Less Is More: Similarity Models for Content-Based Video Retrieval

Patrik Veselý and Ladislav Peška(✉)

Faculty of Mathematics and Physics, Charles University, Prague, Czechia
ladislav.peska@matfyz.cuni.cz

Abstract. The concept of object-to-object similarity plays a crucial role in interactive content-based video retrieval tools. Similarity (or distance) models are core components of several retrieval concepts, e.g. Query by Example or relevance feedback. In these scenarios, the common approach is to apply some feature extractor that transforms the object to a vector of features, i.e., positions it into an induced latent space. The similarity is then based on some distance metric in this space.

Historically, feature extractors were mostly based on some color histograms or hand-crafted descriptors such as SIFT, but nowadays state-of-the-art tools mostly rely on some deep learning (DL) approaches. However, so far there were no systematic study of how suitable are individual feature extractors in the video retrieval domain. Or, in other words, to what extent are human-perceived and model-based similarities concordant. To fill this gap, we conducted a user study with over 4000 similarity judgements comparing over 20 variants of feature extractors. Results corroborate the dominance of deep learning approaches, but surprisingly favor smaller and simpler DL models instead of larger ones.

Keywords: Content-based video retrieval · Similarity models · User study

1 Introduction

In the 21st century, digital video data started to be produced at an unprecedented quantity. Even individual users may produce tens hours of home videos every year, while hundreds hours of video content is being uploaded to YouTube every minute[1]. Nonetheless, while the volume of produced content increases at tremendous speed, the challenge of effective and efficient search and retrieval of this content remains open. Commercially available search engines still mainly focus on metadata-based search, but current research gradually shift towards multimedia content understanding, innovative retrieval models and GUIs.

This is well illustrated e.g. on the recent editions of Video Browser Showdown competitions [7] as well as prototype tools for content-based video retrieval, e.g. [8,12,16]. While the main task of these tools is to search within the video content, this is often simplified by representing videos as sequences of keyframes and

[1] https://www.oberlo.com/blog/youtube-statistics.

D.-T. Dang-Nguyen et al. (Eds.): MMM 2023, LNCS 13834, pp. 54–65, 2023.
https://doi.org/10.1007/978-3-031-27818-1_5

therefore reduced to an image search with some additional information (e.g., audio transcription or the notion of sequentiality). Some of the main utilized retrieval concepts are text search via text-to-image joint embeddings [14,21], using multi-queries with some temporal proximity conditions [8,12,16], various Query-by-Example (QBE) models [8,16], or models incorporating iterative relevance feedback of users [12]. Especially the latter two approaches are based on the concept of object-to-object similarity, which is the main focus of this paper.

Early approaches on similarity modeling utilized handcrafted features such as color histograms [18], texture descriptors [10] or more complex sets of low-level semantic descriptors [11]. In contrast, the vast majority of state-of-the-art solutions use some variant of pre-trained deep networks, mostly convolutional neural networks (CNN) or transformer-based architectures. Deep learning (DL) approaches provided tremendous improvements in many computer vision related tasks, such as image classification, object detection, video segmentation and so on. However, the problem of suitability of (pre-trained) DL approaches to serve as feature extractors for some similarity models was largely neglected so far. More specifically, we are not aware of any systematic evaluation of human-perceived and machine-induced similarity models in the context of video retrieval. We assume that part of the reason is the lack of suitable datasets for the video retrieval domain and its rather costly acquisition due to the necessity to collect human similarity judgements.

We consider this to be a substantial gap, which may hinder the future development of effective video retrieval systems. The problem is not only to evaluate which feature extractors are more suitable for similarity-based retrieval tasks, but also to estimate how much confidence we may have in the conformity of the similarity models. In general, our work focus on the duality between data semantics and its visual similarity [23], where the duality of human- and machine-perceived similarity can be considered as an instance of this generic problem. In particular, the main contributions of this paper are as follows:

- Evaluating the level of concordance with human similarity judgements for over 20 feature extractors including both recent DL and shallow techniques.
- Identifying several contextual features affecting expected levels of agreement.
- Providing a dataset of human similarity judgements for future usage.

