IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 25, 2022, accepted March 30, 2022, date of publication April 11, 2022, date of current version April 15, 2022.

*Digital Object Identifier 10.1109/ACCESS.2022.3165177*

# Exploring the Temporal Cues to Enhance Video Retrieval on Standardized CDVA

**WON JO[1], GUENTAEK LIM[2], JOONSOO KIM[3], JOUNGIL YUN[3], 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

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) and Grant 2021-0-02067 (Next Generation AI for Multi-purpose Video Search).

**ABSTRACT** As the demand for large-scale video analysis increases, video retrieval research is also becoming more active. In 2014, ISO/IEC MPEG began standardizing compact descriptors for video analysis, known as CDVA, and it is now adopted as a standard. However, the standardized CDVA is not easily compared to other methods because the MPEG-CDVA dataset used for performance verification is not disclosed, despite the fact that follow-up studies are underway with multiple versions of the CDVA experimental model. In addition, analyses of modules constituting the CDVA framework are insufficient in previous studies. Therefore, we conduct self-evaluations of CDVA to analyze the impact of each module on the retrieval task. Furthermore, to overcome the obstacles identified through these self-evaluations, we suggest temporal nested invariance pooling, abbreviated as TNIP, which implies temporal robustness realized by improving nested invariance pooling, abbreviated as NIP, one of the features in CDVA. Finally, benchmarks of the existing CDVA and the proposed approach are provided on several public datasets. Through this, we show that the CDVA framework is capable of boosting the retrieval performance if utilizing the proposed approach.

**INDEX TERMS** Content based retrieval, information representation, MPEG standards.

## I. INTRODUCTION

The demand for large-scale video analysis has increased in recent years, and significant research and improvements have been accomplished, particularly in the area of action recognition as part of the video classification problem. These advancements have resulted in an increasing number of classes of trimmed video-based datasets, as well as some untrimmed video-based datasets. However, this tendency causes different interpretations or ambiguities across classes, making it difficult to discern between videos. Accordingly, this has led to an increased need for content-based video retrieval (CBVR) [1] research that does not define classes for specific actions or situations but rather identifies relevant videos.

CBVR uses text sources, audio sources, and visual sources as query types, and many approaches focusing on CBVR are being developed. Among them, video-to-video retrieval (hereinafter referred to as video retrieval) utilizing visual sources is broadening its research scope from near-duplicate video retrieval (NDVR) [2] to fine-grained incident video retrieval (FIVR) [3], requiring more strengthened representations to distinguish between more complicated scenes.

To this end, the moving picture experts group (MPEG) has performed large-scale video analysis through standardization of compact descriptors for video analysis (CDVA) [4], and this approach has been effective for video retrieval through several versions of the CDVA experimental model (CXM). However, the MPEG-CDVA dataset [4] used to measure the performance of CXM is not publicly available, making comparisons with other methods difficult. Furthermore, because there has been a lack of analysis of how each module constituting the CDVA framework affects video retrieval, the advancement of subsequent studies utilizing it has been slowed. There is also a technological limitation in that temporal contexts cannot be encoded because the video retrieval task is cast with a keyframe-based image retrieval task.

In this paper, to make headway with this standardized method, we analyze the impact on the video retrieval task with self-evaluations. We address two modules within the

The associate editor coordinating the review of this manuscript and approving it for publication was Cesar Vargas-Rosales.

framework that are directly involved in video retrieval, as well as details that have not been compared before. We then propose temporal nested invariance pooling (TNIP) to alleviate the temporal context restraint. Inspired by nested invariance pooling (NIP) [5], designed to acquire invariance to several image-based transformations, TNIP is designed to boost distinctiveness via temporal cues.

Furthermore, we present the previously unseen comparisons between the CDVA framework and other state-of-the-art methods on released video retrieval datasets. Also, by comparing the performance changes that occur when replacing NIP with TNIP in the CDVA framework, we demonstrate the effectiveness of utilizing temporal cues via the proposed method.

Our contributions are threefold as summarized below:

1) We self-evaluate the standardized CDVA, which demonstrates how each module affects the video retrieval task.
2) We propose a TNIP, which is enabled to be strengthened via temporal cues by replacing NIP in the CDVA framework.
3) We benchmark the CDVA framework against other state-of-the-art methods on published video datasets, proving the efficacy of the proposed TNIP.

## II. RELATED WORKS

Video-to-video retrieval (abbreviated as video retrieval), referring to the content-based video retrieval using only visual sources, aims to search for the most relevant video for given a query video from database videos. Video retrieval methods are roughly divided into two categories depending on features used to measure the degree of similarity between videos. These can be frame-level features and video-level features.

### A. FRAME-LEVEL FEATURES

These methods generally extract frame-level features using a pre-trained convolutional neural network (CNN) as the backbone network [6]–[10], retrieving videos containing related frames by approximate nearest neighbor retrieval. The main concept of frame-level feature-based video retrieval is to aggregate frame-level similarities into video-level similarity. Temporal hough voting [11], [12] searches for temporal alignments, making use of the relative timestamp between matched frames to determine video level similarity outcomes. The graph-based temporal network (TN) detects the longest shared path, which indicates the location of the copy, in the similarity matrix between the query and the reference video. Another approach is to use dynamic programming (DP) [13] which is applied to extract the largest matched diagonal block from the frame-to-frame similarity matrix and tolerate limited horizontal and vertical movements for flexibility. Video similarity learning (ViSiL) [14] introduces an architecture that considers fine-grained spatio-temporal relations between pairs of videos. These relations are trained by accumulating the refined frame-to-frame similarity obtained from the CNN subnet and then calculating the video-to-video similarity. These frame-level feature-based methods use more information than video-level feature-based methods, leading to relatively high retrieval performance. However, these frame-level feature-based methods commonly disregard the redundancy between successive frames, meaning a higher computation cost, resulting in low retrieval efficiency.

