International Journal on Digital Libraries
https://doi.org/10.1007/s00799-023-00367-0

Gesture retrieval and its application to the study of multimodal communication

Mahnaz Parian-Scherb[1] · **Peter Uhrig**[2] · **Luca Rossetto**[3] · **Stéphane Dupont**[4] · **Heiko Schuldt**[1]

Received: 22 August 2022 / Revised: 7 May 2023 / Accepted: 19 May 2023

Abstract
Comprehending communication is dependent on analyzing the different modalities of conversation, including audio, visual, and others. This is a natural process for humans, but in digital libraries, where preservation and dissemination of digital information are crucial, it is a complex task. A rich conversational model, encompassing all modalities and their co-occurrences, is required to effectively analyze and interact with digital information. Currently, the analysis of co-speech gestures in videos is done through manual annotation by linguistic experts based on textual searches. However, this approach is limited and does not fully utilize the visual modality of gestures. This paper proposes a visual gesture retrieval method using a deep learning architecture to extend current research in this area. The method is based on body keypoints and uses an attention mechanism to focus on specific groups. Experiments were conducted on a subset of the *NewsScape* dataset, which presents challenges such as multiple people, camera perspective changes, and occlusions. A user study was conducted to assess the usability of the results, establishing a baseline for future gesture retrieval methods in real-world video collections. The results of the experiment demonstrate the high potential of the proposed method in multimodal communication research and highlight the significance of visual gesture retrieval in enhancing interaction with video content. The integration of visual similarity search for gestures in the open-source multimedia retrieval stack, *vitrivr*, can greatly contribute to the field of computational linguistics. This research advances the understanding of the role of the visual modality in co-speech gestures and highlights the need for further development in this area.

Keywords Gesture retrieval · Co-speech gestures · Multimodal retrieval · Video archive · Communication studies

1 Introduction

Human communication is inherently multimodal. In its most natural form—face-to-face interaction—we do not simply decode a stream of sounds to create a message; as some simplified models of communication imply, we actually hear so much more in the audio signal, communicated by pitch, intensity, and voice quality, and we see so much more, some related to the audio content (in particular lip movement), some less strongly so but often still highly relevant, gesture, facial expression, or body pose. Consider the following short quote from the Ellen DeGeneres Show [1]:

(1) okay | let's just break this down | first of all I'm not married | I'm married | that's all

✉ Mahnaz Parian-Scherb
mahnaz.parian-scherb@unibas.ch

Peter Uhrig
peter.uhrig@tu-dresden.de

Luca Rossetto
rossetto@ifi.uzh.ch

Stéphane Dupont
stephane.dupont@umons.ac.be

Heiko Schuldt
heiko.schuldt@unibas.ch

[1] Department of Mathematics and Computer Science, University of Basel, Basel, Switzerland

[2] Center for Scalable Data Analytics and Artificial Intelligence (Scads.AI), TU Dresden, Dresden, Germany

[3] Department of Informatics, University of Zurich, Zurich, Switzerland

[4] Department of Computer Science, University of Mons, Mons, Belgium

Published online: 24 July 2023

(NewsScape 2015-01-15 _0000_US_KNBC_The_Ellen_DeGene res_Show, 0:02:16-0:02:24)

From the verbal channel alone [2], the utterance is difficult to fully understand. Even listening to the audio is not enough to understand it, although the audio signal shows something strange, i.e., the first instance of *married* being twice as long the as the second. The context of the utterance is that DeGeneres's words are a reply to an Op-Ed piece in the *Christian Post* by Larry Tomczak [3]:

> Here's how Hollywood is promoting homosexuality right now:
> – [...]
> – "Ellen DeGeneres" celebrates her lesbianism and "marriage" in between appearances of guests like Taylor Swift to attract young girls.

Still, the full meaning of the utterance only becomes clear when we look at the visual mode (scan QR code or click on the url in the footnote)[1]. The use of so-called "*air quotes*" on the first instance of *married* mirrors Tomczak's use of quotation marks in the article, implying that her lesbian marriage is not a real marriage, a position that DeGeneres then goes on to reject when she utters *married* for the second time, accompanied by a shrug-like gesture.

Of course, this is an extreme example, but a computer system that is designed to successfully and seamlessly interact with humans needs to be aware of the influence co-speech gesture has on the meaning of the utterance, just as it needs to be aware of the influence of prosodic cues such as rising intonations that encode certain types of questions. This is particularly the case in a wide range of modern applications in human machine interactions when users interact with systems via several modalities. In this paper, we describe a set of technologies that could be used to help interactive computer systems understand multimodal communication in the future. Such methods can also be used to study multimodal communication in a research context in order to create models and baselines for such systems.

Since the problem of understanding such multimodal communication is a super-set of the problem of speech transcription, as it is currently used for many natural language interfaces for digital libraries or general interactive systems, the computational cost for its analysis is correspondingly larger. Audio-visual retrieval systems [4], which are the other side of the multimedia content, also require sophisticated infrastructure for serving large amounts of data. When dealing with large repositories of video data, being able to query for such non-verbal communication, in addition to already established text-based retrieval mechanisms, can further aid in increasing the accessibility of the contained information.

We thus also describe the creation and continuous improvements of research infrastructures that can be used in this kind of research. As we will see, finding relevant data based on co-speech gestures—which is only the first step of any research endeavor—is already a substantial challenge, especially in the absence of ground truth in the collection.

Data collection for the study of multimodal communication, which can be used in search systems in digital libraries, has always been labor-intensive. Traditionally, it involved either carrying out and recording experiments in a laboratory setting, or sifting through countless hours of video. The situation improved for certain research questions with collections such as the UCLA Library Broadcast NewsScape, which records TV news (in a broad sense) from the USA, but also other places around the globe. The NewsScape project [5] [2] offers various access options (see [6] for a critical evaluation) to the collection, but none of them allow for complex linguistic searches or even searches combining text with audiovisual features. It is for this reason that the research presented in this paper fills a gap that has been slowing down research on multimodal communication considerably.

The contributions of this paper are threefold. First, we introduce the current methods in multimodal computational linguistics to analyze, annotate and search for co-speech gestures with a focus on the visual modality through a novel field of research—gesture retrieval based on computer vision technology. The pipeline proposed in this paper benefits from state-of-the-art deep learning methods to find the most visually similar hand gestures which are articulated by people in multi-person settings recorded from different perspectives. Second, our experimental analysis on in-the-wild recordings establishes the baseline for further research in gesture similarity retrieval and opens up the path to include the visual modality in the study of co-speech gesture in empirical linguistics. Third, we investigate the integration of this visual search pipeline with the open-source multimodal retrieval system, *vitrivr*, and its potential to be used in multimodal human communication studies.

The remainder of the paper is organized as follows: Sect. 2 describes the different modalities of human communication and the existing work in computational linguistics to study co-speech gestures. Section 3 focuses on the visual modality and presents our proposed method to perform visual search in hand gestures. Additionally, we introduce *vitrivr* and how this system can be used for multimodal search in large video collections. In Sect. 4, we present the result of our experiments of the proposed method on one of the largest collections of real-world TV footage. Section 5 discusses the results and future directions and Sect. 6 concludes the paper.

[1] http://go.redhenlab.org/zaa/14

[2] http://newsscape.library.ucla.edu/.

2 Multimodality in communication

Essentially language, gesture, posture, and other non-verbal modalities are used often at the same time when we communicate. Co-speech gestures are nonverbal behaviors that occur simultaneously with speech and play an important role in human communication and cognitive processes [7]. There are several different types of co-speech gestures, including [8]:

- Iconic gestures: These gestures imitate or symbolize an action or object. For example, holding an imaginary steering wheel to indicate driving a car.
- Pointing gestures: These gestures direct the attention of the listener to a specific object or location. For example, pointing to a book to indicate the topic of conversation.
- Metaphoric gestures: These gestures reflect the meaning of the speech and add emphasis or clarify abstract ideas. For example, sweeping hand movements to indicate a large quantity.
- Beat gestures: These gestures serve as a rhythmic accompaniment to speech and help to emphasize certain words or phrases. For example, tapping fingers to emphasize a point being made.

Performing a search in multimodal communication requires a good understanding of each of the individual modalities and the links between them to be able to find co-occurrences of multimodal events. In this section, we describe the role verbal, auditory and visual modality in co-speech gesture search.

2.1 Verbal modality

What we call verbal modality here is the level of the textual representation, separated from the actual communicative event. This is of course an abstraction and idealization, but one that is used in most Natural Language Processing (NLP) applications including human–machine interfaces and in most of linguistic research. For most research purposes, an NLP pipeline consisting of tokenization and sentence splitting, Part-of-Speech tagging, lemmatization and often also syntactic analysis or named entity recognition are performed. Compared to the computational requirements of audio and visual analysis, text processing has a very small footprint. In our research infrastructure, we used the subtitles transmitted with the TV recordings from the NewsScape dataset as textual data. After some cleaning processes, e.g., the removal of non-spoken text such as "*[APPLAUSE]*" and specially developed sentence splitting to take care of the all-uppercase data used in US-American closed captioning, Stanford CoreNLP [9] is used to tokenize, tag, truecase, dependency-parse, and annotate for named entities. The resulting data are then converted into a token-based vertical format (one column per information type, with the word in the first column and word-level annotations in further columns; pseudo-XML tags cover larger spans; see [10] for the file format in general and [6] for this specific implementation), which forms the basis of further processing steps.

