Challenges of Image and Video Retrieval

M.S. Lew[1], N. Sebe[1], and J.P. Eakins[2]

[1] LIACS Media Lab,
Leiden University, The Netherlands,
{mlew, nicu}@liacs.nl
[2] Institute for Image Data Research,
University of Northumbria at Newcastle, UK
john.eakins@unn.ac.uk

What use is the sum of human knowledge if nothing can be found? Although significant advances have been made in text searching, only preliminary work has been done in finding images and videos in large digital collections. In fact, if we examine the most frequently used image and video retrieval systems (i.e. www.google.com) we find that they are typically oriented around text searches where manual annotation was already performed.

Image and video retrieval is a young field which has its genealogy rooted in artificial intelligence, digital signal processing, statistics, natural language understanding, databases, psychology, computer vision, and pattern recognition. However, none of these parental fields alone has been able to directly solve the retrieval problem. Indeed, image and video retrieval lies at the intersections and crossroads between the parental fields. It is these curious intersections which appear to be the most promising.

What are the main challenges in image and video retrieval? We think the paramount challenge is 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.

In this proceedings, several papers [1][2][3][4][5][6][7][8] touch upon the semantic problem and give valuable insights into the current state of the art. Wang et al [1] propose the use of color-texture classification to generate a codebook which is used to segment images into regions. The content of a region is then characterize by its self-saliency which describes its perceptual importance. Bruijn and Lew [2] investigate multi-modal content-based browsing and searching methods for Peer2Peer retrieval systems. Their work targets the assumption

that keyframes are more interesting when they contain people. Vendrig and Worring [3] propose a system that allows character identification in movies. In order to achieve this, they relate visual content to names extracted from movie scripts. Denman et al [5] present the tools in a system for creating semantically meaningful summaries of broadcast Snooker footage. Their system parses the video sequence, identifies relevant camera views, and tracks ball movements. A similar approach presented by Kim et al [8] extracts semantic information from basketball videos based on audio-visual features. A semantic video retrieval approach using audio analysis is presented by Bakker and Lew [7] in which the audio can be automatically categorized into semantic categories such as explosions, music, speech, etc. A system for recognizing objects in video sequences is presented by Visser et al [6]. They use the Kalman filter to obtain segmented blobs from the video, classify the blobs using the probability ration test, and apply several different temporal methods, which results in sequential classification methods over the video sequence containing the blob. An automated scene matching algorithm is presented by Schaffalitzky and Zisserman [4]. Their goal is to match images of the same 3D scene in a movie. Ruiz-del-Solar and Navarrete [9] present a content-based face retrieval system that uses self-organizing maps (SOMs) and user feedback. SOMs were also employed by Oh et al [10], Hussain et al [11], and Huang et al [12] for visual clustering. A ranking algorithm using dynamic clustering for content-based image retrieval is proposed by Park et al [13]. A learning method using the AdaBoost algorithm and a k-nearest neighbor approach is proposed by Pickering et al [14] for video retrieval.

An overview of challenges for content-based navigation of digital video is presented by Smeaton [15]. The author presents the different ways in which video content can be used directly to support the navigation within large video libraries and lists the challenges that still remain to be addressed in this area. An insight into the problems and challenges of retrieval of archival moving imagery via the Internet is presented by Enser and Sandom [16]. The authors conclude that the combination of limited CBIR functionality and lack of adherence to cataloging standards seriously limits the Internet's potential for providing enhanced access to film and video-based cultural resources. Burke [17] describes a research project which applies Personal Construct Theory to individual user perceptions of photographs. This work presents a librarian viewpoint toward content-based image retrieval. A user-centric system for visualization and layout for content-based image retrieval and browsing is proposed by Tian et al [18].

