Are You Watching Closely? Content-based Retrieval of Hand Gestures

Mahnaz Amiri Parian
University of Basel
University of Mons
mahnaz.amiriparian@unibas.ch

Luca Rossetto
University of Zurich
rossetto@ifi.uzh.ch

Heiko Schuldt
University of Basel
heiko.schuldt@unibas.ch

Stéphane Dupont
University of Mons
stephane.dupont@umons.ac.be

ABSTRACT

Gestures play an important role in our daily communications. However, recognizing and retrieving gestures in-the-wild is a challenging task which is not explored thoroughly in literature. In this paper, we explore the problem of identifying and retrieving gestures in a large-scale video dataset provided by the computer vision community and based on queries recorded in-the-wild. Our proposed pipeline, *I3DEF*, is based on the extraction of spatio-temporal features from intermediate layers of an I3D network, a state-of-the-art network for action recognition, and the fusion of the output of feature maps from RGB and optical flow input. The obtained embeddings are used to train a triplet network to capture the similarity between gestures. We further explore the effect of a person and body part masking step for improving both retrieval performance and recognition rate. Our experiments show the ability of *I3DEF* to recognize and retrieve gestures which are similar to the queries independently of the depth modality. This performance holds both for queries taken from the test data, and for queries using recordings from different people performing relevant gestures in a different setting.

CCS CONCEPTS

• **Information systems** → **Specialized information retrieval**; **Similarity measures**; • **Applied computing** → *Arts and humanities*; • **Human-centered computing** → Gestural input.

KEYWORDS

Content-based Gesture Video Retrieval, Deep-neural Embedding

ACM Reference Format:
Mahnaz Amiri Parian, Luca Rossetto, Heiko Schuldt, and Stéphane Dupont. 2020. Are You Watching Closely? Content-based Retrieval of Hand Gestures. In *Proceedings of the 2020 International Conference on Multimedia Retrieval (ICMR '20), June 8–11, 2020, Dublin, Ireland.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3372278.3390723

ICMR '20, June 8–11, 2020, Dublin, Ireland

ACM ISBN 978-1-4503-7087-5/20/06…$15.00
https://doi.org/10.1145/3372278.3390723

1 INTRODUCTION

Gestures are a natural part of human communication since they augment the spoken word and help to convey both meaning and emotion. Gesture recognition has recently gained special attention due to the development of robotic agents with whom humans can interact by means of (hand) gestures. The gestures that such applications would need to understand in addition to certain emblems are co-speech gestures which happen spontaneously to enrich the speech utterance.

Compared to gesture recognition, relatively little work has been done in comparing the similarity of such recordings for the purpose of retrieval. However, this is highly relevant in several use cases such as the support for linguists in analyzing co-speech gestures, or as a complement to other modes of retrieval in video search. In this paper, we therefore present a content-based retrieval approach for videos of hand gestures.

The proposed method, *I3D Embedding Fusion (I3DEF)*, benefits from the intrinsic properties of a deep convolutional neural network with 3D kernels to extract spatio-temporal features [35]. Studies have shown the advantage of using intermediate CNN layers in different architectures for retrieval and recognition tasks [22]. Therefore, we develop our retrieval system based on the intermediate layer's features from Inflated 3D Inception (I3D). The extracted features are used to train a triplet network [12] to learn the similarity between the gesture videos. Although this method is wildly used in image and video retrieval [7, 27], learning the metric to measure and retrieve similar gestures is a challenging task. In our setup, we explore the different configurations to find the best practice to retrieve similar gestures. To test the retrieval performance of *I3DEF*, we selected queries both from test section of the dataset used for training and validating the model, and real-world videos. Our experiments compare the results of the similarity of gestures as rated by users of the system.

The contribution of this paper is threefold: Firstly, *I3DEF*, an efficient feature extraction pipeline is proposed to capture spatio-temporal characteristics of the gesture videos. Secondly, this pipeline is extended to retrieval training to learn similarities between gestures within one class. The representations are put into evaluation with real-world queries and similarity of the retrieved results are assessed by volunteers. Thirdly, to improve the retrieval results, body-part segmentation is added to the feature extraction process, which improved the retrieval results.

The remainder of this paper is organized as follows: Section 2 reviews related work. Section 3 presents *I3DEF* in detail and Section 4 reports on the experiments we have conducted. Section 5 discusses the evaluation results and Section 6 concludes.

