Similarity Management of Data
the DISA Experience

Pavel Zezula

FI DISA, Masaryk University, Brno, Czech Republic

Abstract. As the current data is typically weekly structured or unstructured at all, access to objects is only possible through similarity of the object's salient features or properties. Consequently, similarity approach to searching is increasingly playing more and more important role in development of data processing applications. In the last twenty years, the technology has matured and many centralized, distributed, and even peer-to-peer architectures have been proposed. However, the use of similarity searching in numerous potential applications is still a challenge. In the talk, four research directions in developing similarity search applications at Masaryk University DISA laboratory are to be discussed. First, we concentrate on accelerating large-scale face recognition applications and continue with generic image annotation task for retrieval purposes. In the second half, we focus on execution of similarity query streams and finish the talk with an ambition topic of content-based retrieval in human motion-capture data collections. Applications will be illustrated by online prototype implementations.

1 Introduction

In the future, access to digital media stored in memories and circulating among networked computers will have to follow, or at least get much closer in its form, to the behavior of real life evolution and communication between species. There, recognition, learning, and judgment presuppose an ability to categorize stimuli and classify situations by *similarity*, because any event in the history of organism is, in a sense, unique and the specific case of the *exact/partial* match is marginal. As practically any kind of fact can nowadays become a digital part of the networked media – whatever we see, say, measure, observe, test, or otherwise experience, is or at least can be in digital form – computers must provide access to required data through similarity based operations, because it is the similarity that is in the world "revealing".

Similarity in general is determined by *stimuli* materialized as extractions from digital objects in form of *features* (*descriptors, properties,* etc.), and the way we grade closeness of objects in user defined *operations* able to satisfy information needs. We can calibrate the stimuli and operations from two different points of view: (1) *effectiveness* concerns the way similarity is defined and the extent to which it reflex the human perception of the relation, and (2) *efficiency* regards the processing speed, costs, or an effort needed to get results independently from scale.

Philosophical and psychological theorizing about similarity has been dominated by the geometric model for much of the twentieth century, see e.g. [22]. Though later criticized by several authors [23,7], it has become important paradigm for developing an *extensible* form of similarity search technology. The central assumption of this type of model is that the similarity can be related by a linear or monotonic decreasing function to inter-point distances in the *metric space* [14], that is, the larger the measure of similarity between two objects, the smaller the distance between the co-responding points in the metric space. Though the origins of the topic in computer science are older, the boom started in the 1990s with the M-tree [4] and resulted in many interesting scientific and technological achievements. The metric space paradigm extends the range of indexable similarity measures but at the same time loses the supportive advantage of coordinate systems to define partitioning of search spaces. The main advantage is that such approach is also able to consider data domains, which are not sortable – typical for a majority of contemporary digital data seen through their content descriptors. Since the similarity is in fact measured as a dissimilarity, specifically a distance, the applied techniques are often designated as *distance searching.*

Several key publications summarize achievements in this area. The first survey [3] includes results till the year 2000. It presents known approaches in original taxonomy with the objective to discover core properties that would allow combination of existing principles to form future better proposals. The second survey [6] divides existing methods for handling similarity search into two classes. The first class directly indexes objects based on distances (distance-based indexing), while the second is based on mapping to a vector space (mapping-based approach). However, the main part of this article is dedicated to a survey of distance-based indexing methods, and the mapping-based methods are only outlined. In 2006, a book named Similarity Search: The Metric Space Approach [24] presented the state-of-the-art in developing index structures and supportive technologies for searching complex data modeled as instances of a metric space. The metric searching problems are also considered in the last edition of the encyclopedic book by Hanan Samet [15] called Foundations of Multidimensional and Metric Data Structures.

2 Similarity Search in Applications

Though a lot of progress has been made [1,12] and several interesting similarity search demonstration prototypes, see Fig. 1 for illustration, are already available, [10,11], the fact still is that the extensibility property – one system used for many applications – of the metric space approach to similarity searching, has not yet been fully exploited [13].

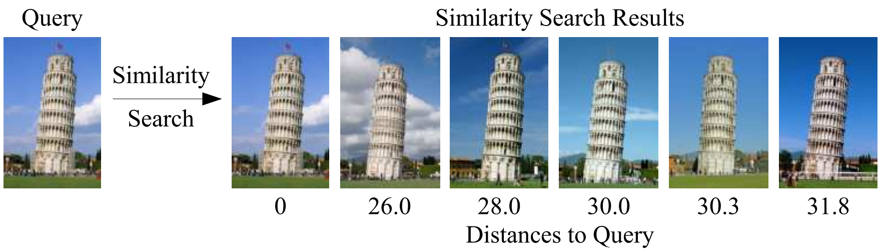

Fig. 1. Image similarity search

An application of similarity searching in several dimensions have been investigated in the Data Intensive Systems and Applications (DISA) Laboratory of Masaryk University, Brno. The first objective is to speedup *face retrieval* in large collections of common photographs. They also develop *image annotation* systems and study mechanisms which should be applied for similarity searching in *data streams*. Finally, they consider a very complex data type, called *motion capture data*, to develop scalable similarity search, filtering and annotation/classification mechanisms. In the following, we shortly outline each of the activities. In the talk, the content-based image similarity search as well as its applications will be demonstrated by on-line demonstrations.