2 Related Work

We are not aware of any directly related work from the video retrieval domain. However, there are some works aiming on the similarity perception in the image domain [6,13,20,22], which is quite close to our research area.

Peterson et al. [20] focused on 6 categories of images (e.g., animals, automobiles, furniture) and human judgement was obtained for within-category pairs on a 10-point rating scale (from "not similar at all" to "very similar"). Authors then compared human similarity judgements with the similarity induced by several CNN networks (AlexNet, VGG, GoogLeNet and ResNet) pre-trained on

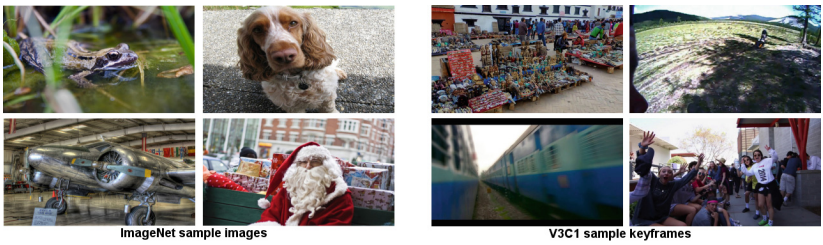

Fig. 1. Samples from ImageNet dataset (left) and V3C1 dataset (right).

the ImageNet dataset [3]. Results indicate varying performance among different image categories, but ResNet architecture [5] achieved the best performance on raw representations (i.e., without any fine-tuning for the particular task).

Roads and Love [22] published the Human Similarity Judgements extension to the ImageNet dataset [3]. In this case, users were asked to select first and second most similar images (out of 8 available) to the query image. Similarity models were subsequently compared w.r.t. triplet accuracy, i.e. whether the candidate image A or B is more similar to the query item Q. Relatively large pool of deep learning architectures were compared including e.g., VGG, ResNet, DenseNet and Inception. Again, ResNet model [5] with 50 layers provided the best triplet accuracy.

In contrast to both related works, we added several "up-to-date" network architectures which emerged recently. Similarly as in [22], we adopted a query-response style of similarity judgements and we also utilized triplet accuracy as our main evaluation metric. However, instead of "select two from eight" scenario, we asked users to only compare the similarity of two candidate items, directly following the triplet scenario. On one hand, this reduces the volume of triplets generated from a single user feedback, but it also minimizes the effects of positional biases [9] and contextual interference of other displayed items. Also, this decision allows us to easily tune the expected difficulty of the task through estimated similarity levels between all three images (see Sect. 3.3). More importantly, there are fundamental differences between underlying datasets. While ImageNet images mostly depict one clearly identifiable object, video keyframes are much more heterogeneous. There can be multiple "main" objects visible, some of them blurry or fuzzy due to the camera motion etc. (see Fig. 1 for examples). Therefore, it is not clear whether the performance on ImageNet or similar image datasets transfer well into the video domain.

It is also worth noting that instead of using pre-trained feature extractors, one can aim to learn feature representations directly from similarity judgements. However, such approaches are extremely demanding on the volume of available data. This is problem in ours as well as many other domains, so the approaches such as [6] (as well as active learning and transformed representations mentioned in [22] and [20] resp.) are not plausible in our scenario.

The volume of similarity judgements we collected is lower than in the related works, but the dataset is large enough to identify the most suitable feature extractors and also to suggest several contextual features that may help to estimate the level of agreement between human-perceived and machine-induced similarity.

3 User Study

3.1 Dataset and Pre-processing

Usually, image datasets consist of images with a focus on the main object or a few main objects. The number of main object types is often limited and even some of the biggest ones such as ImageNet [3] have a limited number of classes. Image similarity is a subjective skill of humans to assess which images are more or less similar. This skill can be used not only on nicely arranged images but in general on any image - even the ones where we can't properly describe the overall scenery. We aim to investigate if we can mimic the said human skill with computer vision and deep learning methods in general. To achieve this, we used the first part of the Vimeo Creative Commons Collection dataset (V3C1) [1]. This dataset contains 7 475 videos of various lengths and topics. We specifically utilized the provided keyframes for each shot, resulting into 1 082 659 images in total. Additional cleaning process was applied to remove single color images and other trivial artefacts (e.g. extremely blurry images). After the cleaning, 1 010 398 keyframes remained in the dataset.