2 RELATED WORK

Video retrieval techniques have gone through many changes from pre-deep learning to date. Many methods rely on classifiers which *tag* the videos with textual labels [23, 25]. These methods vary between those labeling the existing objects in the video key frames [26], or generate an action label for a sequence of frames [4, 9, 26].

Content-based video retrieval is mainly based on feature extraction and similarity calculation. Early feature extractors were based on hand crafted features such as color histograms [11, 15, 30], Local Binary Patterns [15, 29, 39] or key-point descriptors (like SIFT [40] and SURF [2]). More recently, deep learning methods have been developed to produce feature vectors from activation of deep convolutional networks (CNN) [14, 24]. These methods are either based on vectors obtained from the activation of fully connected layers [37] or intermediate layers of CNNs [18], or by using the regional maximum activation of convolutions [32] which is based on the output of Region of Interest (ROI) pooling layers of sampled frames from a video clip [1, 28].

Methods based on learning spatio-temporal features from an entire video exploit the ability of 3D kernels of CNNs to capture the temporal dependencies of video frames, as well as the spatial information, and represent an entire shot in a single vector. C3D [33] and I3D [4] have shown superior performance in video feature extraction. Gesture videos benefit from these methods as well, since the temporal dependencies in the gesture videos are more prominent.

Representation learning is one of the important components of video retrieval. This component is responsible to train a network to favor smaller distances to similar samples. Two of the methods to learn these similarities are Siamese network [10, 13] and triplet networks [36] where the former learns the similarity between pairs, and the latter is based on similarity and dissimilarity between positive and negative pairs. Inspired by the improved performances of video retrieval with triplet networks, *I3DEF* uses triplet loss to learn the similarity between hand gestures. Hand gesture retrieval in videos is scarcely seen in literature. [38] has used low-level edge orientation features to find the best match between the query and the collection and [5] used ensemble attractor networks to retrieve single stroke 2D gestures.

3 METHOD

In this section, the main components of *I3DEF* are described. The feature extraction is generating an embedding to represent the spatio-temproal dependencies of the gesture frames. This embedding is used for the retrieval training component, which adjusts the parameters of the network to learn the similarities of embeddings.

3.1 Feature Extraction

2D kernel convolutional networks have achieved state of the art results in image related tasks in computer vision [16]. However, for videos, these networks disregard the temporal dependencies between frames in a sequence. The base model used in *I3DEF* for the extraction of features is the I3D network [4]. This network essentially has the architecture of Inception, introduced in [31], with the 2D kernels inflated to 3D ones.

To begin with feature extraction, we use the I3D network pre-trained on Imagenet [6] and Kinetics 400 [17]. Although gesture recognition is similar to action recognition in many ways, one major difference is the importance of hand motion and independence of the gestures to the scene, setting, and background. To mitigate the impact of neural network conceptual inference, which in this special task would lead to a misplaced focus, we extract the features not from the last layer of the network, but from the intermediate convolutional layers. This approach leverages the existence of local descriptions in the intermediate features to generate a compact global video representation. These features are expected to be more representative of the local features of the frames. To achieve this objective, we take the features from the output of the fourth module of the I3D network.

To generate the video descriptors, uniform sampling of video frames is applied, to have 40 frames per video. The input to the network is RGB and the optical flow of the videos, where the latter is extracted via the TVl1 [3] algorithm. The two streams of the network are fused after passing through a convolutional layer and then are fed into the fully connected layer. The probability of the classes are obtained via a softmax layer. The newly added layers after fusion are trained separately and then the whole network is fine-tuned to learn the class labels of the training set.

To adapt the output of the network to the retrieval task to have a feature vector per video, we remove the softmax layer to have a 1024 dimensional feature vector per video.

3.2 Retrieval Training

To alter the objective of the network from classification to retrieval, we need to train the network to learn the similarity between the videos. For a given query gesture video, we compute the similarity between the query and each video in the training set, and rank the results based on the maximum similarity between pairs. Ideally, our network can retrieve the most relevant videos to the query from the dataset. To learn the similarity measure, we train the already fine-tuned network with triplet-loss [27].

A collection of triplets $\tau = \{(v_i, v_i^+, v_i^-), i = 1, ..., N\}$ are drawn to start the learning process, where v_i, v_i^+, v_i^- are the feature vectors of the anchor (positive and negative samples, resp.). This triplet construction helps to encode relative similarity between videos.

