frontiers | Frontiers in Imaging

TYPE Specialty Grand Challenge
PUBLISHED 04 August 2022
DOI 10.3389/fimag.2022.951934

OPEN ACCESS

EDITED BY
Jinchang Ren,
Robert Gordon University,
United Kingdom

REVIEWED BY
Deepak Kumar Rout,
Mahousadhi Healthcare Pvt Ltd., India

*CORRESPONDENCE
Guoping Qiu
guoping.qiu@nottingham.ac.uk

SPECIALTY SECTION
This article was submitted to Image Retrieval, a section of the journal Frontiers in Imaging

RECEIVED 24 May 2022
ACCEPTED 27 June 2022
PUBLISHED 04 August 2022

CITATION
Qiu G (2022) Challenges and opportunities of image and video retrieval. *Front. Imaging.* 1:951934. doi: 10.3389/fimag.2022.951934

Challenges and opportunities of image and video retrieval

Guoping Qiu*

School of Computer Science, University of Nottingham, Nottingham, United Kingdom

KEYWORDS

image indexing, image retrieval, object recognition, scene understanding, user interface, information visualisation

1. Introduction

According to some estimations, more than 3 billion images and 700,000 h of video are shared on social media daily. When dealing with such a flood of content, researchers and practitioners are confronted with the challenge of how to efficiently index the image and video data and develop friendly tools to enable users to quickly find what they are looking for. Indexing and retrieval of images and videos are very challenging due to the primitive nature of their raw data representations which lack readily available structural and semantic information. Performing image and video retrieval requires to first process the data to extract discriminative, meaningful, and interpretable features, and then to gain high level understanding at the object, scene, and semantic levels, and finally to develop systems and tools to efficiently index the data and to help users to find what they are looking for intuitively, easily, and accurately.

Therefore, image and video retrieval technologies cover a very broad area of modern information sciences including database, data structure, artificial intelligence, machine learning, computer vision, and visualisation, etc. Despite much progress since IBM first introduced their query by image content (QBIC) system (Flickner et al., 1995) nearly 30 years ago, rapid increase in the volume of image and video data at an unprecedent speed meant that the field is facing new challenges as well as many unsolved existing problems.

2. Discriminative, meaningful, and interpretable features

In their raw forms, image and video data is organized in a 2-dimensional raster format. This makes it very difficult to use the data directly for indexing. It is therefore necessary to perform feature extraction such that image and video can be represented in ways that are not only more compact but also easier to organize to facilitate retrieval. In fact, throughout the history of image analysis and computer vision, feature extraction or image representation has been a major theme, from early days classic transformations to the latest deep learning.

2.1. Classic transformations

One of the earliest transforms for image feature extraction is the Hoteling Transform (HT), also known as Karhunen-Loeve transform (KLT) and Principal Component Analysis (PCA). Well known and successful applications of this transformation

in image analysis include eigenface and active appearance models. Whilst the transformation matrix of PCA is learned from training data to capture important statistical properties of the training dataset, there exist many data independent transforms that have their roots in Fourier Transform and frequency analysis including discrete cosine transform (DCT), Walsh-Hadamard Transform (WHT), Gabor Transform (GT) and Wavelet Transform (WT).

2.2. Statistical features

One of the simplest and sometimes very effective features for image indexing and retrieval is color histogram (Swain and Ballard, 1991). Whilst color histogram treats individual pixel independently, a very popular feature generally referred to as bag of visual words (BVW) processes a group of neighboring pixels together. A codebook of either the raw pixels of local image patches or some feature representations such as scale-invariant feature transform (SIFT) (Lowe, 1999) of the local image patches is designed based on vector quantization (VQ) (Gray, 1984). The codewords are statistically the most representative patterns of the local image patches. An image can then be represented by the frequencies of occurrence of the codewords in the codebook. The idea of BVW was first developed for content-based image indexing and retrieval (Qiu, 2002) and later widely adopted by the computer vision community in a variety of different forms (Sivic and Zisserman, 2003).

2.3. Representation learning

The features discussed above are sometimes referred to as handcrafted features in the sense that the models for extracting the features are designed individually by their designers. The problem with these features is that one feature can only capture one aspects of the image statistics and it is difficult to know how effective it is for a particular task because the design of the features and the task the features are used for are independent. Representation learning (Bengio et al., 2013), more popularly known as deep learning or feature learning, automatically design representation features through training convolutional neural networks (CNNs) based on either supervised or unsupervised learning. These features, sometimes known as deep features have been proven to be more effective than traditional handcrafted features in a variety of applications including image retrieval (Wan et al., 2014).

