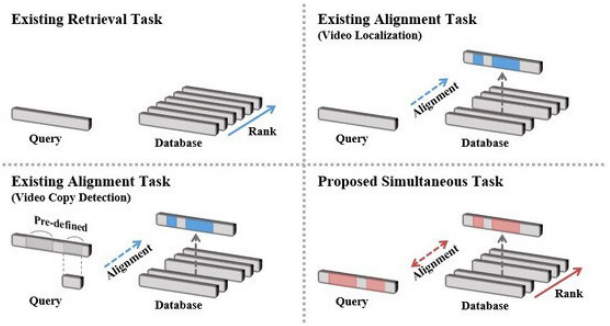


FIGURE 2. Comparison of tasks related to video retrieval and alignment. The proposed method simultaneously conducts retrieval and alignment and is bidirectional for both the query and database in the alignment process.

the same event captured, and Action Video Retrieval (AVR), which searches for videos of a human-centric act, are actively underway on datasets [6], [7], [8]. These studies are divided into two parts, depending on how similarities between videos are determined. One of the two parts utilizes frame-level features for the similarity estimation, such as an approach [9] for matching in the Fourier domain, an approach [10] utilizing a frame-to-frame similarity map, and standard frameworks [11], [12], [13], [14]. The other part utilizes video-level features that describe a single vector per video, such as an approach [15] with two fusion variations from pre-extracted frame-level features and an approach [16] aggregating frame-level features with self-attention. Video alignment is also divided into two parts: video localization and video copy detection, of which the former aims to locate in a semantically similar relationship and the latter in a visually similar relationship. Video localization is focused on the action only, such as an approach [17] to finding boundaries representing predefined action classes and an approach [8] based on cross-gated bilinear matching. In contrast, video copy detection focuses only on the copied scene, such as approaches [18], [19], [20] employing a temporal Hough transform, approaches [21], [22] utilizing a matching kernel, and approaches [23], [24] based on a similarity map.

Even though these many studies have been conducted, the two tasks are usually handled as independent problems. Video retrieval usually disregards the issue of video alignment, so they cannot determine which part of a retrieved video is similar to the query. Video alignment tends to oversimplify its task by localizing only reference videos for a given query video clip. However, it should be considered simultaneously because it is hard to know when and what functions a user will need in a real-world situation. Admittedly, it may seem that certain approaches [4], [5], [9], [13], [23] are already able to handle retrieval and alignment at the same time, even if they are insufficient to claim this. Indeed, the approaches [4], [5], [13], [23] do not take into account the exploration of videos that are directly related to the video-level topic of a query, as they only leverage similarities between keyframes, except for continuous signals in a video (like image retrieval). For example, if Fig. 1(a) is selected as a keyframe, videos containing the scene will be returned to the user, even if the topics are different (i.e., airplane accident, bombing, natural disaster, etc.). For this reason, they are closer to video copy detection and segment-level retrieval. On the other hand, the approach [9] cannot be considered to be handled simultaneously due to its conflicting assumption that alignment requires a set of videos known as a single event in advance, whereas retrieval requires distinguishing relations among videos that contain numerous unknown events.

In addition, tasks handled by other streams [25], [26], and [27] may seem similar to the proposed task. However, the task handled by [25] covers only retrieval between different modalities (especially text and image in this work) and is far from our task, which cover retrieval and alignment between two videos at the same time. In addition, the task handled by [26] differs from our task in that it uses text and deals only with alignment to find a location related to the query on a video when given a text query and an untrimmed video. Finally, although the task handled by [27] considers retrieval and alignment at the same time, it is fundamentally different because the input is a pair of text and video. And unlike our task, which considers alignment in both inputs in a pair, it considers alignment in only one input.

In this paper, we present a novel task named simultaneous video retrieval and alignment to resolve the aforementioned problems for real-world applications. As shown in Fig. 2, the proposed task is an extended video-to-video retrieval system that ranks at the video level while aligning at the segment level. It assumes that the query and database consist of untrimmed videos, which are the majority of real-world applications, and determines the degree of alignment in both directions. In addition, Simultaneous video Retrieval and Alignment framework (SRA) is proposed as a solution for the new task. To effectively represent long and untrimmed videos with multiple contents (for both videos in a database and query videos), SRA consists of two stages. The first stage is the foreground proposal, referring to the selection of only likely meaningful content in an untrimmed video, except for distracting content that is common anywhere or contains editing effects. The second stage consists of downstream tasks, including retrieval and alignment from areas selected in the previous stage. Furthermore, we propose two criteria to support the new task covered by SRA: the metric mAP@J and

the dataset FIVR+A. mAP@J is a new metric that explores similar relationships among videos while assigning relevant locations in a video pair based on existing metrics where only retrieval or alignment is measurable. FIVR+A is a dataset in which hierarchical annotations at the segment level are added to a widely known incident-centric dataset [6], replacing existing datasets that cannot handle both retrieval and alignment efficiently. Finally, multi-pronged analyses of SRA using these two criteria are presented while also comparing their outcomes to those of other methods.

In short, our contributions are threefold, as summarized below:

1) A new task, which is simultaneous video retrieval and alignment, and a method for completing the task, SRA, are proposed.
2) Two criteria, the metric mAP@J and the dataset FIVR+A, are proposed to support the evaluation of the new task.
3) Multi-pronged analyses via extensive experiments demonstrate how our approach manages video retrieval and alignment concurrently or independently.

The rest of the paper is organized as follows: Section II provides related tasks. Section III introduces the new task, simultaneous video retrieval and alignment. Also, two criteria for this new task, the metric mAP@J and the dataset FIVR+A, are described here. Section IV proposes the baseline, Simultaneous video Retrieval and Alignment framework (SRA) as a solution for the new task. Section V explains multi-pronged experiments, including benchmarks with other methods, ablation studies, and analyses. Section VI presents the conclusion that summarizes the previous contents.

II. RELATED WORK

A. VIDEO RETRIEVAL

Most video-to-video retrieval approaches use either frame-level features or video-level features when calculating the degree of video-level similarity. Methods based on frame-level features encode spatial information by describing each frame and encode temporal information by exploring the successive spatial information. On the other hand, methods based on video-level features encode spatio-temporal information using all of the frames in a video at once.

1) FRAME-LEVEL FEATURES

Methods based on frame-level features compare the similarities between features described for each frame in two videos in sequence and estimate a video-level similarity by aggregating them. An approach [4] known as Dynamic Programming (DP) extracts the diagonal pattern from the frame similarity map to find a near-duplicate area. An approach [5] known as Temporal Network (TN) detects the longest path in a graph generated via keypoint frame matching to identify visually equivalent scenes between the two videos. An approach [9] known as Circulant Temporal Encoding (CTE) employs the Fourier transform to encode frame features and compare them in the frequency domain. An approach [10] known as Video Similarity Learning (ViSiL) utilizes metric learning based on a frame-to-frame similarity map. A standard framework [11], known as Compact Descriptors for Video Analysis (CDVA), and its variations [12], [13], [14], [28] describe two types of transformation-resistant keyframe features. In general, these methods outperform methods based on video-level features. However, they usually incur a high computational cost with low retrieval efficiency due to the redundancy between consecutive frames. Theoretically, the upper bound of the time complexity can reach $T_Q \times T_R$ (T_Q and T_R denote the number of frames in a query video and a database video, respectively). A naive way to reduce the time complexity is to sample only some of the frames throughout. However, this type of strategy may result in significant performance degradation.

