Surch: Enabling Structural Search and Comparison for Surgical Videos

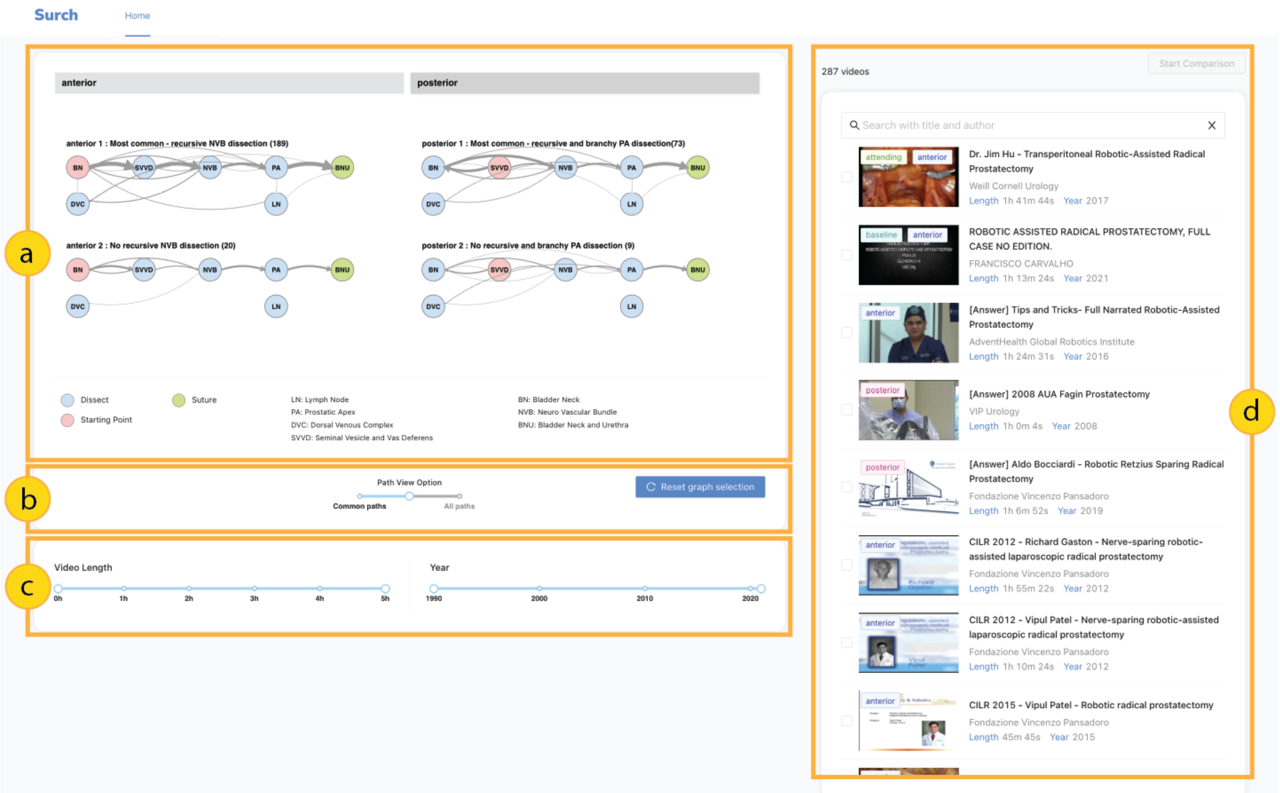
Jeongyeon Kim
imurs4825@gmail.com
Computer Science,
Stanford University
Stanford, CA, USA

Daeun Choi
daeun.choi@kaist.ac.kr
School of Computing, KAIST
Daejeon, Republic of Korea

Nicole Lee
nicolelee2001@gmail.com
School of Computing, KAIST
Daejeon, Republic of Korea

Matt Beane
mattbeane@ucsb.edu
Technology Management, University
of California, Santa Barbara
Santa Barbara, CA, USA

Juho Kim
juhokim@kaist.ac.kr
School of Computing, KAIST
Daejeon, Republic of Korea


Surch
Home
anterior
posterior
anterior 1 : Most common - recursive NVB dissection (189)
posterior 1 : Most common - recursive and branchy PA dissection(73)
anterior 2 : No recursive NVB dissection (20)
posterior 2 : No recursive and branchy PA dissection (9)
BN
SVVD
NVB
PA
BNU
DVC
LN
Dissect
Suture
Starting Point
LN: Lymph Node
PA: Prostatic Apex
DVC: Dorsal Venous Complex
SVVD: Seminal Vesicle and Vas Deferens
BN: Bladder Neck
NVB: Neuro Vascular Bundle
BNU: Bladder Neck and Urethra
Path View Option
Common paths
All paths
Reset graph selection
Video Length
0h
1h
2h
3h
4h
5h
Year
1990
2000
2010
2020
a
b
c
d
287 videos
Start Comparison
Search with title and author
attending
anterior
baseline
posterior

Dr. Jim Hu - Transperitoneal Robotic-Assisted Radical Prostatectomy
Weill Cornell Urology
Length 1h 41m 44s Year 2017

ROBOTIC ASSISTED RADICAL PROSTATECTOMY, FULL CASE NO EDITION.
FRANCISCO CARVALHO
Length 1h 13m 24s Year 2021

[Answer] Tips and Tricks- Full Narrated Robotic-Assisted Prostatectomy
AdventHealth Global Robotics Institute
Length 1h 24m 31s Year 2016

[Answer] 2008 AUA Fagin Prostatectomy
VIP Urology
Length 1h 0m 4s Year 2008

[Answer] Aldo Bocciardi - Robotic Retzius Sparing Radical Prostatectomy
Fondazione Vincenzo Pansadoro
Length 1h 6m 52s Year 2019

CILR 2012 - Richard Gaston - Nerve-sparing robotic-assisted laparoscopic radical prostatectomy
Fondazione Vincenzo Pansadoro
Length 1h 55m 22s Year 2012

CILR 2012 - Vipul Patel - Nerve-sparing robotic-assisted laparoscopic radical prostatectomy
Fondazione Vincenzo Pansadoro
Length 1h 10m 24s Year 2012

CILR 2015 - Vipul Patel - Robotic radical prostatectomy
Fondazione Vincenzo Pansadoro
Length 45m 45s Year 2015

Figure 1: A user can search videos based on structures of procedural knowledge using the search interface of Surch. (a) Procedural Graph: structures of surgical procedures are represented in a graph format. The user can search videos based on graph-based interactions. (b) Path View Option: the user can adjust the level of detail of the graph, displaying all paths or only common paths. (c) Video Filters: the user can filter the videos based on two basic filters, video length and year of upload filters. (d) Video List and Text-Based Search Bar: the user can browse the list of surgical videos and input keywords to search videos.

CHI '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9421-5/23/04…$15.00
https://doi.org/10.1145/3544548.3580772

ABSTRACT

Video is an effective medium for learning procedural knowledge, such as surgical techniques. However, learning procedural knowledge through videos remains difficult due to limited access to procedural structures of knowledge (e.g., compositions and ordering of steps) in a large-scale video dataset. We present Surch, a system that enables structural search and comparison of surgical procedures. Surch supports video search based on procedural graphs generated by our clustering workflow capturing latent patterns within surgical procedures. We used vectorization and weighting schemes that characterize the features of procedures, such as recursive structures and unique paths. Surch enhances cross-video comparison by providing video navigation synchronized by surgical steps. Evaluation of the workflow demonstrates the effectiveness and interpretability (Silhouette score = 0.82) of our clustering for surgical learning. A user study with 11 residents shows that our system significantly improves the learning experience and task efficiency of video search and comparison, especially benefiting junior residents.

CCS CONCEPTS

• **Human-centered computing → Interactive systems and tools**; **Empirical studies in HCI**.

KEYWORDS

Video-Based Learning, Procedural Knowledge Representation, Surgical Learning, Structural Video Search, Cross-Video Interaction

ACM Reference Format:
Jeongyeon Kim, Daeun Choi, Nicole Lee, Matt Beane, and Juho Kim. 2023. Surch: Enabling Structural Search and Comparison for Surgical Videos. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), April 23–28, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3544548.3580772

1 INTRODUCTION

Procedural knowledge is knowledge about how to do something, for example, how to carry out a sequence of operations or actions [74, 105]. Learning procedural knowledge consists of two main steps: grasping a conceptual overview of a procedure's structure and mastering individual steps in that procedure [31, 65]. In this regard, video is a helpful aid because it contains the overall flow of procedural steps and demonstrates intricate details of techniques that are difficult to document or verbally explain [6, 30].

This is particularly true for surgery, a domain where videos are growing in popularity as learning materials for residents (surgeons in training) and medical students as they learn surgical procedures [18, 80, 120]. However, the conceptual structure of procedural knowledge such as compositions and step ordering is latent and not easily accessible in the UI for surgical video. This is because it is costly to extract semantic information from surgical data via domain experts and difficult to capture meaningful structural features from large-scale video data as they become a series of pixels once the video is encoded. In addition, existing video interfaces do not support a focused comparison of critical skills for each surgical step, which is important in learning procedural knowledge.

To address these challenges, we first conducted iterative interviews with six surgeons and thirty residents, aiming to investigate the current practice and challenges of video-based surgery learning. Our results revealed that search and comparison were the most challenging and costly tasks associated with learning surgery through video, corroborating our assessment that current platforms offered limited support for learning procedural knowledge from videos. Residents' main pain point was the limited access to semantic information while searching videos and the lack of support for cross-video reference and comparison. The residents wanted a semantic video search function based on surgical phases and approaches (e.g., anterior or posterior approach in robotic prostatectomy). They also sought to compare the approaches and techniques in a wide range of surgical cases to be prepared for various situations, but existing video interfaces such as YouTube did not support cross-video referencing and comparison. Residents, as a consequence, ended up searching for videos by relying on limited metadata such as video titles and uploader information and then comparing said videos by scrubbing through the timelines in multiple windows.

Accordingly, to enhance the learning experience through surgical videos, we designed and implemented a system named Surch (**Sur**gery + Sear**ch**), a video interface that supports structural search and cross-video comparison for surgery videos. For this initial study, we focused on recordings of robot-assisted radical prostatectomy, the most common robotic procedure [48, 101]. Our system consists of (1) a computational pipeline that automatically translates surgery videos into procedural graphs that represent the structure of procedural sequences, (2) a video search interface that enables structural search through filters based on the semantic structure of surgical procedures, and (3) a video comparison interface that allows cross-video comparison based on surgical phases.

As part of the computational pipeline, we first annotated the surgery video dataset for 296 videos of prostatectomies from college students. We then trained a CNN-LSTM model to automatically detect the surgical phases. On top of that, we devised a clustering workflow tailored to the analysis of procedural knowledge. The workflow captures key features of procedural knowledge structures such as recursive and branching step patterns, thereby enabling the detection of latent important learning points. Existing clustering methods for constructing visual representations of sequential data [24, 68] are not apt to identify meaningful patterns in surgical procedures as identifying patterns in surgical procedures requires consideration for analysis dimensions that are meaningful in surgical learning, such as recursive structures or unique paths in surgeries. To this end, we vectorized procedural data using custom analysis dimensions and designed a weighting scheme with parameter optimization through a grid search [20]. In addition, previous work focuses on interactional support for static content such as natural language text [23] and command log data [24, 68], but comparing procedural knowledge in videos poses a new challenge such as aligning temporal information of multiple videos. Our clustering workflow addresses these issues and enables rich video interactions, including synchronized navigation across multiple videos.

Powered by the automated pipeline, the Surch video interface enables structural video search and comparison (Fig. 1). First, the search interface provides an interactive semantic graph of surgical procedures. This visual representation helps residents understand the conceptual structure of the surgery and supports graph-based interaction for filtering videos. For instance, surgical phases are visualized as nodes in graphs, and residents can click these nodes to watch videos that contain the selected phases. Second, Surch supports cross-video comparison. Residents can quickly reference and switch across videos by watching different clips in multiple windows. They can also synchronize the videos at a phase level via graph-based interactions: clicking a surgical phase included in two surgical procedures synchronizes their video play bars, thereby enabling juxtaposed comparison of that phase in multiple videos.

We evaluated the effectiveness and validity of each part of our system: the clustering workflow and the interface. We demonstrate the robustness of our clustering results by using both quantitative and qualitative measures, using six different evaluation metrics for clustering quality and expert evaluation with senior residents. The clustering workflow achieved a Silhouette score of 0.76 and 0.87 for the dataset of anterior and posterior approaches, and the residents reported that the clusters aligned with what they learned and allowed new discoveries of surgical paths. Our clustering results detected not only the most standard approach (valuable for juniors) but also unique routes (valuable for seniors). To test if our video interface improved video search and comparison in surgery learning, we conducted a user study with 11 residents. The results show that our system significantly increased the learning experience and perceived task efficiency for video search and comparison. Our system enables users to watch more videos in a shorter time, and junior residents submitted slightly more meaningful comparisons across videos when using our system. We also discuss the design implications of search and comparison tools for videos with procedural knowledge.

