Gesture of Interest: Gesture Search for Multi-Person, Multi-Perspective TV Footage

Mahnaz Parian[1,2],Claire Walzer[1],Luca Rossetto[3],Silvan Heller[1],Stéphane Dupont[2],Heiko Schuldt[1]

[1]DBIS Group, University of Basel
Basel, Switzerland

[2] ISIA Lab, University Mons
Mons, Belgium

[3]DDIS Group, University of Zurich
Zurich, Switzerland

***Abstract*—In real-world datasets, specifically in TV recordings, videos are often multi-person and multi-angle, which poses significant challenges for gesture recognition and retrieval. In addition to being of interest to linguists, gesture retrieval is a novel and challenging application for multimedia retrieval. In this paper, we propose a novel method for spatio-temporal gesture retrieval based on visual and pose information which can retrieve similar gestures in multi-person scenes through continuous shots. The attention-aware features, extracted from human pose key-points, together with a sophisticated pre-processing module, alleviate the susceptibility of gesture retrieval to background noise and occlusion. We have evaluated our method on a subset of the NewsScape Dataset. Our experimental results demonstrate the effectiveness of the proposed method in retrieving similar results in occluded scenes as measured by the quality of the top 5 results.**

***Index Terms*—Content-aware Gesture Retrieval, Similarity Search, Video Retrieval, Deep-neural Embedding**

I. Introduction

Human communication consists of non-verbal modalities such as gestures alongside verbal speech (with the exception of sign language). However, our understanding of the co-occurrence of these modalities with spoken language is still limited. The study of the temporal proximity of hand gestures and speech identifies gestural and verbal patterns which happen systematically. While projects such as NewsScape [1] have given researchers access to large volumes of news footage, understanding of gesture-speech co-occurrences is limited by the manual annotations of gestures in order to enable their retrieval and subsequent analysis.

One way to overcome the lack of annotations in large gesture corpora is using a video retrieval method tailored to gestures. Searching in large scale video collections is a challenging task which is being tackled by different video retrieval systems in recent years [2], [3]. However, the capabilities to search for gestures in such collections have not been developed to the same extent. In addition to the challenges of finding a similar gesture articulation, datasets often exhibit multi-person and multi-angle scenes, which pose significant challenges for gesture retrieval, such as gesture articulation potentially continuing over shot transitions (see Fig. 1). Such a loss of information will result in poor gesture feature extraction and thus poor retrieval performance.

In this paper, we present an attention-aware pose based dynamic gesture retrieval method based on visual and pose cues which can retrieve gestures in occluded, multi-person, multi-angle settings. The gestures we are looking at are dynamic movements of the hands and body parts. *Gesture of Interest (GoI)* benefits from state of the art segmentation and feature extraction components to perform gesture-based retrieval. The main contribution of this work is twofold: Firstly, we propose a spatio-temporal pre-processing method which robustly segments human instances in occluded scenes and tracks them through multiple perspectives and generates uninterrupted gesture sequences per each individual. Secondly, we propose a feature extraction and metric learning module inspired by the pose-based attention mechanism for gesture search in multi-person interaction scenes. GoI jointly uses these two components to search in large collections of videos and retrieve similar gestures to the query independent of changes in perspective. The evaluations on the recorded TV footage from the NewsScape dataset show that our method is very robust in retrieving similar gestures to the query in multi-person and occluded scenes.

Fig. 1: Example of a gesture spanning through adjacent shots.

The remainder of the paper is organized as follows: In Section II, we discuss related work and Section III introduces the spatio-temporal segmentation and feature extraction of GoI. Section IV presents the evaluation setup, results and discussions of our user study. Section V concludes.

II. Related work

The area of gesture recognition has benefited substantially from action recognition methods due to the similarity of their approaches in understanding motion and posture. However, pre-prosessing has a more dominant role in gesture analysis.