2) VIDEO-LEVEL FEATURES

Methods based on video-level features encode one video as one feature vector and calculate the video-level similarity by comparing the obtained features. An approach [15] known as Deep Metric Learning (DML) creates a visual codebook for intermediate Maximum Activation of Convolutions (iMAC) features [29] and trains a network through metric learning. An approach known as Temporal Context Aggregation (TCA) [16] combines frame-level features with self-attention to learn a single video vector. An approach [30] known as Hashing Codes (HC) combines and hashes multiple local and global features to handle the accuracy and scalability issues. Given that the entire video with multiple frames is abbreviated into a single feature, these methods are advantageous for describing the semantic information of successive connections and have relatively low computational costs. However, the loss of information increases as the length of the video increases. For this reason, the difficulty of video representation increases, resulting in a decrease in retrieval ability. If the feature representation of these methods is sufficiently improved, video retrieval with high efficiency is possible.

B. VIDEO ALIGNMENT

The task of video alignment aims to localize areas similar to a query clip within a given reference collection. The task is divided into video localization and video copy detection according to the contents to be found.

1) VIDEO LOCALIZATION

Video localization methods seek locations in the reference segment that show action similar to a given query. An approach [17] known as Boundary-Matching Network (BMN) generates action proposals through a boundary-matching mechanism and computes their confidence. An approach [8] known as Video Re-Localization (VReL) proposes cross-gated bilinear matching to predict four types of localization states. In these methods, the query is assumed to be a specific class or a trimmed video that contains only one action, and unidirectional alignment is conducted in reference videos containing at least one action related to the query.

However, this experimental setting is unlike the real-world situations in which most query videos are given untrimmed with unrelated pairs. In addition, unlike video copy detection, these methods focus on the concept of semantic similarity to find similar action contents.

2) VIDEO COPY DETECTION

Video copy detection methods seek to locate the reference segment that is in a copy relationship with a segment of the query. Approaches [18], [19], [20] utilize the Hough histogram resulting from the Hough transform. An approach known as Temporal Matching Kernel (TMK) [21] encodes sequences of frames with periodic kernels. An approach [22] known as Learning to Align and Match Videos (LAMV) trains the feature transform coefficients based on TMK. An approach [23] known as Segment Similarity and Alignment Network (SSAN) lines up videos through segment-level search using keyframes and then generates a similarity map to detect similarity patterns. An approach [24] known as Video Similarity and Alignment Learning (VSAL) conducts a mask-based temporal similarity measurement and a step-based partial alignment in a similarity map. An approach [31] known as Binary Temporal Alignment (BTA) leverages local features to find the left and right borders of a copied segment and concatenate them. An approach [32] known as Fast Partial Video Copy Detection (FPVCD) employs global features and a modified TN [5] for efficient video copy detection. Due to the repeated detection process for each predefined query segment, these methods operate in a unidirectional manner. In addition, they rely significantly on visual similarity in their hunt for identical duplicates.

C. VIDEO DATASETS

Video retrieval datasets contain the query videos and database with video-level relationships, mainly in consideration of near-duplicate situations and/or events, such as CC_WEB_VIDEO [3], EVVE [7], and FIVR [6]. Video alignment datasets comprise segment-level correlations within a collection of videos, primarily for copy detection with an emphasis on visual similarity, such as VCDB [18] and FIVR-200k-PVCD [24]. Meanwhile, ActivityNet, reorganized in earlier work [8], was released to deal with the alignment problem of semantic similarity. In addition, Kordopatis-Zilos et al. [10] attempt the task of retrieval using the reorganized ActivityNet [8]. However, this dataset is confined to action classes and consists only of references containing at least one segment without distractors. We present a dataset to overcome this limitation in Section III-C.

III. A NEW TASK: SIMULTANEOUS VIDEO RETRIEVAL AND ALIGNMENT

A. PROBLEM FORMULATION

A given query video Q and database video R are defined as follows:

$$Q = \{q_t\}_{t=1}^{T_Q}, R = \{r_t\}_{t=1}^{T_R} \tag{1}$$

where q_t and r_t refer to the t-th frame in the query and the database video, respectively. T_Q and T_R denote the total number of frames in the videos. The sets of segments associated with each other between Q and R are denoted as $\psi^Q = \{(q_{start}, q_{end})\}$ and $\psi^R = \{(r_{start}, r_{end})\}$, respectively. q_{start} and q_{end} are correspondingly the starting and ending frames of the segment related to R in Q. Analogously, r_{start} and r_{end} are the starting and ending frames of the segment related to Q in R. The similarity between the pair of videos is denoted as $Score_v$. By collecting these components, the new task F, which performs video retrieval and alignment simultaneously, is defined as follows:

$$(Score_v, \psi^Q, \psi^R) = F(Q, R) \tag{2}$$

Existing video retrieval methods use only the similarity score $Score_v$ to rank videos. Video localization and video copy detection methods find only the same or similar segments, which are only equivalent to ψ^R in the database. Unlike previous attempts to perform video retrieval or alignment, the proposed method handles $Score_v$, ψ^Q, and ψ^R together.

B. EVALUATION METRIC: mAP@J

$$\begin{aligned} mAP@J &= \frac{1}{N_Q}\sum_{m=1}^{N_Q} AP@J_{(m)}, \\ AP@J_{(m)} &= \sum_{n=1}^{N_R} precision_{(m,n)} \Delta recall_{(m,n)} J_{(m,n)} \\ J_{(m,n)} &= \frac{1}{2}\left(\omega^Q_{(m,n)} + \omega^R_{(m,n)}\right) \\ \omega^k_{(m,n)} &= \frac{\left(\psi^k_{(m,n)} \bigcap GT^k_{(m,n)}\right)}{\left(\psi^k_{(m,n)} \bigcup GT^k_{(m,n)}\right)}, \ where\ k \in \{Q, R\} \end{aligned} \tag{3}$$

We introduce a new evaluation metric, the mean average precision at the Jacccard weight, abbreviated as mAP@J, which measures the performance of the simultaneous video retrieval and alignment task. Because previous metrics have verified "retrieval" and "alignment" individually, measuring the connection between the two tasks is difficult. Therefore, both the method and the evaluation metric are not easily classified as in "simultaneous retrieval and alignment". On the other hand, the proposed metric considers "retrieval" and "alignment" as a single task. It is formulated in Eq. 3.

Here, mAP@J is a form in which the Jacccard weight J, as an alignment weight, is multiplied when calculating the average precision (AP) for each query based on the retrieval metric, the mean average precision (mAP). In Eq. 3, N_Q and N_R refer correspondingly to the number of videos in the query (abbreviated Q) and the database (abbreviated R). $J_{(m,n)}$ refers to the Jacccard weight between the n-th video in the database and the m-th query video, where the database is sorted w.r.t. the corresponding query video. This is calculated as the average of $\omega^Q_{(m,n)}$ and $\omega^R_{(m,n)}$, which determines how well the locations are aligned in each of the m-th query and

FIGURE 3. An example of annotation for two related video pairs of FIVR+A. Labeling is done on a segment-by-segment basis. According to their similarity with the video-level topic, the segments are categorized as N, S, or H. The fading effect between event segments is labeled F.

n-th database video pairs. Additionally, $\omega^k_{(m,n)}$ (k is Q or R) is the Jaccard index [33] between the set of predicted aligned segments $\psi^k_{(m,n)}$ and the ground truth segments $GT^k_{(m,n)}$ in each video.