To summarize, our work makes the following contributions:

- An annotated dataset of surgical phases in 296 prostatectomy videos and a CNN-LSTM model for surgical phase detection in prostatectomy
- An automated clustering pipeline that generates procedural graphs for surgical videos
- A design and implementation of Surch, a video interface that provides structural video search and cross-video comparison
- Results of quantitative pipeline evaluation and empirical user study

2 RELATED WORK

We discuss three domains of previous work that our work builds on: (1) learning surgery through video, (2) content-based video search, and (3) cross-video comparison.

2.1 Learning Surgery Through Video

Video is the primary medium for learning surgical training [2, 11, 99]. According to Mota et al. [80], 98.6% of residents and surgical specialists use videos for surgery preparation. Surgery video recordings provide key learning points in surgical training, such as anatomical landmarks and surgical maneuvers [1, 77]. They also accelerate the learning curve for surgical training [56]. As video becomes a predominant medium in learning surgery, several researchers designed supporting tools to enhance video-based surgery learning. The existing work has sought to improve how medical students and professionals learn and communicate through a video medium, including in-video navigation, video summary, and telemonitoring (i.e., remote patient monitoring).

To enhance video navigation in surgical recordings, Hudelist et al. [53] developed a video interface that displays clickable and zoomable keyframes of endoscopic videos. Munzer et al. [82] introduced EndoXplore, a video player that supports content-based video navigation based on phases of surgical operations. Their system automatically extracts the surgical phases from endoscopic videos and displays the clickable thumbnails that allow navigation to each surgical phase. However, their automatic phase detection relied on the instrument types, which do not distinguish main phases using the same instruments. Also, EndoXplore only supports a single video interaction, while residents usually browse and compare multiple procedures per case. Meanwhile, another thread of work investigated collaborative video learning. Surgeons and residents construct a shared understanding of surgery through telementoring that monitors patients remotely [15, 35]. Mentisa et al. [77] investigated how surgeons communicate over laparoscopic videos when they are in remote settings. Avellino et al. [12] introduced a system design for a collaborative video summary tool. They suggested that the requirements for a collaborative surgery summary tool include enabling appropriate division of tasks and management of dependencies between tasks.

Prior work, however, is limited to interactions around a single video due to the two main challenges of interaction support involving multiple videos: difficulties in extracting semantic information from videos and building meaningful links across videos. In this context, content-based video search and comparison, the two main pain points revealed by our formative study, remained unsupported by the existing systems or techniques.

2.2 Content-Based Video Search

A massive amount of videos are produced and uploaded online. In particular, surgical recordings are piling up, driven by the prevalence of endoscopic video recordings and advances in camera technologies [90, 97]. The traditional search method relies on textual metadata input by the uploader (e.g., titles, tags), which does not reflect rich visual features or structural information of video content. One of the key solutions to mediate the video search problem is a video search based on the in-video semantic features such as objects, actions, and relationships between them [25, 26, 89, 102]. To enable semantic video search, it is necessary to retrieve information and knowledge from videos. To this end, there have been attempts to extract semantic information from videos with procedural knowledge, for example, a path of players' movements in sports videos [4, 10, 107], step-by-step structures in how-to videos [28, 40, 66, 116], and surgical phases in surgery videos [29, 39, 43, 67, 95, 112, 121]. Tan et al. also [109] provided a landscape of procedural video datasets whose task includes furniture assembly, makeup, and cooking. The computer vision community improved self-supervised learning [45] and procedural learning [16] for videos that contain procedural tasks by aligning temporal information between multiple videos.

Such semantic information extracted from videos enables content-based video search [51]. For example, a body of existing work displays the semantic information in the format of a sequence of keyframes of videos, which enables quick video browsing. Barthel et al. [17] allowed visual exploration and search of videos by providing related video scenes with high visual similarity. Hurst et al. [54] designed an interface that presents video thumbnails to support large-scale video browsing on mobile devices. Another thread of work tried to represent the information in videos into a hierarchical structure by organizing knowledge in videos into a tree structure [119], hierarchically clustering video shots [81, 110] and providing an overview of non-linear causal relationships [84]. Meanwhile, several video search tools allow various types of queries for video search, including sketches of video frames [87, 92], textual keywords [9, 86], and video clips [44, 61].

However, none of the existing work supports video search based on procedural knowledge structure in videos, which can be the most intuitive way of exploring procedural videos as structural information delivers the steps' connections and their logical order in procedural knowledge [41]. Thus, we propose a video search interface that provides a visual representation of procedural videos. Our system allows viewers to understand the landscape of procedures in myriad videos and filter video clips by using graph-based interactions.

2.3 Cross-Video Comparison

Most surgeons reach an early plateau of average performance maintained for the rest of their career [34]. In contrast, exposure to various surgical techniques by comparing multiple videos can play a key role in overcoming the plateau [120]. Previous research has introduced interfaces for comparing multiple videos. For instance, Balakrishnan et al. [14] suggested a visualization method that overlays highlighted edges of actions' differences in videos to identify the subtle dissimilarities between motions. Bellini et al. [19] proposed MyStoryPlayer, a video interface that supports video comparison based on multi-video views synchronized considering relationships of audiovisual content in video-based learning and training scenarios. Meanwhile, video synchronization is one of the main challenges in designing a video comparison system, which aligns multiple videos in a common temporal line considering audio and visual relationships between videos. Segundo et al. [98] introduced a crowdsourcing technique that manages the convergence of crowds' contributions and distributes videos for video synchronization. On the other hand, Tharatipyakul et al. [111] suggested essential conceptual components for supporting video comparison, which includes playback and synchronization, complexity reduction, and interactivity. In the surgery domain, in particular, Matsuda et al. [73] designed multi-view video interfaces to enable users to watch videos from multiple sources such as endoscopes, x-ray, and experts' eye and hand movements. Hudelist et al. [52] built a video player for surgical video comparison on tablet devices. Their video interface displays two video windows side-by-side, allowing the comparison of the same procedure of different patients.

However, there is limited support for cross-video comparison at a phase level, which was needed by residents to find a feasible technique for each surgical phase based on our interview results. The phase-level interaction, in particular, needs to be supported since the surgical phases are the basic units of surgery procedures, and segmenting surgical procedures into phases plays a vital role in reducing the learners' cognitive load [79, 85]. To fill this gap, our system enables cross-video comparison based on the procedural structure of a surgery. In this way, residents can easily reference various techniques for the same phase, thereby achieving expert performance by engaging in deliberate practice in a focused area of surgery [34].

3 FORMATIVE STUDY

To understand the current practices and needs surrounding video-based surgery learning, we conducted multiple rounds of need-finding interviews over a year with residents and surgeons. We used an iterative interviewing method [36, 114] to investigate the research problems and discover our research direction.

3.1 Method

3.1.1 Participants. We recruited six surgeons and thirty residents in urology in the U.S. by contacting surgeons through email, asking them to recommend colleagues who specialize in robotic surgery. The residents consisted of 6 first Post Graduate Year one (PGY1), 6 PGY2, 7 PGY3, 3 PGY4, 7 PGY5, and 1 PGY6. All of the participants had prior experience learning about surgery through videos. We refer to six surgeons as S1 through S6 and thirty residents as R1 through R30.

3.1.2 Procedures. We conducted remote semi-structured interviews using Zoom and recorded the interviews under consent. All of the authors have received training on the ethical treatment of human subjects, and we did not collect any data from patients, including video recordings, clinical history, or medical images.

The interviews lasted about 50 minutes with two main sessions: current practice and challenges of video-based surgery learning. We used a saturation method [21] to determine the number of participants. We conducted several batches of group interviews with 3-4 participants per each interview session and performed a preliminary analysis of the transcripts for each batch. We stopped conducting more interviews when the analysis stopped revealing new insights.

3.2 Analysis

The interviews were video recorded and transcribed using transcription service [1]. Two of our authors performed thematic analysis [50]. They independently made a codebook for half of the transcripts, using an inductive approach. They then merged and refined the codebook until they reached a consensus. The two authors then coded two randomly selected transcripts using the codebook. To validate the qualitative coding, we computed Cohen's kappa. The average Cohen's kappa score across the entire code was 0.81 with a standard deviation of 0.06. Each author then coded the rest of the interviews independently. After the coding, they met to discuss discrepancies in applying the code set and adjusted the coded data.

3.3 Findings

3.3.1 Context and Practice of Video-Based Surgery Learning. The residents were using three main surgery video platforms: YouTube, Michigan Urological Surgery Improvement Collaborative (MUSIC),

[1] https://otter.ai/.

and one offered by the American Urological Association (AUA). Most of our interviewees watched surgery videos a night before participating in surgical procedures. Their purpose in watching these videos was to remind themselves of certain surgical techniques or the flow of surgical phases. They stressed that they usually watch videos under tight schedules. Meanwhile, their video-based surgery learning involved multiple activities, which included video search, navigation, comparison, bookmarking, note-taking, and reviewing.

3.3.2 *Challenges and User Needs.*

Limited Access to Semantic Information.

Residents' main consideration when searching for surgery videos was the feasibility of a given procedure. They sought videos with procedures that they could perform in practice. They determined the feasibility of a particular procedure by inducing semantic information such as approaches and techniques. Residents pointed out that the existing video platforms did not support such semantic search, so they needed to invest significant time into manually navigating to video segments they wanted to watch.

Lack of Structural Information on Procedural Knowledge.

Residents wanted to search videos using structural information of surgical procedures, for example, the compositions and orders of phases. In particular, residents' need for structural information differed depending on their expertise. Junior residents expressed a need to find videos that contain a standard instance of a given procedure to learn the common structure of phases. In addition to searching for standard procedures, several novice residents wanted to find videos that did not miss or skip any surgical phase for a certain surgery. Several surgeons (S1, 2, 3) noted that surgical phases are one of the important learning units for novice residents. By contrast, senior residents sought videos that could teach them variations and details of each phase. For example, R30 stated that "several surgery videos show interesting portions of surgery which is complex, but I'm a junior level and want to learn the basic steps.". R29, a PGY5 resident, also mentioned that "as a junior resident, you need to stick to your attending's approach, but as you become more independent, you might need to find your own technique." and "I want to go beyond. I would want to see many different versions of that specific variation."

Lack of Support for Cross-Video Reference and Comparison.

We could observe a clear user need for comparing multiple videos. Residents wanted to compare different procedures, approaches (e.g., anterior approach or posterior approach in prostatectomy), and techniques (e.g., dissection using cautery or scissors). Meanwhile, most of our interviewees pointed out the surgical phase as a unit of comparison. R26, for example, said that "I want to know different techniques for the same phase." R30 wished to compare numerous videos for the same phase, stating that "it'd be helpful if I can compare 20 different clips for a specific phase. Doctors do it slightly differently and patients' situations can be different." S3 also mentioned that building links across multiple videos would enhance residents' learning. However, comparing videos for the same phase involved multiple, challenging, and time-consuming steps. Residents had to search for a video, manually advance and rewind to navigate to the phase they were interested in, play that clip, and repeat this process for the next video. They mentioned these inefficiencies repeatedly as they described their process for comparing multiple techniques for the same phases using available video platforms.

Limited Surgery Metadata.

Some residents wanted additional information regarding a particular surgical video, such as operating room situations (e.g., port placement in robotic surgery) and anatomical landmarks, although they were not mentioned by many interviewees. Residents likewise wanted audio narrations to know the anatomical landmarks and distinctions between phases. S1 explained that "understanding anatomy is one of the first steps of learning as residents." They also wanted to know surgeons' expertise (e.g., surgeon volume) and surgery outcomes (e.g., restoring sexual function in prostatectomy), which are not accessible in most surgery recordings.

4 DESIGN GOALS

Guided by the findings from our formative interviews, we derived a set of design goals for a system that supports structural search and cross-video comparison for surgery videos.

D1. Provide a structural landscape of procedural videos

The residents we interviewed wanted a bird's-eye view through which they could identify a standard procedure and its variations. Our system should make the range of actual procedure phase orderings available to users so that they can refer to typical and atypical structures of procedures while learning a surgery.

D2. Support structural video search

Residents wished to search for videos that contain specific surgical approaches and phases. Our system should enable users to find videos using structural information associated with video content.