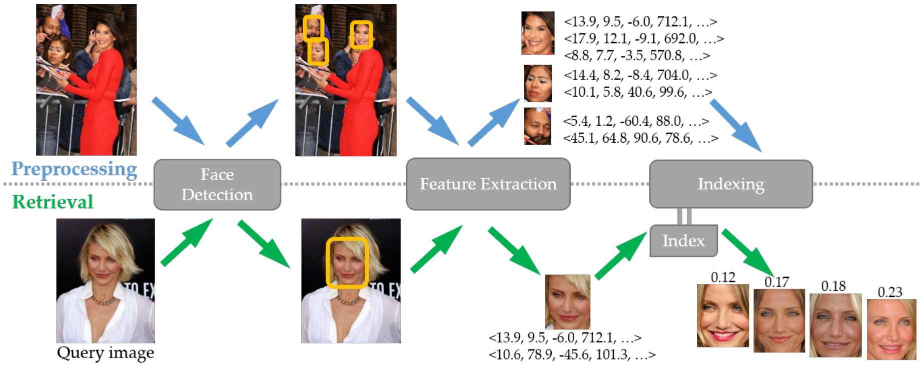

Fig. 2. Face detection and similarity retrieval

2.1 Similarity Searching in Images of Human Faces

Face recognition is a problem of verifying or identifying a face appearing in a given image. In our project, sketched in Fig. 2, we do not develop new face detection or comparison modules but rather demonstrate how a synergic collaboration of existing systems combined with similarity searching can scale into the dimension of large image collections [19]. Specifically, by integrating three OpenCV, NeuroTech and Luxand similarity measures, we achieve high-quality and more stable results, compared to the measures evaluated independently. We also demonstrate, how effectiveness can significantly improve by employing the concept of *multi-face queries* along with optional *relevance feedback*. Finally, a metric indexing structure is applied to achieve scalability of the similarity search.

2.2 Image Annotation

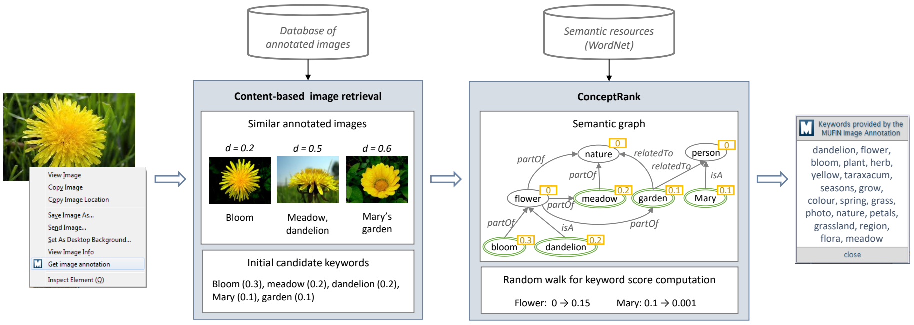
View Image
Copy Image
Copy Image Location
Save Image As...
Send Image...
Set As Desktop Background...
View Image Info
Get image annotation
Inspect Element (Q)
Database of annotated images
Content-based image retrieval
Similar annotated images
d = 0.2
d = 0.5
d = 0.6
Bloom
Meadow, dandelion
Mary's garden
Initial candidate keywords
Bloom (0.3), meadow (0.2), dandelion (0.2), Mary (0.1), garden (0.1)
Semantic resources (WordNet)
ConceptRank
Semantic graph
nature
0
person
0
partOf
relatedTo
relatedTo
partOf
isA
flower
0
meadow
0.2
garden
0.1
Mary
0.1
partOf
partOf
partOf
isA
bloom
0.3
dandelion
0.2
Random walk for keyword score computation
Flower: 0 → 0.15
Mary: 0.1 → 0.001
Keywords provided by the MUFIN Image Annotation
dandelion, flower, bloom, plant, herb, yellow, taraxacum, seasons, grow, colour, spring, grass, photo, nature, petals, grassland, region, flora, meadow
close

