Chapter 1
Semantic Gap in Image and Video Analysis: An Introduction

Halina Kwaśnicka and Lakhmi C. Jain

Abstract The chapter presents a brief introduction to the problem with the semantic gap in content-based image retrieval systems. It presents the complex process of image processing, leading from raw images, through subsequent stages to the semantic interpretation of the image. Next, the content of all chapters included in this book is shortly presented.

1.1 Introduction

The problem of the semantic gap is crucial and is seen in many tasks of image analysis, as Content-Based Image Retrieval (CBIR) or Automatic Image Annotation (AIA). The semantic gap is a lack of correspondence between the low-level information extracted from an image and the interpretation that the image has for a user. How to transform the features computed from raw image data to the high-level representation of semantics carried out by that image is still the open problem. This problem exists despite the observed intensive research with the use of different approaches to solving, or at least narrowing, the semantic gap in image analysis, especially in image retrieval. This gap is perceived as a barrier to image understanding. Some researchers claim that the understanding of how humans perceive images should be helpful [1, 2]. A typical CBIR method is a query-by-example system. In real life application finding an image as an appropriate users query is hard [3]. Easier and more intuitive is to describe the intended image by some keywords. Combining different media, like images, text,

H. Kwaśnicka (✉)
Department of Computational Intelligence, Wroclaw University of Science
and Technology, Wroclaw, Poland
e-mail: halina.kwasnicka@pwr.edu.pl

L. C. Jain
Founder, KES International, Leeds, UK

L. C. Jain
Faculty of Science, Technology and Mathematics, University of Canberra, Canberra, Australia
e-mail: jainlakhmi@gmail.com; jainlc2002@yahoo.co.uk

H. Kwaśnicka and L. C. Jain (eds.), *Bridging the Semantic Gap in Image and Video Analysis*, Intelligent Systems Reference Library 145,
https://doi.org/10.1007/978-3-319-73891-8_1

video, sound, into one application is a subject of Multimedia Information Retrieval. It is also the widely developed field of research.

The output of CBIR systems is a ranked list of images; the images are ordered according to their similarity to the users query image. However, similarity is measured using low-level features extracted from images; this causes that returned images often do not meet users expectations, similarity based on low-level features do not correspond the human perception of similarity. Research on how human perception is working is intensively developed, one can expect that their results will be useful in bridging the semantic gap [4–9]. Authors of [9] try to model of human cortical function aiming simulation of the time-course of cortical processes of understanding meaningful, concrete words. The different parts of the cortex are responsible for general and selective, or category-specic, semantic processing. In [5] authors studied the humans and automatic perception of orientation of color photographic images. They concluded that the interaction with the human observers allows defining sky and people as the most important cues used by humans at various image resolutions. These and other results in the field of understanding human perception can be a hint for the creators of computer systems understanding images. Some researchers focus on developing a computer system that mimics the perceptual ability of people [10]. Such systems try to consider knowledge about the structure and the surrounding environment of a scene.

An analysis of the perception of images by man suggests that computer vision systems should also take into account some knowledge. The computer systems require acquired knowledge at different levels. To explain it let us see on vision systems from three perspectives: knowledge, algorithmic and implementation perspectives. From the implementation perspective, the used programming languages and computer hardware can be considered; this is not interesting for us here. The algorithmic perspective is essential—we have to decide the way of representing the relevant information, also the most suitable algorithms for use. The most interesting perspective is the knowledge perspective. Here, the questions could concern the knowledge that enters a process, the knowledge obtained in the process, constraints determining the process, and others.

An image (a scene) corresponds to basic properties of real-world. The next processing step uses physics, photometry, and so on. Further processing requires models of objects to be recognized, models of situations and common sense knowledge (see Fig. 1.1).

Information derived from primitive features, extracted from images, is the low-level knowledge. The semantic relationships and patterns, gathered by knowledge discovering methods, are the second level of knowledge [10]. Gathering such knowledge requires considering the correlation between the low-level information with the interpretation of concepts related to domain knowledge. Machine learning has to recognize complex structural relations between visual data and the semantic interpreted by human observing the considered scene.