In addition, mAP@J is designed to reduce the effect of video pairs that contribute significantly due to their high video-level rank, even when they are not aligned well. Conversely, it retains the original contribution by imposing fewer penalties on well-aligned video pairs, regardless of their rank. Hence, the proposed metric simultaneously evaluates retrieval and alignment by inducing a structure in which the degree of alignment reorganizes the ranking of the video level.

C. DATASET: FIVR+A

Although some datasets [18], [24], [34] provide segment-level annotations, they are limited to specific topics (action or copy detection), and the degree to which something is related is difficult to determine because it simply appears in a Boolean form. On the other hand, for a dataset [6] containing a hierarchical structure at the video level, segment-level annotation is not provided. A dataset, Fine-grained Incident Video Retrieval plus Alignment (FIVR+A), is proposed to deal with the aforementioned issue in conjunction with the new task. This is based on the FIVR dataset, and segment-level information is added to indicate where a video-level topic appears in a video. Because query videos represent each video-level topic, the segment-level annotation is divided into four hierarchical conditions corresponding to them as follows:

- Normal (**N**) refers to a segment where the temporal span and camera viewpoint are similar, i.e., frames are similar in terms of their visual appearance.
- Soft (**S**) refers to a segment where the temporal span is similar but the camera viewpoint is different.
- Hard (**H**) refers to a segment where the temporal span and camera viewpoint are different, but where it can be inferred semantically as the same topic.
- Fade in/out (**F**) refers to a segment where a fading effect is observed before or after the segments corresponding to the preceding conditions (**N**, **S**, and **H**).

The closer to **N** corresponds to a near-duplicated scene in the copy detection task; the closer to **H** corresponds to a semantically deducible scene containing several changes (e.g., daylighting differences, side views). Furthermore, **F**, the location of the fade effect occurring during the shot conversion process, is provided, which is the first among video retrieval or alignment datasets as far as we know. According to these conditions, 9,960 segments are annotated at FIVR-5K, as shown in Fig. 3. The number of segment pairs between related videos in the database and query videos is 17,619 for all of the above conditions, 5,871 for **NSH** only, 3,996 for **NS** only, and 2,067 for **N** only. The average percentage of each criterion in all videos is as follows: **N** amounts to 22.29%, **S** is 14.72%, **H** is 11.55%, **F** is 1.67%, and the rest equals 49.77%. All annotations are manually performed and thoroughly reviewed by computer vision experts multiple times. These annotations can be obtained by emailing the author.

IV. A BASELINE: SIMULTANEOUS VIDEO RETRIEVAL AND ALIGNMENT SYSTEM

Because untrimmed videos with various content interwoven (meaningful or not) are inevitably found in real life as well as within the task, a method for handling them requires an additional process of suggesting meaningful candidates to reduce complexity. To this end, our baseline, Simultaneous

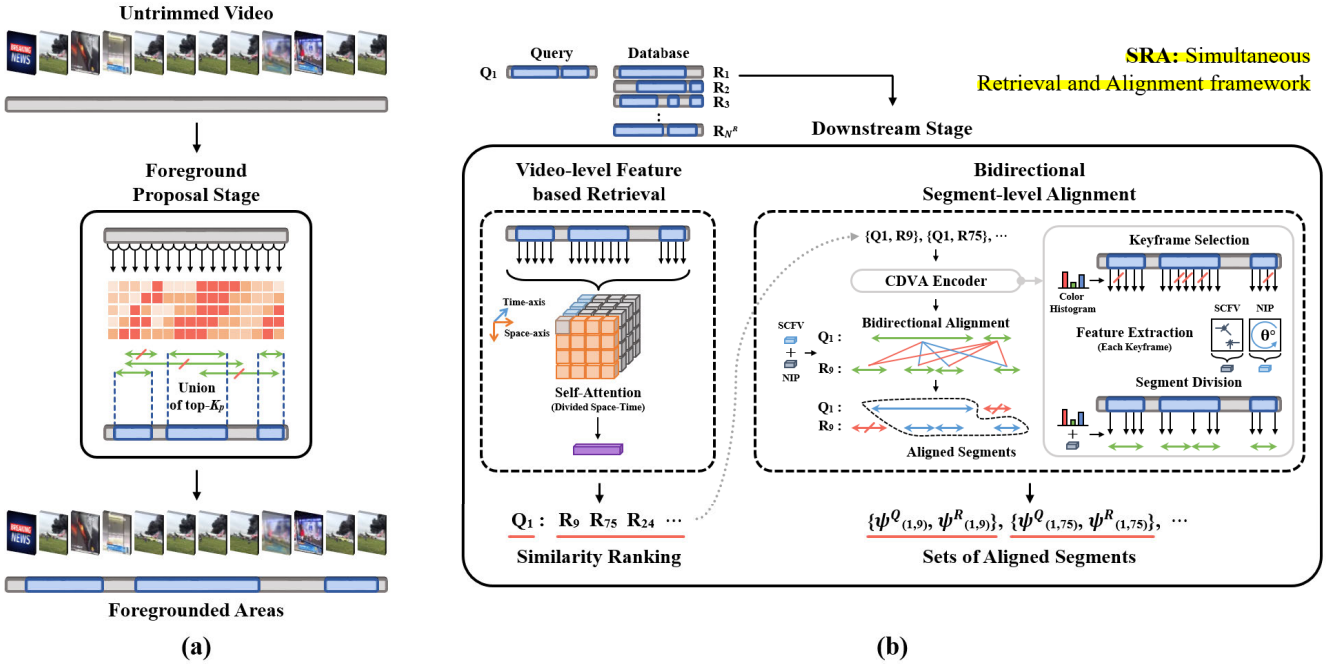


FIGURE 4. Structure of the proposed framework: (a) represents the foreground proposal stage for processing untrimmed video, and (b) represents the downstream stage that handles retrieval and alignment at the same time.

video Retrieval and Alignment framework (SRA), uses two-stages consisting of a foreground proposal stage and a downstream stage. The foreground proposal stage offers areas that appear to be in the foreground, and the downstream stage handles each task within these areas. Consequently, as shown in Fig. 4, the two-stage approach is designed to search for related video pairs and find content-related segment pairs.

A. FOREGROUND PROPOSAL STAGE

The purpose of this stage is to localize likely-meaningful areas in an untrimmed video entangled with multiple contents. In an untrimmed video representing a specific event, unrelated areas induced by editing effects or distractors are also included. For example, an area in which only logos and/or text appear at the introduction of news, as shown in Fig. 1(a), is far from the content implied by a specific event. During a fade-in/fade-out sequence upon a change of scene, as shown in Fig. 1(b), completely different scenes are mixed, complicating the recognition of one event. Also, because one event consists of various interactions between objects and scenes, the content are commonly seen in any video, such as the ''explanation at the news desk'' scene in the video related to ''forest fire'' shown in Fig. 1(c), interfering with distinctive representation.

Our approach, which eliminates ambiguous or non-relevant parts in untrimmed videos and screens likely-meaningful areas, i.e., the foreground, is inspired by the temporal action proposal method. Specifically, the key idea of this stage is to find temporal segments that are either shot transitions or have low relevance to video-level topics, similar to an earlier protocol [17]. Given an untrimmed video, snippet-level features f^{sni} are described from each non-overlapping snippet of size T_{sni} via the backbone network. Proposals are generated through continuous relationships between each snippet-level feature, and the confidence is computed for each proposal. At this time, we follow the aforementioned protocol [17], but confidences are learned to represent the relevance to a specific event by utilizing pre-defined event-dependent information in the target dataset rather than focusing solely on the action, as before. Thereafter, the union of proposals corresponding to the top-K_p of confidence is assigned as the foreground. The foreground indirectly supports the downstream stage to produce the retrieval and alignment outcomes.

