Available online at www.sciencedirect.com

ScienceDirect

Computer Vision and Image Understanding 110 (2008) 309–311

Computer Vision and Image Understanding

www.elsevier.com/locate/cviu

Guest Editorial

# Similarity Matching in Computer Vision and Multimedia

Comparing two images, or an image and a model, is the fundamental operation for any retrieval systems. The similarity matching of two images can reside in the hierarchical levels from pixel-by-pixel level, feature space level, object level, and semantic level. In most systems of interest, a simple pixel-by-pixel comparison will not suffice: the difference that we determine must bear some correlation with the perceptual difference of the two images or with the difference between two adequate semantics associated to the two images. Similarity matching techniques are developed mostly for recognition of objects under several conditions of the distortion while similarity measures, on the other hand, are used in applications like image databases. Matching and dissimilarity measurement are not seldom based on the same techniques, but they differ in emphasis and applications.

Similarity judgments play a central role in theories of human knowledge representation, behavior, and problem solving and as such they are considered to be a valuable tool in the study of human perception and cognition. Tversky [12] describes the similarity concept as “an organizing principle by which individuals classify objects, form concepts, and make generalizations.”

Retrieval by similarity, i.e. retrieving images which are similar to an already retrieved image (retrieval by example) or to a model or schema, is a relatively old idea. From the start it was clear that retrieval by similarity called for specific definitions of what it means to be similar. Smeulders et al. [11] discuss several types of similarity that need to be considered when one is analyzing a pair of images: similarity between features, similarity of object silhouettes, similarity of structural features, similarity of salient features, and similarity at the semantic level. Summarizing their ideas (see also the survey of Datta, et al [4]) one can identify three components that typically every system for retrieval by similarity needs to have:

- Extraction of features or image signatures from the images, and an efficient representation and storage strategy for this precomputed data.
- A set of similarity measures, each of which captures some perceptively meaningful definition of similarity, and which should be efficiently computable when matching an example with the whole database.
- A user interface for the choice of which definition of similarity should be applied for retrieval, presentation of retrieved images, and for supporting relevance feedback.

Gudivada [9] has listed different possible types of similarity for retrieval: color similarity, texture similarity, shape similarity, spatial similarity, etc. Some of these types can be considered in all or only part of one image, can be considered independently of scale or angle or not, depending on whether one is interested in the scene represented by the image or in the image itself.

Representation of features of images - like color, texture, shape, motion, etc. - is a fundamental problem in visual information retrieval. Image analysis and pattern recognition algorithms provide the means to extract numeric descriptors which give a quantitative measure of these features. Computer vision enables object and motion identification by comparing extracted patterns with predefined models.

The research in the area has made evident that (see also [10] and [14]):

- A large number of meaningful types of similarity can be defined. Only part of these definitions are associated with efficient feature extraction mechanisms and (dis)similarity measures.
- Since there are many definitions of similarity and the discriminating power of each of the measures is likely to degrade significantly for large image databases, the user interaction and the feature storage strategy components of the systems will play an important role.
- Visual content based retrieval is best used when combined with the traditional search, both at user interface and at the system level. The basic reason for this is that content based retrieval is not seen as a replacement of parametric (SQL), text, and keywords search. The key is to apply content based retrieval where appropriate, which is typically where the use of text and keywords is suboptimal. Examples of such applications are where visual appearance (e.g. color, texture, motion) is the primary attribute as in stock photo/video, art, etc.

Most of the state of the art research follow these lines and the papers selected for this special issue make no separate

doi:10.1016/j.cviu.2008.04.001

opinion in this respect. In the following we describe briefly the contribution of each accepted paper in the special issue and we conclude with several remarks on the current state of the art in similarity matching and some current trends.