2.4. Challenges and opportunities

Like in other applications, the challenges of feature extraction for image and video indexing and retrieval lies in how to make the features discriminative, meaningful, and interpretable. Discriminative features are difficult to obtain because the pixel space is much larger than the feature space, and feature extraction is a Many-To-One (MTO) mapping, meaning that different pixel space entities can be associated with the same or similar entities. Many features do not have an easily explainable meaning. For example, a CNN is a black box in the sense that features extracted by its hidden layers have no clear meanings. It is unknown what image properties certain deep features represent. Similarly, because many features do not have clear meanings, they are uninterpretable, thus making it hard to use specific features for specific purposes. For example, if our purpose is to retrieve human faces, then there are no "face features" from a CNN that we can directly use. Making features more discriminative, meaningful, and interpretable is a great challenge facing researchers and practitioners in image and video retrieval. It is hoped that in the coming years, we will see papers published in this journal making progress in this and related areas.

3. Understanding at object, scene, and semantic levels

To accurately locate specific image and video contents, it is necessary to understand what objects have appeared in an image or a video frame, what the scene is about and the relations between the objects, and what high level semantics they convey. Therefore, another major challenge for image and video retrieval is gaining understanding at the object, scene, and semantic levels. The various features discussed previously are low level representations of image contents. There exists an object gap in the sense that there is no simple one to one correspondence between feature and object. Similarly, scene gap exists. A scene level understanding represents a higher-level knowledge than that of object level, we not only need to know the objects but also the relations of the various objects. Semantic gap is even more challenging. For example, how can we infer from features exacted from a group of red color pixels that they are from a red patch of paint or from a red rose flower; and features extracted from a scene containing human faces that these people are having dinner or in the middle of a meeting. A good image and video retrieval solution requires understanding the contents at the object, scene, and semantic level.

3.1. Object level

Object recognition is probably one of the most researched topics in computer vision, from regular handwritten digits and human faces to everyday objects and animals, have been extensively studied. Recent progress in deep learning (LeCun et al., 2015) has advanced object recognition performances

significantly. From an image and video retrievals perspective, accurately recognizing specific objects is only the necessary first step, equally important is how we can make use of object recognition systems and results to help users to find the contents they are looking for. Therefore, researchers and practitioners should always bear in mind the ultimate purpose is to help indexing the contents more efficiently and effectively, and help making retrieval more accurately.

3.2. Scene level

After recognizing objects in an image, interpreting, and understanding the scene, and inferring the semantic meaning beyond objects is extremely challenging (Xiao et al., 2013). For example, suppose we have a group of people at a scene, to be able to detect and recognize people is not enough to understand what these people are doing. Whether they are having a meeting or having dinner or something else needs to be inferred from the object level understanding. Whilst humans have this unique ability to reason beyond what can be seen, to equip machines with the same ability is very difficult. The task of scene understanding generally involve analyzing the 3D structure layout of the scene and the spatial, functional, and semantic relationships between objects. Again, in the context of video retrieval, how to best harvest results of scene understanding to make contents easily and readily available needs much more research.

3.3. Semantic/language level

A picture may be worth a thousand words, but from an image retrieval s perspective, is the most useful to be able to find the words that accurately describe a picture. Recent years has seen much progress in image and video tagging and captioning (Hossain et al., 2019) where words and sentences are automatically generated to describe the visual content. This provides the ability of using language to describe and represent the high-level semantic knowledge about the image and video data. As humans are much more accustomed to using language to describe what there are looking for, therefore vision language modelling will play a key role in image and video retrieval. Recently emerged large pretrained vision language models such as VisualGTP (Chen et al., 2021) and CLIP (Contrastive Language-Image Pre-training) (Radford et al., 2021) are likely to play very useful roles in image retrieval.

3.4. Challenges and opportunities

Despite much progress, image and video understanding at the object, scene, and semantic levels remains to be very challenging. The difficulties facing object recognition researchers include viewpoint and lighting variations, occlusion, complex background, intra-class variation, and low image quality. For example, recognizing objects in low-resolution and noisy images inevitably increases difficulty. In addition to the challenges of inferring 3D information from 2D image data and modelling the spatial relations of objects in the 3D scene environment, the difficulty faces scene level understanding includes all those facing object recognition. Automatic image captioning remains challenging despite impressive progress in deep learning-based solutions. One common problem in any application of deep learning is the lack of labeled data to train deep models. How to obtain sufficiently large databases for training object, scene, and semantic level understanding models is also one of the challenges facing researchers in image and video retrieval. Whether image retrieval can help constructing training databases for training deep learning models is worth investigating. Again, how these different levels of understanding models can be turned into retrieval tools and systems to facilitate finding desired information quickly and accurately is the additional major challenge facing image and video retrieval researchers and practitioners.