B. DOWNSTREAM STAGE

1) VIDEO-LEVEL FEATURE-BASED RETRIEVAL

Video-level features are popularly used for the task of video retrieval due to their efficiency and low time complexity, but they should be able to represent a video effectively even when the length of the video is long. With the emergence of self-attention modules, Transformers have been shown to be successful at describing global context information in various computer vision tasks, whereas convolutional neural networks fundamentally fail to capture that information.

Bertasius et al. [35] present a convolution-free approach to video classification built exclusively on self-attention over space and time. TimeSformer [35] adapts the standard transformer architecture for videos by enabling spatio-temporal features learned directly from a sequence of

frame-level patches. However, because the network can only handle pre-defined classes, we redesign the objective to allow distance learning based on similarities by focusing on the internal long-term video representation capability of the model.

In detail, frames as an input are uniformly sampled from an untrimmed video. The frames are fed into the network, i.e., TimeSformer, and video-level intermediate features f^{inter} are extracted. In this way, intermediate features are calculated for a triplet $(f^{inter}_{(\cdot)}, f^{inter}_{(+)}, f^{inter}_{(-)})$ from an anchor, a positive, and a negative, respectively. Based on the anchor in the triplet, the positive is a video representing the same event at the video level, and the negative refers to a video representing the other event. As indicated in Eq. 4, the network is optimized through the triplet ranking loss [36] $\mathcal{L}_v$, where $simf^{inter}_{(\cdot,\cdot)}$ denotes the cosine similarity between two features corresponding to the subscript below, and γ is the margin. Here, $simf^{inter}_{(\cdot,\cdot)}$ denotes the similarity; hence, the greater $simf^{inter}_{(\cdot,-)}$ and smaller $simf^{inter}_{(\cdot,+)}$ both increase the loss. Through the structure of the network, which has less receptive field restriction along the time axis, and with event-wise distance learning, the network can capture a long-term feature representation for an event in a given video. For the inference process, $Score_v$ is calculated as the cosine similarity between video-level features from the query-database pairs.

$$\mathcal{L}_v = \max\left\{0, simf^{inter}_{(\cdot,-)} - simf^{inter}_{(\cdot,+)} + \gamma\right\} \tag{4}$$

2) BIDIRECTIONAL SEGMENT-LEVEL ALIGNMENT

Segment-level alignment mainly intends to find where a particular topic in one video is located within another. Compact Descriptors for Video Analysis (CDVA) [12], [13], [14] from the Moving Pictures Experts Group (MPEG) also supports segment-level alignment, assuming unidirectional matching. CDVA [11] was adopted as a standard due to the effective processes that allow the identification of similar content through two types of frame-level features extracted without additional learning. However, unlike previous one-way alignment methods, as noted in Section I, it should be possible to localize segments representing a video-level topic in both directions between untrimmed video pairs.

To solve the bidirectional segment-level alignment problem, we partially exploit the structure of the standardized CDVA [13]. First, to avoid the redundancy of frame-level features, color histogram-based keyframe selection is performed according to CDVA. After selecting non-duplicated keyframes in the order of time, two types of features, Scalable Compressed Fisher Vector (SCFV) and Nested Invariance Pooling (NIP), are extracted for each keyframe. SCFV is described as accumulating local attributes in a frame to explore details, and NIP is described as aggregating global attributes to be robust to image transformations. Then, through SCFV similarity and color histogram difference between consecutive keyframes, segments composed of the same contents are grouped into a single video. When calculating the segment-level alignment between two videos, segment-level similarities are computed one-by-one between the videos, and whether to align is determined by thresholding those similarities. In this case, the segment-level similarity is obtained as the maximum value of keyframe-level similarities belonging to them, and the keyframe-level similarities are determined by the sum of SCFV and NIP similarities. After that, CDVA determines whether it is aligned or not only in unidirection and produces a single-segment outcome by collecting the aligned segments. On the other hand, our approach employs each individual aligned segment (i.e., ψ^Q and ψ^R) as a result after determining whether it is aligned in both directions.

C. TRAINING AND INFERENCE

The foreground proposal stage and the video-level feature-based retrieval of the downstream stage are both optimized in the manner explained above, and the bidirectional segment-level alignment of the downstream stage is employed without extra learning. In inference, the foreground proposal stage is performed first. The areas, a union of the top-K_p, are allocated as foregrounds. The foregrounds are then handed down to the downstream stage, which proceeds in the sequence of alignment following retrieval. The retrieval process embeds frames sampled from foreground areas for videos in the query and the database to a one-dimensional vector per video via the event-wise distance-learned network, and the videos in the database are ranked by measuring the similarities to each query. During the alignment process, the query-database videos are paired in the order of the relevant ranks, and the segment-level alignment is performed in the manner explained before by selecting keyframes from the foregrounds of the two videos belonging to the pair. This sequence of returning similar videos using temporal cues and assigning related positions between them can not only place two directly relevant videos in a video-level topic into the alignment candidate group but can also converge on optimal efficiency if the range of alignment is constrained by the relevant ranks derived from the retrieval process when considering online.

V. EXPERIMENTS

A. EVALUATION SETUP

1) FIVR

This is an incident-centric dataset [6] containing three retrieval tasks: Duplicate Scene Video Retrieval (DSVR), Complementary Scene Video Retrieval (CSVR), and Incident Scene Video Retrieval (ISVR), with the importance of semantic information increasing as one progresses from DSVR to ISVR. Among this family, FIVR-5K is employed for the retrieval evaluation, which consists of 5,000 videos in the database and 50 query videos with video-level annotations. Also, the proposed FIVR+A-5K, a form in which segment-level annotations between query videos and the database are added to FIVR-5K, is used for the simultaneous evaluation of the retrieval and alignment outcomes. In the former case, how well similar videos are found is assessed via a prominent

metric called mAP [37], and in the latter case, how well the locations of related segments, including similar videos, are also found, is assessed via the proposed metric mAP@J. For the evaluation of alignment, the **N**, **S**, and **H** conditions are considered the ground truth segments.

2) ACTIVITYNET

This is an action localization dataset [34] originally consisting of 10,024 training videos, 4,926 validation videos, and 5,044 test videos with annotations containing exact video segments that correspond to specific actions. Based on the original dataset, 494 videos from the test videos were reorganized for the evaluation [8]. For the simultaneous evaluation of two tasks previously not considered, any pair with one common action label regards that segments representing that label in the pair are aligned while also defining the pair as relevant at the video level. Similar to the previous dataset, mAP and mAP@J are used as the evaluation metrics.

3) VCDB

This is a copy detection dataset [18] consisting of 528 videos classified into 28 query sets with approximately 9,000 copied segment pairs. It is evaluated whether a segment visually similar to a predefined segment of one video is well aligned within another video. Because most segments are easy to detect due to relatively simple spatial and temporal transformations, this dataset is considered to address a simplified video alignment task. To assess the alignment process of the proposed approach, a segment-level best F1-score [18] is used as an evaluation metric. For a reasonable comparison with the same metric, only methods that are identified by paper or contact as measuring the performance with the best F1-score via the segment-level precision-recall curve are included in the benchmark.