3.2 Similarity Models

In this study, we utilized four types of feature extractors (Color-based, SIFT-based, CNN-based and Transformer-based) to acquire feature vectors. For each of these vectors, we used cosine distances to estimate image similarities.

Color-Based Extractors. Simple color-based extractors were commonly used in the pre-deep learning era for content-based image retrieval. Due to their usual simplicity, it worked well in some domains [18] and thanks to the hand-crafted nature these methods usually do not need any training data. We utilized three variants of color-based feature extractors denoted as *RGB Histogram*, *LAB Clustered* and *LAB Positional*.

RGB Histogram computes a pixel-wise histograms of all three color channels (64 and 256 bins per histogram were evaluated). Then the three histograms are concatenated and L_2-normalized. The downside of RGB color space is that it is not uniformly distributed w.r.t. human-perceived color similarity. Thus we also focused on LAB color space [19] that is designed to cope with this non-uniformity. Specifically, *LAB Clustered* extractor computes K-means w.r.t. LAB representation of all pixels. The centroids from K-means are then taken as representatives for the image. Representative colors are sorted w.r.t. hue compound of HSV color space and concatenated to form a final feature vector. The variant

with $K = 4$ clusters was evaluated. Finally, *LAB Positional* extractor exploits individual regions of the image. The image is divided into several chessboard-like regions, which are represented via its mean color. We evaluated region grids of 2×2, 4×4 and 8×8.

SIFT-Based Extractors. People usually do not use only the color when assessing the similarity - shapes and textures can also play a significant role. Such image attributes can be captured by Scale-invariant feature transform (SIFT) [17] keypoints. However, the number of keypoints may vary among images, so additional post-processing is necessary to get a global feature descriptor. We utilize the vector of locally aggregated descriptors (*VLAD*) as proposed in [11]. VLAD is obtained by computing the bag-of-features dictionary for SIFT keypoints and then aggregating the keypoints as residuals from the nearest features from the dictionary (the variant with a dictionary of 64 features was evaluated).

CNN-Based Extractors. In the recent years, convolutional neural networks such as [5,25] have been widely used as state-of-the-art in many fields of computer vision and video/image retrieval. These deep networks need a lot of training data to learn and such large datasets are currently not available for general video keyframes similarity as well as many other domains. Therefore, transfer learning [24] approach is often used. The method uses data from another domain to train a deep network and then use this pre-trained (optionally fine-tuned) network on the target domain. In video retrieval, the fine-tuning step is often skipped and keyframe's features are represented as activations of the last connected layer.

We utilized the following architectures: *ResNetV2* [5] with 50, 101, and 152 layers, *EfficientNet* [25] and its variants B0, B2, B4, B6, and B7 (indicating the size of the network) and *W2VV++* [14]. Both *ResNetV2* and *EfficientNet* were trained on image classification task using the ImageNet [3] dataset. In contrast, *W2VV++* was trained on MSR-VTT [27] and TGIF [15] datasets and its task was to project both images and textual description to a joint vector space.

Transformer-Based Extractors. A novel Transformer architecture [26] was originally proposed for natural language processing, but it was soon adopted in many other domains. The Vision transformers were first applied in [4] and quickly improve over state-of-the-art in many computer vision tasks. We utilized three variants of transformers architecture: Vision transformer (*ViT*) [4], *CLIP* [21] and *ImageGPT* [2]. All Transformer models were taken from the HuggingFace library[2]. Similarly as for CNNs, we used activations of the last connected layer as a feature vector. *ViT* was trained on the ImageNet dataset and we evaluated two sizes: base and large. For the *CLIP* model we used two patch size variants: 16 and 32. Similarly as *W2VV++*, CLIP also aim to create a shared latent space for text and images. Finally, for *ImageGPT* was also trained on the ImageNet dataset and we utilized its small and medium variants. In this case - as per author's suggestions - we used network's middle layers to create feature vectors.

[2] https://huggingface.co/.