4. Intuitive, easy to use and accurate retrieval systems

Ultimately, image and video retrieval algorithms will only be useful if they can be turned into user friendly tools that can genuinely help everyday users finding what they are looking for. Therefore, in addition to the challenges of representing the data at the feature level, gaining understanding of images at the object, scene, and semantic levels, as well as describing the visual contents using words and sentences, researchers and practitioners must also design highly effectively and user-friendly interfaces, develop tools and algorithms to enable users to interact with the system and refine retrieval results, and present the retrieval results in a way that is intuitive and adaptive to viewing environments—mobile, pad, and desktop.

4.1. Retrieval interface

An effective interface design is very important to image and video retrieval (Eakins et al., 2004; Dudley and Kristensson, 2018; Huang et al., 2019). Compared with research in feature extraction and content understanding, retrieval interface design has received relatively less attention and yet it is just as important if not more so. A good interface can compensate for the shortcomings of retrieval algorithms. Yet, there is little research on integrating retrieval algorithms and interface design for image and video retrieval systems. User interface designers should understand how the retrieval algorithms works so that

specific interface features can be designed to take advantage of the algorithms, as well as to compensate for their weaknesses. An early example of integrated design is Qiu et al. (2007) where image features (colors) used in organizing the database images into clusters are directly reflected on the interface, and the indexing keys are used to partition the display areas such that the interface provides an intuitive "mental image" of the database. It is hoped that we will see more such integrated design in the future to advance state of the art in image and video retrieval.

4.2. Interactive tools

Unlike text retrieval, user intentions in image and video retrieval in most cases are rather vague because it is very difficult to describe visual contents precisely (Kofler et al., 2016). The retrieval results are therefore often not precise, and it is necessary to refine the query to obtain more accurate results. Interactive tools that enable users to refine queries should form an integral part of an image and video retrieval system. Similarly, a good interactive tool can compensate for the weaknesses of retrieval algorithms. Therefore, the design of the interactive features should work together with the retrieval algorithms. Again, very little seems to have been done in integrating interactive tools with retrieval algorithms. It is hoped that we will see works integrating interactive tools with retrieval algorithms in the future.

4.3. Presentation and visualisation of results

How the retrieval results are presented to the users will not only affect user experience but also direct determine how quickly users can find what they are looking for. As an image will occupy a larger display area than a word, the number of images that can be displayed on a screen is very limited. It is therefore very important to design effective result visualisation schemes that can facilitate users find what they are looking for. It appears not much research has been directed to the design of visualisation schemes specifically suited for image retrieval. Again, integrating visualisation algorithms with retrieval algorithms is likely help advancing state of the art.

4.4. Challenges and opportunities

Very little research has gone into integrating feature representations and image understanding algorithms with the design of user interface, interactive tool, and result visualisation schemes. The challenge lies in how algorithms running behind the interface, including feature representations and image understanding algorithms can be used to help interface design, and how the design of the interface can compensate for some of the shortcomings of the backend algorithms, such that they work together seamlessly to advance state of the art in image and video retrieval. It is believed that integrated solutions will be a fertile area for innovation and we hope more researchers will pay attention to this direction.

5. Applications of image and video retrieval

The final challenge in the years ahead for image and video retrieval includes improving existing applications and developing new applications. Existing applications include image search on the Internet, organizing personal photo album, and specific domains such as medical imaging (Müller et al., 2004) and visual localization (Sattler et al., 2012). With the development of new technologies such as deep learning, there have already been many works applying deep learning for image retrieval. As deep learning requires huge amount of labeled training data, how image retrieval may be used to help prepare training datasets for improving deep model training, and in training general artificial intelligence model through continuous learning is an intriguing problem. Another interesting application direction for image retrieval would be how it may help advancing scientific research in fields such as biomedicine and environmental sciences.