2.2 Auditory modality

The audio signal on its own cannot be easily searched. Even if there is a full transcript, we can only guess where the relevant position in the audio is, assuming a roughly equal distribution of words across the recording. For very small datasets, audio and text are sometimes aligned manually, but for larger datasets this quickly becomes infeasible.

If our data are from subtitles, we can not only extract the text itself but also the rough alignment provided by the time when a subtitle appears on and disappears from the screen. However, these subtitles can sometimes be displayed with a substantial offset, particularly in live shows where the subtitling necessarily lags behind, so that associations between the words based on the timestamps of the subtitles on the one hand and gestures on the other hand cannot be established. Even if the timing is relatively accurate, TV subtitles tend to be displayed in lines (e.g., of 32 characters in the USA), so that individual words or constructions cannot be picked out anyway. The first version of our corpus research infrastructure made use of this rough alignment only. The state-of-the-art solution to this problem is to submit the transcript/ subtitles and the audio/video recording to automatic forced alignment software (e.g., [11–13]). The success rates vary and are highly dependent on the accuracy of the transcript provided to the software by the user (see [14] for an evaluation). For real-world subtitles, we have observed accuracies between 67% and 90%.

In principle, the same approach can be used with automatic speech recognition, using tools such as kaldi [15], DeepSpeech2 [16], Whisper [17], or YouTube's automatic subtitles. Often, these come with timing information, albeit sometimes only implicit (e.g., YouTube's subtitles are usually quite well aligned but the time codes will be off when someone speaks slowly or makes pauses). Even if the Automatic Speech Recognition (ASR) transcript comes without timestamps, forced alignment can be used. Note, however, that transcripts generated by ASR usually lack features such as capitalization and punctuation marks (except for Whisper), which also means that we do not have sentence boundaries. Of course we can still search such data for words, but many NLP tools that rely on sentence boundaries will suffer from massively degraded performance or fail to run at all. Thus while Part-of-Speech tagging is relatively robust in this respect—for architectural reasons we expect higher error rates mainly around the missing punctuation marks—syntactic parsing fails dramatically because the entire model

is based on individual sentences. This is the reason why in our YouTube-based Russian-language corpus, we have so far worked without syntactic parsing.[3]

It must be stressed here that the alignment of text and audio signal is of course also, by the very nature of audiovisual media, an alignment of text and video. Thus, it acts as a bridge between the verbal and visual modalities. Even if a researcher is not interested in the auditory modality at all and just wants to study the relationship between the verbal and visual modalities, they need to rely on such alignments, no matter whether they are created manually, by means of forced alignment, or by automatic speech recognition.

For our research infrastructure, we deployed a modified version of the Gentle forced aligner [11] to more than 300,000 h of TV news recordings with their subtitles on the High-Performance Computing facilities at Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU). The alignment timestamps for each successfully aligned token were added in extra columns into the vertical files.

2.3 Visual modality

In addition to natural language processing methods, analyzing the videos and extracting information from the media can greatly help research on communicative gestures. This modality is specifically important to bridge the semantic gap and reduce the ambiguity of textual queries when describing a particular gesture. *Query-by-Gesture*, which is a subset of Query-by-Example (QbE), is a type of query formulation which can be used to retrieve the visually similar gestures in a video collection, regardless of their functions.

Visual search in videos is essentially based on encoding the visual information into an embedding which can preserve the similarity between the input samples [18]. In real-world videos, for example in TV recordings, there are numerous challenges that need to be addressed either by the feature extraction method or by preprocessing steps [19]. One of these challenges is the presence of multiple persons in a scene, which requires a mechanism to track the individual who is performing the gesture across different frames. Additionally, unlike activities such as playing the piano, which are bound to certain environments and objects, hand gestures can happen in conversation in any situation and environment. On top of these, there is the complications of the gesture trajectory and dynamics which needs to be represented robustly.

The existing work in gesture retrieval is sparse and is mainly based on gesture datasets with curated and controlled gesture articulations [20, 21]. These datasets often do not represent real-world challenges well enough and the methods which are developed for these datasets cannot be applied to real-world applications. Therefore, in this paper, we propose a pipeline to encode hand gestures and use them to find similar instances to the query video. This method can be used together with other modalities to allow for the search of specific co-speech gestures.

2.4 Text-based research workflow and infrastructure

The research infrastructure currently in use is based on the modalities discussed in Sects. 2.1 and 2.2. In order to offer researchers a performant yet linguistically sound search engine, we opted for CQPweb [22] with custom modifications to display video snippets in the search results. The system inherits many of its strengths and weaknesses from the underlying Open Corpus Workbench [23], whose query language has become sort of an industry standard in corpus linguistics, so that versions of it are also used by the corpus manager of the Sketch Engine [24] and ANNIS [25]. It reads the vertical file format and creates a highly compressed search index, which makes even large collections (of up to 2.1 billion words) searchable within sufficiently short time spans. Due to the forced alignment, we normally have the exact location of the word in the video file and can provide short snippets—the default is 4 s before and after the search expression. Thus, a large set of hits can be screened in short periods of time, the query can then be refined, and the next query results can be screened again. In addition, we offer a special download feature that exports the data in a format usable by the Red Hen Rapid Annotator (see [14]), which is a web-based tool for the fast classification of arbitrary textual, audio, or audiovisual data.

With this, data researchers can now look for words, phrases, or more abstract linguistic constructions and analyze and classify the video data for co-speech gesture. Thus, we can look for the noun *choice* near the preposition *between* and find examples of a gesture co-occurring with it (scan QR code or click on the url in the footnote[4]; example taken from [14]).

In this example, we see that the two options are realized gesturally as two separate strokes, one with each hand. And we can look at many of these examples, in a short period of time, which is a tremendous improvement over previous methods.

[3] Most research questions relevant to this paper however will not require syntactic parsing, but theoretically interesting questions such as the association of abstract grammatical structures with particular gestures become more difficult to study when syntactic annotation is missing.

[4] http://go.redhenlab.org/pgu/0014

However, the process is still less straightforward than one would might think. In the TV news collection, [26] showed that they had to discard more than 80% of the hits found in the text for their timeline expressions (e.g., *from beginning to end*) because the speaker was not on the screen (weather map, outside shot, audience shot, reaction shot, etc.) or because the hands of the speaker were not visible (often in U.S. TV news, we see close-ups of the speaker's face). Thus, having a tool that gets rid of such examples already speeds up the work of the researcher substantially. But there is a second, conceptual problem. With the method outlined above, we can search for words and structures, and find out which gestures co-occur with them, but this is a one-way street. If we were interested to find out where the gesture shown in the example for *choice between* occurs, too, in order to determine whether it has a stable meaning, we actually need to come up with potential linguistic structures to search for. Should we look for the word *alternative(s)*? Or maybe *options*? Is *on the other* hand related? We can never be sure we obtained the full picture of a certain gesture use if we cannot search for similar gestures.

3 Gesture visual similarity search

When considering gestures from the visual perspective, they are usually categorized as a subset of human actions. This allows for methods developed in the action recognition area to be used—directly or indirectly—in a visual gesture understanding and retrieval task. The state-of-the-art in the field of human activity recognition is rich in both non-learning based [27, 28] and learning-based methods [29, 30]. However, the direct application of these methods in extracting information from linguistic gestures is not optimal, mainly for two reasons: Firstly, the independence of communication gestures from the surroundings and from objects the speaker interacts with, which in action recognition task can help the identification of an activity. Secondly, the high temporal dependency between the hand gesture frames in a video, which needs to be modeled carefully during the feature extraction. Therefore, a functional gesture retrieval method needs to address these differences in addition to the existing challenges which appear in any human activity recognition task, such as occlusion and the presence of multiple persons in a scene.

To overcome these challenges, we propose a two stage gesture retrieval method using a deep neural network. The first stage comprises sophisticated preprocessing methods to reduce the dependency of the hand gestures on the background and suppress the occlusion effect. Additionally in order to be able to retrieve the hand gestures articulated by multiple persons in such scenes, we propose a cross-angle spatio-temporal segmentation module, which essentially tracks each individual through different camera shots and by removing the background clutter prepares short clips for the next stage. The second stage is a spatio-temporal gesture feature extraction method which captures the temporal dependencies between the frames and produces discriminative features which can be used for the search. An overview of the different stages of our method can be seen in Fig. 1. In the following sections, we describe our method in detail.

3.1 Cross-angle spatio-temporal human segmentation in multi-person scenes

The preprocessing stage of our method uses two individual modules to perform semantic segmentation and person re-identification to remove the cluttered background, tracks individuals through different shots and detect them in scenes with multiple persons. The human segmentation methods vary in performance based on their architecture and the modality of input they use. A large amount of existing work is based on RGB frames where the human instances are recognized by region-based methods [31, 32], which is based on region of interest (RoI) detection. One of the main drawbacks of these methods is the poor performance in occluded scenes, where the bounding boxes of multiple instances of persons are overlapping with each other. However, the human body parts can be defined as the pose skeleton, which can help locating and estimating the position of the body joints in highly overlapping instances. Using the body joint keypoints to create a skeletal model and segmentation mask of a person—which is referred to as bottom-up approach [33]—is an effective method in contrast to top-down methods where the human skeletons are detected based on the bounding boxes identified in a scene. For this reason, we use Pose2Seg [34] which has a bottom-up structure and with the unique segmentation module, identifies and masks human instances even in complex environments. The segmentation network consists of three main parts: Affine-Align, Skeletal Features and SegModule which are explained below.