3.3 Evaluation Procedure

The user study was conducted during June–August, 2022 via a dedicated website[3]. Participants were recruited via internal channels, mostly from the pool of reliable participants of previous studies conducted by the authors. Participants' task was first to (optionally) provide some personal details (age, education, knowledge of DL) and then they were forwarded to the study itself. At each step, participants were presented with one query keyframe Q and two candidate keyframes A and B. The task was to decide, which of the candidates is more similar to the query[4]. Participants simply clicked on the desired image and confirmed the selection. The volume of iterations was not limited (i.e., participants could continue with the study as long as they wanted). In the initial briefing, users were asked to provide at least 20–30 judgements, but most participants actually provided more feedback (mean: 116, median: 77 judgements per user).

For each iteration of the study, individual tasks were selected at random from the dataset of 66 000 pre-generated triplets. This dataset was selected as follows: first, 100 query items were selected at random. Then, for each feature extractor, we sorted all 1M keyframes from the most to least similar (w.r.t. cosine and Euclidean distances) and clustered them into 5 bins w.r.t. their rank (break points at $[2^4, 2^8, 2^{12}, 2^{16}, 2^{20}]$). Then, for each bin, we selected one candidate from that bin at random and iterated over all same-or-more distant bins, selecting the second candidate from each of them. This produced 30 triplets per extractor and query item. The idea behind this procedure is to include both similar and distant examples w.r.t. all extractors. Naturally, for all judged triplets we evaluated distances and ranks w.r.t. all extractors.

Overall, we collected 4394 similarity judgements from 38 unique participants. The pool of participants leaned a bit towards higher-educated and higher DL knowledge as compared to the general population. Most of participants were 20–35 years old. The sample could be a bit better balanced, which we plan to address in our follow-up work. However, note that all observed age, education and DL knowledge groups were present in the pool of participants.

For all extractors and all triplets with human judgement, we evaluated which of the candidate items (A or B) is closer to the query keyframe Q. Then, we compared this with the actual human judgement. We report on triplet accuracy (*acc*), i.e., the ratio of concordant judgements in the dataset.

4 Results

Figure 2 depicts overall triplet loss results for individual extractors. All extractors performed better than random guessing (i.e., $acc > 0.5$), but performance differences among extractors were rather substantial. Notably, there was a clear gap between shallow (color and SIFT-based) approaches, which were inferior to all deep-learning models. Out of DL models, *ImageGPT* (both versions) was

[3] https://otrok.ms.mff.cuni.cz:8030/user.

[4] The exact prompt was "Which image is more similar to the one on the top?".

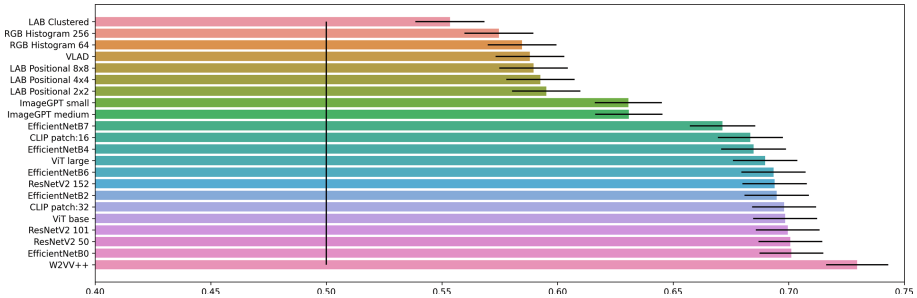

Fig. 2. Triplet accuracy results w.r.t. individual feature extractors. X-axis denote triplet accuracy, black bars denote 95% confidence intervals and horizontal line indicate the accuracy of random guessing.

clearly inferior, while there was another larger gap between the performance of the best extractor (*W2VV++*) and the second best (*EfficientNetB0*).[5]

We also observed an interesting fact that smaller variants of the same architecture almost consistently outperformed the larger variants. To verify this, we compared results of two groups of models including largest and smallest variant of given architectures respectively.[6] Overall, smaller models significantly outperformed the larger ones with $p = 0.016$ according to Fisher exact test. Similarly, transformer-based architectures (*CLIP*, *ViT*, *ImageGPT*) were significantly dominated by traditional CNN architectures (*ResNetV2*, *EfficientNet*, *W2VV++*): $p = 2.2e - 11$ w.r.t. Fisher exact test.