A. Pre-processing Methods

1) Multi-Person Spatial Segmentation: Many person detection methods originate from the object detection tasks. Although Mask R-CNN [4] is one of the popular methods

in instance segmentation, it detects the objects based on bounding boxes which results in sub-optimal segmentation for overlapping and occluded human instances. In contrast to the Mask R-CNN, there are recent human instance segmentation methods [5]–[9] which bene t from pose structure. The top-down algorithms (human detection followed by pose estimation) [10] essentially face the same problem in occlusion as non-pose based methods. In this paper we use Pose2Seg [6] which is a bottom-up (key-points detection followed by pose estimation) [11] approach that generates instance proposals from key-points and estimates human segmentation.

2) Multi-Perspective Person Search:Multi-perspective person search or person tracking is the task of recognizing a target person over several perspectives in different spatial and temporal instances. This task is most often solved by using visual object detectors [12] and by comparing features from the detected person with the collection [13], [14]. Deep learning-based methods are based on extracting features from cropped human instances while using binary veri cation or triplet loss functions for similarity learning [15], [16]. Xiao et al. [17] combine the person search and re-identi cation task in a single CNN. In this paper, we re-purpose [17] to align the continuous sequence of persons in multiple perspectives to capture the entire gestural context.

B. Gesture and Action Feature Extraction

The common approach in gesture and feature extraction is to train a supervised recognition network and to use the trained model to embed the features. However these methods vary in the input type and the architecture of the network.

1) Vision Based:Video feature extraction bene ted a lot from 2D-CNNs. However, they only consider spatial inter-relation of pixels without any temporal context. A common way to add the temporal context is to average the prediction scores obtained from individual frame features [18]. Other methods integrate the temporal structure in the network architecture such as long short-term memory (LSTM) [19]. On the other hand, two stream architectures [20] extract spatial and temporal features using optical ow as a separate input for the temporal context. 3D convolutional networks [21] extract spatio-temporal information directly from raw RGB data. Zhang et al. [22] extracted the 2D spatio-temporal feature maps using 3D convolutions and fed them into a bi-directional convolutional LSTM, for an end-to-end extraction of spatio-temporal features. Furthermore, Carreira et al. [23] proposed the I3D model, which combines the two-stream architecture with the 3D-CNN.

2) Attention Based:After the increased popularity of the attention mechanism in Natural Language Processing (NLP), computer vision tasks also adapted this useful method in their feature extraction pipelines. Soft attention feature extraction is based on weighting the average of features and focusing on different parts of the frame [24]. VideoLSTM [25] learns the sequential features with motion-based attention, which provides better guidance towards relevant spatio-temporal locations. Recent works on the Video Action Transformer

Fig. 2: Architecture diagram of the GoI method

Network [26] uses a modi ed Transformer architecture [27] to classify the action of a target person.

3) Pose Based Feature Extraction:Pose based feature extraction models aim to learn a representation that best preserves the spatio-temporal relations among the joints. In contrast to vision based models, they are more robust against a complex background, motion speed, and changing body scales. Cao et al. [28] presented an attention model, which predicts spatio-temporal key-points in 3D convolutional feature maps. Du et al. [29] divide the human skeleton into ve parts and feed them into ve bi-directional RNNs, which are then fused part by part. Recurring Pose-Attention Network (RPAN) [30] resize the joint coordinates to a 2D map to feed it into a pre-trained CNN. They also consider the correlation of the joints by dividing the human skeleton into semantically correlated parts. Our approach uses the rich human joints-attended features extracted by RPAN [30] for learning the similarity between the gestures.

C. Gesture Retrieval

Gesture retrieval is concerned with nding similar instances of dynamic hand gestures in videos. Earlier works [31] established a method to retrieve people using their pose. Youse et al. [32] used hierarchical scoring of edge-orientation features to nd the best match between the query and the collection. I3DEF [33] incorporated vision-based deep features with a two-stream network. These models fail in occluded, multi-person and multi-perspective scenarios, as they have no mechanism to identify or track individuals.

III. METHOD

The GoI method brings together two main building blocks (Fig 2): i.) multi-angle spatio-temporal person segmentation and tracking and ii.) gesture feature extraction and retrieval.