6. Concluding remarks

Explosive increase in image and video collections has created huge demand for advanced tools to manage the data and to help users to find what they are looking for easily. However, despite much progress in related fields such as computer vision and machine learning, image and video retrieval remains to be very challenging. In addition to continue making progress in traditional areas, an important direction for advancing image retrieval is to integrate data processing algorithms with interface design and interactive tool development. Also recently emerged pretrained large vision language models such as CLIP are likely to have major impact on the image retrieval.

Author contributions

The author confirms being the sole contributor of this work and has approved it for publication.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: a review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* 35, 1798–1828. doi: 10.1109/TPAMI.2013.50

Chen, J., Guo, H., Yi, K., Li, B., and Elhoseiny, M. (2021). Visualgpt: data-efficient adaptation of pretrained language models for image captioning. *arXiv preprint arXiv:2102.10407*. doi: 10.48550/arXiv.2102.10407

Dudley, J. J., and Kristensson, P. O. (2018). A review of user interface design for interactive machine learning. *ACM Trans. Interact. Intell. Syst.* 1, 8. doi: 10.1145/3185517

Eakins, J. P., Briggs, P., and Burford, B. (2004). "Image retrieval interfaces: a user perspective," in *Image and Video Retrieval*, eds P. Enser, Y. Kompatsiaris, N. E. O'Connor, A. F. Smeaton, and Smeulders (Berlin; Heidelberg: Springer Berlin Heidelberg), 628–637.

Flickner, M., Sawhney, H., Niblack, W., Ashley, J., Huang, Q., Dom, B., et al. (1995). Query by image and video content: the qbic system. *Computer* 28, 23–32. doi: 10.1109/2.410146

Gray, R. (1984). Vector quantization. *IEEE Assp Mag.* 1, 4–29. doi: 10.1109/MASSP.1984.1162229

Hossain, M. Z., Sohel, F., Shiratuddin, M. F., and Laga, H. (2019). A comprehensive survey of deep learning for image captioning. *ACM Comput. Surv.* 51, 1–36. doi: 10.1145/3295748

Huang, F., Canny, J. F., and Nichols, J. (2019). "Swire: sketch-based user interface retrieval," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI '19* (New York, NY: Association for Computing Machinery), 1–10.

Kofler, C., Larson, M., and Hanjalic, A. (2016). User intent in multimedia search: a survey of the state of the art and future challenges. *ACM Comput. Surveys* 49, 1–37. doi: 10.1145/2954930

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539

Lowe, D. (1999). "Object recognition from local scale-invariant features," in *Proceedings of the Seventh IEEE International Conference on Computer Vision, Vol. 2* (IEEE). doi: 10.1109/ICCV.1999.790410

Müller, H., Michoux, N., Bandon, D., and Geissbuhler, A. (2004). A review of content-based image retrieval systems in medical applications–clinical benefits and future directions. *Int. J. Med. Inform.* 73, 1–23. doi: 10.1016/j.ijmedinf.2003.11.024

Qiu, G. (2002). Indexing chromatic and achromatic patterns for content-based colour image retrieval. *Pattern Recog.* 35, 1675–1686. doi: 10.1016/S0031-3203(01)00162-5

Qiu, G., Morris, J., and Fan, X. (2007). Visual guided navigation for image retrieval. *Pattern Recog.* 40, 1711–1721. doi: 10.1016/j.patcog.2006.09.020

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., et al. (2021). "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning* (PMLR), 8748–8763.

Sattler, T., Weyand, T., Leibe, B., and Kobbelt, L. (2012). "Image retrieval for image-based localization revisited," in *British Machine Vision Conference, Vol. 1* (Surrey), 4.

Sivic, J., and Zisserman, A. (2003). "Video google: A text retrieval approach to object matching in videos," in *Computer Vision, IEEE International Conference on, Vol. 3* (Nice: IEEE), 1470–1470.

Swain, M. J., and Ballard, D. H. (1991). Color indexing. *Int. J. Comput. Vis.* 7, 11–32. doi: 10.1007/BF00130487

Wan, J., Wang, D., Hoi, S. C. H., Wu, P., Zhu, J., Zhang, Y., et al. (2014). "Deep learning for content-based image retrieval: a comprehensive study," in *Proceedings of the 22nd ACM international conference on Multimedia* (Orlando, FL), 157–166.

Xiao, J., Hays, J., Russell, B., Patterson, G., Ehinger, K., Torralba, A., et al. (2013). Basic level scene understanding: categories, attributes and structures. *Front. Psychol.* 4, 506. doi: 10.3389/fpsyg.2013.00506