Next we focused on how similarity estimations of individual extractors correlate with each other and whether this can be utilized in some form of an ensemble model. Figure 3 denote the ratio of concordant judgements for all pairs of extractors. Different size variants of the same architecture were all highly correlated (0.7–0.95) as well as all DL approaches (0.65–0.85) except for *ImageGPT*. These were, quite surprisingly, more correlated to *RGB Histogram* models than to other DL approaches. Overall, there were no prevalently discordant pairs, but average agreement of pairs not specifically mentioned here was quite small (0.5–0.65).

To evaluate the possibility of creating ensemble models, we focused on how the level of agreement between extractors affect the triplet accuracy scores. Figure 4 depicts how triplet accuracy changes when certain volume of other extractors assume the same ordering of candidates as the current one. For the sake of space we only depict results for the best and the worst extractor. In both cases, the accuracy improves almost linearly with the volume of concordant

[5] All mentioned differences were stat. sign. with $p < 0.05$ w.r.t. Fisher exact test.

[6] The first group included *RGB Histogram 256*, *LAB Positional 8x8*, *ImageGPT medium*, *EfficientNetB7*, *ViT large* and *ResNetV2 152*. The second group included *RGB Histogram 64*, *LAB Positional 2x2*, *ImageGPT small*, *EfficientNetB0*, *ViT base* and *ResNetV2 50*.

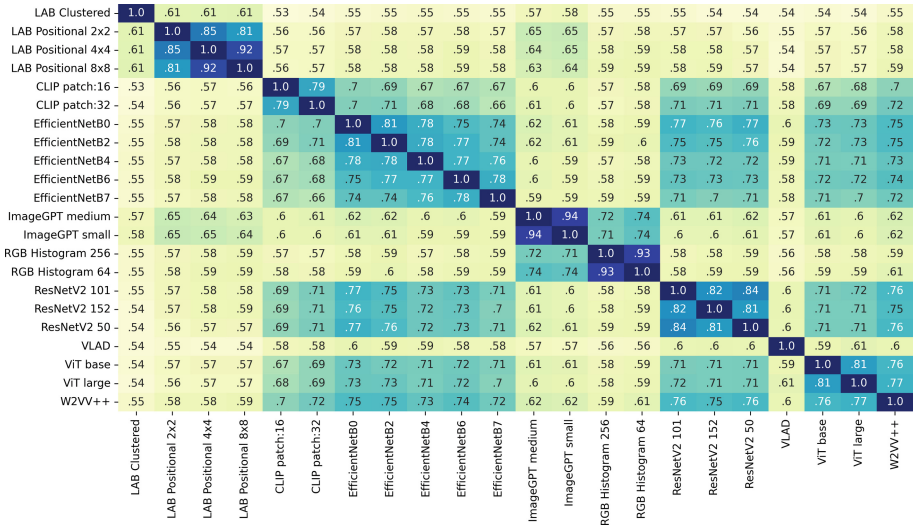

Fig. 3. Level of agreement between individual feature extractors.

extractors. For *W2VV++*, however, the difference is smaller overall and also the situation where only a handful of extractors agree with *W2VV++* is quite rare (as the larger confidence intervals indicate). We also experimented with a variant where only the 6 best-performing extractors are considered. In this case, if none or just one additional extractor agrees with *W2VV++* (in total 10% of cases), the model would perform better with inverted predictions. Therefore, we can conclude that there is some space for improvements via ensemble-based solutions in the future.

Next, we focused on how does the distances between images affect the results. In principle, three variables (and their interplay) should be considered: distance from query image to both candidates and distance between both candidates. In either case, we can focus on the raw values, or some derived statistics, e.g.