D3. Provide on-demand affordances for video search

The formative interviews revealed different user needs for video search depending on residents' expertise. Juniors needed to understand the overall flow of surgeries and basic anatomy, while seniors wanted to learn variations and multiple techniques of performing the same procedure. Our system should provide on-demand affordances such as filters for advanced video search that will not overwhelm novices, yet will satisfy experts' needs on demand.

D4. Enable step-level video comparison We observed a clear user need for comparing the same phase across multiple procedures. Our system should allow users to reference multiple videos at a phase level by supporting cross-video interactions such as synchronization and multi-view switching.

5 VIDEO INTERFACE: STRUCTURAL SEARCH AND COMPARISON

This section introduces Surch's video interface that supports structural search and comparison of surgical procedures.

5.1 Procedural Graph

Goldstein et al. [41] revealed that the graph representation of knowledge can serve as learners' basic memory structure for procedural knowledge. Inspired by the previous work that designed a visual representation for sequence data [24], we use the term "procedural graph" to refer to a graph-based representation for phase sequences in procedural knowledge, which provides a visual overview of the

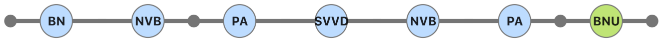


Figure 2: A procedural graph for a single procedure. Nodes indicate surgical phases and edges connect the consecutive phases. The anatomies are written on the nodes and the colors of the nodes represent the surgical actions.

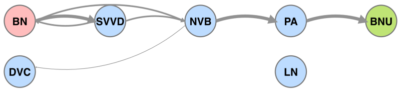


Figure 3: A procedural graph for multiple procedures. The thickness of the arrows indicates the commonality of paths and is proportional to the number of flows from one phase to another.

surgical procedures' distribution. Figure 2 and 3 show an example of a procedural graph for a single procedure and aggregation of multiple procedures, respectively. A node and edge represent a surgical phase and path between phases. More specifically, surgical phases consist of anatomy and surgical action, such as bladder neck dissection. The consecutive phases are connected through edges. The aggregated representation shows the overall structure of multiple procedures. Similar to the Sankey diagram [103], the thickness is proportional to the number of flows from one phase to another included in the Surch data repository. For instance, a thick edge implies that this path is included in many videos, indicating that the path is commonly taken by surgeons. Meanwhile, the color of the nodes is for the surgical actions: blue for dissection and green for suturing. The red indicates the starting point of the procedures. The locations of edges indicate the directions of flows, which often implies the recursive structures of procedures. The edges located above the nodes are for the path from the left phase to the right phase, and the ones below the nodes are for the reverse flow. The small gray circles in the graphs for an individual procedure indicate non-major steps, such as parts where surgeons explain a surgery or operating room situation.

5.2 Interfaces

The procedural graph (Figure 1 and 4) shows the structure of surgical procedures, using a graph representation with nodes and edges (**D1.** Provide a structural landscape of procedural videos). Users can also filter videos by clicking the nodes or edges (**D3.** Provide on-demand affordances for video search). If they click an edge, our system displays a list of videos that contain a certain path (Figure 1 (a)). In the case of a node, the system shows videos, including the selected phase. This graph-based search filter enables video search based on structural information of procedures (**D2.** Support structural video search). Surch also provides the clustering results of procedures generated by our clustering workflow, described in Section 7, with the auto-generated cluster labels as needed. By default, the system displays two graphs, one for the anterior approach and the other for the posterior approach. However, if the user is interested in further details, they can expand the graphs to see the clustering results for each approach (Figure 1 (a)). The user can browse the videos in a specific cluster by interacting with the graph for that cluster.

After referencing the clustering visualization and procedural graph, users can select videos of interest. The selected videos are highlighted in the clustering visualization window. After selecting the videos, clicking the "Start Comparison" button leads users to a video comparison page (Figure 4). The video comparison page supports multi-window videos (Figure 4 (a)) and phase-based video synchronization (Figure 4 (c)) (**D4.** Enable step-level video comparison). Users can watch multiple videos through juxtaposed video windows. Meanwhile, they can synchronize the playbacks of multiple videos using interactive graphs. For example, if they click the node for the "ladder neck dissection" phase, all the videos are navigated to the chosen phase at once so that users can compare multiple videos for the same phase. We intentionally did not synchronize the playback within a phase since residents in our formative study indicated they were overwhelmed by playing multiple videos at once. They rather preferred to play the same phase alternately in juxtaposition. They can also choose a preferred layout for video windows among grid layouts with videos in small grids, a 1:N layout with a single main window and multiple small sub-windows, and a 1:1:N layout with two main windows and small windows for the rest of the videos (Figure 4 (b)).

Surch is a web application built using TypeScript, HTML/CSS, and the D3 library [83].

6 DATASET: LABELING SURGICAL PHASES

To enable structural search and comparison of surgical procedures (Section 5), we segmented videos by surgical phase. No large dataset on surgical phases in prostatectomy was publicly available, however. Furthermore, traditional expert-led methods for labeling videos suffer from the high cost and limited availability of surgeons and residents [8, 117]. To mitigate these challenges, we designed and implemented a data labeling workflow that blends automated approaches with low-level human effort. We recruited college students in premed and biological studies to obtain labeled data at a relatively low cost. We believe this approach has practical value in various domains where expert resources are scarce and expensive. This section describes our workflow for annotating and filtering our surgical video dataset.

6.1 Data Labeling

6.1.1 Participants. We recruited 13 college students majoring in premed, bioengineering, and veterinary medicine by posting on online communities. Although all of their majors are relevant to the medical domain, some of them had no prior knowledge of surgery.

6.1.2 Procedures. Data labeling sessions were conducted remotely. We had a 1-hour instruction session with all labelers using Zoom in which we briefly introduced our experiment and provided learning materials for prostatectomy. The materials included a basic set of surgical phases for prostatectomy that we built based on two papers [49, 76] and one surgery tutorial [59], a list of prostatectomy recordings with their phases labeled, and an introduction to our video annotation tool. The data collection was conducted over two weeks, and each labeler was assigned 20 videos to annotate.

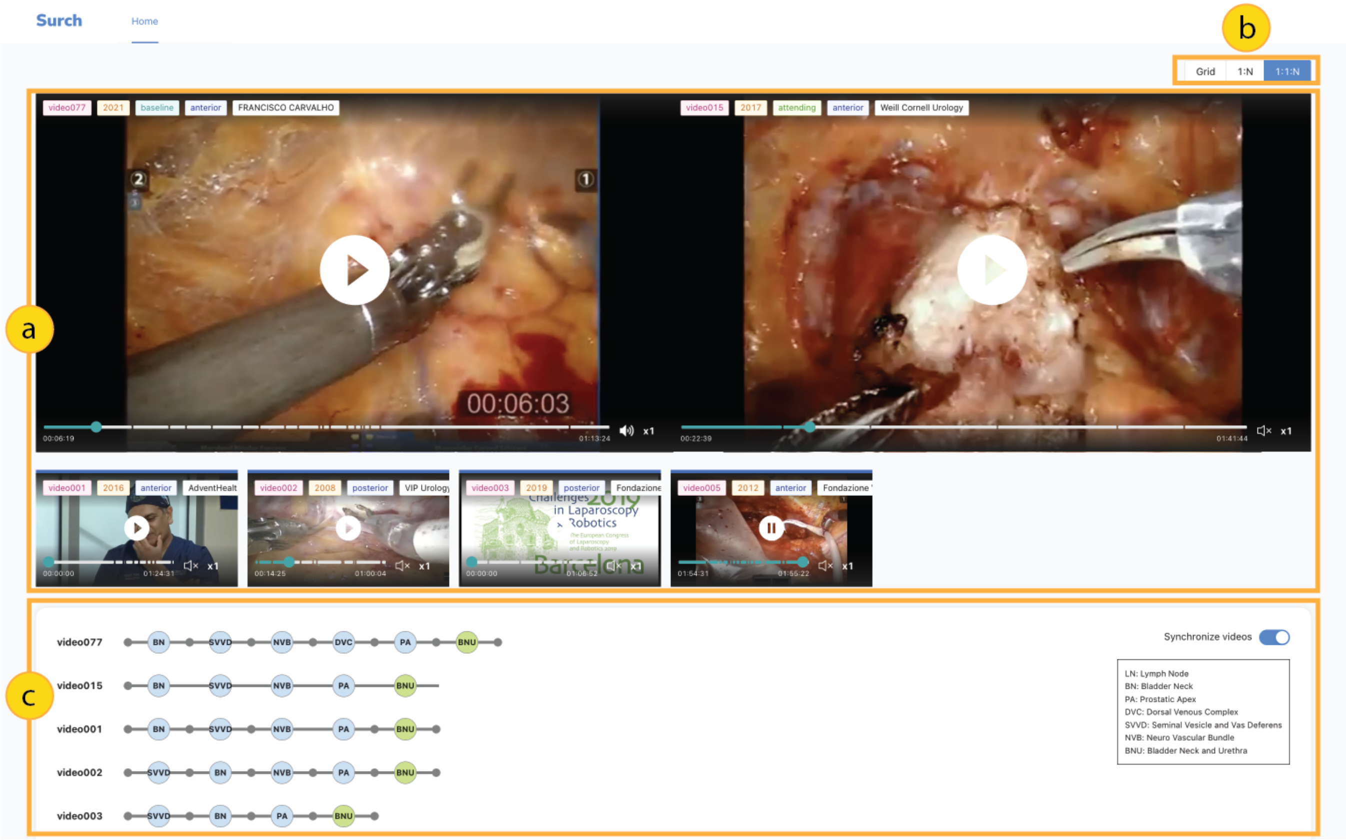


Figure 4: The user can play multiple videos at once and synchronize the playback of videos for a specific phase. (a) Multi-View Videos: the user can watch multiple videos simultaneously in a juxtaposed layout. (b) Layout Option: the user can choose a preferred layout of multiple video windows. (c) Phase-Based Synchronization: the user can sync the playback of multiple videos for specific surgical phases, navigating the videos at the same step in procedures.

They were provided with USD 400 and a bonus of USD 10 for each additional video if they were willing to take on more labeling tasks. We built a video annotation tool for surgical videos. Details of the tool are presented in Supplementary Materials.

6.1.3 Data Filtering. The data labeling process was a two-step process: (1) all labelers first annotated the surgical phases and after the initial labeling, (2) we asked two labelers who were chosen as qualified annotators with high accuracy of labeling to inspect the entire labels.

Given that our labelers were not experts in the surgical domain, we first evaluated each labeler’s level of understanding of prostatectomy by examining the labels for the three example videos. A labeler was verified to contribute labels only if they correctly labeled these videos. When they failed this verification task, we provided additional instructions and learning materials and asked them to re-label the example videos until they annotated them correctly. After the initial labeling, the two qualified labelers inspected the whole dataset. During the inspection process, the two labelers had frequent discussions about using consistent labels with a unified granularity level and terminology. Our team was led by a coauthor who spent 2.5 years engaged in ethnographic research of urologic robotic surgery at 5 hospitals in the U.S. and observed over a thousand hours of procedures. He spent significant time with residents as they prepared for procedures, via surveying videos and practicing via a simulator. This author familiarized the labeling team so that they could reliably label the procedure, and backstopped coding and analysis as needed to ensure accurate ratings. Although we have made efforts to improve the quality of the dataset, careful validation by surgeons would be required, especially for clinical uses such as decision-making support in actual surgery cases. We release this dataset for 296 videos as an open dataset for further research, including the full label set and metadata for video sources [2].

6.1.4 Data Ethics. All videos were from public sources (i.e., YouTube and MUSIC). Each was collected and posted by surgeons or medical residents, who bear full legal and ethical responsibility for the protection of their patients’ identities. Included videos only contain intracorporeal footage and therefore do not disclose patients’ identifying information.

[2] https://github.com/imurs34/robotic_surgery_video_dataset/

7 COMPUTATIONAL PIPELINE: CONVERTING VIDEOS INTO PROCEDURAL GRAPHS

This section describes the technical pipeline that powers the video interface of search and comparison.

7.1 Clustering workflow for surgical procedures

We first classified the surgical procedures based on the surgical approaches — anterior and posterior — depending on the seminal vesicle (SV) and vas deferens (VD) dissection phase, thereby reflecting the most widely used criterion to group the procedures. On top of that, we aimed to reveal latent patterns or unknown structures that separate the large-scale dataset of surgical procedures. Such information can allow discoveries of significant structures of surgeries, ultimately promoting advanced learning of surgical approaches. To this end, we adopted the clustering method, an unsupervised learning algorithm with no need for feeding predetermined group information, which can identify natural structures and latent patterns of surgical procedures.

