The State of the Art in Image and Video Retrieval

N. Sebe[1], M.S. Lew[2], X. Zhou[3], T.S. Huang[4], and E.M. Bakker[2]

[1] University of Amsterdam, The Netherlands
nicu@science.uva.nl
[2] Leiden University, The Netherlands
{mlew, erwin}@liacs.nl
[3] Siemens Corporate Research, USA
xiang.zhou@scr.siemens.com
[4] University of Illinois at Urbana-Champaign, USA
huang@ifp.uiuc.edu

Image and video retrieval continues to be one of the most exciting and fastest-growing research areas in the field of multimedia technology. What are the main challenges in image and video retrieval? Despite the sustained efforts in the last years, we think that the paramount challenge remains bridging the semantic gap. By this we mean that low level features are easily measured and computed, but the starting point of the retrieval process is typically the high level query from a human. Translating or converting the question posed by a human to the low level features seen by the computer illustrates the problem in bridging the semantic gap. However, the semantic gap is not merely translating high level features to low level features. The essence of a semantic query is understanding the meaning behind the query. This can involve understanding both the intellectual and emotional sides of the human, not merely the distilled logical portion of the query but also the personal preferences and emotional subtones of the query and the preferential form of the results.

Another important aspect is that digital cameras are becoming widely available. The combined capacity to generate bits of these devices is not easy to express in ordinary numbers. And, at the same time, the growth in computer speed, disk capacity, and most of all the rapid expansion of the web will export these bits to wider and wider circles. The immediate question is what to do with all the information. One could store the digital information on tapes, CD-ROMs, DVDs or any such device but the level of access would be less than the well-known shoe boxes filled with tapes, old photographs, and letters. What is needed is that the techniques for organizing images and video stay in tune with the amounts of information. Therefore, there is an urgent need for a semantic understanding of image and video.

Creating access to still images is still hard problem. It requires hard work, precise modeling, the inclusion of considerable amounts of a priori knowledge and solid experimentation to analyze the contents of a photograph. Luckily, it can be argued that the access to video is somehow a simpler problem than access to still images. Video comes as a sequence, so what moves together most likely forms an entity in real life, so segmentation of video is intrinsically simpler than

a still image, at the expense of only more data to handle. So the potential to make progress on video in a semantic understanding is there.

Moving from images to video adds several orders of complexity to the retrieval problem due to indexing, analysis, and browsing over the inherently temporal aspect of video. For example, the user can pose a similarity based query of "find a video scene similar to this one." Responding to such a query requires representations of the image and temporal aspects of the video scene. Furthermore, higher level representations which reflect the structure of the constituent video shots or semantic temporal information such as gestures could also aid in retrieving the right video scene.

Several new paradigms have emerged along the themes of image and video understanding. Examples include semantic image video retrieval models, interactive retrieval paradigms, affective and emotional interaction, image and video retrieval based on human perception, human computer interaction issues in image and video retrieval, learning and relevance feedback strategies, and intelligent summaries.

In the proceedings, several papers touch upon the semantic problem and give valuable insights into the current state of the art. Dimitrova [1] summarizes the main research topics in automatic methods for high-level description and annotation. Enser and Sandom [2] present a comprehensive survey of the semantic gap issues in visual information retrieval. Their goal is to provide a better informed view on the nature of semantic information needed and on the representation and recovery of semantic content across the broad spectrum of image retrieval activity. Naphade and Smith [3] argue that semantic understanding of multimedia content necessitates models for semantic concepts, context, and structure. In this context, they propose a hybrid framework that can combine discriminant and generative models for structure and context. Another framework for video content understanding using context and multimedia ontologies is proposed by Jaimes et al [4]. They present an expert system that uses a rule-based engine, domain knowledge, visual detectors, and metadata. Detecting semantic concepts from video using temporal gradients and audio classification is proposed by Rautiainen et al [5]. Sanchez et al [6] analyze different ways of coupling the information from multiple visual features in the representation of visual contents using temporal models based on Markov chains. Similarly, Yan et al [7] present an algorithm for video retrieval that fuses the decisions of multiple retrieval agents in both text and image modalities. An integrated image content and metadata search and retrieval across multiple databases is presented by Addis et al [8]. Audio-assisted video scene segmentation for semantic story browsing is investigated by Cao et al [9].