To prevent the network from learning only the easy similarities, we introduce a margin m. Therefore, the relation of similarity between two pairs is defined as:

$$D(f_\theta(v_i), f_\theta(v_i^+)) - D(f_\theta(v_i), f_\theta(v_i^-)) + m < 0 \tag{1}$$

where $m = 0.5$.

The objective is to optimize the following triplet loss equation to make sure the network creates the embedding in a way that similar videos have lower embedding distance in the vector space.

$$\min_\theta \sum_{i=1}^{k} L_\theta \left(v_i, v_i^+, v_i^-\right) + \lambda \|\theta\|_2^2 \tag{2}$$

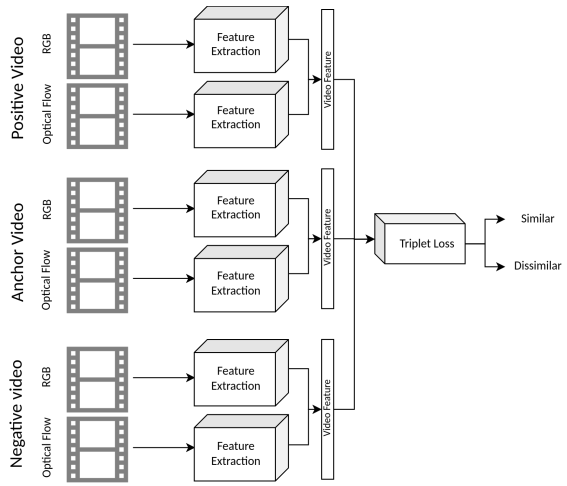

Figure 1: Illustration of the *I3DEF* pipeline with the feature extraction and Triplet loss.

where λ is a regularizer to prevent overfitting and k is the total number of triplets.

To prepare the data for training, we select the anchor class randomly, the anchor and positive samples are taken from the same class, and the negative samples are taken from other classes, except the selected classes for positive (or anchor) samples.

4 EXPERIMENTS

In this section, we explore effect of different variations of our setup and architecture to analyze the performance of the *I3DEF* architecture for gesture retrieval via the Chalearn Isolated gesture benchmark [34]. We measure the generalizability of the retrieval system to the real-world queries and measure the similarity rated by volunteer assessors.

4.1 Network Variations

To have a thorough comparison of different methods to extract features or learn the similarity, we have used three different network variations: The first model uses the feature extractor presented in Section 3.1 trained with triplet loss (*I3DEF*). The second model is using the features extracted from the proposed feature extractor trained with contrastive loss (*I3DEF-C*). The third model is using the masking module, and triplet loss to extract similarity features (*I3DEF-M*). In the following, the last two variations are explained in detail.

4.1.1 Person Masking Mechanism. One variation of our existing setup is to use person masking to mitigate the effect of the background of videos, and to improve the results by guiding the network to focus on the person in the video frames. To further improve the robustness of the system to the background variations and clutter, we use person detection as a masking module to emphasize the person in action and their body parts and reduce the effect of the background. For this purpose, we use the recently published person segmentation method [20].

Before feeding the image to the network, we used the pre-trained model of the aforementioned segmentation algorithm on COCO-2014 [21] which is fine-tuned on the PASCAL person part [8] dataset[1], to extract a feature map with weights according to the location of the person in the frame. As suggested by the authors, we use the multi-scale option to improve the accuracy of the segmentation results. Since in our dataset, the background is still in all the frames and optical flow frames are immune to noisy background, the acquired feature map from the segmentation model is multiplied by the RGB frames only. In our experiment results, this model is referred to as *I3DEF-M*.

4.1.2 Contrastive Loss. Siamese networks also present a method of similarity learning for retrieval purposes which uses a contrastive loss. We change the representation learning component of the *I3DEF* to the Siamese network by introducing similar and dissimilar pairs to the network. This variation of the network, *I3DEF-C*, is using contrastive loss [13] instead of the triplet loss to optimize the similarity distances of the pairs.

In this setup, differently from triplet loss, the batches for training consist of video pairs, which are either similar (i.e., from the same class) or dissimilar (i.e., from different classes).

4.2 Evaluation Protocols

As our retrieval pipeline *I3DEF*, in contrast to networks used for classification tasks, does not produce a label, we cannot use the usual metrics such as best prediction score or accuracy to evaluate our performance. The output of the system consists of embeddings which are used to find the similarity by computing the distance between them. We therefore generate a set of 9 query videos, 5 of which are taken from the test set of the Chalearn Isolated gesture dataset [34], and the other 4 consist of new recordings of different people performing the relevant gestures. These videos were recorded in a similar setting to the ones in the dataset but using a different camera and showing different people, in order to see if the model is able to generalize beyond the specific properties of the dataset.