To further customize the clustering algorithm to our context of quantifying procedural knowledge, we (1) built a set of analysis dimensions for procedural knowledge in surgeries that characterize the procedural structures and (2) designed a clustering workflow with a custom weighting scheme that considers the relative importance between the structural components in procedures. The following sections describe our clustering workflow in detail.

7.1.1 Analysis Dimensions for Procedural Knowledge. We designed customized analysis dimensions to characterize surgical procedures in a large-scale dataset: the number of optional phases, the number of repetitions of phases, the number of branches between phases, and the number of unique paths. Merrill et al. [78]'s three factors that determine the complexity of procedural tasks inspired us in designing analysis dimensions, which include the total number of steps, the number of repetitive sequence structures, and the number of alternate sequence structures. Meanwhile, we co-designed the analysis dimensions with two surgeons to reflect clinical and practical domain expertise. For example, unique paths or routes in a procedure imply uncommon variations of surgical approaches that may be worth noting. Surgeons also found it useful to indicate the optional surgical phases that are not necessarily included in all procedures but are often done by surgeons depending on the distinctive anatomical status and patients' profiles such as age, surgery history, and BMI. We detail these dimensions as follows:

- Number of optional phases: most prostatectomy procedures are standardized and the number of optional phases that are not mandatory in every procedure (e.g., lymph node dissection) characterize the procedure.
- Number of repetitions of phases: repetitive phases such as retouch of seminal vesicle and vas deferens compose repetitive sequence structures.
- Number of branches between phases: several phases have multiple branches diverging into multiple subsequent phases, creating alternative sequence structures.
- Number of unique paths: each branch has different probabilities since some paths are more common than others, which implies that the preceding phase almost always leads to the following phase, while others are unique and rarely observed in the dataset.

We vectorized surgical procedures based on these four dimensions for each phase, which enables the clustering workflow to reflect the main dimensions that represent the structural features of each procedure. To specify, each procedure was converted to a binary vector, quantifying whether an analysis dimension applies to the procedure. The details of the vectorization scheme for each dimension are listed below.

- Number of optional phases: we first determined if a procedure contains optional phases. The list of optional phases consists of DVC dissection and lymph node dissection, which was built based on discussions with the two surgeons. For each optional phase, it was coded as 1 if the phase is contained in the procedure, and 0 if not.
- Number of repetitions of phases: we coded in the same way as the optional phases. It was coded for all seven phases.
- Number of branches between phases: we first used the median split method to convert the procedures into binary vectors. For all seven phases, it was coded as 1 if the number of a phase's branches was larger than the median value of the number of branches of that phase across all the procedures and coded as 0 if not.
- Number of unique paths: we coded in the same way as the number of branches between phrases. It was coded for all seven phases.

7.1.2 Weighting Scheme for Structural Components. Setting proper weights to features is critical in the clustering process, affecting the separation of clusters [5, 27]. We calculated the weights of two factors that were used in vectorization: phases and dimensions for analyzing procedural knowledge. Our weighting scheme aims to reflect the relative importance of each phase and dimension since several phases or dimensions play more decisive roles in characterizing procedures. The details are in the Appendix.

We estimated the weights as a product of two terms: the frequency of phases and dimensions within procedures and the inverse of the frequency of phases and dimensions across procedures. We experimented with the weighting scheme for multiple parameters for optimization. The parameters included logarithmic scale and exponents. To formalize our weighting scheme, we use the following notation. Vectors use arrows, $\vec{x}$, and matrices are in boldface, $\mathbf{X}$. $\vec{x_i}$ represents the vector number i in the set of vectors. $f(x, y)$ is a function of variables x and y. For our dataset of 296 videos, we vectorized the surgical procedures into P using the method in 7.1.1. Each row of $\mathbf{P}$, $\vec{x}$ is a vector expression of a procedure. $\mathbf{W}$ is the weights matrix built based on our weighting scheme for dataset $\mathbf{P}$. Thus, $\mathbf{W}$ has the same dimensions as $\mathbf{X}$. $\mathbf{P}$ is an $m \times n$ matrix, where m is the total number of procedures and n is a multiplication of the number of phases and the number of dimensions. We define a function f as $f(p_i, d) = 1$ if a dimension d applies to a procedure p, and $f(p_i, d) = 0$ otherwise. Meanwhile, the weighting scheme is formalized as in Figure 7, for a weight w.

7.1.3 Weighted Principal Component Analysis. PCA is a classical method to transform high dimensional data into lower dimensional data [62, 63] and a rich body of work revealed that the PCA

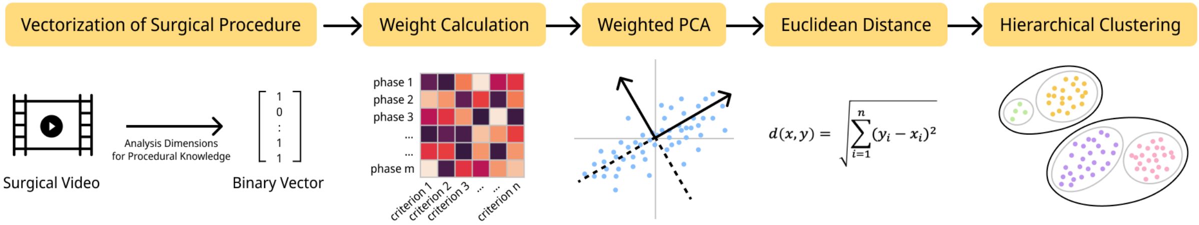


Figure 5: An overview of our clustering workflow. It first vectorizes the surgical procedures based on analysis dimensions for procedural knowledge. It then calculates the weights for the analysis dimensions and phases. The weighted PCA is conducted for the dimension reduction and it runs the hierarchical clustering based on Euclidean distance between procedures.

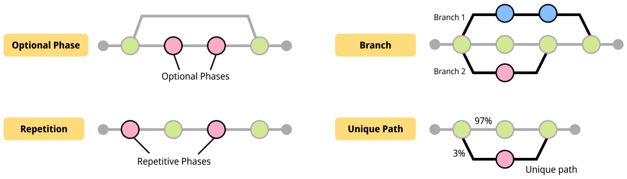


Figure 6: Our four analysis dimensions for surgical procedures: the number of optional phases, the number of repetitions of phases, the number of branches between phases, and the number of unique paths.

enhances the performance of clustering [60]. Our workflow also adopted a weighted PCA that solves the eigenvectors based on expectation maximization [13].

7.2 Experimental Setup

Using our dataset of 296 surgery videos, we experimented with two parameters to optimize the weighting scheme. The parameters consist of scale and exponent. First, we considered linear and logarithmic scales for each term. The logarithmic scale usually dampens the effect of terms, adjusting the effect of a dominant term. We also tested five exponents for each term, between 0 and 1 with an interval of 0.25, creating 25 combinations. The five exponents were evaluated to find an optimal balance between the two terms. Each experiment was repeated 50 times, considering the randomness involved in the clustering procedures. The number of clusters was determined using the elbow method [108].

7.3 Evaluation of Clustering Workflow

The performance of clustering is difficult to evaluate using a single measure, since the goals and contexts of clustering should be considered in the evaluation process [115]. We did not build ground truth labels for clustering, as our goal was to discover latent structural patterns in procedures without assuming predefined answers. Accordingly, we used both quantitative and qualitative methods to evaluate the clustering results. Details of thresholds and the complete evaluation results are in the Supplementary Materials.

7.3.1 Quantitative Measures. We used five internal evaluation measures to assess the quality of clustering from multiple perspectives, including the Silhouette score [100], Davies–Bouldin index [91], the Calinski-Harabasz score [71], the Coefficient of Variation [22], and balance measure [7]. To specify, the Silhouette score indicates the goodness of a clustering result. The Silhouette score ranges from -1 to +1, where a high value indicates that clusters are well apart from each other and clearly distinguished. The Davies-Bouldin score measures the ratio of within-cluster distances (i.e., compactness of the clusters) to between-cluster distances (i.e., cluster separation), where a small value indicates a better separation. The Calinski-Harabasz score is defined as the ratio of the sum of between-cluster dispersion and of within-cluster dispersion, and large values are good ratios. The Coefficient of Variation is the ratio of the standard deviation to the mean and shows the extent of variability in relation to the mean of the population. The higher the score, the greater the dispersion, meaning poorer clustering. The balance measure, on the other hand, represents how the sizes of the clusters are balanced, without predominantly large or small clusters. The balance measure is defined as the maximum cluster size over the minimum cluster size, implying that smaller values with similar cluster sizes mean well-balanced clustering results.

7.3.2 Quantitative Evaluation Results.

In this section, we report the results of the experiment in Section 7.2. Through experimenting with the scale and exponent of various ranges, we sought to validate the impact of dampening each term to better balance the terms.

Scale. Compared to frequency weights of documents and DNA sequences [3], the weights of procedures exhibit large variance since the number of phases in a procedure is smaller than the number of words in a document on average. Thus, applying a logarithmic scale for the first term (i.e., frequency weight term) led to better clustering performance, normalizing the variance. Whereas, the variance in the second term (i.e., the inverse of the frequency weight term) was smaller than the contexts of other domains. The small variance attributed to the standardized structures of procedures. Most surgical phases, except for a couple of them, are essential steps in prostatectomy. This characteristic of the surgery procedure

$$\left(\frac{number\ of\ occurrence(s)\ of\ a\ phase\ in\ a\ procedure}{number\ of\ phase(s)\ of\ a\ procedure}\cdot\frac{total\ number\ of\ procedures}{number\ of\ procedure(s)\ where\ a\ phase\ appears}\right)$$

$$\cdot\left(\frac{number\ of\ phase(s)\ that\ a\ dimension\ applies\ to\ in\ a\ procedure}{number\ of\ phase(s)\ of\ a\ procedure}\cdot\frac{total\ number\ of\ procedures}{number\ of\ procedure(s)\ where\ phase(s)\ that\ a\ dimension\ applies\ to\ appears}\right)$$

$$=\left(\frac{\left|\left\{i\mid c=p_i\right\}\right|}{\left|\vec{p_j}\right|}\cdot\frac{|P|}{\left|\left\{j\mid c=p_j\right\}\right|}\right)\cdot\left(\frac{\left|\left\{i\mid 1=f(p_i,d)\right\}\right|}{\left|\vec{p_j}\right|}\cdot\frac{|P|}{\left|\left\{j\mid 1=f(p_j,d)\right\}\right|}\right)$$

$$for\ phase\ c,\ dimension\ d,\ i{-}th\ phase\ p_i\ in\ j{-}th\ procedure\ \vec{p_j}\ in\ a\ set\ of\ all\ procedures\ P.$$

Figure 7: We designed a weighting scheme for characterizing surgical procedures, a product of two terms: the frequency of phases and dimensions within procedures and the inverse of the frequency of phases and dimensions across procedures.

Exponent of First Term	Exponent of Second Term	Silhouette Score	Coefficient of Variation	Minimum Cluster Size	Balance Measure
0.00	0.00	0.60	6.11	**11.84**	24.18
0.00	0.25	0.69	18.70	5.26	39.43
0.00	0.50	0.68	26.06	**14.02**	14.98
0.00	0.75	0.66	23.30	**14.64**	15.82
0.00	1.00	0.56	**4.91**	**15.70**	11.09
0.25	0.00	0.68	28.41	**13.30**	14.97
0.25	0.25	0.60	8.77	**16.78**	13.04
0.25	0.50	0.65	**3.75**	**14.52**	38.74
0.25	0.75	0.58	**1.65**	5.20	51.70
0.25	1.00	**0.73**	136.21	**11.94**	27.05
0.50	0.00	0.61	5.49	**12.04**	25.37
0.50	0.25	0.61	**0.40**	1.32	175.30
0.50	0.50	**0.74**	14.37	4.52	45.10
0.50	0.75	**0.75**	7.40	4.74	51.29
0.50	1.00	0.64	28.62	2.44	84.68
0.75	0.00	0.68	14.72	**12.52**	17.73
0.75	0.25	0.93	55.63	2.00	111.00
0.75	0.50	0.56	**4.37**	**18.70**	48.59
0.75	0.75	**0.76**	20.22	**17.14**	35.26
0.75	1.00	**0.76**	**2.49**	**19.08**	10.77
1.00	0.00	0.65	25.61	**7.44**	29.44
1.00	0.25	0.61	6.83	**33.58**	7.14
1.00	0.50	**0.83**	71.63	3.84	65.02
1.00	0.75	**0.70**	53.53	**6.92**	42.68
1.00	1.00	0.67	13.62	**6.12**	44.51

Figure 8: Quantitative evaluation results of clustering workflow for the anterior approach with the logarithmic scale for the first term and the linear scale for the second term. The experimental parameters are scale and exponent. The evaluation metrics include the Silhouette score, coefficient of variation, minimum cluster size, and balance measure. The results for Davies–Bouldin index and Calinski-Harabasz score, and the results for the posterior approach are in Supplementary Materials. The values that satisfy the threshold are bolded.

dataset results in a better clustering performance when not using the logarithmic scale for the second term.