The input to the network consists of a sequence of N frames with the dimension of $m \times n$ and sequence of human poses in each frame. The RGB stream of data is used to extract features using a Feature Pyramid Network [35]. The human pose consists of skeleton keypoints which are extracted using OpenPose [36], which takes the raw video frames and extracts the skeletal joint keypoints. The affine-align operation aligns human bodies to template human poses which are a collection of most frequent poses in real-world observations. The pose template is used as a reference to assign a score to each input pose and find an affine transform matrix and apply it to the image features. The skeletal features are formed by the Part Affinity Fields [36] and a confidence map which indicates the pairwise relationship of the body parts and the probability of the existence of the body part in the predicted location, respectively [36]. Finally, the segmentation

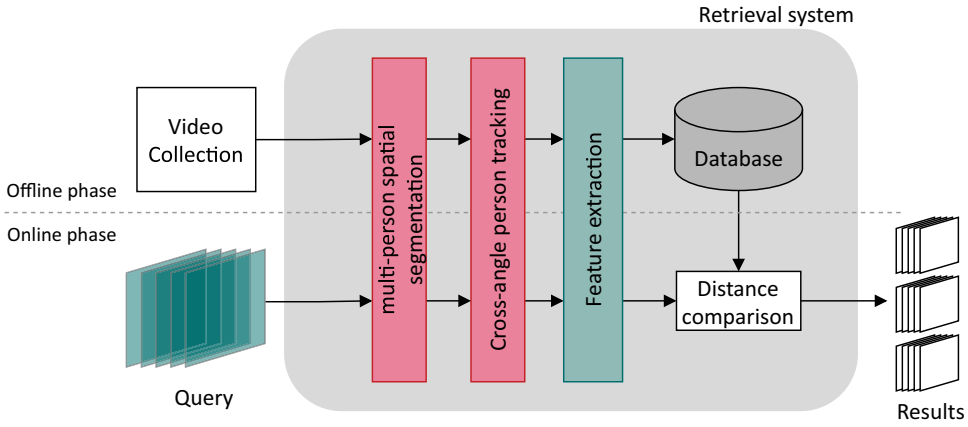


Fig. 1 An overview of the gesture video retrieval pipeline including the cross-angle spatio-temporal preprocessing stage and the feature extraction. In the offline phase, the features of the video clips are extracted and stored in a database and are used to retrieve similar video clips based on the Euclidean distance

is mapped to the actual frame using a convolutional neural network and the original resolution is restored via bilinear up-sampling.

In addition to the occlusion and background clutter, in multi-person scenarios different people occasionally perform gestures simultaneously which cannot be processed by a gesture recognition method. Additionally, while long videos require a form of segmentation into parts which detects abrupt transition effects, these temporal segmentations often lead to a gesture being cut before it ends. To overcome the multi-person processing challenge and guarantee a gesture-friendly video segmentation, we propose a framework to segment the videos based on the presence of each individual in a sequence of frames. The method is inspired by the person re-identification method in [37] and tracks each individual in scenes and trims the video into short clips containing that person. This model is originally based on RoI proposals extracted by a CNN and human feature embeddings obtained from ResNet-50. We adapt the person search network by replacing the RoI extraction component with the output of the instance segmentation method. Replacing the bounding box extraction with the pixel-wise human instance segmentation increases the robustness of the re-identification in occluded settings.

The re-identification is performed by measuring the cosine similarity between the features of person proposals in a frame and the gallery of people from the collection. To track the people appearing in the frames, the similarity measure should take into account a query as well as a reference image to calculate the similarity score. However, in the beginning, there is no such reference gallery available. Therefore, we initialize the gallery with the first instance of a person in the first frame as pid_1. If there are multiple persons in one frame, all are added in the gallery as e.g., $\mathbf{g} = \{pid_1, pid_2\}$. With each newly segmented frame, a feature vector of each masked human instance is extracted and is compared with the gallery as the query image. According to the similarity score, either the person in the frame will get the same pid as the reference or will be added as a new pid_3 to the gallery (see Fig. 2). The similarity score threshold τ is set to 0.5.

After the identification of each instance in all the frames, and adding the queries with similarity score $< \sigma$ to the gallery, we stitch the consecutive frames in which one pid is present. Given $\mathbf{S} = \{\mathbf{fr}_1, \mathbf{fr}_2, \ldots, \mathbf{fr}_N\}$ as the sequences of video frames, the clip containing the pid_i, $\mathbf{Sc}_{n,i}$, is formed as [38]:

$$\mathbf{Sc}_{n,i} = \Big\{\forall \mathbf{fr}_j \in \mathbf{S} \mid pid_i \in \mathbf{fr}_j \wedge (pid_i \in \mathbf{fr}_{j-1} \vee \mathbf{Sc}_{n,i} = \emptyset)\Big\} \tag{1}$$

In other words, the first masked instance of pid_i in a video initiates a short clip onto which the next frame is stacked, in case the pid_i is present in that frame. This process continues until the person instance is not in the frame, and the clip is ended. Therefore, the output of the preprocessing stage will be multiple sequences of clips containing masked instances of each person in the video.

3.2 Pose-based gesture feature extraction

The complex gestural trajectories in multi-perspective scenarios require a robust feature extraction module which can represent discriminative information about hand articulations. The pose modality is an asset in scenes with interactions between people and can help in recognizing and following their hand movements. Additionally, optical flow is a robust motion modeling algorithm when the brightness is consistent and there are no abrupt movements. In scenes where camera cuts exist, this sudden big displacement of pixels would break-down the usability of optical flow. Different methods in hand gesture recognition used combinations of various modalities including depth data [39–41] to extract features from dynamic hand articulations. However, the largest sources of videos, which are TV footage

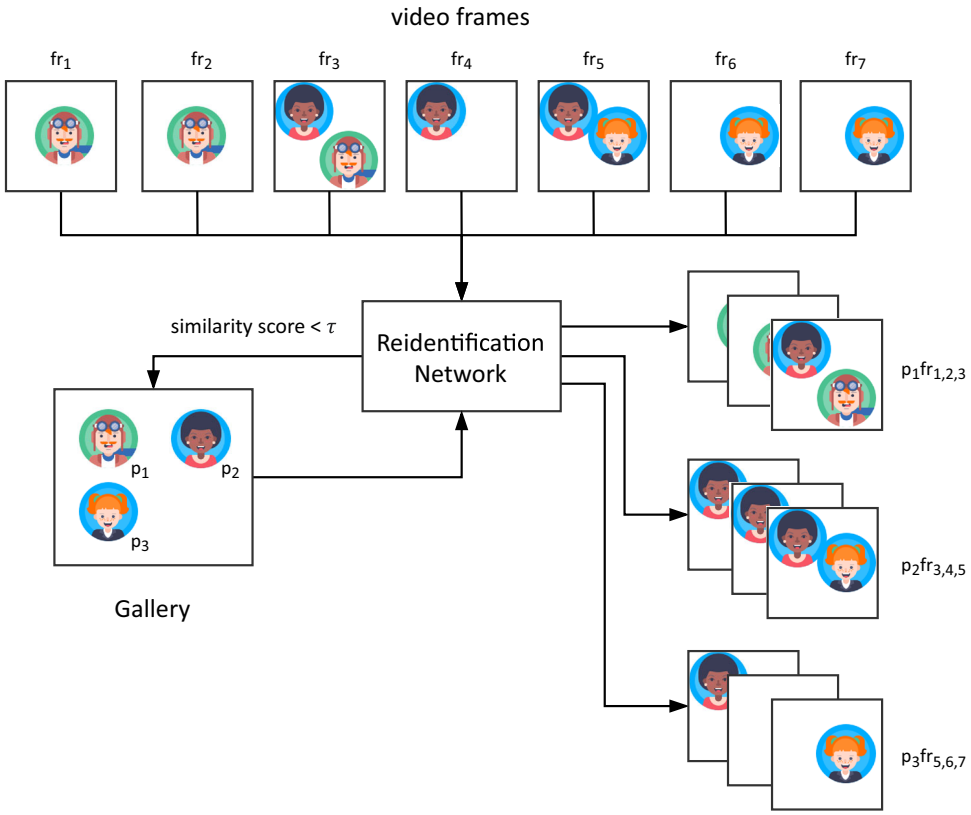


Fig. 2 A simplified illustration of the Cross-angle Spatio-Temporal Human Segmentation in Multi-Person setting. The gallery collects the first instance of each person to use as the reference when the new query image appears. The output are short clips based on the continuous presence of a person in adjacent frames

and online platforms such as YouTube, do not include depth modality, which limits the usage of these methods. Therefore, our proposed model uses the pose keypoints together with an RGB stream of data to create an attention map to learn the similarity of hand movements, which is inspired by RPAN [30]. This method is an end-to-end Recurrent Neural Network (RNN) [42] with a pose-attention mechanism that learns to focus on active human joints. Since the majority of actions in the recorded footage of news or talk shows is hand motions, the attention mechanism would help extracting hand gesture features.

RPAN originally has two streams of RGB and optical flow, where each of them are trained with the human activity recognition objective. Due to the limitations of optical flow in scenes with abrupt movements which are common in talk show footage because of camera cuts, we replace the optical flow stream with pose keypoints streams. The pose information together with the RGB input are used to train the attention module to focus on the active human joints.