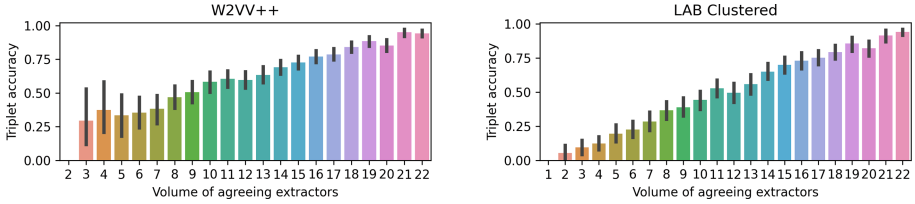

Fig. 4. Dependence of triplet accuracy on the volume of other feature extractors agreeing with the current one. Left: *W2VV++* (best-performing extractor), right: *LAB Clustered* (worst-performing). Note that for *W2VV++*, there were only very few cases where the volume of agreeing extractors was < 7.

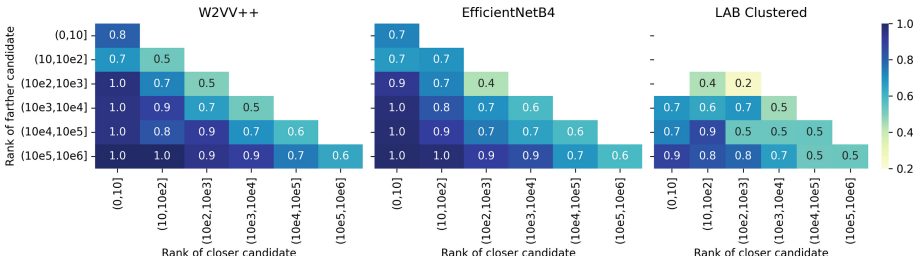

Fig. 5. The effect of distance-based ranking of both candidate items on triplet accuracy. Left: *W2VV++* (best-performing extractor), center: *EfficientNetB4* (average-performing), right: *LAB Clustered* (worst-performing). Note that only bins with 10+ items are depicted.

distance-based ranking. Figure 5 depict how the triplet accuracy depends on the distance-based ranking of both candidate items. We can see a clear decrease of performance towards the diagonal (i.e., where both candidates are approx. equally distant from the query image). In these cases, even the similarity estimation of the best extractors are close to the random guess. For all extractors, accuracy improves with increasing difference in ranks of both candidates (i.e. towards bottom-left corner). Specifically, if one of the candidates is very close or very distant (first column/last row), even average extractors are almost 100% concordant with the human judgement. We further performed a Gini gain analysis for the best extractor. Seemingly, the most informative single variable is the difference in distances of both options to the query. Additional gain can be achieved by also adding the distance between both candidates, but the improvements were small compared to the previous case. For the sake of completeness, Fig. 6 (left) depicts dependence of accuracy on the relative difference in candidate's distances (normalized by the distance to the farther candidate), where a similar pattern can be observed for all extractors.

Finally, we observed whether some features derived from the user's behavior can suggest the level of concordance between similarity perception. As Fig. 6 (right) indicate, if users needed shorter *time to decide* about the similarity, their decision was more concordant with the similarity based on feature extractors. Supposedly, this would mostly indicate simpler cases, but using *time to decide* as additional factor along with the difference in query-candidate distances and the distance between both candidates, further Gini gain was achieved. As such, user's feedback features can provide valuable additional information.

Note that it is not realistic to assume that either query-candidate distances or user's time to decision can be manipulated on purpose to achieve better concordance of similarity perception. Instead, this information can be transformed to the level of "trust" we have in our model of similarity perception. The level of trust can e.g. affect the strictness of update steps in relevance feedback, or introduce certain level of randomness in QBE approach.