Exponent. We could find optimal exponents for each surgical approach based on the results of experiments - 0.75 and 1.0 for the anterior approach and 0.25 and 0.5 for the posterior approach. According to literature [42, 55], we set thresholds for the Silhouette score, coefficient of variation, and minimum cluster size as 0.7, 5.0, and 5% of the dataset size, respectively. After filtering using the thresholds, we chose the optimal exponent for which the results have the highest Silhouette score, Calinski-Harabasz score, minimum cluster size, and the lowest Davies–Bouldin index, coefficient of variation, balance measure.

	Silhouette Score	Balance Measure
Surch	0.82	11.61
No vectorization	0.50	5.50
No weighting	0.70	114.70

Figure 9: A result of a comparative experiment with three conditions, Surch, the workflow with no vectorization, the workflow with no weighting. The Surch workflow achieved the highest Silhouette score of and a reasonably low balance measure compared to the other two conditions.

Findings. The experimental results indicate that both frequency weight and the inverse term should be considered in calculating the importance of phases and analysis dimensions, showing low Silhouette scores when either of the exponents is zero, in other words, when one of the terms is not considered. The evaluation also demonstrates the importance of considering multiple evaluation criteria. For example, a high Silhouette score does not necessarily indicate well-structured clustering, sometimes having imbalanced cluster sizes such as when exponents are 1.0, 0.5 for each term. We also found that normalization only depending on the scale is not sufficient, but further optimization is needed by experimenting with different exponents. Meanwhile, the optimal exponents differ depending on the dataset, anterior or posterior. The first term was more important in the anterior approach, where the variance within procedures is low, which indicates that the local features play more roles in characterizing procedures. Another trend we could observe is that small exponents for the first terms led to too scattered clustering results, not sufficiently reflecting local features of individual procedures. A possible reason is that the effect of the first terms has already been dampened by applying the logarithmic scale, thereby causing a double reduction with small exponents.

7.3.3 Comparative Experiment of Quantitative Evaluation. We validated our clustering workflow by comparing it to the workflow without a vectorization and workflow without a weighting scheme. For the condition without the vectorization, we used the workflow introduced from the existing work [23] that extracts the notable patterns of components and actions from the procedures. For the condition without the weighting scheme, from our workflow, we replaced our weighting scheme with the common frequency weighting scheme [57]. In consideration of the randomness of clustering algorithms, we conducted 50 repeated experiments.

As summarized in Figure 9, the Surch workflow achieved the highest Silhouette score of 0.82 and a reasonably low balance measure of 11.61 compared to the other two conditions. The workflow with no vectorization achieved the Silhouette score of 0.5 and the balance measure of 5.5, whose low Silhouette Score indicates that it does not detect substantial structure from data. Meanwhile, the workflow with no weighting scheme achieved the Silhouette score of 0.7 and the balance measure of 114.7. The high balance measure implies that it generates imbalanced clusters with small clusters with a few instances and a big cluster including most of the instances.

Based on our observation of the clustering results, the instances are too spread out, not forming discernible groups without the vectorization. This result implies that the vectorization using our analysis dimensions allows for forming compact clusters, reducing noise in the procedure data by focusing on representative structural features instead of insignificant phases or orders. Meanwhile, with no weighting scheme, the clustering results in one major cluster containing most instances except for one of two instances. This imbalanced result indicates that the consideration of the relative importance of components in the procedures by our weighting scheme leads to clear separation between clusters.

In summary, the vectorization decreases intra-cluster distances and the weighting scheme increases the inter-cluster distance, which leads to compact clustering with little overlap.

7.3.4 Qualitative Measure. We conducted an expert evaluation to complement the limitations of quantitative evaluation by asking about experts' opinions on the quality of the produced clusters. We collected feedback from seven senior residents (PGY 3, 4, and 5) [R1-7] on the effectiveness and interpretability of the clustering results during the user study. Residents highlighted two sides of the clustering results: findability and discoverability [72]. From the perspective of findability that supports goal-oriented seeking for specific information, they found it useful to know the common approaches by referring to the largest cluster. All residents explained that the grouping aligns with what they learned, for example, R5 (PGY 5) mentioned the re-touch of a neurovascular bundle is one of the key differences between surgeons' approaches. On the other hand, others noticed interesting patterns from the clustering results, elaborating that "I've never seen somebody doing other steps between the bladder neck dissection and prostatic apex dissection. It's interesting to see these routes in this group." (R2 - PGY 4). Some residents pointed out that cluster names were not intuitive. Our system automatically generated cluster labels such as "recursive NVB dissection" based on the analysis dimensions for surgical procedures, but the residents explained that they were not familiar with such terms and suggested alternative cluster labels that emphasize more contextual features of procedures in clusters. For example, one of the auto-generated cluster labels was "branchy PA dissection", meaning that there exist divergent paths after the PA dissection phase, but one resident (R5 - PGY5) proposed "lymph node dissection after PA dissection" as an alternative. He added that the name could reflect clinically meaningful features of procedures.

7.3.5 Model Training. We aimed to achieve two goals by training the model on our dataset: (1) validating the quality of the dataset by testing if the model is able to learn from the data and generalize well and (2) increasing the generalizability of the Surch pipeline to afford new video dataset without labels.

The class includes seven surgical phases, which are SVVD (Seminal vesicle and vas deferens Dissection), LN (Lymph node Dissection), BN (Bladder neck Dissection), NVB (Neurovascular Bundle Dissection), DVC (Dorsal Venous Complex Dissection), PA (Prostate Apex Dissection), and BNU (Bladder neck and urethra Suture).

First, to validate the annotated dataset (Section 6), we trained a basic CNN-LSTM model [58, 70] without any fine-tuning or data preprocessing and achieved an accuracy of 48% (sequence length: 2, model layer: 3, learning rate: .01, .001, .0001, batch size:128). Considering that the baseline model without fine-tuning or architecture change showed accuracies of 41% [88], 69% [106], and 53% [94] for the existing surgical phase dataset (e.g., Cholec80, M2CAI16-workflow), our accuracy of 48% that is comparable to the ones for baseline conditions from previous research indicates that the labels are valid enough to train models [46, 47].

After testing the baseline, we fine-tuned the model on our dataset. We utilized focal loss [69] to mitigate the class imbalance issue, and adopted an ensemble method [93] to improve the overall performance. We compared CNN, CNN-LSTM hybrid, and average ensemble techniques, and the ensemble showed the highest performance. Meanwhile, we designed a CNN-LSTM layer followed by a single fully connected layer and softmax layer. We pretrained the CNN using ImageNet [32] due to the lack of a large-scale prostatectomy dataset. We used an 80/20 split for training and test dataset. The details of hyperparameters are provided in Supplementary Materials. We achieved an accuracy of 78.4% with an overall precision of 67.3%. In classifying surgical phases, high precision is more desirable than high recall, as incorrect labeling can cause confusion to residents, especially novices. To the best of our knowledge, there exists no surgical phase detection model for prostatectomy that is publically available. 78.4% was enough for us to power the system, but we expect to collaborate with the computer vision community to further improve the accuracy. Even with the improvement, however, achieving 100% accuracy is difficult, while the incorrect detection can especially mislead novice residents. To reduce the danger of providing inaccurate classification results, the confidence level of the algorithm's results can be displayed so that residents are aware of how reliable the system is. We release the model and code as opensource [3], and expect our model to be utilized by future researchers as a baseline for surgical phase detection in prostatectomy.

8 USER STUDY

We conducted a controlled user study that demonstrated the effects of our system for surgical learners. The goals of this study were to evaluate the effectiveness of Surch in video search and comparison, to investigate how Surch affected surgical learning, and to explore the potential of structural search and comparison

[3] https://github.com/imurs34/surch-model/

for procedural video. The study was a within-subjects design, and participants used two different video interfaces: baseline YouTube video interface and Surch, which supported structural video search and comparison through procedural graphs. The order of conditions and tasks was counterbalanced across participants. Our video library contained 296 video recordings of robotic prostatectomy. 246 were from YouTube and 50 were from the MUSIC library.

8.1 Participants

We recruited 11 participants [P1-P11] (2 female and 9 male) from the two U.S. medical schools by contacting surgeons through email. They consist of 2 PGY1, 2 PGY2, 2 PGY3, 2 PGY4, and 3 PGY5 residents in Urology. All of them had prior experience using video to learn about surgery and surgical technique. They received 150 USD for up to a 1-hour study.

8.2 Procedure

The study was conducted using remote conferencing software, and video-recorded with consent. The participants were first given a 5-minute instruction session that introduced the features and interfaces of Surch. They then familiarized themselves with Surch. After this exploration, they were asked to complete two tasks in a real-world video-based surgical learning environment. The tasks were designed based on discussions with two robotic Urologic surgeons and four residents at top US teaching hospitals to have pedagogical value and reflect natural learning practice. The first task instructed the participants to "Find three different approaches to the procedure and highlight 2-3 ways they differ." for junior residents and "Given your attending's surgery recording, find a video showing an approach to the procedure that is most contrasting and the one that is most similar to your attending's approach." for senior residents The second task asked, "Given the approach to the seminal vesicle and vas deferens dissection in the video XX, find 2 videos that contain markedly different approaches to the seminal vesicles." After the task sessions, the participants completed a questionnaire on confidence in understanding a surgery, such as its landscape and techniques, perceived learning experience and learning efficiency, cognitive load, willingness to use, and ease of use. We then followed up with a post-session interview that probed their perception of each video interface and the reasons behind their usage behaviors. We scored their task submissions based on discussions with surgeons and the following criteria: the number of (1) differences and similarities of procedures and (2) meaningful comparison points, including surgical techniques and approaches. We also evaluated the performance of tasks using quantitative measures that include time taken for searching videos, navigating to surgical phases, and comparing multiple videos.

8.3 Results

8.3.1 User Behavior. All participants actively used search and comparison features supported by Surch. In particular, they searched for videos depending on the approaches and filtered videos based on the structures of the procedure. They used both surgical phases and paths for search, manipulating nodes and edges in the procedural graph. Several participants (4/11) completed the tasks without watching the videos at all, saying that the structural information provided by Surch was enough to notice the comparison points between the surgeries. Meanwhile, three participants played multiple videos sequentially and the rest of them played them at the same time. One participant (P8) even played five different videos synchronously. In contrast, the participants using the baseline player attempted to use text keywords to search for surgical approaches and phases, such as anterior and SV. No participant used the text-based search bar on Surch. One participant mentioned that Surch's structural search feature eliminated the need for the search bar associated with the baseline video search interface.

8.3.2 Task Performance. The participants watched more videos when using Surch (M = 8.5, std = 4.8) than using the baseline player (M = 6.2, std = 2.5) with $p < 0.05$. It took slightly less time to complete the tasks when using Surch (M = 10.9 minutes, std = 6.3) than via the baseline player (M = 12.0, std = 7.8). However, there was no large difference in the number of videos watched per minute (Surch: M = 0.8, std = 0.2, baseline: M = 0.7, std = 0.5) with p = 0.27.

8.3.3 Response Quality. The results of task scoring showed that the participants submitted slightly better responses for the given tasks of searching and comparing surgical procedures when using Surch (M = 3.7, std = 0.6) than the baseline player (M = 3.6, std = 0.7) with no significant difference using the paired t-test (p = 0.32). Meanwhile, we could observe that junior residents showed a bigger difference in response quality (baseline: M = 2.8, std = 1, Surch: M = 3.5, std = 1) than seniors (baseline: M = 3.4, std = 0.7, Surch: M = 3.6, std = 0.8), indicating that Surch could be more beneficial to junior or novices with less prior knowledge regarding the quality of comparison points they found.

8.3.4 User Perception. Overall, participants found Surch enhances search and comparison for surgery learning. Figure 10 summarizes the survey responses. To analyze the survey responses, we used a Wilcoxon signed-rank test.

Enhance surgical learning.

On a 7-point Likert scale question (1: strongly disagree, 7: strongly agree), the participants reported that they could better understand the landscape of prostatectomy ($p < 0.01$). They showed significantly higher satisfaction with their learning experience when using Such ($p < 0.005$). The junior residents, in particular, appreciated the structural representation of the procedures highlighting the educational value of the procedural graph. For example, P5 (PGY1) said, "as an early trainee, it is hard for me to discern the surgical steps without the graph. The graph helped me think through each stage of the procedure." P1 (PGY1) also stated, "(Using the procedural graph) I like that I could see overarching steps of prostatectomies and nuanced differences between videos.". Meanwhile, several participants mentioned that Surch helps them achieve their learning objectives. P9 (PGY5) mentioned that Surch "allows surgeons and trainees to seek out teachable moments in a surgery." P6 (PGY 5) also added that "we, the trainees want to watch a lot of videos for a specific step with different ways. This (video player) would help do this."