The spatial features from the RGB data input are extracted using a ResNet [43] and generate a convolutional cube with the size of $K_1 \times K_2 \times n$ based on aggregating n feature maps with the dimension of $K_1 \times K_2$. Each element in this convolutional cube is accessed by $\boldsymbol{C}_t(\cdot) \in \mathbb{R}^n$. Therefore, for each frame t, the convolutional cube contains a feature vector $\boldsymbol{C}_t(k)$ at each location k where $k = 1, ..., K_1 \times K_2$ [30]. To capture the temporal dependencies between the video frames, Long Short-Term Memory (LSTM) units are used.

We follow the pose attention mechanism based on a human part configuration from Du et al. [30]. The part configuration is a clustering of joints into parts which are usually involved in an activity together. Having the focus on these parts would enable the model to learn the part-specific features.

The pose attention mechanism is defined by an attention heatmap $\alpha_t^J(k)$ for each feature vector $\boldsymbol{C}_t(k)$ for each joint (J) which together with semantically relevant joints, form a body part structure (P):

$$\alpha_t^J(k) = \frac{\exp\left(\boldsymbol{v}^J \tanh\left(\boldsymbol{A}_h^P \boldsymbol{h}_{t-1} + \boldsymbol{A}_c^P \boldsymbol{C}_t(k) + \boldsymbol{b}^P\right)\right)}{\sum_k \exp\left(\boldsymbol{v}^J \tanh\left(\boldsymbol{A}_h^P \boldsymbol{h}_{t-1} + \boldsymbol{A}_c^P \boldsymbol{C}_t(k) + \boldsymbol{b}^P\right)\right)} \tag{2}$$

Here, $\boldsymbol{h}_{t-1}$ is the LSTM hidden state of the previous frame, $\boldsymbol{v}^J$, $\boldsymbol{A}_h^P$, $\boldsymbol{A}_c$, $\boldsymbol{b}^P$ are the attention parameters, all of which except $\boldsymbol{v}^J$, are shared between the joints $(J \in P)$. Based on this attention heatmap, the human-part feature is extracted:

$$\boldsymbol{F}_t^P = \sum_{J \in P} \sum_k \alpha_t^J(k) \boldsymbol{C}_t(k) \tag{3}$$

After extracting all the human-part features from Eq. 3, they are fused together by a pooling layer to generate pose-related features, $\boldsymbol{SF}_t$. To capture the temporal dimension of

the movements, the features are then fed to an LSTM and the $(\boldsymbol{h}_t)$.

To learn the similarity metric between the samples so that we can use this method in a retrieval setting, we train the network using a triplet loss. For this reason, we use the output of the LSTM unit after processing the entire video and map it to a feature vector (fv) using two fully connected layers. To measure the similarity between the samples we use the Euclidean distance. To train the network, a collection of triplets $\sigma = (\boldsymbol{fv}_i, \boldsymbol{fv}_i^+, \boldsymbol{fv}_i^-), i = 1, ..., k$ are drawn, where $\boldsymbol{fv}_i$, $\boldsymbol{fv}_i^+$, $\boldsymbol{fv}_i^-$ are the feature vectors of the anchor, positive and negative samples, respectively. To encourage the network to learn diverse similarities, we introduce a margin μ. Therefore, the relation of similarity between two pairs is defined as:

$$D(f_\theta(\boldsymbol{fv}_i), f_\theta(\boldsymbol{fv}_i^+)) - D(f_\theta(\boldsymbol{fv}_i), f_\theta(\boldsymbol{fv}_i^-)) + \mu < 0 \tag{4}$$

where $\mu = 0.5$. The triplet loss is defined as following:

$$\mathcal{L}_\theta(\boldsymbol{fv}_i, \boldsymbol{fv}_i^+, \boldsymbol{fv}_i^-) = max(\boldsymbol{D}(f_\theta(\boldsymbol{fv}_i), f_\theta(\boldsymbol{fv}_i^+)) - D(f_\theta(\boldsymbol{fv}_i), f_\theta(\boldsymbol{fv}_i^+)) + \mu, 0) \tag{5}$$

In order to fulfill the objective of the retrieval task, to ensure that similar samples with smaller distance are mapped onto closer embeddings, we minimize the loss function in Eq. 5:

$$\min_\theta \sum_{i=1}^{k} \mathcal{L}_\theta \left(\boldsymbol{fv}_i, \boldsymbol{fv}_i^+, \boldsymbol{fv}_i^-\right) \tag{6}$$

where k is the total number of triplets. To sample the triplets, we select the anchor class randomly, and positive and anchor samples share the same class label. To mine the negative samples, we consider all the classes except the anchor class.

The feature similarity learning network is trained in an end-to-end fashion and produces features which have a smaller distance in the embedding for similar samples. After training, the feature vectors of the entire collection are extracted and stored in the database. When querying, the query video undergoes the same preprocessing and feature extraction stages, and a 2,048 dimensional feature vector is obtained. This feature vector is then compared with the feature vectors stored in the database according to their Euclidean distance and a ranked list of the most similar videos is retrieved (online-phase).

3.3 vitrivr

vitrivr [44] is a general-purpose, open-source multimedia retrieval stack which supports a variety of media types, such as images, audio, video, and 3D models concurrently [45, 46]. While being designed as a general-purpose multimedia retrieval solution [47], vitrivr's modular architecture enables its customization for special use cases, such as gesture retrieval. The following will briefly introduce the vitrivr stack and its components and outline how they can be used in the context of the gesture retrieval work presented in this paper.

The vitrivr stack consists of three primary components: the persistent storage layer *Cottontail DB* [48], the feature extraction and query processing engine *Cineast* [49] and the visual user interface *vitrivr-ng* [46]. The storage layer is responsible for persistently storing all *feature* information extracted from the media documents and for providing efficient comparison operations between the stored representations and a query of the same form. While such features can be represented in multiple ways, we will focus here on the *vector* representation, as it is the most relevant for this use case. Such a feature vector $fv \in \mathbb{R}^n$ encompasses the relevant information of a media document or part thereof in such a way that *similar* documents are transformed into vectors which are *close* to each other, given some distance measure. The storage layer offers functionalities which enable efficient comparison between such vectors in order to retrieve the closest n vectors in the database that correspond to the most similar previously encountered media documents, according to the selected feature.

These feature vectors are generated by Cineast, both during an offline extraction phase, where media documents are indexed in order to make them searchable, as well as during an online retrieval phase, where the same transformations are applied to relevant query objects and compared to the previously seen ones. To do this, Cineast contains a multitude of *feature modules*, each of which implements a different feature transformation, targeting one or more of the supported media modalities and representing a different notion of similarity. Which feature modules are used and how their results are to be combined is configurable and modules representing specialized notions of similarity, such as the ones described above, can easily be added, in order to tailor a deployment to any use case. Since for media documents with a temporal component, such as audio and video, it is commonly not particularly useful to have an entire document as the unit of retrieval, Cineast performs temporal segmentation and provides the generated segments to the feature modules which produce embeddings of these individual segments during feature extraction, in order to achieve higher temporal resolution during retrieval. In order to perform retrieval on a query, the modules are tasked with performing their transformation on the query object and comparing the resulting vector with the previously generated ones. The result of this comparison is a list of media items or their segments, each with an associated similarity score $s \in [0, 1]$, where a higher value indicates a more similar element. The results of multiple modules can

be combined using a weighed average of the scores, using L_1-normalized weights.

Queries are expressed via the user interface *vitrivr-ng* which then also presents the query results. The interface is a browser-based application and offers several query formulation methods [50], including text, audio, color- and semantic-sketches [51], pose [52], etc. The most relevant of these query formulation methods for this use case are *QbE* and *more like this (mlt)*. The QbE mode enables users to specify a query using a media document external to the indexed collection. In our case, this would consist of a short video clip showing a particular gesture, which can either be uploaded or directly recorded using a webcam. In contrast, the mlt mode uses previously retrieved results for *relevance feedback*, in order to enable a user to expand a result set in any particular direction.

The multimodality and modularity of vitrivr put it in a ideal position to be used for the verbal/vision modality retrieval in videos with linguistic content. It is in particular the combination of text and video retrieval that will improve the results obtained for multimodal patterns and increase the overall retrieval performance. The integration of query by gesture in *vitrivr-ng* as a subtype of QbE without irrelevant visual clutter would additionally allow the search by example as well as the textual description of the gestures. Such textual descriptions could be added by the user upon retrieval of clusters of relevant results, e.g. when the user finds a set of Palm up, open hand (PUOH) gestures, they can assign the label and future searches for that label will retrieve similar instance, even if they had not been seen and labeled by the user yet.

4 Experiments

To evaluate the performance of the developed method for visual gesture retrieval, we conduct an experiment using a large, real-world video collection.

4.1 Evaluation setup

We trained the feature extraction module of our pipeline on Chalearn Isolated gesture dataset with 249 classes and approximately 36,000 training videos containing one gesture per video. We used a subset of 259 videos from the NewsScape dataset, specifically from the *Ellen DeGeneres show*, which covers the entire year 2017 and is provided to us by Distributed Little Red Hen Lab.[5] This is a specifically interesting dataset due to the various challenges present in talk shows, where people use hand and body gestures naturally. The dataset exhibits various sources of occlusions on the hands such as objects, persons and banners or subtitles on the scene. Due to the nature of the talk shows, usually there is more than one person in the scene, sometimes the audience is also shown. We ran the entire two stage pipeline of our method to extract features of the 259 h of videos prior to the evaluation. The extracted features from these clips were stored in the database and were used for the retrieval.