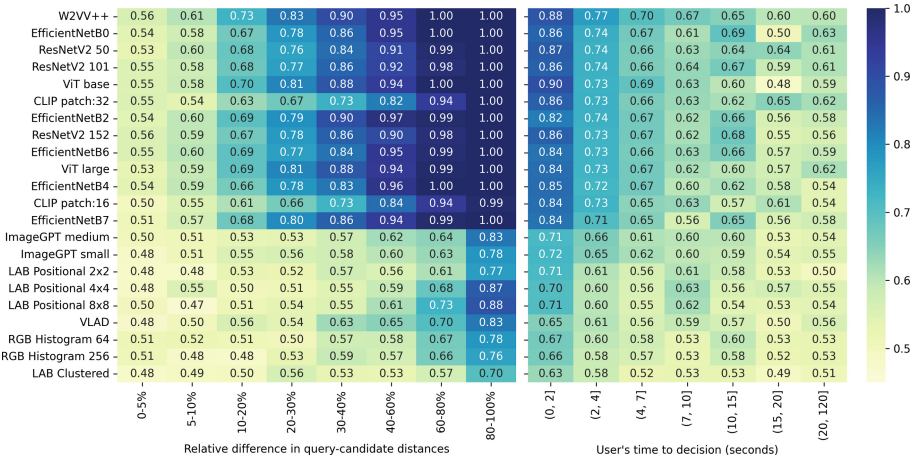

Fig. 6. The dependence of triplet accuracy on the relative differences in query-candidate distances (left) and user's time to decision (right). Extractors are sorted w.r.t. their overall performance.

5 Discussion and Conclusions

In this paper we focused on evaluating the agreement between human-perceived and machine-induced models of similarity for video retrieval. Specifically, we collected a dataset of over 4000 human similarity judgements and compared it with similarity judgements based on 22 variants of feature extractors. The best performing extractor was W2VV++ [14] and in general DL approaches outperformed all shallow models by a large margin. Surprisingly, smaller DL architectures and traditional CNN-based architectures outperformed larger models and those based on transformers respectively. We also identified several variables that can help to estimate reliability of our similarity models, e.g., difference in query-candidate distances, user's time to decision or the level of agreement among extractors.

Nonetheless, this initial study has several limitations. First, participants were forced to make decision even when they were unsure about it. In the future we plan to enhance the study design to allow users to provide graded feedback and therefore reflect the level of uncertainty in their decisions. Related to this, we did not yet focus on an inherent level of noise in human judgements, i.e., what is the level of agreement if multiple users evaluate the same triplets. Both of these enhancements could provide additional context on differences in similarity perception, i.e., are there cases where humans are highly certain and unanimous, yet discordant with machine-induced similarity judgements? Furthermore, so far we did not consider the context of evaluated keyframes. In theory, one extractor may provide well-aligned similarity for, e.g., landscape images, while fail in the similarity of people. Such evaluation would, however, require larger data samples than currently collected. Therefore, we plan to conduct a larger follow-up study

focusing on uncertainty in human judgements as well as various sub-spaces of the dataset.

In a different line of the future work, the study showed some potential for ensemble approaches as well as variables allowing to estimate reliability of the similarity models. These findings should be exploited to improve both used similarity models as well as their application in retrieval tasks. We do not assume that larger quantities of similarity judgement data would be readily available for realistic content-based video retrieval tasks. However, another line of the future work may focus on how to collect such data from the causal usage of video retrieval systems and e.g. propose on-line improvements of similarity judgements through reinforcement learning.

Finally, in a related line of our research, we focus on the problem of constructing artificial visualizations for originally non-visual data for the purpose of similarity-based retrieval [23]. We plan to utilize the findings on human- vs. machine- based similarity compliance to derive an automated visualization quality assessment, which would allow to test a wider range of visualization approaches off-line, without the need for expensive user studies.

Acknowledgments. This paper has been supported by Czech Science Foundation (GAČR) project 22-21696S and Charles University grant SVV-260588. Computational resources were supplied by the project "e-Infrastruktura CZ" (e-INFRA CZ LM2018140) supported by the Ministry of Education, Youth and Sports of the Czech Republic. Source codes and raw data are available from https://github.com/Anophel/image_similarity_study.

References

1. Berns, F., Rossetto, L., Schoeffmann, K., Beecks, C., Awad, G.: V3C1 dataset: an evaluation of content characteristics. In: ICMR 2019, pp. 334–338. ACM (2019)
2. Chen, M., et al.: Generative pretraining from pixels. In: ICML 2020. PMLR (2020)
3. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: CVPR 2009, pp. 248–255. IEEE (2009)
4. Dosovitskiy, A., et al.: An image is worth 16 × 16 words: transformers for image recognition at scale. arXiv (2020)
5. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9908, pp. 630–645. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46493-0_38
6. Hebart, M.N., Zheng, C.Y., Pereira, F., Baker, C.I.: Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. Nat. Hum. Behav. **4**(11), 1173–1185 (2020)
7. Heller, S., Gsteiger, V., Bailer, W., et al.: Interactive video retrieval evaluation at a distance: comparing sixteen interactive video search systems in a remote setting at the 10th video browser showdown. Int. J. Multimed. Inf. Retr. **11**(1), 1–18 (2022). https://doi.org/10.1007/s13735-021-00225-2
8. Hezel, N., Schall, K., Jung, K., Barthel, K.U.: Efficient search and browsing of large-scale video collections with vibro. In: Þór Jónsson, B., et al. (eds.) MMM 2022. LNCS, vol. 13142, pp. 487–492. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-98355-0_43