Application-specific issues are discussed in several papers. Liu and Kender [10] present a new approach for content-analysis and semantic summarization of instructional videos, while Miura et al [11] investigate cooking videos application. Domain-specific information is also used by Lay and Guan [12] for the retrieval of artworks by color artistry concepts. The documentary video application is covered by Velivelli et al [13]. They observed that the amount of information

from the visual component alone was not enough to convey a semantic context but the audio-video fusion conveyed a much better semantic context. Audio-visual synchrony is also used by Nock et al [14] for speaker localization. The news video application is discussed by Pickering et al [15]. They describe their system which captures television news with accompanying subtitles and identifies and extracts news stories from the video. Lexical chain analysis is used then to provide a summary of each story and the important entities are highlighted in the text. The area of sports video is covered by Barcelo et al [16], Miyamori [17], and Baillie and Jose [18].

An important segment of papers discusses the human computer interaction issues in visual information retrieval. Cohen et al [19] present an evaluation of facial expression recognition techniques. They focus on the design of the classifiers used for emotion recognition in two types of settings: static and dynamic classification. Video retrieval of human interactions using model-based motion tracking and multilayer finite stare automata is discussed by Park et al [20]. User studies are performed by Goodrum et al [21] and Hughes et al [22]. Goodrum et al [21] study the search moves made by the users as they transition from one search state to another. Their goal is to identify patterns of search state transition used and the overall frequency of specific state transitions. Hughes et al [22] present an eyetracking study on how people view digital video surrogates. Their subjects were eyetracked to determine where, when, and how long they looked at text and image surrogates. The subjects looked at and fixated on titles and descriptions more than on the images. Also, most people used the text as an anchor from which to make judgments about the search results and used the images as confirmatory evidence for their selection. Sawahata and Aizawa [23] discuss the problems related to the indexing of personal video captured by a wearable imaging system. Accessing and organizing home videos present technical challenges due to their unrestricted content and lack of story line. In this context, Odobez et al [24] propose a spectral method to group video shots into scenes based on their visual similarity and temporal relations. Similarly, Mulhem and Lim [25] propose the use of temporal events for organizing and representing home photos using a structured document formalism and hence a new way to retrieve photos of an event using both image content and temporal context.

An important challenge in visual information retrieval comes from the dynamic interpretation of images under different circumstances. In other words, the perceptual similarity depends upon the application, the person, and the context of usage. Therefore, the machine not only needs to learn the associations, but also has to learn them on-line with a user in the loop. Several papers address learning and relevance feedback issues in image and video retrieval. Relevance feedback with multilevel relevance judgment is discussed by Wu et al [26]. They consider relevance feedback as an ordinal regression and present a relevance feedback scheme based on a support vector learning algorithm for ordinal regression. Similarly, a constructive learning algorithm-based RBF neural network for relevance feedback is proposed by Qian et al [27]. Several effective learning algorithms using global image representations are used for region-based image

retrieval by Jing et al [28]. Howe [29] has a closer look at boosting for image retrieval and classification. The author, performs a comparative evaluation of several top algorithms combined in two different ways with boosting. Learning optimal representation for image retrieval application is investigated by Liu et al [30]. The authors use a Markov chain Monte Carlo stochastic gradient for finding representations with optimal retrieval performance on given datasets. The selection of the best representative feature and membership assignment for content-based image retrieval is discussed in [31].

An overview of challenges for content-based navigation of digital video is presented by Smeaton and Over [32]. The authors present a summary of the activities in the TREC Video track in 2002 where 17 teams from across the world took part. A fast video retrieval technique under sparse training data is proposed by Liu and Kender [33]. Observing that the moving objects' trajectories play an important role in content-based retrieval in video databases, Shim and Chang [34] present an efficient similar trajectory-based retrieval algorithm for moving objects in video databases. A robust content-based video copy identification scheme dedicated to TV broadcast is presented in [35]. The recognition of similar videos is based upon local features extracted at interest points. Similarly, Shao et al [36] extract local invariant descriptors for fast object/scene recognition based on local appearance. Video similarity detection issues are also discussed in [37].