### B. VIDEO-LEVEL FEATURES

These methods encode videos at the video level and search for the neighbors nearest the query video in the feature embedding space. To this end, video-level feature-based methods extract one vector for each video and use dot product or the Euclidean distance to compute the degree of similarity between videos. ER3, the abbreviation for event retrieval, recognition, and recounting, [15] developed a unified framework using the video imprint for the entire video based on a feature alignment procedure that exploits the temporal correlations and removes feature redundancies across frames. Deep metric learning (DML) [16] trains a network using a triplet loss scheme to learn embeddings that minimize the distance between related videos while also maximizing it between irrelevant videos. Temporal context aggregation (TCA) [17] adopts the transformer [18], [19] encoder to improve the representation embedding feature in video retrieval with contrastive learning. Because the video is represented as a single vector, video-level feature-based video retrieval has lower computational complexity than frame-level feature-based methods However, these video-level feature-based methods mostly perform worse than frame-level feature-based methods, mainly due to the fact that a single vector cannot readily capture the entire spatio-temporal structure in a video sufficiently.

The standardized CDVA in this paper is a type of frame-level feature-based method. The section that follows explains how the CDVA framework works.

## III. PRELIMINARY: WHAT IS CDVA?

Compact descriptors for video analysis (CDVA), standardized with compact video descriptors, has advanced through several moving picture experts group (MPEG) meetings. Throughout several meetings, the CDVA framework evolved with the CDVA experimental model (CXM) which aims to verify video retrieval performance. We focus on the latest CXM version, which is version 6.5 [4]. The CDVA framework can be divided into four modules, as shown in Fig. 1. First, keyframes are selected from an input video (section III-A). Second, features are extracted in units of selected keyframes (section III-B). Third, the features are encoded by adding standardized bitstreams in units of segments (section III-C). Finally, the similarity between two given videos is calculated (section III-D). The entire framework is separated into three operating points 16KBps, 64KBps, and 256KBps, depending on how dense frames are. In addition, all parameters mentioned below according to the operating points used can be found in Table 1.

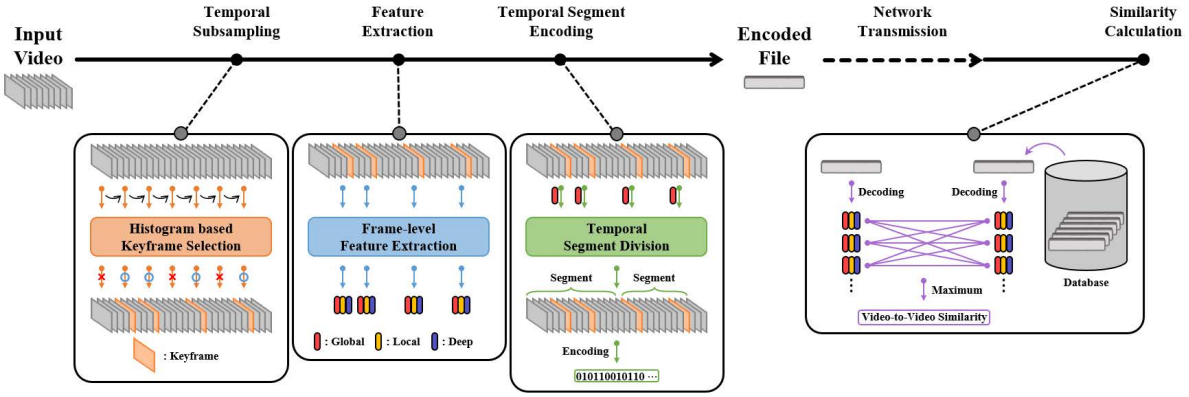


**FIGURE 1.** **Illustration of the CDVA framework for video retrieval. A video as a query is provided to three processes of the CDVA framework for encoding. The encoded file is transmitted to measure the similarity to the database.**

**TABLE 1.** **Parameters details according to the operation points used in the CDVA framework.**

| Parameters | | | | |
|---|---|---|---|---|
| Operating Point | *kfTH* | *stepI* | *segTh* | *verTh* |
| 16KBps | 0.6 | 8 | 18 | 1.98 |
| 64KBps | 0.5 | 6 | 18 | 1.98 |
| 256KBps | 0.4 | 4 | 18 | 1.98 |

### A. TEMPORAL SUBSAMPLING

The process of selecting keyframes from an input video is addressed first. Existing methods that use frame-level features employ uniform sampling to decrease the cost incurred when all frames are used. However, even if there are redundant scenes in the video, the benefit in this case is modest in terms of the cost, as the process causes the frames to be selected repeatedly. For this reason, the CDVA framework determines incidences of duplication through a color histogram and selects unique frames called keyframes. In more detail, keyframe selection begins with uniform sampling. First, frames from the input video are selected uniformly at *stepI* intervals. Thereafter, only frames with a difference from the color histogram of the previous frame greater than *kfTh* are selected as keyframes. This allows for dynamic sampling based on the variance of the scene in the video.

### B. FEATURE EXTRACTION

Next, the process of extracting frame-level features in units of selected keyframes is described. For each keyframe, three image-based features in total are extracted, in this case a handcrafted local and global feature, and a deep feature based on deep learning. Specifically, with regard to the handcrafted features, compressed SIFT [20], used as the local feature, is applied with low complexity transform coding according to the standardization of compact descriptors for visual search (CDVS) [21]. A scalable compressed fisher vector (SCFV) [22] generated by aggregating local features is utilized as the global feature. Moreover, nested invariance pooling (NIP) [5] serves as the deep feature to provide invariance to the representation of the backbone network. At this time, if the successive pooling operations of NIP are configured differently, it is an example of hybrid nested invariance pooling (HNIP) [23]. Given that the two approaches have nearly identical structures, hereinafter we refer to HNIP as NIP. After all of the extraction processes, SCFV and NIP are used during the process of measuring the similarity between videos.