A. Multi-Perspective Spatio-Temporal Person Segmentation

The spatio-temporal person segmentation module is a pre-processing step to reduce background clutter and handle occlusion, multi-person and multi-perspective scenarios. This module combines pose-based person segmentation [6] and person tracking [17] where the latter is adapted to our application. The initial step is the skeleton key-point extraction which takes the raw video frames and extracts the joint key-points. The skeletal pose is essentially a list of vectors which are mapped to pose templates from the dataset. The segmentation model [6] uses a pre-trained base network (ResNet-50 [34]) to

extract the spatial features from the detected human instances in the video frame and skeletal features from pose key-points. The skeletal features are formed by the Part Affinity Fields and confidence map. The skeletal and visual features are then used in a network to reverse the affine operation and predict the pixel-wise person segmentation. Using the skeletal features, the model can segment multiple persons in a frame and is robust against occlusion.

The output of the segmentation model is used in the re-identification model to track specific people and temporally segment the video regardless of changes of camera perspective. The person tracking model we use [17] is a robust method based on visual cues which is used mainly in pedestrian scenarios. This method is originally based on region-of-interest (RoI) proposals extracted by a CNN and human feature embeddings obtained from ResNet-50. 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. We replace RoI proposals with the human segmentation output from the previous step which is robust to occlusion in multi-person scenarios. To track the people reappearing in a sequence of frames, the similarity score between a query and a reference instance is calculated. Since in the beginning there is no such reference gallery available, we initialize the gallery with the first instance of a person in the video. If the similarity score is samller than, the query object is added as a reference to the gallery and is used to re-identify the queries in the next frames. Otherwise, the person is labeled with the same ID as the one in the gallery.

B. Gesture Feature Extraction and Retrieval

1) Gesture Feature Extraction: The complexity of tracking gestural trajectories in multi-angle scenarios requires a robust feature extraction module which can represent discriminative information about hand articulations. After having segmented the entire video spatio-temporally, there will be multiple sequences of frames per person, based on their continuous appearance in the video. Each of these short clips are used for feature extraction and retrieval. The feature extraction module is based on RPAN [30] which is an end-to-end RNN with a pose-attention mechanism that learns to focus on active human joint parts. The spatial features are extracted from a Temporal Segmentation Network (TSN) [35] with only its spatial stream and form a convolutional cube with the size of $K_1 \quad K_2 \quad n$ based on aggregating n feature maps.

To learn the dynamics of an activity and the fine-grained movements which comprise a gesture, it is important to include another level of supervision other than the gestural categories. Du et al. [30] proposed a pose attention mechanism that together with LSTM units can extract pose features to model the temporal dependencies of activities with regard to the body part involvement. This mechanism is defined by an attention heatmap $^{J}_{t}(k)$ for each feature vector of a convolutional cube for each joint (J) which together with semantically relevant joints, form a body part structure (P) [30]:

$$^{J}_{t}(k) = \frac{\exp\{v^{J}\tanh(A^{P}_{h}h_{t-1} + A^{P}_{c}C_{t}(k) + b^{P})\}}{\sum_{k}\exp\{v^{J}\tanh(A^{P}_{h}h_{t-1} + A^{P}_{c}C_{t}(k) + b^{P})\}} \quad (1)$$

where h_{t-1} is the hidden state of the previous frame, v^{J}, A^{P}_{h}, A_{c}, b^{P} are the attention parameters, all of which except v^{J} are shared between the joints ($J \in P$). The human-part features are $F^{P}_{t} = \sum_{J \in P}\sum_{k} {}^{J}_{t}(k)C_{t}(k)$ and fused together by a pooling layer to generate pose-related features S_{t}. To capture the temporal dimension of the movements, the features are then fed to an LSTM (h_{t}) to generate a prediction vector based on the probability for each class label at each frame of video (y_{t}). RPAN is trained end-to-end and its total loss function is the sum of a cross entropy loss L_{action} and a pose-loss L_{pose} (the latter being the Euclidean distance between the ground truth pose annotations and the attention heatmap as in Eq. 1).

2) Gesture Retrieval: As the RPAN is trained for a classification task, the output of the network is a probability score for each class. To suit our purpose of retrieval, we train the network to extract features which represent the similarity between the clips of the same gesture. For this, the similarity between the query and each video clip in the collection is computed, and the results are retrieved based on the maximum similarity between the query and the collection videos.