Retrieval was performed using these 9 queries on each of the three network variants and the obtained videos were independently rated with respect to their similarity to the query by 10 independent assessors using a 4 point Likert scale to assign a similarity score to the retrieved results (*'very good'* = 1, *'good'* = $\frac{2}{3}$, *'ok'* = $\frac{1}{3}$, *'bad'* = 0).

The *Fleiss' Kappa* (κ) was computed for each of the assessments of the results for each of the 9 queries in order to quantify the agreement between the assessors. The resulting values for κ range from 0.09 to 0.38 with a mean of 0.26 which shows that there was some difference in how the different assessors judged the similarity of the retrieved results. This issue arises specifically in queries in which the direction or the details of articulation of a gesture considered to be different by assessors, but in the dataset they are labeled as the same. Therefore, the perceived similarity by the assessors is not necessarily in line with the similarity that would be derived from the labels of the original dataset. For the subsequent evaluation of the retrieval characteristics, we compute the median of all assigned scores.

[1] https://github.com/kevinlin311tw/CDCL-human-part-segmentation

Table 1: Maximum, mean and median dcg and precision at 5, 10 and 20 for the three presented network variants.

model		I3DEF-M	I3DEF	I3DEF-C
dcg	max	4.25	2.69	**5.08**
	mean	**1.87**	1.25	1.62
	median	**2.38**	1.32	1.18
p@5	max	**0.8**	**0.8**	**0.8**
	mean	**0.33**	0.24	0.24
	median	**0.4**	0.2	0.2
p@10	max	**0.5**	0.4	**0.7**
	mean	**0.25**	0.16	0.15
	median	0.3	0.2	0.1
p@20	max	0.6	0.3	**0.7**
	mean	**0.23**	0.13	0.16
	median	**0.25**	0.1	0.15

To evaluate the retrieval performance of the different network variants, in addition the *precision* for several result sets, we compute the *Discounted Cumulative Gain* as defined in Equation 3, where s is the list of scores corresponding to the retrieved results, using the aggregated scores as assigned by the assessors. A result is considered to be relevant if its median score is $\geq \frac{2}{3}$.

$$dcg(s) = \sum_{i=1}^{N} \frac{2^{s_i} - 1}{log_2(i+1)} \tag{3}$$

4.3 Results

Table 1 shows the maximum, mean and median discounted cumulative gain as well as the precision for the top 5, 10 and 20 results per network variant, with the best results highlighted in boldface. It can be seen that while the *I3DEF-C* model has the highest single value for each measure, the *I3DEF-M* model consistently outperforms the other two for all the mean and median aggregated measures.

Table 2 shows the mean of all the measures per model, but separates the queries which were taken from the test set of [34], designated as *'test'* in the table and the newly recorded ones, labelled *'new'*. The best result per row is again printed in boldface. It can be seen that there are large performance differences between the two query sources for the *I3DEF* and *I3DEF-C* models, while the differences for the *I3DEF-M* model are considerably smaller. While the I3DEF-M model is slightly outperformed in the *dcg* as well as the *p@10* and *p@20* metrics when only considering the *'test'* queries, it clearly dominates in all categories when only the *'new'* queries are considered.

5 DISCUSSION

From the results presented in Tables 1 and 2, it can be seen that the masking filter used in the *I3DEF-M* model variant increases the retrieval performance substantially, especially for the queries which were not part of the original dataset the model was trained on, since it causes the model to ignore the background which leads to a better generalization with respect to the setting of the scene in which the gesture is performed. Conversely, the results suggest that the models without the masking module rely too much on visual information within the video which is entirely independent of the gesture, despite the effort of the designers of [34] to counter this.

Table 2: Mean dcg and precision at 5, 10 and 20 for the three presented network variants with respect to if the queries were from the test set of [34] (*test*) or newly recorded (*new*).

model		I3DEF-M	I3DEF	I3DEF-C
dcg	test	1.85	1.75	**2.38**
	new	**1.89**	0.61	0.67
p@5	test	**0.36**	**0.36**	**0.36**
	new	**0.3**	0.1	0.1
p@10	test	0.22	**0.24**	**0.24**
	new	**0.3**	0.08	0.05
p@20	test	0.21	0.16	**0.25**
	new	**0.25**	0.09	0.06

Our analysis on the basis of the Chalearn ISO dataset shows an imbalance in the number of samples per class. Such a large variation in the number of samples per class in the training set biases the feature extractor to learn more of a certain class [19]. To alleviate this problem, it is recommended to use additional datasets, which exhibit a large number of classes.