### C. TEMPORAL SEGMENT ENCODING

After extracting features from all keyframes in one video, the next step is the process of encoding by attaching features to standardized bitstreams. Before doing this, keyframes are grouped into temporal segments according to two criteria, and features belonging to each segment are encoded. In detail, the first segment is initialized to the first keyframe on the time axis of the video, after which subsequent keyframes are grouped into the same segment, when the difference in the color histogram relative to the previous keyframe is less than or equal to *segTh*. Also, even if this is not the case, it is grouped into the same segment where the SCFV similarity to the corresponding keyframe exceeds *verTh*. If neither of these is the case, a new segment is created and the remaining keyframes are repeatedly assigned under the two conditions above. Subsequently, to encode in units of segments, the sum of the SCFV similarities is calculated for all other frames belonging to the same segment. The keyframe with the largest sum is selected as a representative keyframe in the segment, and the encoding order of the remaining frames is sorted in the order of that sum. In this order, SCFV and NIP are encoded through adaptive binary arithmetical coding (ABAC) [24], which uses feature differences between the representative frame and the others. On the other hand, the order closest to the representative frame in time is used for encoding the handcrafted local features, and ABAC is also conducted in the same way. All encoded features are added as blocks to the CDVA bitstreams and are decoded when used to run computations between two given videos.

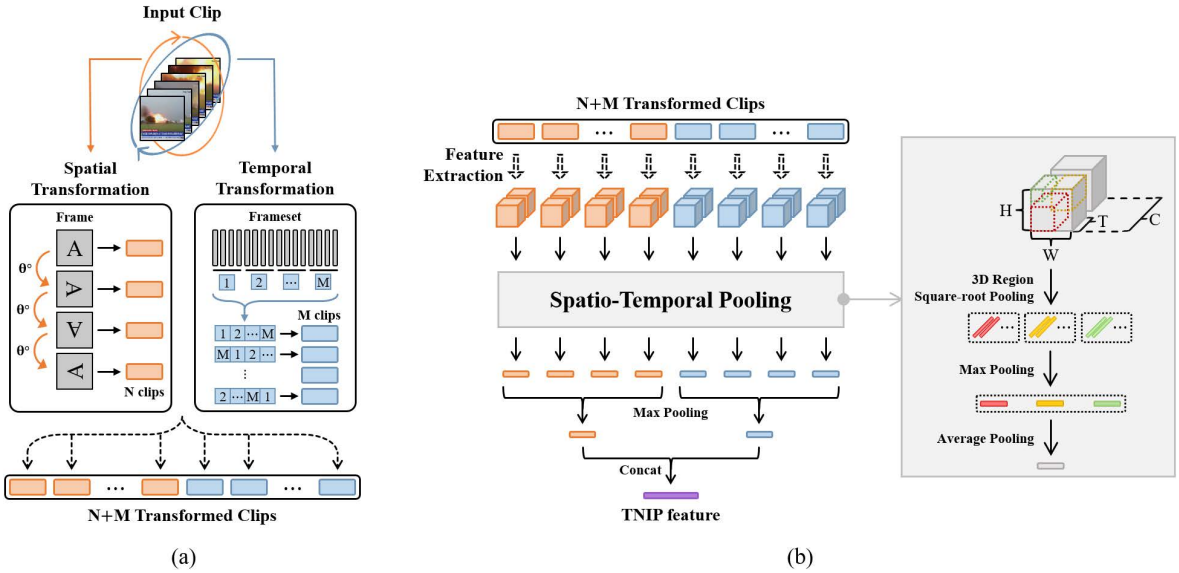


**FIGURE 2.** **Illustration of TNIP. (a) the process of creating a clip set by transforming a clip for each keyframe, and (b) the process of generating a TNIP feature using this clip set.**

### D. SIMILARITY CALCULATION

Here, we describe the process of comparing the similarity with videos in the database after sending the encoded query video to the server from the previous processes. As shown in Eq. 1, the video-level similarity $V$ between two videos $X$ and $Y$ is determined by the maximum value of the fused frame-level similarity $\hat{K}$ between $x_k$ and $y_k$, which are the keyframes of the two videos $X$ and $Y$, respectively. The similarity $\hat{K}$ is computed as the weighted sum between the NIP similarity $K_c$ and the SCFV similarity $K_h$, with alpha set to 0.75. In this way, when a query video is given, a one-to-N manner is used to rank similarities with all videos in the database.

$$\begin{aligned} \hat{K}(x_k, y_k) &= \alpha \cdot K_c + (1-\alpha) \cdot K_h, \\ V(X, Y) &= \max_{x_k \in X, y_k \in Y} \hat{K}(x_k, y_k) \end{aligned} \tag{1}$$

Optionally, after calculating the similarity between the query and the database, re-ranking can be conducted using the handcrafted local feature. This is performed between the corresponding keyframes with the goal of geometric verification.

## IV. PROPOSED METHOD

In the upcoming section VI-A, which deals with self-evaluations, it is discovered that the compact descriptors for the video analysis (CDVA) framework achieve retrieval efficiency through histogram-based keyframe selection. However, the two features used for the similarity comparison commonly tend to worsen as the relevance to be found becomes more complex. In particular, because nested invariance pooling (NIP) is designed according to frame-level transformation, it has less distinctiveness compared to the scalable compressed fisher vector (SCFV), which suitably finds the same scene by aggregating local features. To compensate for this shortcoming, we redesign the NIP into temporal nested invariance pooling, a.k.a. TNIP. It is strengthened by temporal cues to enhance complementarity with SCFV. Details are covered below.