B. IMPLEMENTATION DETAILS

1) FIVR

In the foreground proposal stage, the snippet-level feature f^{sni} is extracted through the backbone, TSP [38]. T_{sni} denotes 16 frames with two-frame intervals. For training, segment-level annotations in the database that exclude the queries used for the test of FIVR+A-5K are utilized and considered to be known in the real world. This means that only 1,931 of the 5,000 videos in the database are used to discern the foreground candidates. For the evaluation, foregrounds are also predicted, including all unseen videos with queries that are not used for learning. For video-level feature-based retrieval in the downstream stage, parameters pre-trained from earlier work [39] are employed without additional fine-tuning for a reasonable comparison with other approaches. In addition, as an input to the network, frames from a single video are selected based on a frame rate of 1/32, and then 32 of them are linearly sampled. If used with the foreground proposal stage, frames that do not belong to the foreground are excluded while selecting frames in a video with the same frame rate. The remaining frames are then arranged in time-axis order, and 32 of them are linearly sampled as an input to the network. For bidirectional segment-level alignment, all detailed settings follow when the operating point of CDVA is 256 KBps without extra training. If used with the foreground proposal stage, the color histogram-based keyframe selection works only for frames that belong to the foreground.

TABLE 1. A benchmark for the video retrieval task on FIVR-5K and ActivityNet. * is the reimplemented performance of other methods.

	Video Retrieval				
		FIVR-5K			ActivityNet
	Method	mAP			mAP
		DSVR	CSVR	ISVR	
frame	VReL [8]	-	-	-	0.209
	DP [4]	-	-	-	0.621
	TN [5]	-	-	-	0.648
	ViSiL$_f$ [10]	0.838	0.832	0.739	0.652
	ViSiL$_{sym}$ [10]	0.830	0.823	0.731	**0.745**
	ViSiL$_v$ [10]	**0.880**	**0.869**	**0.777**	0.710
	TCA$_f$ [16]	0.844	0.834	0.763	-
	TCA$_{sym}$ [16]	0.763	0.766	0.711	-
	TCA$_v$ [16]	0.726	0.735	0.701	-
	CDVA [13]	0.813	0.781	0.673	0.235
	CDVA$_{\text{TNIP}}$ [28]	**0.880**	0.862	0.744	-
video	DML [15]	0.391*	0.399*	0.380*	0.705
	TCA$_c$ [16]	0.609	0.617	0.578	-
	SRA(ours)	**0.700**	**0.719**	**0.705**	**0.800**

TABLE 2. A benchmark for the video alignment task on VCDB.

Video Alignment	
	VCDB
Method	F1-score
Jiang *et al.* [19]	0.650
TMK [21]	0.674
LAMV [22]	0.687
BTA [31]	0.755
FPVCD [32]	**0.861**
SRA(ours)	**0.792**

2) ACTIVITYNET

In the foreground proposal stage and the downstream stage, the structure of the network and settings are identical to those in FIVR. For bidirectional segment-level alignment of the downstream stage, no further learning is performed as above. All others, however, are fine-tuned in the aforementioned training set [8]. This is identical to the process used in other approaches, as it guarantees unseen action classes of the test set, which are not overlaid with those of the training set.

3) VCDB

In this dataset, performances are measured only by the bidirectional segment-level alignment of the downstream stage, which does not require learning because all video pairs with segment-level annotations are utilized for evaluation without a separate training set. All other detailed settings are the same as above.

TABLE 3. A benchmark for the new task that simultaneously addresses video retrieval and alignment on FIVR-5K and ActivityNet. $SRA_{\{\cdot\}}$ indicates that other retrieval approaches replace our retrieval approach in the SRA. It is the reimplemented performance of other approaches for which the source code was disclosed, and the non-disclosed approaches were excluded for fairness.

Video Retrieval & Alignment					
		FIVR+A-5K			ActivityNet
	Method	mAP@J			mAP@J
		DSVR	CSVR	ISVR	
frame	$SRA_{\{ViSiL_f\}}$	0.646	0.639	0.570	-
	$SRA_{\{ViSiL_{sym}\}}$	-	-	-	0.206
	$SRA_{\{ViSiL_v\}}$	**0.675**	**0.667**	**0.590**	-
	CDVA	0.629	0.606	0.516	0.091
video	$SRA_{\{TCA_c\}}$	0.496	0.504	0.459	-
	SRA(ours)	**0.551**	**0.566**	**0.537**	**0.213**

TABLE 4. Ablations in the ISVR task on FIVR-5K and FIVR+A-5K: (a) whether the foreground proposal is used, (b) the top-K_p used to select the foreground, and (c) our multi-segment alignment approach changed from CDVA.

K_p	mAP
1	**0.705**
3	0.688
5	0.663

(a) K_p

Foreground	mAP
-	0.505
✓	**0.705**

(b) Foreground

Segment	mAP@J
Single	0.535
Multi	**0.537**

(c) Multi-Segment

C. COMPARISON WITH OTHER METHODS

1) VIDEO RETRIEVAL

The proposed method, SRA, is evaluated to benchmark the retrieval performance on the FIVR-5K and the reorganized ActivityNet. First, the proposed method shows the best results when compared to methods using video-level features on the two datasets, as indicated in Table 1. This can be seen as reducing the difficulty of representing a single video as a feature by sampling frames only in areas that may be meaningful in an untrimmed video and using the transformer structures to cover temporally large areas. For these reasons, when compared to methods using frame-level features, our method matches the performance of several methods on the FIVR-5K, even achieving a state-of-the-art outcome on the ActivityNet. In addition, unlike these frame-level methods, which calculate multiple features per video, only one is computed in our case. Hence, if the actual search situation is assumed, it can benefit from the relatively improved cost efficiency with comparable retrieval capabilities.

2) VIDEO ALIGNMENT

Table 2 describes how well the alignment process of the proposed method allocates on the VCDB. Although only one of the bidirectional outcomes is used for a reasonable evaluation, the proposed method is placed second, suggesting that the outcomes are well aligned. And, unlike the state-of-the-art (FPVCD) [32], which requires learning, the proposed method does not require that, so it performs the best among this type of method. Therefore, this indicates that the proposed alignment process can efficiently support the entire SRA framework, which deals with various tasks.

3) VIDEO RETRIEVAL AND ALIGNMENT

We show the first simultaneous video retrieval and alignment benchmark in Table 3 because this is the first proposed; $SRA_{\{\cdot\}}$, which replaces other retrieval approaches in our retrieval approach, is intended to show a variety of comparative examples. In addition, the performance of CDVA retrieval, which is based on frame-level features, and CDVA matching, which generates only one segment as a localization output, is also included in the benchmark for comparison. At this time, in order to obtain bidirectional localization outputs from CDVA, we repeat the experiment by changing the position of the pair of entered inputs. Overall, because it is weighted by the alignment approach of SRA, it shows a similar tendency when only retrieval is evaluated. If a method for dealing with retrieval is proposed while resolving the bidirectional alignment in future work, these examples will be used for comparison.