As briefly discussed earlier, the masking used during feature extraction reduces the sensitivity of the network to the background and increases the independence of the features with respect to the environment and causes them to focus more on the person and articulations of the hands. However, the network is not designed to consider multi-scale figures. This artifact can be avoided by using a multi-scale feature extractor which makes the features tolerant towards differences in scaling and a person's distance to the camera.

One of the most observed phenomena during the evaluation of the retrieval results by the assessors is the unclear boundary between dissimilar gestures and similar gestures. According to general belief, filliped trajectories of the gestures are considered to have different meaning and therefore are assessed as dissimilar. However, such gestures have the same label in the dataset. Moreover, the gesture articulation varies from person to person, and sometimes, these differences, make similar gestures, look dissimilar, even though they are from the same category.

6 CONCLUSION

In this paper, we explore methods for the identification and retrieval of gestures using RGB frames and optical flow which is extracted from the RGB data without the additional need for specialized recording equipment such as depth-cameras. We proposed the extraction of features from intermediate layers of an I3D network. We comparatively evaluate three variants of the feature extraction network based on manual assessment of the retrieved results. The performed experiments show that the combination of a triplet loss with masking module focusing on the human body performs best among the tested variants. They also indicate that the body-part-masking can increase the generalization performance of a network, making it more applicable to different datasets showing humans in different settings. More work in this area and especially larger, more diverse and balanced datasets are needed in order to advance the possibilities for gesture retrieval in videos in the future.

REFERENCES

[1] Lorenzo Baraldi, Matthijs Douze, Rita Cucchiara, and Hervé Jégou. 2018. LAMV: Learning to align and match videos with kernelized temporal layers. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7804–7813.

[2] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. 2006. Surf: Speeded up robust features. In *European conference on computer vision*. Springer, 404–417.

[3] Thomas Brox, Andrés Bruhn, Nils Papenberg, and Joachim Weickert. 2004. High accuracy optical flow estimation based on a theory for warping. In *European conference on computer vision*. Springer, 25–36.

[4] Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6299–6308.

[5] Carlos Dávila, Mario González, Jorge-Luis Pérez-Medina, David Dominguez, Ángel Sánchez, and Francisco B. Rodriguez. 2019. Ensemble of Attractor Networks for 2D Gesture Retrieval. In *Advances in Computational Intelligence*, Ignacio Rojas, Gonzalo Joya, and Andreu Catala (Eds.). Springer International Publishing, Cham, 488–499.

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

[7] Yajiao Dong and Jianguo Li. 2018. Video retrieval based on deep convolutional neural network. In *Proceedings of the 3rd International Conference on Multimedia Systems and Signal Processing*. 12–16.

[8] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision* 88, 2 (2010), 303–338.

[9] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. 2016. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1933–1941.

[10] Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, Vol. 2. IEEE, 1735–1742.

[11] Yanbin Hao, Tingting Mu, Richang Hong, Meng Wang, Ning An, and John Y Goulermas. 2016. Stochastic multiview hashing for large-scale near-duplicate video retrieval. *IEEE Transactions on Multimedia* 19, 1 (2016), 1–14.

[12] Elad Hoffer and Nir Ailon. 2015. Deep metric learning using triplet network. In *International Workshop on Similarity-Based Pattern Recognition*. Springer, 84–92.

[13] Chong Huang, Qiong Liu, Yan-Ying Chen, et al. 2017. Local Feature Descriptor Learning with Adaptive Siamese Network. *arXiv preprint arXiv:1706.05358* (2017).

[14] Yu-Gang Jiang and Jiajun Wang. 2016. Partial copy detection in videos: A benchmark and an evaluation of popular methods. *IEEE Transactions on Big Data* 2, 1 (2016), 32–42.

[15] Weizhen Jing, Xiushan Nie, Chaoran Cui, Xiaoming Xi, Gongping Yang, and Yilong Yin. 2019. Global-view hashing: harnessing global relations in near-duplicate video retrieval. *World wide web* 22, 2 (2019), 771–789.

[16] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. 2014. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 1725–1732.

[17] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950* (2017).

[18] Giorgos Kordopatis-Zilos, Symeon Papadopoulos, Ioannis Patras, and Yiannis Kompatsiaris. 2017. Near-duplicate video retrieval by aggregating intermediate CNN layers. In *International conference on multimedia modeling*. Springer, 251–263.