### A. CLIP-LEVEL FEATURE REPRESENTATION

The SCFV is designed to reflect visual detail well from local features, as demonstrated later in section VI-A. However, the NIP leads to poor performance because it works differently. As a solution to this, we are inspired by the fact that NIP shows slight complementarity with SCFV, as indicated in section VI-A. Thus, we redesign this so that NIP supports SCFV with a strong expression of the context. The first modification in this case is to convert the manner of extracting frame-level features from every keyframe into clip-level features. For clarification with regard to dynamic video, the scene flow between the front and rear frames is more critical than the information in one scene. For example, the hidden part when viewed simply through one frame can be grasped by reference to the front and rear scenes. In this way, we attempt to increase the clarity of the context information by utilizing temporal cues that have not been applied in the existing CDVA framework.

The given video $X$ is defined as, $X = \{x_i\}_{i=1}$. In this formula, $i$ denotes the index of the frame in $X$. Thereafter, the selected keyframes are expressed as $x_{k_j}$ during histogram-based selection of the CDVA framework. Here, $k_j$ denotes the index of the $j$-th keyframe from $X$. The existing NIP, described as a frame-level process, embeds $x_{k_j}$ using a backbone network $\Phi_2$ composed of 2D convolution. On the other hand, the method of changing NIP to a clip-level feature is as follows. The input clip of the $j$-th keyframe $c_j = \left\{x_{k_j - \frac{t \times I}{2}}, \cdots, x_{k_j}, \cdots, x_{k_j + \frac{t \times I}{2}}\right\}$ is set by specifying a window $t \times I$ with $t$ frames and an interval of $I$ around every keyframe, embedding it with a backbone network $\Phi_3$

composed of 3D convolution. It is simple, but it is the most intuitive way to use temporal cues.

### B. TEMPORAL NESTED INVARIANCE POOLING

In one video, a frameset (or also known as a clip) that appears in a certain scene has a local context-invariant property. This means that even if the frameset is tangled in time, the represented content should be inferred equally. It is because dynamic changes of objects are considered more important than the corresponding dynamic direction for understanding the content, assuming that the content in a scene is continuously photographed with one camera. For example, for a clip that shows ''running'' even in reverse order, it can be inferred as of the same topic sufficiently due to the dynamic changes in runners. Moreover, when there is a clip that demonstrates ''playing the violin'' even in a tangled order, it can be judged as the same content owing to dynamic changes in the gestures of those playing the violin. With this intuition, we propose temporal nested invariance pooling, a.k.a. TNIP. With the aim of invariance to the local context, TNIP strengthens the sequential connection of temporal cues by the spatial and temporal transformation from the clip-level feature. Moreover, the clip-level feature to which TNIP is applied implies a scene taken with one camera, as the clip is generated on the keyframe extracted via the histogram-based selection of the CDVA framework.

In detail, the clip of the $j$-th keyframe $c_j$ is transformed before embedding it into the backbone network $\Phi_3$. The transformations consist of spatial and temporal transform processes. With regard to spatial transformation, the frame rotates $N$ times by every $\theta^\circ$ angle, similar to what was applied to NIP. All frames in $c_j$ are converted to the same angle at once, creating clip set $C'_j = \left\{c'_{j(1)}, \cdots, c'_{j(N)}\right\}$ consisting of $N$ spatially transformed clips $c'_j$. For temporal transformation, the frame order in $c_j$ is split into $M$ divisions and then intertwined sequentially, creating clip set $C''_j = \left\{c''_{j(1)}, \cdots, c''_{j(M)}\right\}$ consisting of $M$ temporally transformed clips $c''_j$, as given in Fig. 2-(a). After this process, $N + M$ transformed clip sets $C_j = \left\{C'_j, C''_j\right\}$ in total are obtained for clip $c_j$. Thereafter, each clip of the transformed clip set $C_j$ is embedded through $\Phi_3$ as the input to extract feature map set $F_j = \left\{F'_j, F''_j\right\}$; here, $F'_j$ and $F''_j$ denote feature sets comprised by feature map $f'_j$ and $f''_j \in \mathbb{R}^{H\times W\times C\times T}$ from $C'_j$ and $C''_j$ respectively. Each extracted feature map $f'_j, f''_j$ has a temporal axis of $\mathbb{R}^T$ and a spatial axis of $\mathbb{R}^{H\times W}$ in the form of a four-dimensional tensor. To vectorize by applying a spatio-temporal pooling operation to these feature maps, 3D region square-root pooling is conducted on $f'_j$ and $f''_j$, which is motivated by the improved performance demonstrated in a previous study [23] when 2D region square-root pooling was applied to a feature map in the form of a three-dimensional tensor. This broadens the receptive field of both the temporal and spatial axes. To be more specific, a feature map of size $\mathbb{R}^{H_s\times W_s\times C\times T_s}$ is obtained, resulting from the square-root pooling operation with one stride for each size of a three-dimensional region. At this time, there are $S$ types of regions, with $H_s$, $W_s$ referring to the spatial resolution of the output derived from each area and $T_s$ referring to the corresponding temporal resolution. Subsequently, feature map of size $\mathbb{R}^{H_s\times W_s\times C\times T_s}$ is stretched to generate $H_s \times W_s \times T_s$ vectors of size $\mathbb{R}^C$. In the existing NIP, a total of $S \times H_s \times W_s \times T_s$ vectors of size $\mathbb{R}^C$ acquired from different types of regions are averaged, resulting in one vector of size $\mathbb{R}^C$. However, the larger the region size is, the smaller the output size becomes, with proportionally fewer vectors of size $\mathbb{R}^C$. For this reason, the contribution to each scale is not the same, meaning that the process is not sufficiently invariant to the scale. To make it more invariant to the scale, we create one vector of size $\mathbb{R}^C$ by employing max pooling for each of the $H_s \times W_s \times T_s$ vectors of size $\mathbb{R}^C$ obtained for each region; $S$ vectors of size $\mathbb{R}^C$ for all regions are then averaged for the same contribution, resulting in one vector of size $\mathbb{R}^C$. The spatio-temporal pooling operations described above are performed on the feature set $F_j$ to calculate the $N$ and $M$ vectors, respectively. To ensure the strength of each transformation, max pooling is utilized between vectors derived by each transformation, resulting in two vectors. Finally, these vectors are normalized and concatenated to generate a TNIP feature. The TNIP feature yielded through this procedure replaces the NIP for each keyframe in the CDVA framework.