D. ABLATION STUDIES

1) TOP-K_p

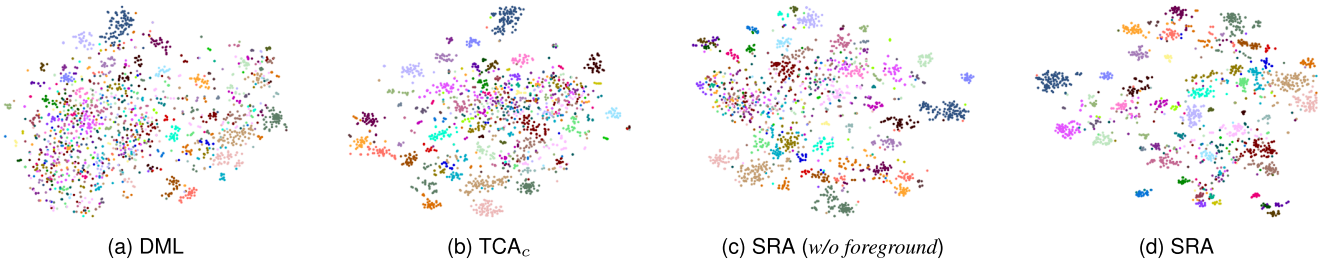
The performance according to top-K_p to determine the foreground is presented in Table 4(a). Because the foreground consists of a union of top-K_p ordered by confidence in the proposals, which are assumed to be meaningful areas, the larger K_p is, the more likely the foreground is to include scenes in which editing effects occur (e.g., fade in/out) or that generally appear in any video. Hence, the probability of containing scenes as a distractor increases, which can interfere with discrimination and cause a drop in the performances. With these results, we set K_p equal to 1, ensuring that the most reliable foreground is provided in the downstream stage.

2) FOREGROUND

Table 4(b) illustrates numerical benefits gained from the foreground proposal stage for efficiently processing untrimmed videos. Utilizing the foreground in our framework yields an improvement of 0.2 in terms of retrieval capability, suggesting that ambiguous or unrelated regions are successfully eliminated from untrimmed videos as intended.

3) MULTI-SEGMENT

Table 4(c) shows the effect of individual segment alignment (called multi-segment) in SRA alignment approach, unlike the existing CDVA framework, which uses per-segment localization. When adopting the change-previous approach (not multi-segment), the process is repeated as described in the benchmark above Section V-C3 to obtain bidirectional results. The conventional approach resulting in one segment involves more noise than the proposed approach resulting in a multi-segment outcome because it forces unaligned regions



(a) DML (b) TCA$_c$ (c) SRA (*w/o foreground*) (d) SRA

FIGURE 5. t-SNE results for the video-level feature in the ISVR task on the FIVR-5K: (a) and (b) visualized results of video-level feature-based state-of-the-art methods, DML and TCA$_c$, respectively; (c) visualized result of the video-level feature used by the retrieval approach of our SRA without the foreground proposal stage; and (d) the result with that stage included.

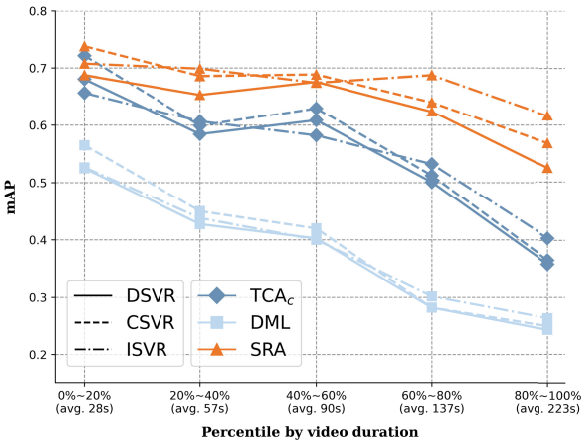


FIGURE 6. According to the video duration, the retrieval capacities of methods using video-level features on FIVR-5K. The x-axis represents the case when dividing the database into five subsets depending on the percentile of the video duration and using each of them as a database.

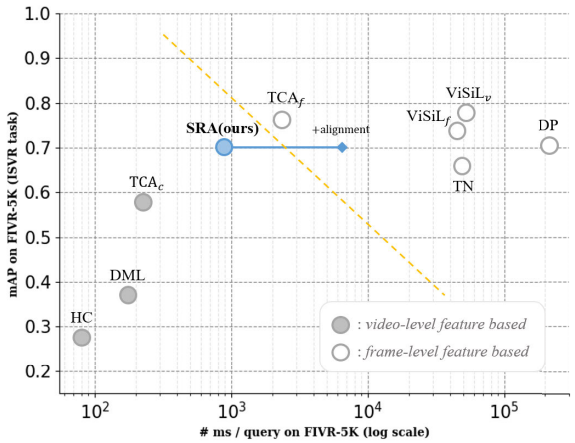


FIGURE 7. Trade-off between retrieval performance and efficiency during the ISVR task on the FIVR-5K. The inference time for the query shown on the x-axis is referenced from earlier work [16].

in one video to be allocated within the output. For this reason, our approach exhibits higher performance due to the relatively high Jaccard weight, as it contains less of this type of noise.

E. ANALYSES

1) LONG-TERM FEATURE REPRESENTATION

In this section, the long-term representation capability of SRA is analyzed. SRA goes through the foreground proposal stage for processing untrimmed videos that contain unnecessary contents, thereby aiming for robustness on the distractor. In addition, video-level features are extracted via the transformer structure with less receptive field restriction on the time axis, aiming to boost the long-term representation for the retrieval.

Fig. 5, which describes the visualization results of the video-level feature, shows that the proposed approach qualitatively achieved those goals. Compared to the two current state-of-the-art methods, DML and TCA$_c$, SRA employing only the transformer structure (without the foreground proposal stage) displays comparable representations in Fig. 5(c). For the intact SRA, which includes the foreground proposal stage as shown in Fig. 5(d), it reveals remarkable discrimination compared to the others.

Another experiment shown in Fig. 6 illustrates the ability to retrieve according to the video duration. In this experiment, the database in the FIVR-5K is divided into five subsets according to the percentile of the video duration, and retrieval performances are measured using each of these subsets as a database. Unlike DML and TCA$_c$, where a large drop of approximately 0.1 over the highest performance begins from the 20% to 40% percentile interval (an average of 57 seconds), SRA shows a similar drop in the 60% to 80% percentile interval (an average of 137 seconds), in which longer videos belong. In addition, compared to the performance in the 0% to 20% percentile interval (an average of 28 seconds), where the shortest videos belong, the performance in the 80% to 100% percentile interval (an average of 223 seconds), where the longest videos belong, falls by about 0.3 in the two other approaches, whereas SRA case shows a smaller drop of about 0.15.

As a result of the above two experiments, with the proposed approach, longer videos are effectively handled by the foreground proposal process and convolution-free architecture for the retrieval task.

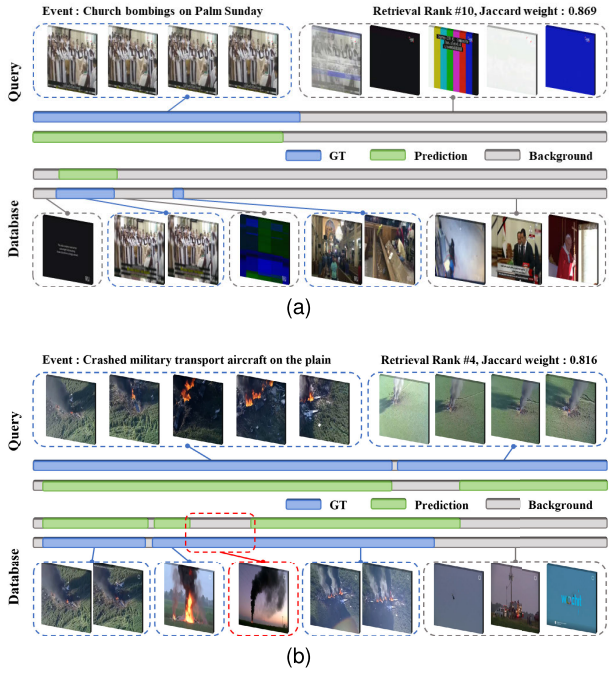


FIGURE 8. Qualitative examples of alignment on FIVR+A-5K.