Efficient video search and comparison.

The participants responded that our system made the video search and comparison more efficient (p = 0.005). P5 (PGY 5) explained that "filtering by the graph makes it easy to find what I'm looking for from the videos.". P8 (PGY 3) highlighted that "the fact

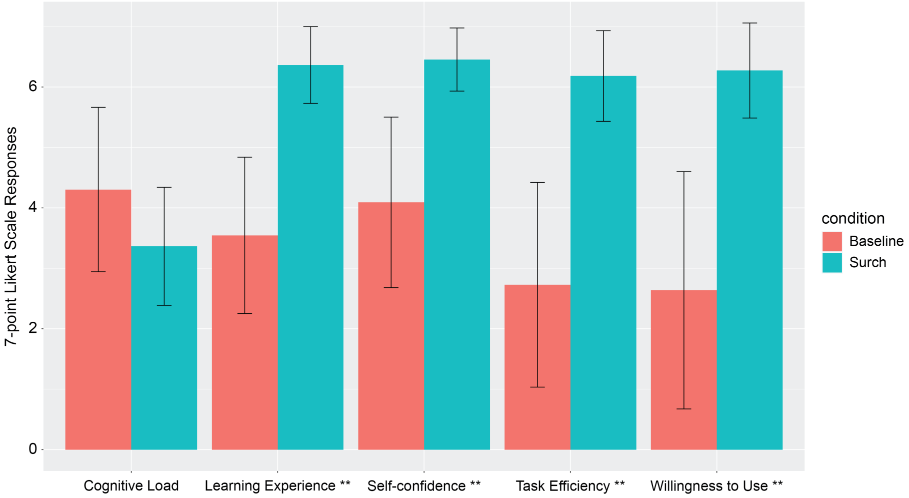


Figure 10: A summary of survey responses of 7-point Likert scale for cognitive load, learning experience, self-confidence, task efficiency, and willingness to use. * p < 0.05, ** p < 0.01. The error bars are defined by one standard deviation with the 68% confidence interval.

that you can sync it up the same phase of the surgery across different videos saves us a lot of time when we're trying to review a lot of videos, focusing on nuances of approaches.". Also, they found it useful to know the commonality of procedures based on the thickness of the edges of the graph.

Feasible system.

There was no significant difference in cognitive load for both video interfaces (p = 0.12). On the other hand, the participants reported a significantly higher willingness to use (p < 0.01) Surch. They suggested various use cases of Surch in their daily learning. For example, P6 (PGY 5) said that "the ability to search videos by approach and key steps would be useful, especially when reviewing or evaluating different ways to approach surgeries." P9 (PGY 5) noted that "I really like how the graph helps the video search because surgery can be done in so many different ways. I sometimes try to find a similar approach to faculty or fellow, and this can help better anticipate the progression of steps in doing the cases." However, some participants (P 3, 4, 6, 7) pointed out that the procedural graph seemed complicated. P8 (PGY 3) mentioned information overload caused by the multiple paths included in the graph.

9 DISCUSSION AND FUTURE WORK

This section discusses the design implications of search and comparison tools for videos with procedural knowledge.

9.1 Generalizability and Domain Specificity

9.1.1 Broad User Needs for Procedural Knowledge. Our approach is generalizable across surgical procedures such as OB-Gyn, Colorectal, General surgery, given that most robotic laparoscopic surgeries share phases and challenges such as blunt dissection of fascia, retraction of organs and tissues, dissection of cancerous masses. Furthermore, during our interviews with surgeons and residents, we observed user needs that overlapped with those associated with a broad range of videos containing procedural knowledge such as cooking, make-up, and assembly. This occurred in two ways. The first was that procedural knowledge often includes the steps in the process of performing a task or skill [37, 38]. A tool such as Surch could thus facilitate learning, as video content in these domains must also be segmented into semantic phases to make information related to step structures available to viewers. The benefits of offering this capability seem clear, as such information has been shown to enhance the video-watching experience in learning to apply make-up [113] and use software tools [66], as it has with surgery videos [85]. Reinforcing this, residents in our studies perceived surgical phases as an essential unit of the stepwise structure associated with a procedure and wanted to search, navigate, and compare videos based on these phases. We expect other populations would have similar learning interests in other procedural knowledge domains.

The second way our work may generalize is in its value for facilitating the acquisition of tacit knowledge. The tacit knowledge residents carefully attend to in videos includes how much traction is applied to various tissues and how bleeders (small capillary injuries caused by the surgeon) are cauterized. Our interviewees emphasized the importance of comparing multiple cases to see the differences in these intricate nuances. While such tacit knowledge is critical for competence in the surgery domain [18, 75], it is evident in a wide range of skilled human activity, such as cyclists' breathing methods and carpenters' hammering skills that they unconsciously perform [96, 104]. A system such as Surch thus has promise for facilitating tacit knowledge acquisition in procedural knowledge domains beyond the surgical.

9.1.2 User Needs Specific to Surgery Domain. These potential broader contributions aside, we found and addressed user needs specific to surgical education. First, there were distinct, strong needs for video search and comparison that varied based on the expertise of residents. Junior residents (PGY1, 2, 3) expressed their need to understand procedures' overall phase structure by watching videos of standard procedures. On the other hand, senior residents (PGY4, 5, 6) focused on variations and details within each phase. These more advanced students who generally sought to compare multiple techniques and approaches to develop their surgical flexibility. Attending surgeons independently verified these differing needs.

Meanwhile, learning surgery differs from acquiring general procedural knowledge because of the complexity of the task. Surgery— for example dissecting a bladder neck as part of the removal of a cancerous prostate— involves far greater uncertainty, dynamism, and risk than many procedural knowledge domains such as food preparation and programming. In particular, the former task involves more complicated considerations such as identifying anatomical landmarks, applying traction to deformable, delicate and interconnected tissues, and cleaning up bleeding, all while being ready to respond to potentially catastrophic failure modes. This complexity requires deep and diverse skills and drives residents' need for exposure to multiple techniques and approaches to be prepared for a wide range of situations.

Lastly, the unique characteristics of surgery recordings pose challenges for video search and navigation. Most surgery videos are quite long (several hours on average), contain numerous scenes that are ostensibly very similar, and lack audio or textual annotation. This lack of semantic indicators for video content poses a significant challenge for video-based learning, in particular, making video search and comparison time-consuming.

9.2 Facilitating interpretation of clustering results

Residents indicated that our clustering workflow output was valuable in two ways, evident in prior work [72]: findability and discoverability. Findability relates to the ability to find content that users already know or assume, and thus supports goals-directed search. On the other hand, discoverability relates to unexpectedly encountering interesting information, which flows from loosely-directed exploration. Some residents explained that our clustering results aligned with what they had already learned and therefore helping them to search for videos they were looking for. Other residents relied on our clustering results to notice new surgical approaches they had not seen before. Our study thus shows that procedural graphs can allow interesting discoveries as well as support the search for specific information.

Residents' perceptions of cluster labels differed by their level of experience. Junior residents took labels as read due to their lack of experience and prior knowledge. One junior resident from the user study used cluster labels to describe the noticeable difference between procedures while doing the comparison task, for example. On the other hand, senior residents generated their own interpretation, relying more on observation of our graph representation than the cluster labels. They wished to see cluster labels containing clinically meaningful features of each cluster, while our auto-generated labels reflected the characteristics of procedures' structures. In future work, we expect to involve experts to generate more informative and interpretable clustering results with labels that are more valuable to advanced learners.

A future system could first perform grouping of large-scale data with procedural knowledge according to their structural features and then experts can participate in label generation in natural language for the clusters, which would better reflect teachable and medically meaningful points. Residents' feedback about the value of discoverability also indicates that a richer human-machine collaboration to create more sensible clustering can also promote surgical learning, as it will provide experts with exposure to new surgical approaches.

10 CONCLUSION

This paper introduces Surch, a video interface that enables structural search and comparison of surgical procedures. Surch allows users to search videos based on a procedural graph generated from a custom dataset and clustering workflow. Our clustering workflow captures unknown patterns of surgical procedures by using vectorization and weighting schemes that characterize the structural features of procedural knowledge, such as recursive structures and unique paths. Surch also enables cross-video comparison by providing synchronized video navigation to the same surgical steps. Our evaluation with quantitative and qualitative measures demonstrated the effectiveness of the clustering results for surgical learning. A user study (N = 11) revealed that our system significantly improves the learning experience and perceived task efficiency of video search and comparison.

ACKNOWLEDGMENTS

This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2021-0-01347, Video Interaction Technologies Using Object-Oriented Video Modeling). This work was also supported by KAIST End Run project.

REFERENCES

[1] Jad M Abdelsattar, TK Pandian, Eric J Finnesgard, Moustafa M El Khatib, Phillip G Rowse, EeeLN H Buckarma, Becca L Gas, Stephanie F Heller, and David R Farley. 2015. Do you see what I see? How we use video as an adjunct to general surgery resident education. *Journal of surgical education* 72, 6 (2015), e145–e150.

[2] Akgul Ahmet, Kus Gamze, Mustafaoglu Rustem, and Karaborklu Argut Sezen. 2018. Is video-based education an effective method in surgical education? A systematic review. *Journal of surgical education* 75, 5 (2018), 1150–1158.

[3] Akiko Aizawa. 2003. An information-theoretic perspective of tf–idf measures. *Information Processing & Management* 39, 1 (2003), 45–65.

[4] Ihab Al Kabary and Heiko Schuldt. 2013. SportSense: using motion queries to find scenes in sports videos. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. 2489–2492.

[5] Muna Al-Razgan and Carlotta Domeniconi. 2006. Weighted clustering ensembles. In *Proceedings of the 2006 SIAM International Conference on Data Mining*. SIAM, 258–269.

[6] David Alderson. 2010. Developing expertise in surgery. *Medical teacher* 32, 10 (2010), 830–836.

[7] Carlos Alzate and Johan AK Suykens. 2008. Multiway spectral clustering with out-of-sample extensions through weighted kernel PCA. *IEEE transactions on pattern analysis and machine intelligence* 32, 2 (2008), 335–347.

[8] Laith Alzubaidi, Muthana Al-Amidie, Ahmed Al-Asadi, Amjad J Humaidi, Omran Al-Shamma, Mohammed A Fadhel, Jinglan Zhang, J Santamaría, and Ye Duan. 2021. Novel transfer learning approach for medical imaging with limited labeled data. *Cancers* 13, 7 (2021), 1590.

[9] Soheyla Amirian, Khaled Rasheed, Thiab R Taha, and Hamid R Arabnia. 2021. Automatic generation of descriptive titles for video clips using deep learning. In *Advances in Artificial Intelligence and Applied Cognitive Computing*. Springer, 17–28.

[10] Jurgen Assfalg, Marco Bertini, Carlo Colombo, and Alberto Del Bimbo. 2002. Semantic annotation of sports videos. *IEEE MultiMedia* 9, 2 (2002), 52–60.

[11] Knut Magne Augestad, Khayam Butt, Dejan Ignjatovic, Deborah S Keller, and Ravi Kiran. 2020. Video-based coaching in surgical education: a systematic review and meta-analysis. *Surgical endoscopy* 34, 2 (2020), 521–535.

[12] Ignacio Avellino, Sheida Nozari, Geoffroy Canlorbe, and Yvonne Jansen. 2021. Surgical Video Summarization: Multifarious Uses, Summarization Process and Ad-Hoc Coordination. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–23.

[13] Stephen Bailey. 2012. Principal component analysis with noisy and/or missing data. *Publications of the Astronomical Society of the Pacific* 124, 919 (2012), 1015.

[14] Guha Balakrishnan, Frédo Durand, and John Guttag. 2015. Video diff: Highlighting differences between similar actions in videos. *ACM Transactions on Graphics (TOG)* 34, 6 (2015), 1–10.

[15] Garth H Ballantyne. 2002. Robotic surgery, telerobotic surgery, telepresence, and telementoring. *Surgical Endoscopy and Other Interventional Techniques* 16, 10 (2002), 1389–1402.

[16] Siddhant Bansal, Chetan Arora, and CV Jawahar. 2022. My View is the Best View: Procedure Learning from Egocentric Videos. In *European Conference on Computer Vision*. Springer, 657–675.