In real-life use of CBIR systems, often a user can have a problem with finding a query image that matches the user's intent [11]. Finding the perfect image from a collection could be an example of such situation. It would then be much easier to

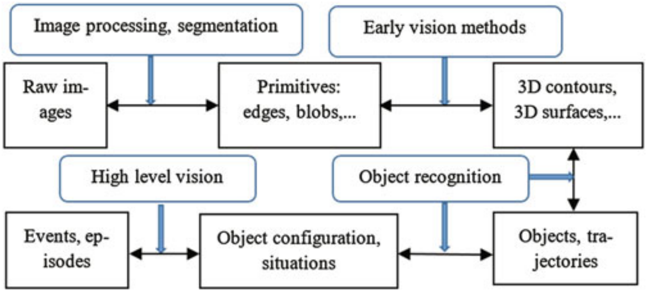
Image processing, segmentation
Early vision methods
Raw images
Primitives: edges, blobs,...
3D contours, 3D surfaces,...
High level vision
Object recognition
Events, episodes
Object configuration, situations
Objects, trajectories

Fig. 1.1 From raw image to image understanding—a schema of processing

describe the desired image using text. The authors of [12], distinguish four scenarios depending on available information for creating CBIR: caption; annotation, tag, keyword; full MPEG-7 annotation. The potential scenarios are: only images; images with captions; images with captions and annotations, tags, keyword; and images with all mentioned descriptions. The authors propose different corresponding tasks for these scenarios such as rule induction for semantic class refinement, use Knowledge-Based System to infer object association or structural projection MPEG-7 representation and index building.

Multimodal CBIRs, i.e., taking into consideration visual, textual and audio features are growing in popularity. How to exploit the visual content of images in the CBIR systems is strongly developed, but there are other subjects worth the attention of researchers. Li et al. present a survey of researches on three problems connected with the semantic gap bridging: image tag assignment, refinement, and tag-based image retrieval [13]. The tag relevance to the visual content of an image hardly influences the quality of CBIR.

As it was mentioned earlier, the subject of semantic gap in the field of content-based image retrieval is intensively studied. The very interesting survey is presented in [14]. Authors comprehensively present achievements in particular steps of the CBIR systems, starting from the framework of CBIR, by image preprocessing, feature extraction, learning system, benchmark datasets, similarity matching, relevance feedback, up to the evaluation of performance and visualization. The authors also indicate some key issues that influence the CBIR. They pointed out as still open problems: representation of images with a focus on local descriptors; automatic image annotation; image indexing to reduce dimensionality; deep learning approach; description of ideal image datasets; re-ranking approaches as post-processing; visualization aspects.

An interesting approach is presented in [15]. The authors extend the latent semantic word and object models, to the latent semantic word, object and part models. The premise of this approach was the fact that not only similarity of semantic of words and semantic of images is important to the CBIR task. Also complex semantic relations within each modality, e.g., there are similar relations in the text to the relation between objects: *object A is a part of object B* and *object B is an instance of object*

C. They developed models able to learn these types of semantic relations across and within modalities simultaneously, using ImageNet and WordNet sources.

Variety of approaches have been developed to improve the CBIR systems that would be able to return the most relevant images with maximum user satisfaction [16–19]. Also, numerous papers containing a survey of the CBIR systems have been published, i.e., [13, 14, 20, 21]. In this book, some chapters present interesting approaches at the different level of CBIR systems and one chapter dedicated to applications of deep learning to bridge semantic gap. We have noticed a lack of survey dedicated to this new learning paradigm applied to image understanding, and the last chapter fills this gap.

1.2 Chapters Included in the Book

Chapter 2 presents a comparative study of the most used and popular low-level local feature extractors in a smart image and video analysis. An overview of different extractors is the first part of the chapter. The authors highlighted the main theoretical differences among the different extractors. A comprehensive study has been performed with use the Freiburg-Berkeley Motion Segmentation (FBMS-59) dataset. The robustness and behavior of compared extractors are discussed. The observations about the matching process are also outlined.