The preprocessing step produced 3,093,022 video clips based on the presence of persons in each camera shot and number of people present in the scene. The clips' lengths vary between fractions of a second and 74 s with an average of 1.45 seconds. The very short clips are artifacts of the mis-re-identification during the preprocessing stage and the very long ones usually are the solo presence of the host talking without the presence of another person. For the evaluations, we removed the ultra short video clips (shorter than 2 s/60 frames). After this filtering, 1,501,037 video clips with an average length of 3.29 seconds were available.

Due to the absence of labeled gestures for the *Ellen DeGeneres show*, we performed a user study to analyze the quality of the results by collecting judgements of the perceived similarity from different people. We asked two groups of non-linguists and linguists who had linguistic or cognitive science backgrounds to participate in the survey separately. We ran the survey on two different servers for *linguists* and *non-linguists* assessors who were asked to rate the *formal* similarity of the gestures (which refers to form of the hands and pattern of movement rather than the context) and the *visual* similarity between the gestures respectively. Although both refer to the same concept, the vocabulary was adjusted to decrease the disparity of results due to misunderstanding.

In total, 76 people participated in our survey, 14 of whom were linguists and 62 had a non-linguistic background. The survey was designed to assign a random query based on the lowest number of results to each participants. At the end of the evaluations, each query result collected 30 scores on average. The participants were given a brief introduction in the evaluation survey and used their own personal devices to access the server. They were asked to use a four point Likert scale to assign a similarity score to the retrieved results (*'very good'* = 4, *'good'* = 3, *'ok'* = 2, *'bad'* = 1).

To asses how well our method generalizes to out-of-the-dataset samples, we selected two types of queries:

- seven videos are taken from the dataset, with four of them representing co-speech gestures, two of them gestures such as clapping and waving and one query represented the act of sitting of the performer.
- three videos were performed by the first author in a setting which has a large difference to the videos from the dataset. This is specifically used to measure the ability of the method to generalize to different samples. The three

[5] http://www.redhenlab.org.

queries aimed to re-create co-speech gestures that occur while talking.

4.2 Evaluation metrics

The retrieval metrics were calculated by considering the result as *relevant* when the normalized score was equal or greater than $\frac{2}{3}$. In addition to the statistical measures, such as mean and variance of the ratings, we use the following metrics in our study as well: ***Fleiss' Kappa*** to measure the inter-rater's reliability, which indicates the degree of agreement between the assessors

Discounted Cumulative Gain to measure the effectiveness of our method in gesture similarity retrieval; defined as

$$dcg(s) = \sum_{i=1}^{N} \frac{2^{s_i} - 1}{log_2(i+1)} \tag{7}$$

where s is the list of scores corresponding to the retrieved results, using the aggregated scores as assigned by the assessors.

Precision to measure the performance of a retrieval system by result relevance. Precision (P) is defined as the number of true positives T_p over the number of all the returned results:

$$P = \frac{T_p}{T_p + F_p} \tag{8}$$

where F_p is the false positives. Precision is usually calculated at k which indicates the proportion of relevant items on top-k results.

4.3 Evaluation results

Table 1 shows the maximum, mean and median discounted cumulative gain dcg as well as the precision for the top 5, 10 and 20 results according to different assessors, with the best results highlighted in boldface.

The high value for the dcg in Table 1, specially from linguistics participants indicates that the ranking of the results are effective, and higher ranked results appeared more similar to the query according to the linguist assessors. The same pattern is visible among the non-linguistic assessors as well, with a marginally lower score. According to precision@5, 10 and 20 from the same table, we can clearly see the high number of similar gestures according to the assessors appeared in the top 5 and 10 results. However, a stable precision at different number of results (P@5, P@10 and P@20) can be also due to a large number of videos in the database. To find a point where the precision starts to degrade and the diversity among the retrieved results increases, further experiments with higher number of results (possibly 50 or 100) would be beneficial. However, with a user study setup, it is unreasonable to expect assessors to rate 100 results of one query.

Table 1 Maximum, mean and median dcg and precision at 5, 10 and 20 of our method based on user study split between linguists, non-linguists and total scores. The best result per column for each category is printed in boldface

	Participants	P@5	P@10	P@20	dcg
mean	overall	0.38	0.4	0.395	5.85
	Linguists	0.4	0.41	0.41	**6.55**
	Non linguists	**0.44**	**0.47**	**0.44**	5.86
median	overall	0.4	0.35	0.375	5.64
	Linguists	0.4	0.35	0.4	5.58
	Non linguists	0.4	**0.6**	**0.5**	**5.70**
max	overall	1	0.9	0.8	8.71
	Linguists	1	1	**0.85**	**12.79**
	Non linguists	0.8	0.8	0.7	8.33

Table 2 Mean precision at 5, 10 and 20 according to the overall scores with respect to if the queries were from the video collection of NewsScape (Dataset) or newly recorded (Custom)

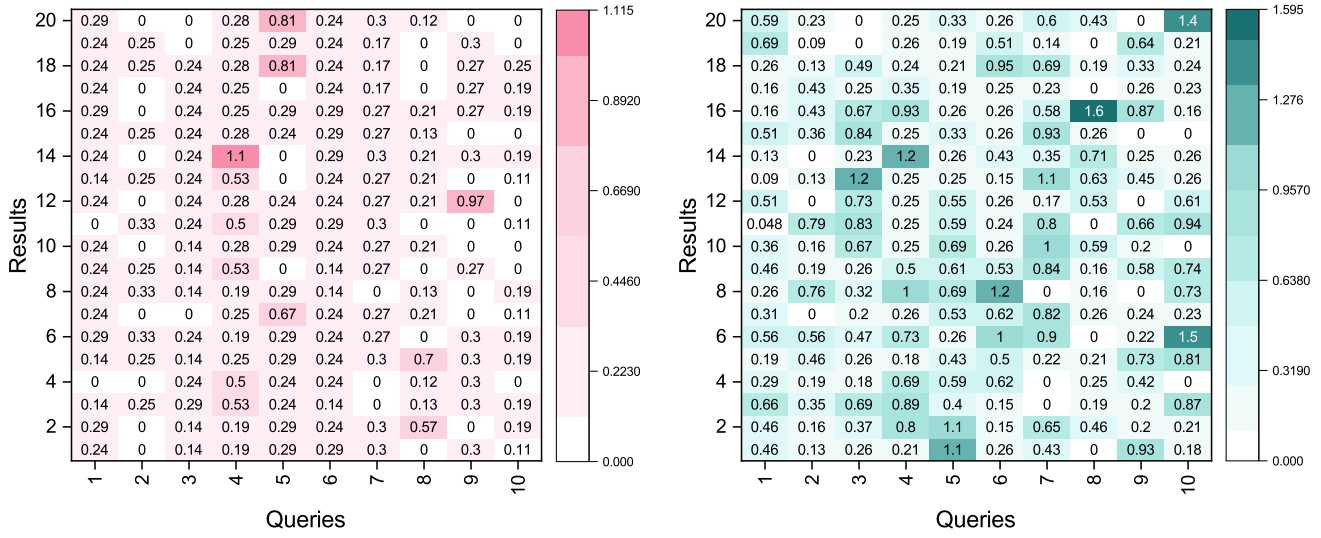
	Dataset			Custom		
	P@5	**P@10**	**P@20**	**P@5**	**P@10**	**P@20**
Overall	0.4	0.5	0.5	0.35	0.25	0.23

Comparing the statistics we gathered over the results, we observed a high number of average inter-rater agreement over all the scores, with $\kappa = 0.67$. This value is higher for linguists, with $\kappa = 0.81$. We can observe this also in the variances of the scores assigned by different assessors to the results (see Fig. 3). The heatmap indicates that the linguist assessors' scores have more consistency, and the variations between the scores are lower. On the other hand, from the heatmap it can be observed that the non-linguist assessors do not share the same consistency, which could be the result of a lack of deeper knowledge about certain co-speech gestures.[6]

We additionally analyzed the mean score each group of assessors assigned to different queries. Our analysis shows that query 8 was rated the worst both according to the linguists and non-linguists. This query is one of the custom made queries which have a significant difference in setting, lighting condition and camera angle to the rest of the data. The gesture performed in this query is shown in Fig. 4. The detailed statistical analysis and the in-depth view to the score of query 8 are shown in Figs. 5 and 6.

The results of the statistical analysis on different queries led us to perform an analysis of the precision with separate

[6] Note however that we expect a larger variance in the non-linguists due to a larger sample size, which might explain some of the difference visible in Fig. 3.