2) PERFORMANCE VS EFFICIENCY

The trade-off between the performance and efficiency of SRA is explained in Fig. 7. Compared to TCA_c, a video-level feature-based method, SRA ranks highest with a margin of 0.127 in terms of retrieval performance, although the inference speed is nearly four times slower. In contrast, compared the frame-level feature-based methods, the outcome for SRA is as low as 0.058 compared to that of TCA_f, but with a faster inference speed by about 1.5 times and nearly 60 times faster than the state-of-the-art method, $ViSiL_v$, although the result was as low as 0.072. In particular, even when the proposed approach goes through the bidirectional alignment process, the speed is eight times faster than that of $ViSiL_v$. These results imply that the proposed SRA is in the most efficient position in terms of the trade-off relationship between performance and speed.

3) QUALITATIVE EXAMPLES OF ALIGNMENT

Qualitative examples of alignment, included in SRA framework, are described in Fig. 8. As mentioned earlier, the bidirectional alignment is performed in the query and database, which consist of untrimmed videos, which subsequently penalizes video-level ranking with the Jaccard weight. Fig. 8(b), one of the examples, shows the alignment outcomes in a query-database pair connected based on the video-level topic "Church bombings on Palm Sunday". Because the two videos in the pair contain a lot of content that is not relevant to the video-level topic, the bidirectional video alignment is required to determine which parts of the videos are related. The examples show that the proposed approach can filter these distractors very finely. Fig. 8(b), another example, shows the alignment outcomes in a query-database pair connected based on the video-level topic "Crashed military transport aircraft on the plane". In this case, our approach successfully allocates segments even when a scene representing the event is filmed at an angle different from that of the query. However, for a scene captured at a different time and angle from the query in the red-dotted area, our approach fails to grasp the semantic relationship. This implies that it is needed to understand a more semantic relationship in future studies.

VI. CONCLUSION

In this paper, we present a novel simultaneous video retrieval and alignment task that retrieves videos relevant to a given query and aligns related locations at the segment level, as well as a solution to it. Because previous approaches focused on either the video-to-video retrieval problem without considering how to localize long and untrimmed videos with multiple contents or on the video alignment problem while oversimplifying without considering the numerous practically irrelevant videos in the database, their applicability to real-world situations was inevitably limited. To evaluate and compare the performances of the proposed solution with those of other methods on this task, we present a metric and a dataset with necessary annotations that is now open to the public. We believe that our work defines the complex real-world video retrieval problem more practically and presents a viable solution while also demonstrating outstanding performance.

REFERENCES

[1] T. Acosta, P. Acosta-Vargas, J. Zambrano-Miranda, and S. Lujan-Mora, "Web accessibility evaluation of videos published on YouTube by world-wide top-ranking universities," *IEEE Access*, vol. 8, pp. 110994–111011, 2020.
[2] S. Iqbal, A. N. Qureshi, and A. M. Lodhi, "Content based video retrieval using convolutional neural network," in *Proc. SAI Intell. Syst. Conf.*, 2018, pp. 170–186.
[3] X. Wu, C.-W. Ngo, A. G. Hauptmann, and H.-K. Tan, "Real-time near-duplicate elimination for web video search with content and context," *IEEE Trans. Multimedia*, vol. 11, no. 2, pp. 196–207, Feb. 2009.
[4] C.-L. Chou, H.-T. Chen, and S.-Y. Lee, "Pattern-based near-duplicate video retrieval and localization on web-scale videos," *IEEE Trans. Multimedia*, vol. 17, no. 3, pp. 382–395, Mar. 2015.
[5] H.-K. Tan, C.-W. Ngo, R. Hong, and T.-S. Chua, "Scalable detection of partial near-duplicate videos by visual-temporal consistency," in *Proc. 17th ACM Int. Conf. Multimedia*, Oct. 2009, pp. 145–154.
[6] G. Kordopatis-Zilos, S. Papadopoulos, I. Patras, and I. Kompatsiaris, "FIVR: Fine-grained incident video retrieval," *IEEE Trans. Multimedia*, vol. 21, no. 10, pp. 2638–2652, Oct. 2019.
[7] J. Revaud, M. Douze, C. Schmid, and H. Jegou, "Event retrieval in large video collections with circulant temporal encoding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2459–2466.
[8] Y. Feng, L. Ma, W. Liu, T. Zhang, and J. Luo, "Video re-localization," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 51–66.
[9] M. Douze, J. Revaud, J. Verbeek, H. Jégou, and C. Schmid, "Circulant temporal encoding for video retrieval and temporal alignment," *Int. J. Comput. Vis.*, vol. 119, no. 3, pp. 291–306, Sep. 2016.
[10] G. Kordopatis-Zilos, S. Papadopoulos, I. Patras, and Y. Kompatsiaris, "ViSiL: Fine-grained spatio-temporal video similarity learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6351–6360.
[11] *Call for Proposals for Compact Descriptors for Video Analysis*, ISO/IEC JTC1/SC29/WG11/N15339, 2015.

[12] Y. Lou, Y. Bai, J. Lin, S. Wang, J. Chen, V. Chandrasekhar, L.-Y. Duan, T. Huang, A. C. Kot, and W. Gao, "Compact deep invariant descriptors for video retrieval," in *Proc. Data Compress. Conf. (DCC)*, Apr. 2017, pp. 420–429.

[13] J. Lin, L.-Y. Duan, S. Wang, Y. Bai, Y. Lou, V. Chandrasekhar, T. Huang, A. Kot, and W. Gao, "HNIP: Compact deep invariant representations for video matching, localization, and retrieval," *IEEE Trans. Multimedia*, vol. 19, no. 9, pp. 1968–1983, Sep. 2017.

[14] L.-Y. Duan, Y. Lou, Y. Bai, T. Huang, W. Gao, V. Chandrasekhar, J. Lin, S. Wang, and A. C. Kot, "Compact descriptors for video analysis: The emerging MPEG standard," *IEEE MultimediaMag.*, vol. 26, no. 2, pp. 44–54, Apr. 2019.

[15] G. Kordopatis-Zilos, S. Papadopoulos, I. Patras, and Y. Kompatsiaris, "Near-duplicate video retrieval with deep metric learning," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 347–356.

[16] J. Shao, X. Wen, B. Zhao, and X. Xue, "Temporal context aggregation for video retrieval with contrastive learning," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 3268–3278.

[17] T. Lin, X. Liu, X. Li, E. Ding, and S. Wen, "BMN: Boundary-matching network for temporal action proposal generation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3889–3898.

[18] Y.-G. Jiang, Y. Jiang, and J. Wang, "VCDB: A large-scale database for partial copy detection in videos," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 357–371.

[19] Y.-G. Jiang and J. Wang, "Partial copy detection in videos: A benchmark and an evaluation of popular methods," *IEEE Trans. Big Data*, vol. 2, no. 1, pp. 32–42, Mar. 2016.

[20] M. Douze, H. Jegou, and C. Schmid, "An image-based approach to video copy detection with spatio-temporal post-filtering," *IEEE Trans. Multimedia*, vol. 12, no. 4, pp. 257–266, Jun. 2010.