We train the RPAN with a triplet loss function [36] with the objective of learning the representations in a way that similar videos have a smaller distance, and the non-similar pairs of videos have a larger distance than. To meet this objective, the RPAN output is modified to extract the features before the softmax layer, to obtain the feature vector which contains the pose information. The training process includes a collection of triplets from the extracted features from each video $\{h_{T}; h^{+}_{T}; h_{T}\}$, where $h_{T}; h^{+}_{T}; h_{T}$ are the feature vectors of the anchor, positive and negative samples, respectively. To generalize the training of a triplet loss network, we introduce a margin, m, set to an experimentally determined value of $m = 0.3$ which results in separate clusters of the embeddings. Finally, the training process includes the minimization of the $L_{triplet} = \max(D(h_{T}; h^{+}_{T}) \quad D(h_{T}; h_{T}) + m; 0)$ to generate features with smaller distance for similar videos and larger distance for dissimilar ones.

IV. Experiments

In this section, we explain the implementation details of the network and evaluation setup to perform gesture retrieval for real-world data, followed by the evaluation results.

A. Dataset

RPAN is originally trained on an action recognition dataset. In our application, we first train the network on a large-scale gesture recognition dataset, Chalearn, and evaluate it on a part of NewScape available via the UCLA Library.

1) Chalearn dataset: The Chalearn Iso dataset [37] is a large-scale gesture dataset containing 249 types of gestures performed by 21 individuals. All the 47 933 videos in the dataset come in RGB and depth video format.

2) NewsScape dataset:The UCLA Library Broadcast NewsScape dataset [1] contains more than 400 000 television shows recorded since 2005. This dataset is gathered from news and talk shows all around the world and comes with RGB videos and the transcribed context. As the videos in this dataset are not hand picked for research, they require robust action and gesture recognition methods to overcome the challenges in the real-world data. The multi-perspective and multi-person setting of these videos, which are an inherent characteristic of TV shows, are two examples of these challenges. We use a subset of this dataset containing roughly 3 hours of recordings split into 30 videos of 4–12 minutes from The Ellen DeGeneres Showprovided to us by the Redhen Lab.

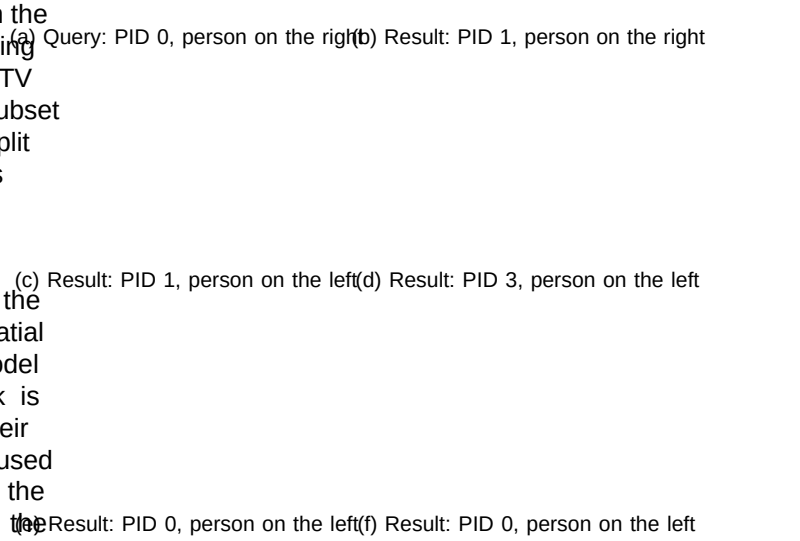
B. Implementation Details

The pose information needed for pre-processing and the feature extraction is obtained using OpenPose [11]. The spatial person segmentation is based on the of cial pre-trained model released by the authors of Pose2Seg [6]. The network is fed with the sequences of video frames along with their extracted key-points, and the segmentation masks are used to separate each individual person as well as to remove the background. The background removal especially causes the feature extraction network to be independent of the clutter.