(a) Linguists (b) Non-linguists

Fig. 3 The heatmap showing the variance of the scores per query rated by linguists and non-linguists assessors (ratings are between 1-4)

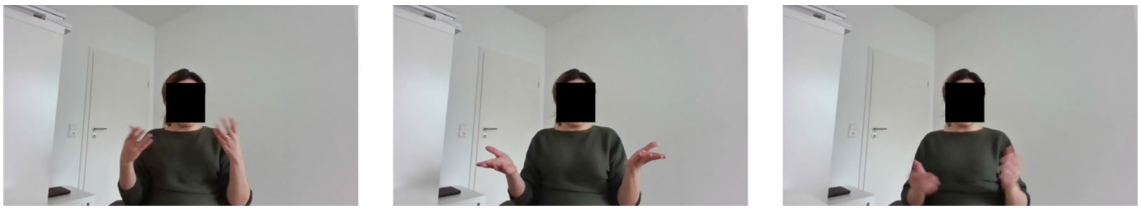
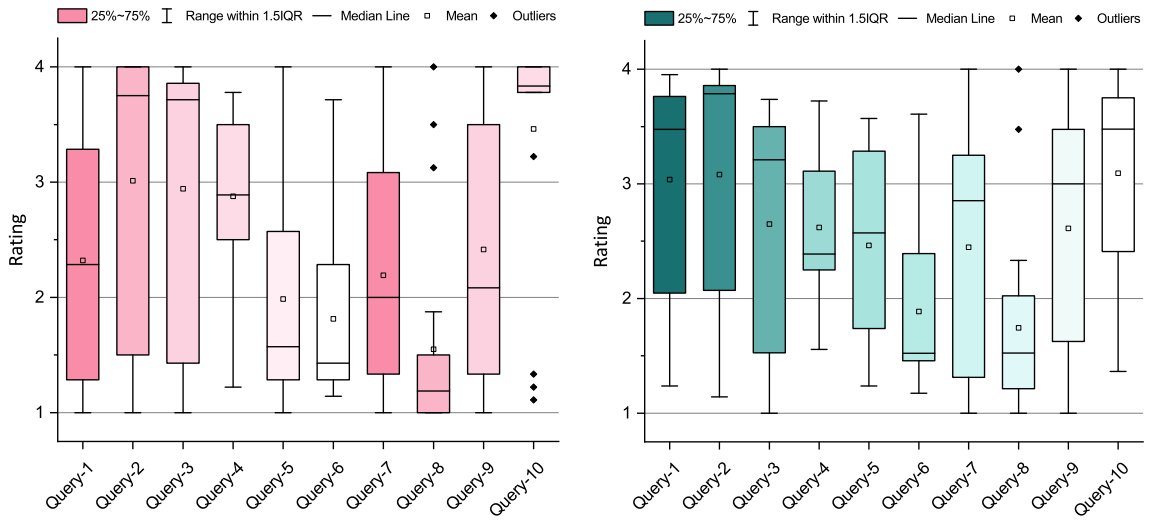
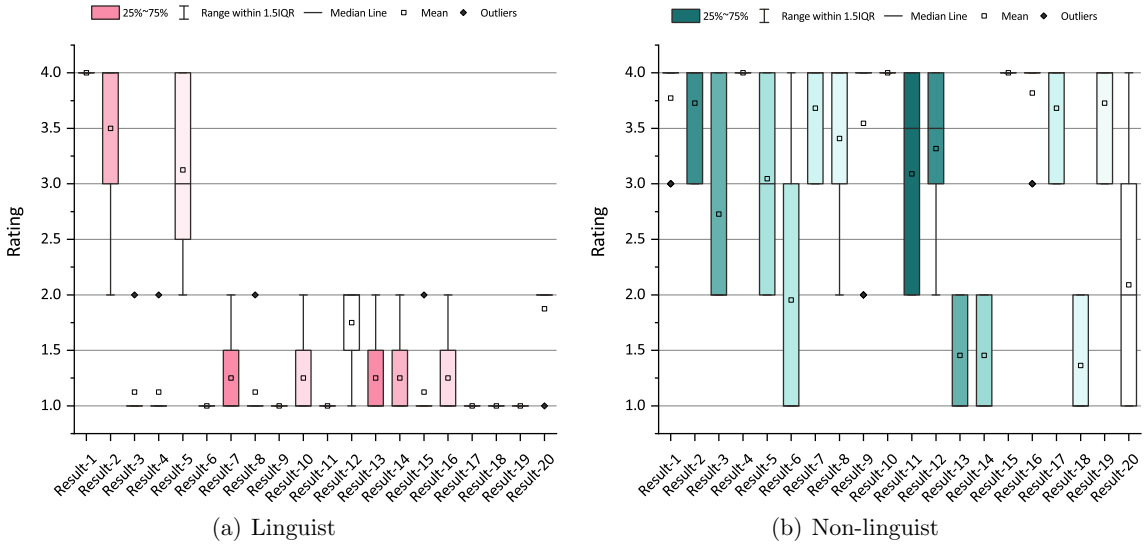


Fig. 4 Sample frames from query video 8 recorded by the first author and used as custom query



(a) Linguist (b) Non-linguist

Fig. 5 The descriptive statistical analysis of the mean score given by linguist and non-linguist assessors for all the queries



(a) Linguist (b) Non-linguist

Fig. 6 The detail scores given to custom query 8 separately by linguist and non-linguist assessors. The linguists consistently rated this query's results with low scores, while there is disparity of the ratings between non-linguist assessors

types of queries, to see how well the method can generalize. For this purpose, we computed the precision at different k separately for dataset and custom queries as can be seen in Table 2. We can clearly see that the custom queries have lower precision than the queries which are selected from the collection. We expected this to some extent due to the considerable difference between the custom query and the collection videos. Since the method also relies on the visual cues of the videos, change of colors and angle of the camera can be a reason for the difference of the performance.

Beside the quantitative studies, one of the queries and results obtained from our method are shown in Fig. 7.

5 Discussions and future directions

The results of the visual gesture retrieval on a real-world video dataset containing numerous instances of gestures gave us several insights into how this modality can help the multimodal search in large-scale video collections for analyzing co-speech gesture in linguistics.

5.1 Perceptive similarity of gestures

The evaluations performed in this paper further support the point made in previous works [20, 53, 54] about the notion of similarity and the unclear boundaries between dissimilar and similar gestures.

In linguistics, where hand gestures are defined with different components such as form and function, there is not *one* single notion of similarity which is generally applicable. For example, one of the selected hand gestures in the evaluation is showing the host clapping with only the palms of the hand (Fig. 8). The retrieved results to this query video are very similar to the act of clapping. However when listening to the host speaking, she is explicitly referring to this type of clapping (with palms only) as not fulfilling the objective of clapping:

> Ellen: Okay. So she was – and that was not making any noise. If you do that, it hardly does anything at all. The whole point is to make noise. (NewsScape: 2017-07-20_2200_US_KNBC_The_Ellen_DeGeneres_Show, 3:00-3:07)

Therefore, linguistically speaking, the retrieved results, which show the 'normal' form of clapping, do not have a similar *function*, therefore are not entirely *"similar"*. Since our proposed method is entirely independent of the audio signal, the results cannot reflect such functional similarity.

Taking a look at the transcript of this example, we clearly see that the gesture of clapping is not explicitly mentioned in the speech, which is another example of how useful the retrieval of visual instances can help analyze co-speech gesture.

The results obtained by separate groups of assessors showed that people with no linguistics background rate the similarity between the query video and the results more critically than the linguistic assessors. This can be the result of focusing on different elements such as facial expression, detailed trajectory of the gestures and the different camera angle in the recorded videos. For example, generally filliped trajectories of gestures are considered to have different meaning and therefore are assessed as dissimilar. However, occasionally such gestures share the same functionality in

Fig. 7 Sample frames from the query two persons hugging (top) and three top results retrieved from our method

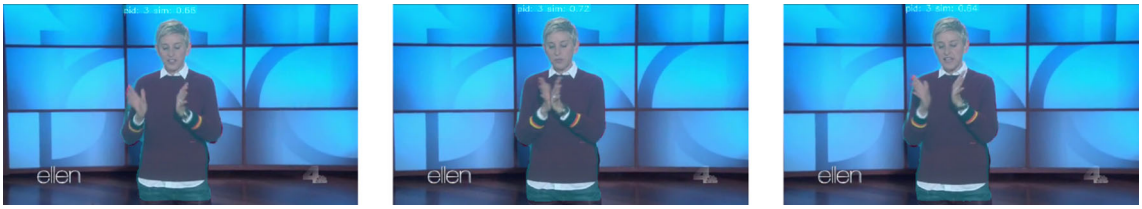


Fig. 8 Sample frames from the clapping query from the NewsScape dataset with the different function of the intended

linguistics. Additionally, the gesture articulation varies from person to person and this difference can potentially result in two similar gestures to appear dissimilar.

5.2 Multimodal gesture retrieval

Although our visual gesture retrieval pipeline has obtained good rating from assessors, this method is entirely independent of speech. Therefore, the results cannot reflect the functional similarity between the gestures. The combination of speech/text and vision and to extract the features from a co-embedding of these modalities can potentially improve the result of search both by addressing the function of the gestures and reducing the number of false positives retrieved by the visual similarity search.

Additionally, the integration of visual similarity search combined with the audio transcript, even though they may not share the same semantics (as shown in the example of clapping) can improve the search for instances where either of the modalities fails to find a specific instance of a multimodal pattern. In addition to the modalities explained in this paper, other types of modalities can be of help when searching for gestures in real-world videos. When talking about co-speech gestures, emotions and facial expressions can also play a role in narrowing down the search. The research in analyzing the emotion modality as a part of text [55], speech [56] or vision [57] has shown the importance of considering this modality in human interactions.

It is worth noting that the developed visual gesture retrieval method is entirely independent of the depth modality. The state of the art in gesture recognition in the largest isolated hand gesture dataset [58], which is a collection of 21 individuals performing pre-defined gestures from 249 classes, are mainly obtaining the high recognition rate using the depth modality. However, an enormous amount of video recordings does not come with depth data, and the perfor-

mance of these methods degrade, when this modality is removed [20, 41]. Since our proposed method is independent of the depth modality, it is sufficiently flexible to be used on any video recording obtained from various sources (mobile recordings, web videos, etc.).