In addition, new techniques are presented for a wide range of retrieval problems, including object matching [38], shape-based retrieval [39], searching in large-scale image databases [40], hierarchical clustering in multidimensional index structures [41], k-d trees for database indexing [42], image retrieval based on fractal codes [43], as well as applications in areas of historical watermarks [44], trademarks [45], and web image retrieval [46].

In order for image and video retrieval to mature, we will need to understand how to evaluate and benchmark features, methods, and systems. An efficiency comparison of two content-based retrieval systems is presented in [47]. Similarly, a performance comparison between different similarity models for CBIR with relevance feedback is presented by Heesch et al [48].

References

1. Dimitrova, N.: Multimedia content analysis: The next wave. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2728, Springer (2003) 8–17
2. Enser, P., Sandom, C.: Towards a comprehensive survey of the semantic gap in visual image retrieval. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2728, Springer (2003) 279–287
3. Naphade, M., Smith, J.: A hybrid framework for detecting the semantics of concepts and context. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2728, Springer (2003) 188–197
4. Jaimes, A., Tseng, B., Smith, J.: Modal keywords, ontologies, and reasoning for video understanding. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2728, Springer (2003) 239–248

5. Rautianen, M., Seppanen, T., Pentilla, J., Peltola, J.: Detecting semantic concepts from video using temporal gradients and audio classification. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2728, Springer (2003) 249–258
6. Sanchez, J., Binefa, X., Kender, J.: Combining multiple features in temporal models for the representation of visual contents in video. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2728, Springer (2003) 208–217
7. Yan, R., Hauptmann, A., Jin, R.: Multimedia search with pseudo-relevance feedback. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2728, Springer (2003) 229–238
8. Addis, M., Boniface, M., Goodall, S., Grimwood, P., Kim, S., Lewis, P., Martinez, K., Stevenson, A.: Integrated image content and metadata search and retrieval across multiple databases. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2728, Springer (2003) 88–97
9. Cao, Y., Tavanapong, W., Kim, K.: Audio-assisted scene segmentation for story browsing. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2728, Springer (2003) 428–437
10. Liu, T., Kender, J.: Spatial-temporal semantic grouping of instructional video content. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2728, Springer (2003) 348–357
11. Miura, K., Hamada, R., Ide, I., Sakai, S., Tanaka, H.: Associating cooking video segments with preparation steps. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2728, Springer (2003) 168–177
12. Lay, J., Guan, L.: Concept-based retrieval of art documents. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2728, Springer (2003) 368–377
13. Velivelli, A., Ngo, C.W., Huang, T.: Detection of documentary scene changes by audio-visual fusion. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2728, Springer (2003) 218–228
14. Nock, H., Iyengar, G., Neti, C.: Speaker localisation using audio-visual synchrony: An empirical study. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2728, Springer (2003) 468–477
15. Pickering, M., Wong, L., Rüger, S.: ANSES: Summarisation of news video. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2728, Springer (2003) 408–417
16. Barcelo, L., Oriols, X., Binefa, X.: Spatio-temporal decomposition of sport events for video indexing. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2728, Springer (2003) 418–427
17. Miyamori, H.: Automatic annotation of tennis action for content-based retrieval by integrated audio and visual information. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2728, Springer (2003) 318–327
18. Baillie, M., Jose, J.: Audio-based event detection for sports video. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2728, Springer (2003) 288–297
19. Cohen, I., Sebe, N., Sun, Y., Lew, M., Huang, T.: Evaluation of expression recognition techniques. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2728, Springer (2003) 178–187