Chapter 3 is dedicated to image segmentation. The author claims that reliable segmentation algorithms, extracting as accurately as possible, regions with a certain level of semantic uniformity significantly improve the automatic annotation of an image. The developed segmentation technique is based on scale-insensitive maximally stable extremal regions (SIMSER) features a generalization of the popular MSER features, which is rather useless in semantic image segmentation. The chapter describes the experimental study of relations between semantics based image annotation and SIMSER features, focusing on color images.

Chapter 4 shows a generalization of known active contour technique, namely active partitions. The proposed approach can be applied to more sophisticated image content representations than raw pixel data. The reduction of search space enables to use evolutionary computations, less sensitive or invariant to the choice of initial solutions. The author demonstrates the flexibility of the proposed approach; it can be applied to both global and local image analysis.

Chapter 5 deals with 3D object recognition in RGB-D images, in indoor autonomous robotics. The proposed framework integrates solutions for: generic object representation; trainable transformations between abstraction levels; reasoning under uncertain and partial data; optimized model-to-data matching; efficient search strategies. As such, the framework is an application-independent generic model based. It was verified in robot vision scenarios. The approach allows to identify what kind of knowledge is needed and to utilize existing meta-level knowledge to learn concept types instead of memorizing individual instances. An interesting feature of the proposed framework is decomposition of an object into simpler elements, named parts. The authors confirmed experimentally that the approach might easily be adapted to multiple scenarios.

Chapter 6 concerns efficient automated mechanisms for processing video contents. The vast gap between what humans can comprehend based on cognition, knowledge, and experience, and what computer systems can obtain from signal processing, causes the subject very difficult. On the other hand, the increasing popularity and ubiquity of videos need efficient automated mechanisms for processing video contents. The spatiotemporal annotation of complex video scenes, in the form interpretable for machines, can be obtained by fusion of structured descriptions with textual and audio descriptors. This annotation can be used in scene interpretation, video understanding, and content-based video retrieval.

Chapter 7 focuses on how deep learning can be used in bridging the semantic gap in the content-based image retrieval. The chapter briefly presents the traditional approaches and introduces into deep learning, methods and deep models useful in CBIR. The authors distinguished three basic structure levels for scene interpretation using deep learning; they are feature level, common sense knowledge level, and inference level. The chapter presents the applications of deep learning at the particular levels of CBIR. Finally, the application deep models in bridging the semantic gap are summed in a table, and the growing popularity of DL in image analysis is shown.

1.3 Conclusion

The chapter provides some problems connected with a gap between automatic image interpretation and how human perceive the semantic content of an image. Steps of image processing from raw image to semantic image interpretation are presented. Each step influences the result of CBIR systems. From the semantic gap bridging point of view, the most interesting seems to be a knowledge level of image analysis. However, it strongly depends on the lower levels. A raw image reflects basic real-world properties. Features extracted from a raw image strongly influence the further process, and by this, the final results. Deep models are becoming increasingly popular and are rapidly developed. They deal with complicated tasks such as choosing the suitable set of features. Instead, they learn the feature. Deep models release a human from the need to define features and algorithms of image processing; they are worth developing.

References

1. Alzubaidi, M.A., Narrowing the semantic gap in natural images. In: 5th International Conference on Information and Communication Systems (ICICS), Irbid, 2014, pp. 1–6 (2014). https://doi.org/10.1109/IACS.2014.6841972