[17] Kai Uwe Barthel, Nico Hezel, and Radek Mackowiak. 2015. Graph-based browsing for large video collections. In *International Conference on Multimedia Modeling*. Springer, 237–242.

[18] Matthew Beane. 2019. Shadow learning: Building robotic surgical skill when approved means fail. *Administrative Science Quarterly* 64, 1 (2019), 87–123.

[19] Pierfrancesco Bellini, Paolo Nesi, and Marco Serena. 2015. MyStoryPlayer: experiencing multiple audiovisual content for education and training. *Multimedia Tools and Applications* 74, 18 (2015), 8219–8259.

[20] James Bergstra and Yoshua Bengio. 2012. Random search for hyper-parameter optimization. *Journal of machine learning research* 13, 2 (2012).

[21] H Russell Bernard and Harvey Russell Bernard. 2013. *Social research methods: Qualitative and quantitative approaches*. Sage.

[22] Charles E Brown. 1998. Coefficient of variation. In *Applied multivariate statistics in geohydrology and related sciences*. Springer, 155–157.

[23] Minsuk Chang, Léonore V Guillain, Hyeungshik Jung, Vivian M Hare, Juho Kim, and Maneesh Agrawala. 2018. Recipescape: An interactive tool for analyzing cooking instructions at scale. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–12.

[24] Minsuk Chang, Ben Lafreniere, Juho Kim, George Fitzmaurice, and Tovi Grossman. 2020. Workflow graphs: A computational model of collective task strategies for 3d design software. (2020).

[25] Shih-Fu Chang, William Chen, Horace J Meng, Hari Sundaram, and Di Zhong. 1997. VideoQ: an automated content based video search system using visual cues. In *Proceedings of the fifth ACM international conference on Multimedia*. 313–324.

[26] Shih-Fu Chang, Wei-Ying Ma, and Arnold Smeulders. 2007. Recent advances and challenges of semantic image/video search. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, Vol. 4. IEEE, IV–1205.

[27] Hao Cheng, Kien A Hua, and Khanh Vu. 2008. Constrained locally weighted clustering. *Proceedings of the VLDB Endowment* 1, 1 (2008), 90–101.

[28] Pei-Yu Chi, Sally Ahn, Amanda Ren, Mira Dontcheva, Wilmot Li, and Björn Hartmann. 2012. MixT: automatic generation of step-by-step mixed media tutorials. In *Proceedings of the 25th annual ACM symposium on User interface software and technology*. 93–102.

[29] Tobias Czempiel, Magdalini Paschali, Daniel Ostler, Seong Tae Kim, Benjamin Busam, and Nassir Navab. 2021. Opera: Attention-regularized transformers for surgical phase recognition. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 604–614.

[30] Martin Dawes and Marko Lens. 2007. Knowledge transfer in surgery: skills, process and evaluation. *The Annals of The Royal College of Surgeons of England* 89, 8 (2007), 749–753.

[31] Giuseppe Delmestri and Peter Walgenbach. 2005. Mastering techniques or brokering knowledge? Middle managers in Germany, Great Britain and Italy. *Organization Studies* 26, 2 (2005), 197–220.

[32] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[33] Yannick Dupraz. 2013. Using weights in Stata. *Accessed on August* 10 (2013), 2017.

[34] K Anders Ericsson. 2004. Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. *Academic medicine* 79, 10 (2004), S70–S81.

[35] Simon Erridge, Derek KT Yeung, Hitendra RH Patel, and Sanjay Purkayastha. 2019. Telementoring of surgeons: a systematic review. *Surgical innovation* 26, 1 (2019), 95–111.

[36] Sam Fujisaka. 1986. Informal Structured Iterative Interview: A Method and Short Training Course for the central Visayas Research Consortium. *Philippine quarterly of culture and society* 14, 3 (1986), 263–274.

[37] Robert M Gagne. 1988. Mastery learning and instructional design. *Performance Improvement Quarterly* 1, 1 (1988), 7–18.

[38] Robert M Gagne, LJ Briggs, and WW Wager. 1988. Instructional design. *New York: Holt, Rinehart, Winston* (1988).

[39] Carly R Garrow, Karl-Friedrich Kowalewski, Linhong Li, Martin Wagner, Mona W Schmidt, Sandy Engelhardt, Daniel A Hashimoto, Hannes G Kenngott, Sebastian Bodenstedt, Stefanie Speidel, et al. 2021. Machine learning for surgical phase recognition: a systematic review. *Annals of surgery* 273, 4 (2021), 684–693.

[40] Chrysoula Gkonela and Konstantinos Chorianopoulos. 2014. VideoSkip: event detection in social web videos with an implicit user heuristic. *Multimedia Tools and Applications* 69, 2 (2014), 383–396.

[41] Ira P Goldstein. 1979. The genetic graph: a representation for the evolution of procedural knowledge. *International Journal of Man-Machine Studies* 11, 1 (1979), 51–77.

[42] E Guang-Xin, Bai-Gao Yang, Yan-Bin Zhu, Xing-Hai Duang, Wang-Dui Basang, Xiao-Lin Luo, and Tian-Wu An. 2020. Genome-wide selective sweep analysis of the high-altitude adaptability of yaks by using the copy number variant. *3 Biotech* 10, 6 (2020), 1–6.

[43] Annetje CP Guédon, Senna EP Meij, Karim NMMH Osman, Helena A Kloosterman, Karlijn J van Stralen, Matthijs Grimbergen, Quirijn AJ Eijsbouts, John J van den Dobbelsteen, and Andru P Twinanda. 2021. Deep learning for surgical phase recognition using endoscopic videos. *Surgical Endoscopy* 35, 11 (2021), 6150–6157.

[44] Pranabjyoti Haloi and MK Bhuyan. 2021. Video Searching and Retrieval using Scene Classification in Multimedia Databases.. In *2021 2nd International Conference for Emerging Technology (INCET)*. IEEE, 1–7.

[45] Sanjay Haresh, Sateesh Kumar, Huseyin Coskun, Shahram N Syed, Andrey Konin, Zeeshan Zia, and Quoc-Huy Tran. 2021. Learning by aligning videos in time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5548–5558.

[46] Monica Haurilet, Ziad Al-Halah, and Rainer Stiefelhagen. 2019. Spase-multi-label page segmentation for presentation slides. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 726–734.

[47] Monica Haurilet, Alina Roitberg, Manuel Martinez, and Rainer Stiefelhagen. 2019. Wise—slide segmentation in the wild. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 343–348.

[48] S Duke Herrell and Joseph A Smith Jr. 2005. Robotic-assisted laparoscopic prostatectomy: what is the learning curve? *Urology* 66, 5 (2005), 105–107.

[49] András Hoznek, Laurent Salomon, Leif Eric Olsson, Patrick Antiphon, Fabien Saint, Antony Cicco, Dominique Chopin, and Clément-Claude Abbou. 2001. Laparoscopic radical prostatectomy. *European urology* 40, 1 (2001), 38–45.

[50] Daniel J Hruschka, Deborah Schwartz, Daphne Cobb St. John, Erin Picone-Decaro, Richard A Jenkins, and James W Carey. 2004. Reliability in coding open-ended data: Lessons learned from HIV behavioral research. *Field methods* 16, 3 (2004), 307–331.

[51] Weiming Hu, Nianhua Xie, Li Li, Xianglin Zeng, and Stephen Maybank. 2011. A survey on visual content-based video indexing and retrieval. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 41, 6 (2011), 797–819.

[52] Marco A Hudelist, Sabrina Kletz, and Klaus Schoeffmann. 2016. A Multi-Video Browser for Endoscopic Videos on Tablets. In *Proceedings of the 24th ACM international conference on Multimedia*. 722–724.

[53] Marco A Hudelist, Sabrina Kletz, and Klaus Schoeffmann. 2016. A tablet annotation tool for endoscopic videos. In *Proceedings of the 24th ACM international conference on Multimedia*. 725–727.

[54] Wolfgang Hürst, Rob van de Werken, and Miklas Hoet. 2015. A storyboard-based interface for mobile video browsing. In *International Conference on Multimedia Modeling*. Springer, 261–265.

[55] Sadiq Hussain, LJ Muhammad, FS Ishaq, Atomsa Yakubu, and IA Mohammed. 2019. Performance evaluation of various data mining algorithms on road traffic accident dataset. In *Information and Communication Technology for Intelligent Systems*. Springer, 67–78.

[56] Andrew M Ibrahim, Oliver A Varban, and Justin B Dimick. 2016. Novel uses of video to accelerate the surgical learning curve. *Journal of Laparoendoscopic & Advanced Surgical Techniques* 26, 4 (2016), 240–242.

[57] Tasha R Inniss, John R Lee, Marc Light, Michael A Grassi, George Thomas, and Andrew B Williams. 2006. Towards applying text mining and natural language processing for biomedical ontology acquisition. In *Proceedings of the 1st international workshop on Text mining in bioinformatics*. 7–14.

[58] Md Zabirul Islam, Md Milon Islam, and Amanullah Asraf. 2020. A combined deep CNN-LSTM network for the detection of novel coronavirus (COVID-19) using X-ray images. *Informatics in medicine unlocked* 20 (2020), 100412.

[59] Kenneth Jacobsohn, Margaret Mulligan, Lance Hampton, and Scott Johnson. 2015. Tutorial for Robotic Prostatectomy. *MedEdPORTAL* 11 (2015), 10307.

[60] Mohammad Jafarzadegan, Faramarz Safi-Esfahani, and Zahra Beheshti. 2019. Combining hierarchical clustering approaches using the PCA method. *Expert Systems with Applications* 137 (2019), 1–10.

[61] Anil K Jain, Aditya Vailaya, and Xiong Wei. 1999. Query by video clip. *Multimedia systems* 7, 5 (1999), 369–384.

[62] Liping Jing, Michael K Ng, and Joshua Zhexue Huang. 2007. An entropy weighting k-means algorithm for subspace clustering of high-dimensional sparse data. *IEEE Transactions on knowledge and data engineering* 19, 8 (2007), 1026–1041.

[63] Ian T Jolliffe and Jorge Cadima. 2016. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374, 2065 (2016), 20150202.

[64] Jae-Yoon Jung and Joonsoo Bae. 2006. Workflow clustering method based on process similarity. In *International Conference on Computational Science and Its Applications*. Springer, 379–389.

[65] Diwas Singh Kc and Bradley R Staats. 2012. Accumulating a portfolio of experience: The effect of focal and related experience on surgeon performance. *Manufacturing & Service Operations Management* 14, 4 (2012), 618–633.

[66] Juho Kim, Phu Tran Nguyen, Sarah Weir, Philip J Guo, Robert C Miller, and Krzysztof Z Gajos. 2014. Crowdsourcing step-by-step information extraction to enhance existing how-to videos. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 4017–4026.

[67] Daichi Kitaguchi, Nobuyoshi Takeshita, Hiroki Matsuzaki, Hiroaki Takano, Yohei Owada, Tsuyoshi Enomoto, Tatsuya Oda, Hirohisa Miura, Takahiro Yamanashi, Masahiko Watanabe, et al. 2020. Real-time automatic surgical phase recognition in laparoscopic sigmoidectomy using the convolutional neural network-based deep learning approach. *Surgical endoscopy* 34, 11 (2020), 4924–4931.

[68] Nicholas Kong, Tovi Grossman, Björn Hartmann, Maneesh Agrawala, and George Fitzmaurice. 2012. Delta: a tool for representing and comparing workflows. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1027–1036.

[69] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*. 2980–2988.

[70] Qingshan Liu, Feng Zhou, Renlong Hang, and Xiaotong Yuan. 2017. Bidirectional-convolutional LSTM based spectral-spatial feature learning for hyperspectral image classification. *Remote Sensing* 9, 12 (2017), 1330.

[71] Szymon Łukasik, Piotr A Kowalski, Małgorzata Charytanowicz, and Piotr Kulczycki. 2016. Clustering using flower pollination algorithm and Calinski-Harabasz index. In *2016 IEEE congress on evolutionary computation (CEC)*. IEEE, 2724–2728.

[72] Stephann Makri, Yi-Chun Chen, Dana McKay, George Buchanan, and Melissa Ocepek. 2019. Discovering the unfindable: The tension between findability and discoverability in a bookshop designed for serendipity. In *IFIP Conference on Human-Computer Interaction*. Springer, 3–23.

[73] Akira Matsuda, Toru Okuzono, Hiromi Nakamura, Hideaki Kuzuoka, and Jun Rekimoto. 2021. A Surgical Scene Replay System for Learning Gastroenterological Endoscopic Surgery Skill by Multiple Synchronized-Video and Gaze Representation. *Proceedings of the ACM on Human-Computer Interaction* 5, EICS (2021), 1–22.