20. Park, S., Park, J., Aggarwal, J.: Video retrieval of human interactions using model-based motion tracking and multi-layer finite state automata. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2728, Springer (2003) 378–387
21. Goodrum, A., Bejune, M., Siochi, A.: A state transition analysis of image search patterns on the web. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2728, Springer (2003) 269–278
22. Hughes, A., Wilkens, T., Wildemuth, B., Marchionini, G.: Text or pictures? An eyetracking study of how people view digital video surrogates. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2728, Springer (2003) 259–268
23. Sawahata, Y., Aizawa, K.: Indexing of personal video captured by a wearable imaging system. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2728, Springer (2003) 328–337
24. Odobez, J.M., Gatica-Perez, D., Guillemot, M.: Spectral structuring of home videos. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2728, Springer (2003) 298–307
25. Mulhem, P., Lim, J.H.: Home photo retrieval: Time matters. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2728, Springer (2003) 308–317
26. Wu, H., Lu, H., Ma, S.: Multilevel relevance judgment, loss function, and performance measure in image retrieval. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2728, Springer (2003) 98–107
27. Qian, F., Zhang, B., Lin, F.: Constructive learning algorithm-based RBF network for relevance feedback in image retrieval. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2728, Springer (2003) 338–347
28. Jing, F., Li, M., Zhang, L., Zhang, H.J., Zhang, B.: Learning in region-based image retrieval. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2728, Springer (2003) 198–207
29. Howe, N.: A closer look at boosted image retrieval. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2728, Springer (2003) 58–67
30. Liu, X., Srivastva, A., Sun, D.: Learning optimal representations for image retrieval applications. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2728, Springer (2003) 48–57
31. Uysal, M., Yarman-Vural, F.: Selection of the best representative feature and membership assignment for content-based fuzzy image database. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2728, Springer (2003) 138–147
32. Smeaton, A., Over, P.: TRECVID: Benchmarking the effectiveness of information retrieval tasks in video. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2728, Springer (2003) 18–27
33. Liu, Y., Kender, J.: Fast video retrieval under sparse training data. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2728, Springer (2003) 388–397
34. Shim, C.B., Chang, J.W.: Efficient similar trajectory-based retrieval for moving objects in video databases. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2728, Springer (2003) 158–167

35. Joly, A., Frelicot, C., Buisson, O.: Robust content-based video copy identification in a large reference database. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2728, Springer (2003) 398–407
36. Shao, H., Svoboda, T., Tuytelaars, T., van Gool, L.: HPAT indexing for fast object/scene recognition based on local appearance. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2728, Springer (2003) 68–77
37. Hoi, C.H., Wang, W., Lyu, M.: A novel scheme for video similarity detection. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2728, Springer (2003) 358–367
38. Kim, S., Park, S., Kim, M.: Central object extraction for object-based image retrieval. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2728, Springer (2003) 38–47
39. Arica, N., Yarman-Vural, F.: A compact shape descriptor based on the beam angle statistics. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2728, Springer (2003) 148–157
40. Ye, H., Xu, G.: Fast search in large-scale image database using vector quantization. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2728, Springer (2003) 458–467
41. Chen, Z., Ding, J., Zhang, M., Tavanapong, W.: Hierarchical clustering-merging in multidimensional index structures. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2728, Springer (2003) 78–87
42. Scott, G., Shyu, C.R.: EBS k-d tree: An entropy balanced statistical k-d tree for image databases with ground-truth labels. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2728, Springer (2003) 448–457
43. Pi, M., Tong, C., Basu, A.: Improving fractal codes-based image retrieval using histogram of collage errors. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2728, Springer (2003) 118–127
44. Riley, K.J., Edwards, J., Eakins, J.: Content-based retrieval of historical watermark images: II - electron radiographs. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2728, Springer (2003) 128–137
45. Eakins, J., Riley, K.J., Edwards, J.: Shape feature matching for trademark image retrieval. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2728, Springer (2003) 28–37
46. Park, G., Baek, Y., Lee, H.K.: Majority based ranking approach in web image retrieval. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2728, Springer (2003) 108–117
47. Rummukainen, M., Laaksonen, J., Koskela, M.: An efficiency comparison of two content-based image retrieval systems, GIFT and PicSOM. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2728, Springer (2003) 478–487
48. Heesch, D., Yavlinski, A., Rüger, S.: Performance comparison of different similarity models for CBIR with relevance feedback. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2728, Springer (2003) 438–447