An important segment of papers discusses the challenges of using different models of human image perception in visual information retrieval. Cox and de Jager [19] developed a statistical model for the image pair and used it to derive a minimum-error hypothesis test for matching. An online Bayesian formulation for video summarization and linking is proposed by Orriols and Binefa [20]. A model-based approach for detecting pornographic images is presented by Bosson et al [21]. Based on the idea that the faces of persons are the first information looked for in an image, Viallet and Bernier [22] propose a face detection system which automatically derives face summaries from a video sequence. A model for

edge-preserving smoothing is presented by Smolka and Plataniotis [23]. Their algorithm is based on the combined forward and backward anisotropic diffusion with incorporated time dependent cooling process. The authors report that their method is able to efficiently remove image noise while preserving and enhancing image edges. Ko and Byun [24] present a model for content-based image retrieval based on regions-of-interest and their spatial relationship. A vision-based approach to mobile robot localization that integrates an image retrieval system with Monte-Carlo localization is presented by Wolf et al [25]. A layered-based model for image retrieval is proposed by Qiu [26].

Object recognition and detection is one of the most challenging problems in visual information retrieval. Several papers present the advances in this area [27][28][29][30]. Obdržálek and Matas [27] present a method that supports recognition of objects under a very wide range of viewing and illumination conditions and is robust to occlusion and background clutter. A hierarchical shape descriptor for object-based retrieval is proposed by Leung and Chan [29]. Brucale, et al [30] propose a class of topological-geometrical shape descriptors called size functions. Sebe and Lew [28] investigate the link between different metrics and the similarity noise model in an object-based retrieval application. They conclude that the prevalent Gaussian distribution assumption is often invalid and propose a Cauchy model. Furthermore, they explained how to create a maximum likelihood metric based on the similarity noise distribution and showed that it consistently outperformed all of the analytical metrics. Based on the idea that the distribution of colors in an image provides a useful cue for image retrieval and object recognition, Berens and Finlayson [31] propose an efficient coding of three dimensional color distribution for image retrieval.

In addition, new techniques are presented for a wide range of retrieval problems, including 3-D object matching [32] and compressed-domain image searching [33], as well as applications in areas as diverse as videos of snooker broadcasts [5], images of historical watermarks [34], funeral monuments [35], and low quality fax images [36].

In order for image and video retrieval to mature, we will need to understand how to evaluate and benchmark features, methods, and systems. Several papers which address these questions are [37], [38], and [39]. While Sebe et al [39] perform an evaluation of different salient point extraction techniques to be used in content-based image retrieval, Black et al [38] propose a method for creation of a reference image set in which the similarity of each image pair is estimated using "visual content words" as a basis vector that allows the multidimensional content of each image to be represented with a content vector. The authors claim that the similarity measure computed with these content vectors correlates with the subjective judgment of human observers and provides a more objective method for evaluating and expressing the image content. Müller et al [37] compare different ways of evaluating the performance of content-based retrieval systems (CBIRSs) using a subset of the Corel images. Their aim is to show how easy it is to get differing results, even when the same image collection, CBIRS, and performance measures are used.

References

1. Wang, W., Song, Y., Zhang, A.: Semantics-based image retrieval by region saliency. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2383, Springer (2002) 27–35
2. de Bruijn, W., Lew, M.: Atomsnet: Multimedia peer 2 peer for file sharing. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2383, Springer (2002) 130–138
3. Vendrig, J., Worring, M.: Multimodal person identification in movies. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2383, Springer (2002) 166–175
4. Schaffalitzky, F., Zisserman, A.: Automated scene matching in movies. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2383, Springer (2002) 176–185
5. Denman, H., Rea, N., Kokaram, A.: Content based analysis for video from snooker broadcasts. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2383, Springer (2002) 186–193
6. Visser, R., Sebe, N., Bakker, E.: Object recognition for video retrieval. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2383, Springer (2002) 250–259
7. Bakker, E., Lew, M.: Semantic video retrieval using audio analysis. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2383, Springer (2002) 260–267
8. Kim, K., Choi, J., Kim, N., Kim, P.: Extracting semantic information from basketball video based on audio-visual features. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2383, Springer (2002) 268–277
9. Ruiz-del-Solar, J., Navarrete, P.: FACERET: An interactive face retrieval system based on self-organizing maps. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2383, Springer (2002) 149–156
10. Oh, K., Zaher, A., Kim, P.: Fast k-nn image search with self-organizing maps. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2383, Springer (2002) 288–297
11. Hussain, M., Eakins, J., Sexton, G.: Visual clustering of trademarks using the self-organizing map. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2383, Springer (2002) 139–148
12. Huang, J., Umamaheswaran, D., Palakal, M.: Video indexing and retrieval for archeological digital library, CLIOH. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2383, Springer (2002) 278–287
13. Park, G., Baek, Y., Lee, H.K.: A ranking algorithm using dynamic clustering for content-based image retrieval. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2383, Springer (2002) 316–324
14. Pickering, M., Rüger, S., Sinclair, D.: Video retrieval by feature learning in key frames. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2383, Springer (2002) 298–306
15. Smeaton, A.: Challenges for content-based navigation of digital video in the Físchlár digital library. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2383, Springer (2002) 203–215