[74] Robert McCormick. 1997. Conceptual and procedural knowledge. *International journal of technology and design education* 7, 1 (1997), 141–159.

[75] PJ McLeod, Yvonne Steinert, Tim Meagher, Lambertus Schuwirth, Diana Tabatabai, and AH McLeod. 2006. The acquisition of tacit knowledge in medical education: learning by doing. *Medical Education* 40, 2 (2006), 146–149.

[76] Mani Menon, Ashok K Hemal, and VIP team. 2004. Vattikuti Institute prostatectomy: a technique of robotic radical prostatectomy: experience in more than 1000 cases. *Journal of endourology* 18, 7 (2004), 611–619.

[77] Helena M Mentis, Yuanyuan Feng, Azin Semsar, and Todd A Ponsky. 2020. Remotely Shaping the View in Surgical Telementoring. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[78] Paul F Merrill. 1987. Job and task analysis. *Instructional technology: foundations* (1987), 141–173.

[79] Roxana Moreno. 2007. Optimising learning from animations by minimising cognitive load: Cognitive and affective consequences of signalling and segmentation methods. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition* 21, 6 (2007), 765–781.

[80] Paulo Mota, Nuno Carvalho, Emanuel Carvalho-Dias, Manuel Joao Costa, Jorge Correia-Pinto, and Estevão Lima. 2018. Video-based surgical learning: improving trainee education and preparation for surgery. *Journal of surgical education* 75, 3 (2018), 828–835.

[81] Anastasia Moumtzidou, Konstantinos Avgerinakis, Evlampios Apostolidis, Fotini Markatopoulou, Konstantinos Apostolidis, Theodoros Mironidis, Stefanos Vrochidis, Vasileios Mezaris, Ioannis Kompatsiaris, and Ioannis Patras. 2015. Verge: a multimodal interactive video search engine. In *International Conference on Multimedia Modeling*. Springer, 249–254.

[82] Bernd Muenzer, Klaus Schoeffmann, and Laszlo Böeszöermenyi. 2017. EndoXplore: A Web-Based Video Explorer for Endoscopic Videos. In *2017 IEEE International Symposium on Multimedia (ISM)*. IEEE, 366–367.

[83] Scott Murray. 2017. *Interactive data visualization for the web: an introduction to designing with D3.* " O'Reilly Media, Inc.".

[84] Megha Nawhal, Jacqueline B Lang, Greg Mori, and Parmit K Chilana. 2019. VideoWhiz: Non-Linear Interactive Overviews for Recipe Videos.. In *Graphics Interface*. 15–1.

[85] Tahmina Nazari, Floyd W van de Graaf, Mary EW Dankbaar, Johan F Lange, Jeroen JG van Merriënboer, and Theo Wiggers. 2020. One step at a time: step by step versus continuous video-based learning to prepare medical students for performing surgical procedures. *Journal of Surgical Education* 77, 4 (2020), 779–787.

[86] Shi-Yong Neo, Huanbo Luan, Yantao Zheng, Hai-Kiat Goh, and Tat-Seng Chua. 2008. Visiongo: bridging users and multimedia video retrieval. In *Proceedings of the 2008 international conference on Content-based image and video retrieval*. 559–560.

[87] Phuong Anh Nguyen, Jiaxin Wu, Chong-Wah Ngo, Danny Francis, and Benoit Huet. 2020. VIREO@ video browser showdown 2020. In *International Conference on Multimedia Modeling*. Springer, 772–777.

[88] Shoichi NISHIO, Belayat HOSSAIN, Manabu NII, Naomi YAGI, Takafumi HIRANAKA, and Syoji KOBASHI. 2020. Surgical Phase Recognition with Wearable Video Camera for Computer-aided Orthopaedic Surgery-AI Navigation System. *International Journal of Affective Engineering* 19, 2 (2020), 137–143.

[89] Amy Pavel, Colorado Reed, Björn Hartmann, and Maneesh Agrawala. 2014. Video digests: a browsable, skimmable format for informational lecture videos. In *Proceedings of the 27th annual ACM symposium on User interface software and technology*. 573–582.

[90] Maya Peled-Raz, Nadav Willner, Dan Shteinberg, Keren Or-Chen, and Tova Rainis. 2019. Digital recording and documentation of endoscopic procedures: physicians' practice and perspectives. *Israel journal of health policy research* 8, 1 (2019), 1–12.

[91] Slobodan Petrovic. 2006. A comparison between the silhouette index and the davies-bouldin index in labelling ids clusters. In *Proceedings of the 11th Nordic workshop of secure IT systems*, Vol. 2006. Citeseer, 53–64.

[92] Luca Rossetto, Ivan Giangreco, Heiko Schuldt, Stéphane Dupont, Omar Seddati, Metin Sezgin, and Yusuf Sahillioğlu. 2015. IMOTION—a content-based video retrieval engine. In *International Conference on Multimedia Modeling*. Springer, 255–260.

[93] Omer Sagi and Lior Rokach. 2018. Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 8, 4 (2018), e1249.

[94] Manish Sahu, Anirban Mukhopadhyay, Angelika Szengel, and Stefan Zachow. 2016. Tool and phase recognition using contextual CNN features. *arXiv preprint arXiv:1610.08854* (2016).

[95] Manish Sahu, Angelika Szengel, Anirban Mukhopadhyay, and Stefan Zachow. 2020. Surgical phase recognition by learning phase transitions. *Current Directions in Biomedical Engineering* 6, 1 (2020).

[96] Kjeld Schmidt. 2012. The trouble with 'tacit knowledge'. *Computer supported cooperative work (CSCW)* 21, 2 (2012), 163–225.

[97] Klaus Schoeffmann, Heinrich Husslein, Sabrina Kletz, Stefan Petscharnig, Bernd Muenzer, and Christian Beecks. 2018. Video retrieval in laparoscopic video recordings with dynamic content descriptors. *Multimedia Tools and Applications* 77, 13 (2018), 16813–16832.

[98] Ricardo Mendes C Segundo, Marcello N de Amorim, and Celso Alberto S Santos. 2017. CrowdSync: User generated videos synchronization using crowdsourcing. In *2017 International Conference on Systems, Signals and Image Processing (IWSSIP)*. IEEE, 1–5.

[99] Naomi M Sell, Douglas J Cassidy, Sophia K McKinley, Emil Petrusa, Denise W Gee, Mara B Antonoff, and Roy Phitayakorn. 2021. A needs assessment of video-based education resources among general surgery residents. *Journal of Surgical Research* 263 (2021), 116–123.

[100] Ketan Rajshekhar Shahapure and Charles Nicholas. 2020. Cluster quality analysis using silhouette score. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 747–748.

[101] NL Sharma, NC Shah, and David Edgar Neal. 2009. Robotic-assisted laparoscopic prostatectomy. *British journal of cancer* 101, 9 (2009), 1491–1496.

[102] Cees Snoek, Kvd Sande, OD Rooij, Bouke Huurnink, J Uijlings, M van Liempt, M Bugalhoy, I Trancosoy, F Yan, M Tahir, et al. 2009. The MediaMill TRECVID 2009 semantic video search engine. In *TRECVID workshop*. University of Surrey.

[103] Kamal Soundararajan, Hiang Kwee Ho, and Bin Su. 2014. Sankey diagram framework for energy and exergy flows. *Applied energy* 136 (2014), 1035–1042.

[104] John—Christopher Spender. 1993. Competitive Advantage from Tacit Knowledge? Unpacking the Concept and Its Strategic Implications.. In *Academy of Management Proceedings*, Vol. 1993. Academy of Management Briarcliff Manor, NY 10510, 37–41.

[105] Jason Stanley and Timothy Willlamson. 2001. Knowing how. *The Journal of Philosophy* 98, 8 (2001), 411–444.

[106] R Stauder, A Okur, and N Navab. 2014. Detecting and analyzing the surgical workflow to aid human and robotic scrub nurses. In *7th Hamlyn Symposium on Medical Robotics. London*.

[107] Fabio Sulser, Ivan Giangreco, and Heiko Schuldt. 2014. Crowd-based semantic event detection and video annotation for sports videos. In *Proceedings of the 2014 international ACM workshop on crowdsourcing for multimedia*. 63–68.

[108] MA Syakur, BK Khotimah, EMS Rochman, and Budi Dwi Satoto. 2018. Integration k-means clustering method and elbow method for identification of the best customer profile cluster. In *IOP conference series: materials science and*

engineering, Vol. 336. IOP Publishing, 012017.

[109] Hui Li Tan, Hongyuan Zhu, Joo-Hwee Lim, and Cheston Tan. 2021. A comprehensive survey of procedural video datasets. *Computer Vision and Image Understanding* 202 (2021), 103107.

[110] Cuneyt Taskiran, Jau-Yuen Chen, Alberto Albiol, Luis Torres, Charles A Bouman, and Edward J Delp. 2004. ViBE: A compressed video database structured for active browsing and search. *IEEE Transactions on Multimedia* 6, 1 (2004), 103–118.

[111] Atima Tharatipyakul and Hyowon Lee. 2018. Towards a better video comparison: comparison as a way of browsing the video contents. In *Proceedings of the 30th Australian Conference on Computer-Human Interaction*. 349–353.

[112] Claudio AB Tiellet, André Grahl Pereira, Eliseo Berni Reategui, José Valdeni Lima, and Teresa Chambel. 2010. Design and evaluation of a hypervideo environment to support veterinary surgery learning. In *Proceedings of the 21st ACM Conference on Hypertext and Hypermedia*. 213–222.

[113] Anh Truong, Peggy Chi, David Salesin, Irfan Essa, and Maneesh Agrawala. 2021. Automatic Generation of Two-Level Hierarchical Tutorials from Instructional Makeup Videos. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

[114] Daniel W Turner III. 2010. Qualitative interview design: A practical guide for novice investigators. *The qualitative report* 15, 3 (2010), 754.

[115] Ulrike Von Luxburg, Robert C Williamson, and Isabelle Guyon. 2012. Clustering: Science or art?. In *Proceedings of ICML workshop on unsupervised and transfer learning*. JMLR Workshop and Conference Proceedings, 65–79.

[116] Sarah Weir, Juho Kim, Krzysztof Z Gajos, and Robert C Miller. 2015. Learnersourcing subgoal labels for how-to videos. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*. 405–416.

[117] Martin J Willemink, Wojciech A Koszek, Cailin Hardell, Jie Wu, Dominik Fleischmann, Hugh Harvey, Les R Folio, Ronald M Summers, Daniel L Rubin, and Matthew P Lungren. 2020. Preparing medical imaging data for machine learning. *Radiology* 295, 1 (2020), 4–15.

[118] Ho Chung Wu, Robert Wing Pong Luk, Kam Fai Wong, and Kui Lam Kwok. 2008. Interpreting tf-idf term weights as making relevance decisions. *ACM Transactions on Information Systems (TOIS)* 26, 3 (2008), 1–37.

[119] Guangnan Ye, Yitong Li, Hongliang Xu, Dong Liu, and Shih-Fu Chang. 2015. Eventnet: A large scale structured concept library for complex event detection in video. In *Proceedings of the 23rd ACM international conference on Multimedia*. 471–480.

[120] Andrew Yee, William M Padovano, Ida K Fox, Elspeth JR Hill, Amanda G Rowe, L Michael Brunt, Amy M Moore, Alison K Snyder-Warwick, Lorna C Kahn, Matthew D Wood, et al. 2020. Video-based learning in surgery: establishing surgeon engagement and utilization of variable-duration videos. *Annals of surgery* 272, 6 (2020), 1012–1019.

[121] Odysseas Zisimopoulos, Evangello Flouty, Imanol Luengo, Petros Giataganas, Jean Nehme, Andre Chow, and Danail Stoyanov. 2018. Deepphase: surgical phase recognition in cataracts videos. In *International conference on medical image computing and computer-assisted intervention*. Springer, 265–272.

A APPENDIX

A.1 Weighting Scheme for Structural Components

One of the simplest weighting schemes commonly used for sequence data in various domains such as information retrieval [3] and statistics [33] is the frequency weight, which assigns weights proportional to the frequencies of data entities. Similarly, frequently observed phases or dimensions should be considered more important in representing the procedures. On the other hand, the weights of entities that occur very frequently should be adjusted since too commonly observed across data can lose a determinative effect, not representing the characteristics of the individual procedure [118]. In other words, we need to weigh down the frequent entities while scaling up the rare ones. The performance of the weighting scheme considering the frequency weights and the inverse of frequency weights was validated in various tasks such as analyzing documents (e.g., tf-idf) and workflows (e.g., af-ipf, tf-ipf) [3, 64].