[21] S. Poullot, S. Tsukatani, A. P. Nguyen, H. Jégou, and S. Satoh, "Temporal matching kernel with explicit feature maps," in *Proc. 23rd ACM Int. Conf. Multimedia*, Oct. 2015, pp. 381–390.

[22] L. Baraldi, M. Douze, R. Cucchiara, and H. Jegou, "LAMV: Learning to align and match videos with kernelized temporal layers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7804–7813.

[23] C. Jiang, K. Huang, S. He, X. Yang, W. Zhang, X. Zhang, Y. Cheng, L. Yang, Q. Wang, F. Xu, T. Pan, and W. Chu, "Learning segment similarity and alignment in large-scale content based video retrieval," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 1618–1626.

[24] Z. Han, X. He, M. Tang, and Y. Lv, "Video similarity and alignment learning on partial video copy detection," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 4165–4173.

[25] L. Zhen, P. Hu, X. Wang, and D. Peng, "Deep supervised cross-modal retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10394–10403.

[26] H. Zhang, A. Sun, W. Jing, L. Zhen, J. T. Zhou, and R. S. M. Goh, "Natural language video localization: A revisit in span-based question answering framework," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 8, pp. 4252–4266, Aug. 2022.

[27] H. Zhang, A. Sun, W. Jing, G. Nan, L. Zhen, J. T. Zhou, and R. S. M. Goh, "Video corpus moment retrieval with contrastive learning," in *Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2021, pp. 685–695.

[28] W. Jo, G. Lim, J. Kim, J. Yun, and Y. Choi, "Exploring the temporal cues to enhance video retrieval on standardized CDVA," *IEEE Access*, vol. 10, pp. 38973–38981, 2022.

[29] G. Kordopatis-Zilos, S. Papadopoulos, I. Patras, and Y. Kompatsiaris, "Near-duplicate video retrieval by aggregating intermediate CNN layers," in *Proc. Int. Conf. Multimedia Modeling* Cham, Switzerland: Springer, 2017, pp. 251–263.

[30] J. Song, Y. Yang, Z. Huang, H. T. Shen, and J. Luo, "Effective multiple feature hashing for large-scale near-duplicate video retrieval," *IEEE Trans. Multimedia*, vol. 15, no. 8, pp. 1997–2008, Dec. 2013.

[31] Y. Zhang and X. Zhang, "Effective real-scenario video copy detection," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 3951–3956.

[32] W. Tan, H. Guo, and R. Liu, "A fast partial video copy detection using KNN and global feature database," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2022, pp. 2191–2199.

[33] P. Jaccard, "The distribution of the flora in the Alpine zone. 1," *New Phytologist*, vol. 11, no. 2, pp. 37–50, Feb. 1912.

[34] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles, "ActivityNet: A large-scale video benchmark for human activity understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 961–970.

[35] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?" in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2021, pp. 1–4.

[36] V. Balntas, E. Riba, D. Ponsa, and K. Mikolajczyk, "Learning local feature descriptors with triplets and shallow convolutional neural networks," in *Proc. Brit. Mach. Vis. Conf.*, 2016, pp. 1–3.

[37] M. Zhu, "Recall, precision and average precision," Dept. Statist. Actuarial Sci., Univ. Waterloo, Waterloo, ON, Canada, 2004.

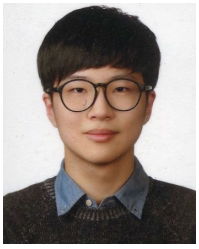
[38] H. Alwassel, S. Giancola, and B. Ghanem, "TSP: Temporally-sensitive pretraining of video encoders for localization tasks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 3173–3183.

[39] A. Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev, and J. Sivic, "HowTo100M: Learning a text-video embedding by watching hundred million narrated video clips," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2630–2640.

WON JO received the B.S. degree from the Department of Intelligent Mechatronics, Sejong University, Seoul, Republic of Korea, in 2021, and the M.S. degree from the Department of Artificial Intelligence, Sejong University, in 2023. He is currently a Researcher with the Robotics and Computer Vision (RCV) Laboratory, Sejong University. His current research interests include computer vision, multimedia, and machine learning, with a particular emphasis on representation learning for video retrieval and alignment. He is also interested in active learning and self-supervised learning, which are efficient learning solutions.

GEUNTAEK LIM received the B.S. degree from the Department of Intelligent Mechatronics, Sejong University, Seoul, Republic of Korea, in 2023, where he is currently pursuing the M.S. degree with the Robotics and Computer Vision (RCV) Laboratory. His current research interests include computer vision and deep learning, with a particular emphasis on representation learning for video understanding.

YUJIN HWANG received the B.S. degree from the Department of Intelligent Mechatronics, Sejong University, Seoul, Republic of Korea, in 2022, where she is currently pursuing the M.S. degree with the Robotics and Computer Vision (RCV) Laboratory. Her research interests include computer vision, active learning, and self-supervised learning.

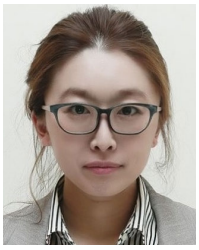


GWANGJIN LEE received the B.S. degree from the Department of Computer and Information Security, Sejong University, Seoul, Republic of Korea, in 2022. He is currently pursuing the M.S. degree with the Robotics and Computer Vision (RCV) Laboratory, Department of Intelligent Mechatronics, Sejong University. His current research interests include computer vision, machine learning, and video retrieval and alignment.

JOONSOO KIM received the B.S. and Ph.D. degrees from the Department of Electrical Engineering, Seoul National University (SNU), Seoul, Republic of Korea, in 2012 and 2017, respectively. He has been with the Electronics and Telecommunications Research Institute (ETRI), since 2017, where he is currently a Senior Member of the Research Staff with the Immersive Media Research Section. His current research interests include light field displays, autostereoscopic 3-D displays, and virtual reality systems.

JOUNGIL YUN received the B.S. degree from the Department of Control and Instrumentation Engineering, Jeonbuk National University, Jeonju, Republic of Korea, in 1996, and the M.S. and Ph.D. degrees from the Department of Mechatronics, Gwangju Institute of Science and Technology (GIST), Gwangju, Republic of Korea, in 1998 and 2005, respectively. Since 2005, he has been with the Electronics and Telecommunications Research Institute (ETRI), Daejeon, Republic of Korea. He is currently a Principal Researcher with the Media Research Division, ETRI. His current research interests include immersive media processing and video coding for machines.

JIYOUNG JUNG received the B.S., M.S., and Ph.D. degrees from the Department of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Republic of Korea, in 2008, 2010, and 2016, respectively. She was with NAVER Labs, South Korea, for a year and joined the Department of Software Convergence, Kyung Hee University, Seoul, Republic of Korea, as an Assistant Professor, in 2017. She has been an Assistant Professor with the Department of Artificial Intelligence, University of Seoul, Seoul, since 2021. Her research interests include understanding 3-D environments, color and depth sensor systems, photometric stereo, and light modeling.



YUKYUNG CHOI (Member, IEEE) received the B.S. degree from the Department of Information and Communication Electronics Engineering, Soongsil University, Seoul, Republic of Korea, in 2006, the M.S. degree from the Department of Electrical and Electronic Engineering, Yonsei University, Seoul, in 2008, and the Ph.D. degree in electrical and electronic engineering/robotics program from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Republic of Korea. She has been an Assistant Professor with the Department of Intelligent Mechatronics Engineering, Sejong University, Seoul, since 2018. She is currently the Director of the Robotics and Computer Vision (RCV) Laboratory. Her research interests include computer vision and robotics.