## V. EVALUATION SETUP

### A. DATASETS

#### 1) FIVR-5K

The goal of this dataset is to address the fine-grained video retrieval, a.k.a. FIVR, problem. It is a subset configured by selecting the 50 most difficult queries from the previously proposed FIVR-200K dataset, consisting of a total of 4,999 videos and 50 queries with associations between videos. It includes three retrieval tasks: duplicate scene video retrieval (DSVR), complementary scene video retrieval (CSVR), and incident scene video retrieval (ISVR). The DSVR task is similar to near-duplicated video retrieval with regard to the search for videos that include nearly visually identical incidents taken in the same time zone. Regarding the CSVR task, the purpose is to search for videos taken at different times in the same place, and this is more difficult than DSVR due to scene changes according to time changes. The ISVR task is more challenging than the previous two tasks because the goal in this case is to search for videos taken at different places and times, but of the same incident. Given these variations of relevance, this dataset is used for self-evaluation as well as for benchmarking.

#### 2) CC_WEB_VIDEO

This dataset was released to solve the near-duplicated video retrieval, a.k.a. NDVR, problem. Specifically, it aims to find a video including frames that are geometrically transformed from or visually associated with frames of a particular query video. It contains 24 query sets with 13,129 videos, and four-way evaluations are mainly performed.

**TABLE 2. Self-evaluation: Sampling. Comparison of retrieval efficiency between the histogram-based sampling of CDVA and uniform-based sampling, a general approach, in FIVR-5K.**

| FIVR-5K | | | | | |
|---|---|---|---|---|---|
| Selection Method | Operating Point | mAP | | | Compression Ratio |
| | | DSVR | CSVR | ISVR | |
| Histogram based | 16KBps | **0.814** | 0.780 | 0.664 | **5561.3** |
| | 64KBps | 0.809 | 0.776 | 0.662 | 4384.9 |
| | 256KBps | 0.813 | **0.781** | **0.673** | 3347.7 |
| Uniform based | 16KBps | 0.832 | 0.799 | 0.683 | **2359.8** |
| | 64KBps | **0.838** | **0.806** | **0.689** | 2156.2 |
| | 256KBps | 0.820 | 0.788 | 0.678 | 1876.9 |

**TABLE 3. Self-evaluation: Representation. Comparison of representations between handcrafted global feature and deep feature of CDVA in FIVR-5K.**

| FIVR-5K | | | | |
|---|---|---|---|---|
| Method | Operating Point | mAP | | |
| | | DSVR | CSVR | ISVR |
| SCFV only | 16KBps | 0.851 | 0.809 | 0.652 |
| | 64KBps | 0.862 | 0.815 | 0.661 |
| | 256KBps | **0.870** | **0.824** | **0.666** |
| NIP only | 16KBps | 0.594 | 0.580 | 0.517 |
| | 64KBps | 0.588 | 0.573 | 0.512 |
| | 256KBps | **0.597** | **0.585** | **0.524** |
| SCFV + NIP | 16KBps | **0.814** | 0.780 | 0.664 |
| | 64KBps | 0.809 | 0.776 | 0.662 |
| | 256KBps | 0.813 | **0.781** | **0.673** |

When calculating the average precision for each query, it is divided into two cases depending on whether only the video included in each query set or the entire video is used. Also, it is divided into two cases depending on whether the existing or ''cleaned'' annotation is used. Four evaluations are conducted by combining these conditions.

### B. IMPLEMENTATION DETAILS

The temporal nested invariance pooling (TNIP) feature is extracted on each keyframe in the video, with $t = 16$ frames and an interval of $I = 2$. The backbone network $\Phi_3$ for clip feature extraction is used as R3D and R(2 + 1)D [8], which are pretrained on the Kinetics-400 dataset. Between them, R(2 + 1)D is set as $\Phi_3$ for the benchmark. $N = 4$ clips are converted through 90° degree rotation each via spatial transformation, and $M = 4$ entangled clips are converted via temporal transformation. The regions used in 3D region square-root pooling are (3, 3, 2), (5, 5, 2), and (7, 7, 2) with a total of $S = 3$; the preceding two indices in the tuple are the size of the spatial axis, and the last index is the size of the temporal axis. All experiments are conducted on NVIDIA V100 GPUs.

## VI. EXPERIMENTS

In this section, we demonstrate a self-evaluation of modules related to retrieval in the compact descriptors for video analysis (CDVA) framework, after which we conduct benchmarks of the CDVA framework on the publicly released video datasets FIVR-5K and CC_WEB_VIDEO, which remains unaccomplished up to this point. Furthermore, the nested invariance pooling (NIP) of the CDVA framework is replaced with the proposed temporal nested invariance pooling (TNIP), and the efficacy outcomes between them are compared. Finally, the effectiveness is examined through ablation studies of TNIP.