5.3 Preprocessing effect in visual search

Our proposed method uses different preprocessing modules on the input data prior to feeding them to the feature extraction module, to specifically overcome the challenges inherent in the real-world data. The effect of segmenting the hand gesture prior to the feature extraction to diminish the background clutter effect has been discussed in [20]. This is particularly important in multi-person settings such as talk shows, where there are numerous sources of occlusion (by objects, another person or the banner or the subtitle) which can greatly degrade the search results. Additionally, our cross-angle temporal video trimming using the re-identification of people has a very important role in long, multi-person videos where the gesture of interest is articulated by a person among multiple people preforming some sort of hand gestures.

The drastic camera movements and change of angles in the existing footage from talk shows causes confusion in the process of projecting the hand gestures into a 2D plane and then extracting features. As long as a large amount of articulation is from one view, the results are not degraded severely. However, when the hand gesture is recorded from two or more points of view, the gesture feature will not be semantically aligned with the reality of the gesture. A possible solution is to use view independent feature embedding methods [59], which can be trained by 2D projections of 3D poses, based on multi-view frames. The advantage of this method over the current embedding is the probabilistic nature of it, which allows to extract view-invariant features from the video scenes and use them to retrieve similar gestures.

5.4 Impact on linguistic studies

The results gathered from linguists who participated in the evaluation survey of our method have shown that the retrieved videos meet the formal similarity expectations of the experts in the field. Despite the lack of functional similarity retrieval in our proposed method, our current pipeline can be used together with *vitrivr* as a preliminary gesture suggestion to reduce the manual annotation effort for co-speech gesture analysis. Additionally, by grouping simple atomic descriptors of hand movements, as in the following example (scan QR code or click on the url in the footnote)[7],

Gesture for "I"; See the Ellen's (blonde lady on the right) gesture:
handedness: both; handshape: flat; orientation: palm lateral, toward body;
location: center; movement type: straight; movement direction: toward body[8]

we can annotate the instances of the gestures in the video and also make them searchable via text-based or menu-based retrieval. The pose-based nature of our visual modality retrieval pipeline can be used to search for individual components of the gestures as described using an annotation tool such as ELAN (both hands, palms lateral, movement toward body). The result of this search can be annotated with the related textual label. This textual label together with the visual modality in vitrivr can be used for effective search in large video collections of gestures to aid linguists in analyzing co-speech gestures. In the long run, such annotated datasets can be used as training material for interactive systems, which can then learn which meanings are associated with which types of gesture.

6 Conclusion

With increasing prevalence of ubiquitous computing, interaction via natural language interfaces become increasingly important. While the usage of text and audio modalities has seen tremendous growth in the past decade in the study of multimodal communication and digital libraries, the visual modality has a very small share in this field. The visual modality and the tools to use it can bridge the semantic gap in textual retrieval systems which are prominently being used in the area of computational corpus linguistics. This paper has introduced a pipeline to retrieve co-speech gestures in real-world scenarios using a pose-based similarity learning method. The pipeline includes two separate stages to preprocess the data and another one to encode the information in the video frames into a feature vector. The features are then used to find visually similar gesture instances from the video collection. The performance of our proposed method has been assessed by two groups of linguists and non-linguists, evaluating the quality of the results based on their similarity to the query video. Our method is very robust to occlusion from different sources (object, person or subtitles), can

[7] http://go.redhenlab.org/pgu/0015

[8] The authors would like to thank Suwei Wu for providing the manual annotation shown here.

extract discriminative features from the videos, and retrieve results similar to custom query gestures. We have seen that the existing search infrastructure from the linguistic side and the computer vision side complement one another, but we are planning to create an integrated search engine for more powerful query options, combining multiple modalities (text, audio, video) simultaneously. The partial independence of the auditory and visual modalities, the undefined notion of perceived similarity between gestures, and the variation in camera angles are some of the remaining challenges which we discussed above to give directions for future research in multimodal interaction research, irrespective of whether it is human–human or human–machine interaction.

Acknowledgements We thank the Distributed Little Red Hen Lab for providing the NewsScape dataset used to carry out the evaluations for this paper. Additionally, the authors would like to express their gratitude to all the assessors who participated in the user-study. Part of this research was carried out using the High-Performance Computing facilities at FAU Erlangen-Nürnberg, supported by the KONWIHR grant *Robot Hen* to the second author.

Funding Open access funding provided by University of Basel This work was partly funded by the Hasler Foundation (contract no. 16074). The work in this paper was done while the first author had a double affiliation with the University of Basel, Switzerland, and the Numediart Institute at the University of Mons, Belgium.

References

1. Uhrig, P.: Multimodal research in linguistics. Z. Angl. Am. **68**(4), 345–349 (2020)
2. Kibrik, A.A., Fedorova, O.V.: Language production and comprehension in face-to-face multichannel communication. In: Komp'juternaja Lingvistika i Intellektual'nye Tehnologii, pp. 305–316 (2018)
3. Tomczak, L.: Are you aware of the avalanche of gay programming assaulting your home? The Christian Post **8** (2015)
4. Mittal, A., Gupta, S.: Automatic content-based retrieval and semantic classification of video content. Int. J. Digit. Libr. **6**(1), 30–38 (2006)
5. Joo, J., Steen, F.F., Turner, M.: Red hen lab: Dataset and tools for multimodal human communication research. Künstliche Intell. **31**(4), 357–361 (2017). https://doi.org/10.1007/s13218-017-0505-9
6. Uhrig, P.: Newsscape and the distributed little red hen lab - a digital infrastructure for the large-scale analysis of tv broadcasts. In: Anne-Julia Zwierlein, K.B. Jochen Petzold, Decker, M. (eds.) Anglistentag 2017 in Regensburg: Proceedings. Proceedings of the Conference of the German Association of University Teachers of English, pp. 99–114. Wissenschaftlicher Verlag Trier, Trier (2018)
7. Krauss, R.M., Hadar, U.: The role of speech-related arm/hand gestures in word retrieval. Gesture, speech, and sign **93** (1999)
8. Krauss, R.M., Chen, Y., Gotfexnum, R.F.: 13 lexical gestures and lexical access: a process model. Language and gesture **2**, 261 (2000)
9. Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J.R., Bethard, S., McClosky, D.: The stanford corenlp natural language processing toolkit. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, System Demonstrations, pp. 55–60. The Association for Computational Linguistics, Baltimore, MD (2014)
10. Evert, S.: The IMS open corpus workbench (CWB)–Corpus Encoding Tutorial. http://cwb.sourceforge.net/files/CWB_Encoding_Tutorial.pdf (2016)
11. Ochshorn, R.M., Hawkins, M.: Gentle: A robust yet lenient forced aligner built on Kaldi. https://lowerquality.com/gentle/ (2017)
12. Schiel, F., Kipp, A.: Probabilistic analysis of pronunciation with" maus" (1997)
13. Kisler, T., Schiel, F., Sloetjes, H.: Signal processing via web services: the use case webmaus. In: Digital Humanities Conference 2012 (2012)
14. Uhrig, P.: Large-scale multimodal corpus linguistics – the big data turn
15. Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P.: The kaldi speech recognition toolkit. In: IEEE 2011 Workshop on Automatic Speech Recognition and Understanding (2011). IEEE Signal Processing Society
16. Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Chen, J., Chrzanowski, M., Coates, A., Diamos, G., Elsen, E., Engel, J.H., Fan, L., Fougner, C., Hannun, A.Y., Jun, B., Han, T., LeGresley, P., Li, X., Lin, L., Narang, S., Ng, A.Y., Ozair, S., Prenger, R., Qian, S., Raiman, J., Satheesh, S., Seetapun, D., Sengupta, S., Wang, C., Wang, Y., Wang, Z., Xiao, B., Xie, Y., Yogatama, D., Zhan, J., Zhu, Z.: Deep speech 2 : End-to-end speech recognition in english and mandarin. In: Balcan, M., Weinberger, K.Q. (eds.) Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016. JMLR Workshop and Conference Proceedings, vol. 48, pp. 173–182. JMLR.org, New York City, NY (2016)
17. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. CoRR arXiv:2212.04356 (2022) https://doi.org/10.48550/arXiv.2212.04356
18. Hoffer, E., Ailon, N.: Deep metric learning using triplet network. In: International Workshop on Similarity-based Pattern Recognition, pp. 84–92 (2015). Springer
19. Mühling, M., Meister, M., Korfhage, N., Wehling, J., Hörth, A., Ewerth, R., Freisleben, B.: Content-based video retrieval in historical collections of the german broadcasting archive. Int. J. Digit. Libr. **20**(2), 167–183 (2019). https://doi.org/10.1007/s00799-018-0236-z
20. Amiri Parian, M., Rossetto, L., Schuldt, H., Dupont, S.: Are you watching closely? content-based retrieval of hand gestures. In: Gurrin, C., Jónsson, B.Þ., Kando, N., Schöffmann, K., Chen, Y.P., O'Connor, N.E. (eds.) Proceedings of the International Conference on Multimedia Retrieval, ICMR 2020, pp. 266–270. ACM, Dublin, Ireland (2020). https://doi.org/10.1145/3372278.3390723
21. Zhang, C.: Dynamic gesture retrieval: searching videos by human pose sequence (2020)
22. Hardie, A.: CQPweb-combining power, flexibility and usability in a corpus analysis tool. Int. J. Corpus linguist. **17**(3), 380–409 (2012)