2. Alzubaidi, M.A.: A new strategy for bridging the semantic gap in image retrieval. Int. J. Comput. Sci. Eng. (IJCSE) **14**(1) (2017)
3. Jaimes, A., Christel, M., Gilles, S., Sarukkai, R., Ma, W.-Y.: Multimedia information retrieval: what is it, and why isn't anyone using it? In: Proceeding MIR 2005, Proceedings of the 7th ACM SIGMM International Workshop on Multimedia Information Retrieval, Hilton, Singapore, pp. 3–8 (2005)
4. Luke, K.-K, Liu, H.-L, Wai, Y.-Y., Wan, Y.-L., Tan, L.H.: Functional anatomy of syntactic and semantic processing in language comprehension. Hum. Brain Mapp. **16**(3), 133–145 (2002)
5. Luo, J., Crandall, D., Singhal, A., Boutell, M., Gray, R.T.: Psychophysical study of image orientation perception. Spat. Vis. **16**(5), 429457 (2003)
6. Friedrich R.M., Friederici A.D.: Mathematical logic in the human brain: semantics. PLoS ONE **8**(1), e53699 (2013). https://doi.org/10.1371/journal.pone.0053699
7. Rommers, J., Dijkstra, T., Bastiaansen, M.: Context-dependent semantic processing in the human brain: evidence from idiom comprehension. J. Cogn. Neurosci. **25**(5), 762–776 (2013)
8. Mitchell, D.J., Cusack, R.: Semantic and emotional content of imagined representations in human occipitotemporal cortex. Sci. Rep. **6**, 20232 (2016). https://doi.org/10.1038/srep20232
9. Tomasello, R., Garagnani, M., Wennekers, T., Pulvermller, F.: Brain connections of words, perceptions and actions: a neurobiological model of spatio-temporal semantic activation in the human cortex. Neuropsychol. **98**, 111–129 (2017)
10. Shrivastava, P., Bhoyar, K.K., Zadgaonkar, A.S.: Bridging the semantic gap with human perception based features for scene categorization. Int. J. Intell. Comput. Cybern. **10**(3), 387–406 (2017)
11. Colombino, T., Martin, D., Grasso, A., Marchesotti, L.: Reformulation of the semantic gap problem in content-based image retrieval scenarios. In: Lewkowicz, M. et al. (eds.) Proceedings of COOP 2010, Computer Supported Cooperative Work, Springer (2010)
12. Li, Y., Leung, C.H.C.: Multi-level semantic characterization and re-finement for web image search. Procedia Environ. Sci. **11**, 147–154 (2011). https://doi.org/10.1016/j.proenv.2011.12.023. (Elsevier Ltd.)
13. Li, X., Uricchio, T., Ballan, L., Bertini, M., M. Snoek, C.G., Bimbo, A.D.: Socializing the semantic gap: a comparative survey on image tag assignment, refinement, and retrieval. ACM Comput. Surv. **49**(1), 14 (2016)
14. Alzu'bi, A., Amira, A., Ramzan, N.: Semantic content-based image retrieval: a comprehensive study. J. Vis. Commun. Image Represent. **32**, 20–54 (2015)
15. Mesnil, G., Bordes, A., Weston, J., Chechik, G., Bengio, Y.: Learning semantic representations of objects and their parts. Mach. Learn. **94**(2), 281–301 (2014)
16. Singh, S., Sontakke, T.: An effective mechanism to neutralize the semantic gap in Content Based Image Retrieval (CBIR). Int. Arab J. Inf. Technol. **11**(2) (2014)
17. Montazer, G.A., Giveki, D.: Content based image retrieval system using clustered scale invariant feature transforms. Optik—Int. J. Light Electron Opt. **126**(18), 1695–1699 (2015)
18. Srivastava, P., Khare, A.: Integration of wavelet transform, Local Binary Patterns and moments for content-based image retrieval. J. Vis. Commun. Image Represent. **42**, 78–103 (2017)
19. Dong, H., Yu, S., Wu, C., Guo, Y.: Semantic Image Synthesis via Adversarial Learning. Accepted to ICCV 2017, Subjects: Computer Vision and Pattern Recognition (cs.CV), arXiv:1707.06873v1 [cs.CV] (2017)
20. Yasmin, M., Mohsin, S., Sharif, M.: Intelligent image retrieval techniques: a survey. J. Appl. Res. Technol. **12**(1), 87–103 (2014)
21. Khodaskar, A., Ladhake, S.: Semantic image analysis for intelligent image retrieval. Procedia Comput. Sci. **48**, 192–197 (2015)