### A. SELF-EVALUATION ON THE CDVA

We describe self-evaluations of two modules that directly affect video retrieval in the CDVA framework scheme. Self-evaluations of this standard are essential for various subsequent CDVA studies. However, these self-evaluations were covered insufficiently in a series of previous papers dealing with CDVA. Prior to these self-evaluations, content based video retrieval (CBVR) addressed mainly performance issues without post-processing after ranking video similarity levels; hence, all experiments are assumed to omit re-ranking.

#### 1) SELF-EVALUATION 1: FRAME SAMPLING

In the methods that use frame-level features, the quality of video similarity varies depending on how frames are selected. For this reason, we analyze how the histogram-based sampling included in the CDVA framework affects the video similarity. First, uniform sampling is chosen as a comparison target to determine whether histogram-based sampling is superior. At this time, the sampling interval is set to be identical to *stepI* according to operating points in Table 1. Under these conditions, the retrieval performance and compression ratio are verified. The former is judged to be the mean average precision (mAP), and the latter is judged to be a reduced memory ratio compared to the average memory of the query video. According to results such as those in Table 2, when comparing the highest performance for each task in terms of retrieval, uniform sampling performs better on average by roughly 0.018. However, in terms of the compression ratio, uniform sampling is compressible up to about 2359.8 times, while histogram-based sampling shows a higher compression rate of up to 5561.3 times. Therefore, histogram-based sampling as included in the current CDVA framework is shown to be a strategy for increasing the efficiency in the trade-off relationship between retrieval capability and the compression ratio.

#### 2) SELF-EVALUATION 2: FEATURE REPRESENTATION

In the CDVA framework, two features are employed to determine the degree of video similarity: handcrafted global feature SCFV, an abbreviation for the scalable compressed fisher vector, and deep feature NIP. Specifically, the two features take the form of a late fusion approach in which each instance of similarity is measured and then mixed. This procedure, however, makes it difficult to determine which feature influences the retrieval performance. We explain this retrieval contribution through the self-evaluation results for each feature. First, as shown in Table 3, we discuss the overall representation abilities. In all cases on FIVR-5K, NIP is less

**TABLE 4. Benchmarking on CC_WEB_VIDEO and FIVR-5K. Video retrieval benchmarking on CC_WEB_VIDEO and FIVR-5K results through a mAP comparison between the proposed method and other state-of-the-arts methods. SCFV+NIP and SCFV+TNIP refer to the existing and proposed CDVA framework, respectively, and the subscript below them in the table refers to the operating point. In the CC_WEB_VIDEO evaluation, (*) denotes evaluation on the entire dataset and subscript *c* indicates that the cleaned version of the annotations was used. The rotated *video* and *frame* correspondingly indicate that those regions are video-level and frame-level feature-based methods.**

| | Method | CC_WEB_VIDEO | | | | FIVR-5K | | |
|---|---|---|---|---|---|---|---|---|
| | | mAP | | | | mAP | | |
| | | cc_web | cc_web* | cc_web$_c$ | cc_web$^*_c$ | DSVR | CSVR | ISVR |
| *video* | DML | 0.971 | 0.941 | 0.979 | 0.959 | - | - | - |
| | TCA$_c$ | 0.973 | 0.949 | 0.983 | 0.965 | 0.609 | 0.617 | 0.578 |
| *frame* | DP | 0.975 | 0.958 | 0.990 | 0.982 | - | - | - |
| | TN | 0.978 | 0.965 | 0.991 | 0.987 | - | - | - |
| | ViSiL$_f$ | 0.984 | 0.969 | 0.993 | 0.987 | 0.838 | 0.832 | 0.739 |
| | ViSiL$_{sym}$ | 0.982 | 0.969 | 0.991 | 0.988 | 0.830 | 0.823 | 0.731 |
| | ViSiL$_v$ | **0.985** | **0.971** | **0.996** | **0.993** | **0.880** | **0.869** | **0.777** |
| | TCA$_f$ | 0.983 | 0.969 | 0.994 | 0.990 | 0.844 | 0.834 | 0.763 |
| | TCA$_{sym}$ | 0.982 | 0.962 | 0.992 | 0.981 | 0.763 | 0.766 | 0.711 |
| | SCFV+NIP$_{16}$ | 0.972 | **0.954** | 0.976 | **0.961** | **0.814** | 0.780 | 0.664 |
| | SCFV+NIP$_{64}$ | 0.972 | 0.952 | 0.976 | 0.958 | 0.809 | 0.776 | 0.662 |
| | SCFV+NIP$_{256}$ | **0.973** | 0.953 | **0.976** | 0.959 | 0.813 | **0.781** | **0.673** |
| | **SCFV+TNIP$_{16}$** | 0.978 | 0.968 | 0.982 | 0.974 | 0.874 | 0.857 | 0.734 |
| | **SCFV+TNIP$_{64}$** | 0.977 | 0.967 | 0.981 | 0.973 | 0.878 | 0.860 | 0.736 |
| | **SCFV+TNIP$_{256}$** | **0.978** | **0.969** | **0.983** | **0.975** | **0.880** | **0.862** | **0.744** |