In the temporal person segmentation module, to prevent clutter in the gallery, we only process frames with fewer than 5 people in them. The base network to extract the person features for re-identi cation is a pre-trained ResNet-50 provided by the authors of [17]. The features extracted from the query and the gallery entries are compared with the cosine similarity measure. Each query will get a similarity score which determines the person ID of that instance. To capture the continuous gesture articulation, a dictionary with the person ID, and the stack of frames which have this ID are constructed. The frames will be concatenated to the sequence as long as the same person ID is detected in consecutive frames. Once the frame does not contain the speci c person ID, a new stack will be created with the new person ID.

The backbone is a ResNet-50 to extract convolutional cubes with the size of $(B \quad T; 224; 224; 3)$ where the batch size is $B = 3$ and $T = 8$ frames from a video sequence. The extracted key-points are formed into 5 body parts. We train the network once on the Chalearn gesture dataset with classi cation objective. During training, we use the Stochastic Gradient Descent optimizer with the momentum $= 0:9$ and the learning rate $= 10^{-5}$.

When the network is initialized with the classi cation weights, we remove the softmax layer and train it again with the triplet loss layer to learn the similarity. The features from the last layer of the LSTM with size $(B; T; 512)$ are used for each $T = 8$ frames and are averaged to create a single feature vector for a video clip $(B; 512)$. In order to minimize the loss, we used the Adam optimizer with a learning rate $= 10^{-5}$. For the triplet loss training, we use the batch all method and we select all valid triplets and average the loss on the hard and semi-hard triplets. We use all the videos as anchors and



(a) Query: PID 0, person on the right (b) Result: PID 1, person on the right

(c) Result: PID 1, person on the left (d) Result: PID 3, person on the left

(e) Result: PID 0, person on the left (f) Result: PID 0, person on the left

Fig. 3: An examples of the evaluation query and its 5 retrieved results. (a) represent the queries.

form the triplets with 3 positives from the same category of the anchor and 4 negatives from the other categories. After training the triplet network, we extract the features for the subset of the NewsScape dataset for retrieval.

GoI extracted 8 177 gesture sequences from the 30 videos of the NewsScape dataset from which the queries were selected randomly. When querying, the distance between the query and the feature instances saved in the database are compared. After obtaining the $R = 5$ closest features, we collected the corresponding output frames. Along with the ranked output recordings, the user receives information about which person instance is assigned as the target person (in case of a multi-person setting), video source and the distance.

C. Evaluation Setup

To evaluate the GoI retrieval method, we use video clips from The Ellen DeGeneres Show. This dataset is speci cally interesting as it includes recorded scenes of multiple persons talking, where the co-speech gestures often happen. This is ideal to measure the performance of our approach. The evaluation was performed in a user study with 20 participants without any specialized linguistic knowledge, who were asked to rate the similarity of the results according to a 4-point Likert scale. The participants were asked to rate the most similar results with 4 and the least with 1. The similarity is concerned with the gesture articulations, and not limited to the person or the point of view of the recording. For the evaluation, 15 predetermined queries with the top-5 results were used.

TABLE I: How often the video in average was rated with for all queries

	1: dissimilar	2: slightly dissimilar	3: slightly similar	4: similar
vid_1	1,71 2,51	4,19 2,99	2,32 2,32	11,37 6,11
vid_2	1,26 1,94	6,35 2,52	3,61 3,12	8,18 3,91
vid_3	3,34 3,18	6,09 2,52	4,57 3,14	5,60 5,32
vid_4	1,38 2,43	4,66 2,58	2,94 2,78	10,66 5,95
vid_5	3,55 3,67	5,79 2,87	4,82 3,21	5,53 5,12

D. Evaluation metrics

To measure the retrieval performance, we use the Fleiss kappa to assess the inter-rater reliability, which is the degree of agreement among the participants when assigning a rating, to see how reliable our retrieved results are.

Another statistical metric we considered for our evaluations is mean reciprocal rank (MRR) score, which is measuring the ranking quality. This metric is de ned as $\frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank_i}$, where Q refers to the number of queries and $rank_i$ is the position of the rst relevant result to the query.