16. Enser, P., Sandom, C.: Retrieval of archival moving imagery - CBIR outside the frame? In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2383, Springer (2002) 194–202
17. Burke, M.: Personal construct theory as a research tool for analysing user perceptions of photographs. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2383, Springer (2002) 363–370
18. Tian, Q., Moghaddam, B., Huang, T.: Visualization, estimation, and user-modeling for interactive browsing of image libraries. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2383, Springer (2002) 7–16
19. Cox, G., de Jager, G.: A linear image-pair model and the associated hypothesis test for matching. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2383, Springer (2002) 56–63
20. Orriols, X., Binefa, X.: Online bayesian video summarization and linking. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2383, Springer (2002) 325–334
21. Bosson, A., Cawley, G., Chan, Y., Harvey, R.: Non-retrieval: Blocking pornographic images. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2383, Springer (2002) 46–55
22. Viallet, J., Bernier, O.: Face detection for video summaries. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2383, Springer (2002) 335–342
23. Smolka, B., Plataniotis, K.: On the coupled forward and backward anisotropic diffusion scheme for color image enhancement. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2383, Springer (2002) 64–73
24. Ko, B., Byun, H.: Multiple regions and their spatial relationship-based image retrieval. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2383, Springer (2002) 74–83
25. Wolf, J., Burgard, W., Burkhardt, H.: Using an image retrieval system for vision-based mobile robot localization. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2383, Springer (2002) 102–111
26. Qiu, G., Lam, K.M.: Spectrally layered color indexing. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2383, Springer (2002) 93–101
27. Š. Obdržálek, Matas, J.: Local affine frames for image retrieval. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2383, Springer (2002) 307–315
28. Sebe, N., Lew, M.: Robust shape matching. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2383, Springer (2002) 17–26
29. Leung, M.W., Chan, K.L.: Object-based image retrieval using hierarchical shape descriptor. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2383, Springer (2002) 157–165
30. Brucale, A., d'Amico, M., Ferri, M., Gualandri, M., Lovato, A.: Size functions for image retrieval: A demonstrator on randomly generated curves. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2383, Springer (2002) 223–232
31. Berens, J., Finlayson, G.: An efficient coding of three dimensional colour distributions for image retrieval. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2383, Springer (2002) 233–240

32. de Alarcón, P., Pascual-Montano, A., Carazo, J.: Spin images and neural networks for efficient content-based retrieval in 3D object databases. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2383, Springer (2002) 216–222
33. Feng, G., Jiang, J.: JPEG image retrieval based on features from DCT domain. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2383, Springer (2002) 112–120
34. Riley, K., Eakins, J.: Content-based retrieval of historical watermark images: I-Tracings. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2383, Springer (2002) 241–249
35. Howell, A., Young, D.: Image retrieval methods for a database of funeral monuments. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2383, Springer (2002) 121–129
36. Fauzi, M., Lewis, P.: Query by fax for content-based image retrieval. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2383, Springer (2002) 84–92
37. Müller, H., Marchand-Maillet, S., Pun, T.: The truth about Corel - Evaluation in image retrieval. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2383, Springer (2002) 36–45
38. Black, J., Fahmy, G., Panchanathan, S.: A method for evaluating the performance of content-based image retrieval systems based on subjectively determined similarity between images. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2383, Springer (2002) 343–352
39. Sebe, N., Tian, Q., Loupias, E., Lew, M., Huang, T.: Evaluation of salient point techniques. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2383, Springer (2002) 353–362