distinctive than SCFV. This implies that the transformation of the image-level included in NIP is ambiguous compared to the detailed feature description of the local area included in SCFV, resulting in insufficient discrimination. For this reason, it shows that the performance is comparably lower when the two features are used concurrently compared to when only SCFV is used. Second, when each of the two features are used, the representation ability according to the task change is tested. As the complexity of the task grows, each of the two features shows performance drops, notably SCFV. This indicates that there is a limit to distinguishing scenes of complex occurrences with only two features are describing at the frame level, and this limitation is more pronounced in SCFV, which focuses on a narrower area within the frame and does not provide contextual information between the frames. Third, when the two features are used at the same time, the representation ability according to the task change is assessed. As previously analyzed, the performance falls as the task becomes more challenging in this case. However, at ISVR, this type of simultaneous usage shows slightly higher performance than when solely SCFV is used. In addition, this simultaneous combination shows better results compared to the high drop from DSVR to ISVR seen when only SCFV is used. This indicates that the two features slightly complement the robustness of each other. In summary, the two features are described only at the frame level, indicating that the more complicated the task becomes, the less distinctive the result is, but a complementary relationship also exists. Accordingly, this implies that if the ambiguity of NIP showing relatively insufficient representation is improved, the overall framework can be boosted by maximizing the complementarity of the two features. Thus, we present the solution to this in the aforementioned section IV.

### B. BENCHMARKING WITH OTHER METHODS

#### 1) BENCHMARKING WITH CDVA

The video retrieval performance of the existing CDVA framework only is reported on the MPEG-CDVA dataset, which is not publicly available. For comparison with other state-of-the-art methods, we provide the benchmarking performance of the CDVA framework using the released datasets used in video retrieval in Table 4. Both the video-level and frame-level feature-based retrieval methods are compared. First, similar to the tendency seen in other frame-level feature-based methods, the original CDVA framework is comparable to video-level feature-based methods on CC_WEB_VIDEO and outperforms them by about 0.1 to 0.2 on FIVR-5K. However, when compared to other frame-level feature-based methods, it performs relatively low on these two datasets. This occurs due to the lagging discrimination capacity of NIP, as mentioned previously in section VI-A. In addition, unlike other methods that entail inter-frame operations, SCFV and NIP for measuring the similarity of conventional CDVA perform within only a single frame, resulting in low performance in ISVR on FIVR-5K.

#### 2) PROPOSED DEEP FEATURE

We will discuss benchmarks with TNIP when it is employed in CDVA instead of NIP, as presented in Table 4. First, when compared to the existing CDVA framework using NIP, our suggestion shows an enhancement on both datasets. In particular, the enhancement is noticeably on FIVR-5K, which even includes instances with high difficulty such as ISVR. Specifically, the average performance of all operating points is improved by about 0.065, 0.081, and 0.072 for the DSVR, CSVR, and ISVR tasks. Moreover, the proposed method offers benefit in terms of quality on CC_WEB_VIDEO,

**TABLE 5. Ablation: Components constituting TNIP. This ablation study is based on the clip-level feature obtained from R(2 + 1)D and the operating point of the CDVA framework to 16KBps. Each "Spatial", "Each Region", and "Temporal" refers to the following items: spatial transformation, operation for each region in 3D region square-root pooling, and temporal transformation.**

| TNIP | | | mAP | | |
|---|---|---|---|---|---|
| Spatial | Each Region | Temporal | DSVR | CSVR | ISVR |
| ✓ | - | - | 0.868 | 0.850 | 0.724 |
| ✓ | ✓ | - | 0.873 | 0.853 | 0.722 |
| ✓ | ✓ | ✓ | **0.874** | **0.857** | **0.734** |

where four evaluations exist, and it exhibits more of an increase of roughly 0.015 when using the entire videos. This implies that TNIP distinguishes different contents within this dataset more robustly. Next, the proposed method reveals competitive performance compared to other state-of-the-art methods on both datasets. It performs similarly to others by improving the performance of the CDVA framework on CC_WEB_VIDEO, albeit at a slightly lower level. Furthermore, this outcomes provides evidence of the excellence of TNIP on FIVR-5K while displaying the same effect in DSVR as ViSiL$_v$, which demonstrates the highest performance. From these results, the proposed TNIP, which strengthens the temporal cues to ensure local context-invariant attributes, performs well with excellent complementarity with SCFV in video retrieval tasks.

### C. ABLATION STUDY

#### 1) COMPONENTS CONSTITUTING TNIP

As indicated in Table 5, we conduct an ablation study on FIVR-5K to investigate the components of TNIP. ''Spatial'' indicates that spatial transformation is used to encode features. ''Each Region'' refers to the 3D region square-root pooling procedure for increased scale invariance, and ''Temporal'' refers to when temporal transformation is involved in the clip. As shown in the first row of Table 5, although only the ''Spatial'' component is selected as the existing NIP, it shows adequate performance because the temporal cues are used through the clip-level feature. Then, as the ''Each Region'' component is added, as in the second row, it shows improvements of 0.005 and 0.003 in DSVR and CSVR, respectively. Because it focuses more on the visual form visible in the area within the frame as it goes from ISVR to DSVR, these improvements imply that our approach, designed for better scale invariance, is effective in the two tasks. On the other hand, a slight decline is displayed in ISVR since it focuses only on the area within the frame until the corresponding row. When all components including the ''Temporal'' are added, it demonstrates enhancements of 0.001, 0.004, and 0.012 in DSVR, CSVR, and ISVR as shown in the third row. In particular, this outcome is most noticeable in ISVR, which focuses on the semantic relationships between consecutive frames. This indicates that our approach, which attempts sequentially tangled transformations to connote a local context-invariant property in the feature, effectively reinforces the temporal cues. Consequently, these results prove that TNIP is designed to work more robustly with visual and semantic information.