Fig. 3. ConceptRank with search based image annotation

The objective of image annotation is to associate binary images with descriptive metadata that would allow application of text search for content based image retrieval – metadata can also be used for image categorization. Such task has a long tradition in the machine learning field, which approaches the problem by training statistical models for prespecified set of categories. State-of-the-art classification methods of this sort achieve very high accuracy, but their utilization is costly in terms of learning time and requires large amounts of reliably-labeled training data [21]. The proposed strategy, called the *concept-rank* [2], combines information provided by efficient and effective similarity search with semantic information obtained from linguistic resources. Specifically, we first select a set of initial candidate keywords from descriptions of similar already annotated images and give them a probability score proportional to their frequency and the similarity of the respective images to the annotated (query) image. Next, we search for links between these keywords using several semantic relationships defined by the WordNet lexical database, in particular we apply the *hypernymy*, *hyponymy*, *meronymy*, and *holonymy*. We also include new related keywords in consideration for annotation. After the identification of semantic relationships, we run a random-walk-based algorithm over the graph of candidate keywords and their relationships to determine the final probabilities of individual candidates. All the process is illustrated in Fig. 3.

2.3 Stream Processing

The problem of fast executing streams of similarity queries is serious for a large number of applications. For example when annotating large collections of images or when publish-subscribe systems consider incoming documents as queries and test them against user profiles to access the level of agreement between the profile and the processed document (image). The acceleration strategy is based on a pragmatic observation that in large collections the response time to process two random queries is significantly higher that the time needed to process two queries representing similar images [9]. The more similar the queries the shorter the query execution time because a larger portion of the database can be pre-cached from a previous query execution. Then by a clever reordering of queries so that the similarity of two consequent queries is as high as possible, the throughput can significantly be improved. The concept is illustrated in Fig. 4. Further performance improvements can be obtained by a parallel execution for example in cloud computer platforms [8].

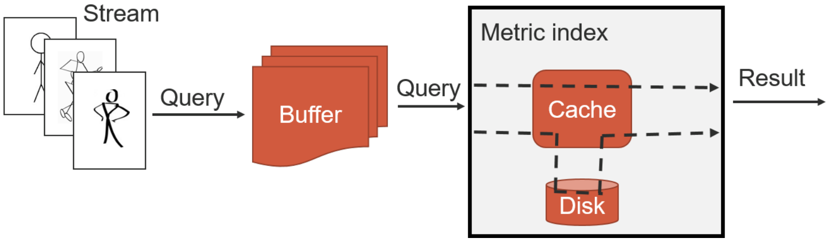

Fig. 4. Similarity query stream processing

2.4 Similarity Searching in Motion Capture Data

Motion capture data is a good example of complex unstructured data. This spatio-temporal data digitally represents human movements in form of 3D trajectories of tracked human body joints. With the recent advances and availability of motion capturing technologies, there is a strong requirement for intelligent management of such data, which has a great potential to be utilized in a number of applications.

To make the content-based management of motion data possible, effective features need to be extracted from 3D skeleton sequences. We propose a 4,096-dimensional vector representation that preserves significant characteristics of original motion sequences [17]. As illustrated in Fig. 5, this representation normalizes and transforms an input sequence into a 2D motion image which is then processed by a convolutional neural network. The last hidden layer of the network is considered as the output feature. The extracted features demonstrate very convenient properties of being (1) of a fixed size, (2) efficiently comparable by the Euclidean distance, and (3) tolerant to a considerable degree of segmentation errors, which is particularly useful for sub-sequence matching.

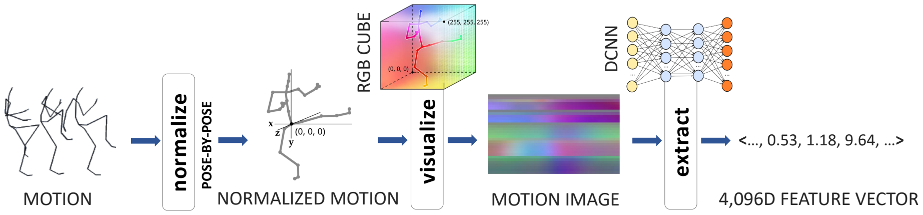

Fig. 5. Feature extraction from 3D human motion sequences.

The extracted features can describe the content of relatively short skeleton sequences taking in order of seconds. This is mainly suitable for recognizing the class of semantic actions using k-nearest neighbor classifiers [20]. In case input sequences are long, the partitioning principle needs to be applied to identify short segments that are better describable by the features. We propose to partition the input sequence into segments of different sizes in a way that an arbitrary sub-sequence overlaps with at least one segment in the majority of frames [16]. This property enables to consider a short query as a single segment and perform efficient sub-sequence searching on a large scale [18]. A similar idea is also used for real-time annotation of pseudo-infinite skeleton sequences [5].