E. Evaluation results

In this section, we explore the results derived from the user study experiment and the ratings. Fig.4 shows an overview of the rating results by the users. At rst glance, almost 70% on average were rated with similar and slightly similar, which indicates GoI is effective in retrieving similar results from real world human interaction scenes.

1) User Agreement: There is always a level of disagreement when asking users to rate the similarity of gestures, depending on how they perceive this similarity. To examine this phenomenon, which can indicate how well the retrieved result comply with user's expected similarity, we measure the Fleiss' kappa coef cient. It ranges from 0.13 to 0.43, the mean of 0.23 indicates only a fair agreement between the participants.

2) Quality of ranking: With a close look to the user ratings, we can examine the distribution of the most similar videos in the retrieved results. According to Fig. 5, 73% of the rst and second ranked results gained highest ratings and 50% of the highest ratings was given to these two ranks, which is an indicator that the ranking strategy was rather optimal as well. Additionally, we calculate the MRR score for all queries by considering the rank of the rst position of the result which has the average rating of more than 3. The measured MRR is 0.85, which states that the rst similar result appeared mostly on the rst place. Fig. 6 shows the distribution of the standard deviation of the rating per each result. It can be seen that there are some absolute agreements on similar results and very low deviation of ratings for the dissimilar results.

F. Discussion

The evaluation results show the effectiveness of GoI in retrieving similar videos to a query. An examples of the query (the action of sitting) and retrieved result examples are depicted in Fig. 3. In some cases the re-appearance of the same person as in the query in the result could be the result of an assumption, suggesting that each individual has a xed gesture vocabulary, which tends to repeat subconsciously. This repetition would result in two instances of the same gesture performed by the same person to be more similar to each other than two instances of the same gesture performed by different individuals. Unfortunately, it is impossible to verify this assumption based on the current data and would need dedicated investigations.

Closer examinations of the query results provided insights on the stability of the method across different perspectives. Although not all query results cover different angles, there are some results which are hand movements of a close up and full body scenes with different actors across multiple camera shot transitions. Additionally, the results show the successful spatial segmentation of people in multi-person scenes. This can be seen prominently in the results where only one of the multiple people in the frame has the same gesture articulation.

Despite the clear instructions to the users on how two gestures could be similar, the distribution of rating illustrated in Fig. 6 shows that the rst query results were judged more critically and towards the end of the evaluation, the users tend to agree more on similarity of gestures. The divergence of the rating indicates the ill-de ned nature of gesture similarity

From a computational point of view, training the feature extraction and pre-processing are the most computationally expensive procedures. Training with triplet loss on Chalearn dataset with 41 662 videos took about 17 days on a NVIDIA 1080Ti with 8gb RAM while feature extraction was done in 30 minutes for the 175 minutes of the subset of NewsScape dataset. However, the pre-processing step including the pose estimation, spatial and temporal segmentation took roughly about one hour for a 6 minutes video. Processing a 6 seconds video query in inference mode takes about 14 seconds.

V. Conclusion

In this paper, we presented our spatio-temporal gesture search method, GoI, consisting of the pre-processing and feature extraction components. The method addresses the multi-person and multi-angle similarity search challenges posed by TV footage. GoI bene ts from two key elements in retrieving similar hand gestures of individuals across the shot boundaries in occluded scenes: 1) human pose information and 2) attention mechanism. The former, together with visual cues, enable robust human instance segmentation in occluded scenes. The latter improves the feature extraction to focus on the human parts.

Acknowledgement

The authors would like to thank the Red Hen Lab for access to the UCLA NewsScape Library of Television News.

References

[1] J. Joo, F. F. Steen, and M. Turner, "Red hen lab: Dataset and tools for multimodal human communication research," Kunstliche Intelligenz, vol. 31, no. 4, 2017.

[2] L. Rossetto, R. Gasser, J. Lokoc, W. Bailer, K. Schoeffmann, B. Muenzer, T. Soucek, P. A. Nguyen, P. Bolettieri, A. Leibetseder, et al., "Interactive video retrieval in the age of deep learning-detailed evaluation of vbs 2019," IEEE Transactions on Multimedia, 2020.