23. Evert, S., Hardie, A.: Twenty-first century corpus workbench: Updating a query architecture for the new millennium (2011)
24. Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlỳ, P., Suchomel, V.: The sketch engine: ten years on. Lexicography **1**(1), 7–36 (2014)
25. Krause, T., Zeldes, A.: Annis3: a new architecture for generic corpus query and visualization. Dig. Scholar. Humanities **31**(1), 118–139 (2016)
26. Pagán Cánovas, C., Valenzuela, J., Alcaraz Carrión, D., Olza, I., Ramscar, M.: Quantifying the speech-gesture relation with massive multimodal datasets: Informativity in time expressions. PLoS ONE **15**(6), 0233892 (2020)
27. Stenger, B.: Template-based hand pose recognition using multiple cues. In: Asian Conference on Computer Vision, pp. 551–560 (2006). Springer
28. Wang, H., Schmid, C.: Action recognition with improved trajectories. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3551–3558 (2013)
29. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6299–6308 (2017)
30. Du, W., Wang, Y., Qiao, Y.: RPAN: an end-to-end recurrent pose-attention network for action recognition in videos. In: IEEE International Conference on Computer Vision, ICCV 2017, pp. 3745–3754. IEEE Computer Society, Venice, Italy (2017). https://doi.org/10.1109/ICCV.2017.402
31. Girshick, R.: Fast R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448 (2015)
32. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2961–2969 (2017)
33. Pishchulin, L., Insafutdinov, E., Tang, S., Andres, B., Andriluka, M., Gehler, P.V., Schiele, B.: Deepcut: Joint subset partition and labeling for multi person pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4929–4937 (2016)
34. Zhang, S., Li, R., Dong, X., Rosin, P.L., Cai, Z., Han, X., Yang, D., Huang, H., Hu, S.: Pose2Seg: Detection Free Human Instance Segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, pp. 889–898. Computer Vision Foundation / IEEE, Long Beach, CA, USA (2019). https://doi.org/10.1109/CVPR.2019.00098
35. Lin, T., Dollár, P., Girshick, R.B., He, K., Hariharan, B., Belongie, S.J.: Feature pyramid networks for object detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, pp. 936–944. IEEE Computer Society, Honolulu, HI, USA (2017). https://doi.org/10.1109/CVPR.2017.106
36. Cao, Z., Simon, T., Wei, S., Sheikh, Y.: Realtime multi-person 2D pose estimation using part affinity fields. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, pp. 1302–1310. IEEE Computer Society, Honolulu, HI, USA (2017). https://doi.org/10.1109/CVPR.2017.143
37. Xiao, T., Li, S., Wang, B., Lin, L., Wang, X.: Joint detection and identification feature learning for person search. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, pp. 3376–3385. IEEE Computer Society, Honolulu, HI, USA (2017). https://doi.org/10.1109/CVPR.2017.360
38. Parian-Scherb, M.: Gesture similarity learning and retrieval in large-scale real-world video collections. PhD thesis, University of Basel (2021)
39. Neverova, N., Wolf, C., Taylor, G., Nebout, F.: Moddrop: adaptive multi-modal gesture recognition. IEEE Trans. Pattern Anal. Mach. Intell. **38**(8), 1692–1706 (2015)
40. Wang, H., Wang, L.: Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 499–508 (2017)
41. Narayana, P., Beveridge, R., Draper, B.A.: Gesture recognition: focus on the hands. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5235–5244 (2018)
42. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997). https://doi.org/10.1162/neco.1997.9.8.1735
43. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, pp. 770–778. IEEE Computer Society, Las Vegas, NV, USA (2016). https://doi.org/10.1109/CVPR.2016.90
44. Rossetto, L., Giangreco, I., Tanase, C., Schuldt, H.: vitrivr: A flexible retrieval stack supporting multiple query modes for searching in multimedia collections. In: Hanjalic, A., Snoek, C., Worring, M., Bulterman, D.C.A., Huet, B., Kelliher, A., Kompatsiaris, Y., Li, J. (eds.) Proceedings of the 2016 ACM Conference on Multimedia Conference, MM 2016, pp. 1183–1186. ACM, Amsterdam, The Netherlands (2016). https://doi.org/10.1145/2964284.2973797
45. Gasser, R., Rossetto, L., Schuldt, H.: Multimodal multimedia retrieval with vitrivr. In: El-Saddik, A., Bimbo, A.D., Zhang, Z., Hauptmann, A.G., Candan, K.S., Bertini, M., Xie, L., Wei, X. (eds.) Proceedings of the 2019 on International Conference on Multimedia Retrieval, ICMR 2019, pp. 391–394. ACM, Ottawa, ON, Canada (2019). https://doi.org/10.1145/3323873.3326921
46. Gasser, R., Rossetto, L., Schuldt, H.: Towards an all-purpose content-based multimedia information retrieval system. CoRR arXiv:1902.03878 (2019)
47. Lokoč, J., Veselỳ, P., Mejzlík, F., Kovalčík, G., Souček, T., Rossetto, L., Schoeffmann, K., Bailer, W., Gurrin, C., Sauter, L.: Is the reign of interactive search eternal? findings from the video browser showdown 2020. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 17(3), 1–26 (2021)
48. Gasser, R., Rossetto, L., Heller, S., Schuldt, H.: Cottontail DB: an open source database system for multimedia retrieval and analysis. In: Chen, C.W., Cucchiara, R., Hua, X., Qi, G., Ricci, E., Zhang, Z., Zimmermann, R. (eds.) Proceedings of the 28th ACM International Conference on Multimedia, MM '20, pp. 4465–4468. ACM, Virtual Event / Seattle, WA, USA (2020). https://doi.org/10.1145/3394171.3414538
49. Rossetto, L., Giangreco, I., Schuldt, H.: Cineast: A multi-feature sketch-based video retrieval engine. In: 2014 IEEE International Symposium on Multimedia, ISM 2014, pp. 18–23. IEEE Computer Society, Taichung, Taiwan (2014). https://doi.org/10.1109/ISM.2014.38
50. Heller, S., Sauter, L., Schuldt, H., Rossetto, L.: Multi-stage queries and temporal scoring in vitrivr. In: 2020 IEEE International Conference on Multimedia & Expo Workshops, ICME Workshops 2020, pp. 1–5. IEEE, London, UK (2020). https://doi.org/10.1109/ICMEW46912.2020.9105954
51. Rossetto, L., Gasser, R., Schuldt, H.: Query by semantic sketch. CoRR arXiv:1909.12526 (2019)
52. Heller, S., Arnold, R., Gasser, R., Gsteiger, V., Parian-Scherb, M., Rossetto, L., Sauter, L., Spiess, F., Schuldt, H.: Multi-modal interactive video retrieval with temporal queries. In: Jónsson, B.Þ., Gurrin, C., Tran, M., Dang-Nguyen, D., Hu, A.M., Binh, H.T.T., Huet, B. (eds.) Proceedings of the 28th International Conference on MultiMedia Modeling, Part II, MMM 2022. Lecture Notes in Computer Science, vol. 13142, pp. 493–498. Springer, Phu Quoc, Vietnam (2022). https://doi.org/10.1007/978-3-030-98355-0_44
53. Rossetto, L.: Multi-modal video retrieval. PhD thesis, University of Basel (2018)
54. Parian-Scherb, M., Walzer, C., Rossetto, L., Heller, S., Dupont, S., Schuldt, H.: Gesture of interest: Gesture search for multi-person,

multi-perspective tv footage. In: Proceedings of the Content-Based Multimedia Indexing, CBMI. IEEE, Lille, France (2021)
55. Wang, S., Maoliniyazi, A., Wu, X., Meng, X.: Emo2vec: learning emotional embeddings via multi-emotion category. ACM Trans. Internet Techn. **20**(2), 13–11317 (2020). https://doi.org/10.1145/3372152
56. Majumder, N., Poria, S., Hazarika, D., Mihalcea, R., Gelbukh, A.F., Cambria, E.: Dialoguernn: An attentive RNN for emotion detection in conversations. In: The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, pp. 6818–6825. AAAI Press, Honolulu, HI, USA (2019). https://doi.org/10.1609/aaai.v33i01.33016818
57. Zhu, T., Xia, Z., Dong, J., Zhao, Q.: A sociable human-robot interaction scheme based on body emotion analysis. Int. J. Control Autom. Syst. **17**(2), 474–485 (2019)
58. Wan, J., Li, S.Z., Zhao, Y., Zhou, S., Guyon, I., Escalera, S.: Chalearn looking at people RGB-D isolated and continuous datasets for gesture recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2016, pp. 761–769. IEEE Computer Society, Las Vegas, NV, USA (2016). https://doi.org/10.1109/CVPRW.2016.100
59. Sun, J.J., Zhao, J., Chen, L., Schroff, F., Adam, H., Liu, T.: View-invariant probabilistic embedding for human pose. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J. (eds.) Proceedings of the 16th European Conference on Computer Vision, Part V, ECCV 2020. Lecture Notes in Computer Science, vol. 12350, pp. 53–70. Springer, Glasgow, UK (2020). https://doi.org/10.1007/978-3-030-58558-7_4