[19] Bartosz Krawczyk. 2016. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence* 5, 4 (01 Nov 2016), 221–232. https://doi.org/10.1007/s13748-016-0094-0

[20] Kevin Lin, Lijuan Wang, Kun Luo, Yinpeng Chen, Zicheng Liu, and Ming-Ting Sun. 2019. Cross-Domain Complementary Learning with Synthetic Data for Multi-Person Part Segmentation. *arXiv preprint arXiv:1907.05193* (2019).

[21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.

[22] C. Ma, J. Huang, X. Yang, and M. Yang. 2019. Robust Visual Tracking via Hierarchical Convolutional Features. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 11 (Nov 2019), 2709–2723. https://doi.org/10.1109/TPAMI.2018.2865311

[23] Foteini Markatopoulou, Damianos Galanopoulos, Vasileios Mezaris, and Ioannis Patras. 2017. Query and keyframe representations for ad-hoc video search. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*. 407–411.

[24] Luca Rossetto, Ivan Giangreco, Ralph Gasser, and Heiko Schuldt. 2018. Competitive video retrieval with vitrivr. In *International Conference on Multimedia Modeling*. Springer, 403–406.

[25] Luca Rossetto, Ivan Giangreco, Claudiu Tanase, and Heiko Schuldt. 2016. vitrivr: A flexible retrieval stack supporting multiple query modes for searching in multimedia collections. In *Proceedings of the 24th ACM international conference on Multimedia*. 1183–1186.

[26] Luca Rossetto, Mahnaz Amiri Parian, Ralph Gasser, Ivan Giangreco, Silvan Heller, and Heiko Schuldt. 2019. Deep learning-based concept detection in vitrivr. In *International Conference on Multimedia Modeling*. Springer, 616–621.

[27] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 815–823.

[28] Omar Seddati, Stéphane Dupont, Saïd Mahmoudi, and Mahnaz Parian. 2017. Towards good practices for image retrieval based on CNN features. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 1246–1255.

[29] Lifeng Shang, Linjun Yang, Fei Wang, Kwok-Ping Chan, and Xian-Sheng Hua. 2010. Real-time large scale near-duplicate web video retrieval. In *Proceedings of the 18th ACM international conference on Multimedia*. 531–540.

[30] Jingkuan Song, Yi Yang, Zi Huang, Heng Tao Shen, and Richang Hong. 2011. Multiple feature hashing for real-time large scale near-duplicate video retrieval. In *Proceedings of the 19th ACM international conference on Multimedia*. 423–432.

[31] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1–9.

[32] Giorgos Tolias, Ronan Sicre, and Hervé Jégou. 2015. Particular object retrieval with integral max-pooling of CNN activations. *arXiv preprint arXiv:1511.05879* (2015).

[33] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*. 4489–4497.

[34] Jun Wan, Yibing Zhao, Shuai Zhou, Isabelle Guyon, Sergio Escalera, and Stan Z Li. 2016. Chalearn looking at people rgb-d isolated and continuous datasets for gesture recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 56–64.

[35] Jindong Wang, Yiqiang Chen, Shuji Hao, Xiaohui Peng, and Lisha Hu. 2019. Deep learning for sensor-based activity recognition: A survey. *Pattern Recognition Letters* 119 (2019), 3–11.

[36] Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. 2014. Learning fine-grained image similarity with deep ranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1386–1393.

[37] Gengshen Wu, Jungong Han, Yuchen Guo, Li Liu, Guiguang Ding, Qiang Ni, and Ling Shao. 2018. Unsupervised deep video hashing via balanced code for large-scale video retrieval. *IEEE Transactions on Image Processing* 28, 4 (2018), 1993–2007.

[38] Shahrouz Yousefi and Haibo Li. 2015. 3D Hand Gesture Analysis through a Real-Time Gesture Search Engine. *International Journal of Advanced Robotic Systems* 12, 6 (2015), 67. https://doi.org/10.5772/60045

[39] Guoying Zhao and Matti Pietikainen. 2007. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE transactions on pattern analysis and machine intelligence* 29, 6 (2007), 915–928.

[40] Wan-Lei Zhao and Chong-Wah Ngo. 2009. Scale-rotation invariant pattern entropy for keypoint-based near-duplicate detection. *IEEE Transactions on Image Processing* 18, 2 (2009), 412–423.