Acknowledgements. This research is supported by the Czech Science Foundation project No. GA19-02033S.

References

1. Batko, M., Novak, D., Falchi, F., Zezula, P.: Scalability comparison of peer-to-peer similarity search structures. Future Generation Comp. Syst. 24(8), 834–848 (2008)
2. Budíková, P., Batko, M., Zezula, P.: Conceptrank for search-based image annotation. Multimedia Tools Appl. 77(7), 8847–8882 (2018), https://doi.org/10.1007/s11042-017-4777-8

3. Chávez, E., Navarro, G., Baeza-Yates, R., Marroquín, J.: Searching in metric spaces. ACM Comput. Surv. 33(3), 273–321 (2001)
4. Ciaccia, P., Patella, M., Zezula, P.: M-tree: An efficient access method for similarity search in metric spaces. In: VLDB. pp. 426–435. Morgan Kaufmann (1997)
5. Elias, P., Sedmidubsky, J., Zezula, P.: A real-time annotation of motion data streams. In: 19th International Symposium on Multimedia. pp. 154–161. IEEE Computer Society (2017)
6. Hjaltason, G., Samet, H.: Index-driven similarity search in metric spaces. ACM Trans. Database Syst. 28(4), 517–580 (2003)
7. Krumhansl, C.: Concerning applicability of geometric models to similarity data: The interrelationship between similarity and spatial density. Psychological Review 85 pp. 445–461 (1978)
8. Nalepa, F., Batko, M., Zezula, P.: Model for performance analysis of distributed stream processing applications. In: Database and Expert Systems Applications - 26th International Conference, DEXA 2015, Valencia, Spain, September 1-4, 2015, Proceedings, Part II. pp. 520–533 (2015)
9. Nalepa, F., Batko, M., Zezula, P.: Enhancing similarity search throughput by dynamic query reordering. In: Database and Expert Systems Applications - 27th International Conference, DEXA 2016, Porto, Portugal, September 5 – 8. p. 15 (2016)
10. Novak, D., Batko, M., Zezula, P.: Generic similarity search engine demonstrated by an image retrieval application. In: Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Boston, MA, USA, July 19-23. p. 840 (2009)
11. Novak, D., Batko, M., Zezula, P.: Large-scale image retrieval using neural net descriptors. In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, Chile, August 9-13, 2015. pp. 1039–1040 (2015)
12. Novak, D., Batko, M., Zezula, P.: Large-scale similarity data management with distributed metric index. Inf. Process. Manage. 48(5), 855–872 (2012)
13. Novak, D., Kyselak, M., Zezula, P.: On locality-sensitive indexing in generic metric spaces. In: Third International Workshop on Similarity Search and Applications, SISAP 2010, 18-19 September 2010, Istanbul, Turkey. pp. 59–66 (2010)
14. O'Searcoid, M.: Metric Spaces. Springer (2006)
15. Samet, H.: Foundations of Multidimensional And Metric Data Structures. Series in Data Management Systems, Morgan Kaufmann (2006)
16. Sedmidubsky, J., Elias, P., Zezula, P.: Similarity searching in long sequences of motion capture data. In: 9th International Conference on Similarity Search and Applications (SISAP). pp. 271–285. Springer International Publishing, Cham (2016)
17. Sedmidubsky, J., Elias, P., Zezula, P.: Effective and efficient similarity searching in motion capture data. Multimedia Tools and Applications 77(10), 12073–12094 (2018), https://doi.org/10.1007/s11042-017-4859-7
18. Sedmidubsky, J., Elias, P., Zezula, P.: Searching for variable-speed motions in long sequences of motion capture data. Information Systems 80, 148–158 (2019)
19. Sedmidubsky, J., Mic, V., Zezula, P.: 8th International Conference on Similarity Search and Applications (SISAP 2015), chap. Face Image Retrieval Revisited, pp. 204–216. Springer International Publishing (2015), http://dx.doi.org/10.1007/978-3-319-25087-8_19
20. Sedmidubsky, J., Zezula, P.: Probabilistic classification of skeleton sequences. In: 29th International Conference on Database and Expert Systems Applications (DEXA). pp. 50–65. Springer International Publishing, Cham (2018)

21. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S.E., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015. pp. 1–9 (2015)
22. Torgerson, W.S.: Multidimensional scaling of similarity. Psychometrika 30 pp. 379–393 (1965)
23. Tversky, A.: Features of similarity. Psychological Review 84 pp. 327–354 (1977)
24. Zezula, P., Amato, G., Dohnal, V., Batko, M.: Similarity Search: The Metric Space Approach, Advances in Database Systems, vol. 32. Springer (2006)