9. Hofmann, K., Schuth, A., Bellogín, A., de Rijke, M.: Effects of position bias on click-based recommender evaluation. In: de Rijke, M., et al. (eds.) ECIR 2014. LNCS, vol. 8416, pp. 624–630. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-06028-6_67
10. Huang, P., Dai, S.: Image retrieval by texture similarity. Pattern Recogn. **36**(3), 665–679 (2003)
11. Jégou, H., Douze, M., Schmid, C., Pérez, P.: Aggregating local descriptors into a compact image representation. In: CVPR 2010, pp. 3304–3311. IEEE (2010)
12. Kratochvíl, M., Veselý, P., Mejzlík, F., Lokoč, J.: SOM-hunter: video browsing with relevance-to-SOM feedback loop. In: Ro, Y.M., et al. (eds.) MMM 2020. LNCS, vol. 11962, pp. 790–795. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-37734-2_71
13. Křenková, M., Mic, V., Zezula, P.: Similarity search with the distance density model. In: Skopal, T., Falchi, F., Lokoč, J., Sapino, M.L., Bartolini, I., Patella, M. (eds.) SISAP 2022. LNCS, vol. 13590, pp. 118–132. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-17849-8_10
14. Li, X., Xu, C., Yang, G., Chen, Z., Dong, J.: W2VV++ fully deep learning for ad-hoc video search. In: ACM MM 2019, pp. 1786–1794 (2019)
15. Li, Y., et al.: TGIF: a new dataset and benchmark on animated GIF description. In: CVPR 2016, pp. 4641–4650 (2016)
16. Lokoč, J., Mejzlík, F., Souček, T., Dokoupil, P., Peška, L.: Video search with context-aware ranker and relevance feedback. In: Þór Jónsson, B., et al. (eds.) MMM 2022. LNCS, vol. 13142, pp. 505–510. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-98355-0_46
17. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004). https://doi.org/10.1023/B:VISI.0000029664.99615.94
18. Lu, T.C., Chang, C.C.: Color image retrieval technique based on color features and image bitmap. Inf. Process. Manag. **43**(2), 461–472 (2007)
19. McLaren, K.: The development of the CIE 1976 (L*a*b*) uniform colour-space and colour-difference formula. J. Soc. Dyers Colour. **92**, 338–341 (2008)
20. Peterson, J.C., Abbott, J.T., Griffiths, T.L.: Evaluating (and improving) the correspondence between deep neural networks and human representations. Cogn. Sci. **42**(8), 2648–2669 (2018)
21. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: ICML 2019, pp. 8748–8763. PMLR (2021)
22. Roads, B.D., Love, B.C.: Enriching ImageNet with human similarity judgments and psychological embeddings. In: CVPR 2021, pp. 3547–3557. IEEE/CVF (2021)
23. Skopal, T.: On visualizations in the role of universal data representation. In: ICMR 2020, pp. 362–367. ACM (2020)
24. Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., Liu, C.: A survey on deep transfer learning. In: Kůrková, V., Manolopoulos, Y., Hammer, B., Iliadis, L., Maglogiannis, I. (eds.) ICANN 2018. LNCS, vol. 11141, pp. 270–279. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01424-7_27
25. Tan, M., Le, Q.: EfficientNet: rethinking model scaling for convolutional neural networks. In: ICML 2019, pp. 6105–6114. PMLR (2019)
26. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
27. Xu, J., Mei, T., Yao, T., Rui, Y.: MSR-VTT: a large video description dataset for bridging video and language. In: CVPR 2016, pp. 5288–5296 (2016)