**TABLE 6. Ablation: Generality according to the backbone. This ablation study is based on the operating point of the CDVA framework to 16KBps. NIP is used when TNIP is not selected.**

| Backbone | Clip-level Feature | TNIP | mAP | | |
|---|---|---|---|---|---|
| | | | DSVR | CSVR | ISVR |
| R3D | ✓ | | 0.869 | 0.846 | 0.721 |
| | ✓ | ✓ | **0.873** | **0.854** | **0.734** |
| R(2+1)D | ✓ | | 0.868 | 0.850 | 0.724 |
| | ✓ | ✓ | **0.874** | **0.857** | **0.734** |

#### 2) TNIP ACCORDING TO THE BACKBONE NETWORK

As indicated in Table 6, we perform an ablation study of TNIP according to the backbone network. R3D and R(2 + 1)D are used as the backbone networks in this case. We compare when only clip-level features are employed to utilize temporal cues and when TNIP is employed to reinforce them. First, when R3D is selected as the backbone network, it presents improvements of 0.004, 0.008, and 0.013 due to TNIP. Even when R(2 + 1)D is selected as the backbone network, it presents a similar tendency with improvements of 0.006, 0.007, and 0.010. As a result, TNIP shows enhanced representation strength with the other two backbones as well, and despite the highest difficulty of ISVR, it exhibits a rather considerable boost.

## VII. CONCLUSION

In this paper, our key contributions are as follows. First, we analyze the modules of CDVA that influence video retrieval (abbreviation for video-to-video retrieval, content-based video retrieval that uses only visual sources) through self-evaluations, for future expanded research on the standardized compact descriptors for video analysis (CDVA) framework. Second, based on analyses of these self-evaluations, temporal nested invariance pooling (TNIP), which replaces nested invariance pooling (NIP) in the CDVA framework, is proposed. TNIP connotes robustness to visual semantic information by exploring the temporal cues. Finally, we also provide benchmarks of the CDVA framework, which has never been done for a public video retrieval dataset. All of our experiments demonstrate that TNIP can significantly boost video retrieval performance outcomes. For future work, we plan to investigate a video-level feature-based method that can benefit from both trade-off relationships between the efficient computational complexity and robust retrieval capability.

## REFERENCES

[1] B. V. Patel and B. B. Meshram, "Content based video retrieval systems," *Int. J. Ubicomp*, vol. 3, no. 2, p. 13, 2012.
[2] X. Wu, A. G. Hauptmann, and C.-W. Ngo, "Practical elimination of near-duplicates from web video search," in *Proc. 15th ACM Int. Conf. Multimedia*, New York, NY, USA, 2007, pp. 218–227.

[3] 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. 2018.
[4] 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.
[5] 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)*, 2017, pp. 420–429.
[6] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015, *arXiv:1409.1556*.
[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[8] K. Hara, H. Kataoka, and Y. Satoh, "Learning spatio-temporal features with 3D residual networks for action recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2017, vol. 2, no. 3, p. 4.
[9] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 6299–6308.
[10] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2015, pp. 4489–4497.
[11] 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, Dec. 2010.
[12] Y.-G. Jiang, Y. Jiang, and J. Wang, "VCDB: A large-scale database for partial copy detection in videos," in *Proc. ECCV*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham, Switzerland: Springer, 2014, pp. 357–371.
[13] 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.
[14] 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.
[15] Z. Gao, G. Hua, D. Zhang, N. Jojic, L. Wang, J. Xue, and N. Zheng, "ER3: A unified framework for event retrieval, recognition and recounting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2253–2262.
[16] 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.
[17] J. Shao, X. Wen, B. Zhao, and X. Xue, "Temporal context aggregation for video retrieval with contrastive learning," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 3268–3278.
[18] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017, pp. 5998–6008.
[19] A. Dosovitskiy *et al.*, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021.
[20] D. G. Lowe, "Distinctive image features from scale-invariant key points," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Mar. 2004.
[21] L.-Y. Duan, F. Gao, J. Chen, J. Lin, and T. Huang, "Compact descriptors for mobile visual search and MPEG CDVS standardization," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, Dec. 2013, pp. 885–888.
[22] F. Perronnin, Y. Liu, J. Sánchez, and H. Poirier, "Large-scale image retrieval with compressed Fisher vectors," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Dec. 2010, pp. 3384–3391.
[23] J. Lin, L. Duan, S. Wang, Y. Bai, and Y. Lou, "HNIP: Compact deep invariant representations for video matching, localization, and retrieval," *IEEE Trans. Multimedia*, vol. 19, no. 9, pp. 1968–1983, Sep. 2017.
[24] D. Marpe, H. Schwarz, and T. Wiegand, "Context-based adaptive binary arithmetic coding in the H.264/AVC video compression standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7, pp. 620–636, Jul. 2003.

**WON JO** received the B.S. degree from the Department of Intelligent Mechatronics, Sejong University, Seoul, Republic of Korea, in 2021, where he is currently pursuing the M.S. degree with the Robotics and Computer Vision Laboratory, Department of Artificial Intelligence. 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 solutions for learning.

**GUENTAEK LIM** was born in Seoul, South Korea, in 1998. He is currently pursuing the bachelor's degree with Sejong University. He is working at the Robotics and Computer Vision Laboratory, Sejong University. His current research interests include machine learning, computer vision, and video understanding.

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

**JOUNGIL YUN** received the B.S. degree in control and instrumentation engineering from Jeonbuk National University, Jeonju, Republic of Korea, in 1996, and the M.S. and Ph.D. degrees in mechatronics from the Gwangju Institute of Science and Technology, 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 working as a Principal Researcher at the Media Research Division, ETRI. His current research interests include immersive media processing and video coding for machines.

**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 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; and the Director of the Robotics and Computer Vision (RCV) Laboratory. Her research interests include computer vision and robotics.