## 1. Papers in the special issue

The eight papers selected for the *Similarity Matching in Computer Vision and Multimedia* special issue of *Computer Vision and Image Understanding* follow roughly the research lines presented in the previous section. Several articles discuss low-level similarity involving shape [5,7], interest points [2], or motion features [1]. Others proposed algorithms for learning the similarity measures for image alignment [3], visual data clustering and retrieval [6], and image category retrieval [8]. Finally, multiple modalities such as text and visual information are employed in cross-lingual broadcast news analysis [13].

- Demirci et al. [5] address the issue of indexing multimedia databases by representing the entries in the database as graphs structures. They represent the topology of the graph and its corresponding subgraphs as vectors whose components correspond to their sorted laplacian eigenvalues. The novelty comes from the powerful representation of the graph topology and the successful adoption of the concept of spectral integral variation in an indexing algorithm. To evaluate their approach the authors perform a number of experiments using an extensive set of recognition trials in the domain of 2D and 3D object recognition showing robustness and efficacy of the proposed solution compared to two different graph-based object representations.
- Shape matching algorithms using skeletons are proposed in [7] which describes several strategies that can be employed to improve the performance of existing part-based shape description and matching algorithms. The author advocates that the ligature-sensitive information should be incorporated into the part decomposition and shape matching process. Additionally, the part decomposition should be treated as a dynamic process in which the selection of the final decomposition of a shape is deferred until the shape matching stage. Finally, both local and global measures should be considered when computing the shape dissimilarity and the skeletal segments must be weighted accordingly. The experimental results show that shape-based retrieval can be significantly improved by incorporating these strategies.
- The task of finding point correspondences between two images of the same scene or object is part of many computer vision applications. In this context, Bay et al. [2] present a novel scale- and rotation-invariant detector and descriptor, coined SURF, which not only can be very efficiently computed but also has comparable performance compared to other existing schemes with respect to repeatability, distinctiveness, and robustness. The framework is tested in two challenging applications: camera calibration treated as a special case of image registration and object recognition.
- A novel framework for matching video sequences using the spatiotemporal segmentation of videos is presented by Basharat et al. [1]. Point trajectories are computed using the SIFT operator and then these trajectories are clustered to form motion segments by analyzing their motion and spatial properties. The temporal correspondence between the estimated motion segments is established based on the common SIFT correspondences. Spatiotemporal volumes are extracted using the consistently tracked motion segments and subsequently, low-level features including color, texture, motion, and SIFT descriptors are extracted to represent a volume. The experiments run on a variety of videos show promising results and prove the effectiveness of the proposed framework.
- Brooks et al [3] note that the need to align or register two images is one of the basic problems of computer vision. This process is often optimizing a similarity measure between images but this may prove too slow in time-critical applications. This observation points to the need of tradeoffs between the amount of computation and the accuracy of the results which has been known in the field of real-time artificial intelligence as deliberation control problem. This paper presents the anytime algorithm framework for optimization of two common similarity measures: mean-squared difference and mutual information. When tested against existing techniques, the proposed method achieves comparable quality and robustness with significantly less computation.
- The idea of similarity metric learning for visual data clustering and retrieval is addressed by Fu et al. [6]. The authors propose a Locally Embedded Analysis (LEA) framework for clustering and retrieval which reveals the essential low-dimensional manifold structure of the data by preserving the local nearest neighbor affinity and allowing a linear subspace embedding through solving a graph embedded eigenvalue decomposition problem. Based on the LEA approach, the authors propose an algorithm for visual data clustering and another one for local similarity metric learning for robust video retrieval. Simulation results demonstrate the effective performance of the proposed solutions in both accuracy and speed aspects.
- Gosselin et al. [8] present a search engine architecture aiming at retrieving complex categories in large image databases. The authors present a scheme based on two-step quantization process for computing visual codebooks. The similarity between images is represented in a kernel fashion. Experiments with a real scenario applied on the Corel database demonstrate the efficiency and the relevance of the proposed architecture.
- Different measures of novelty and redundancy detection for cross-lingual news stories are presented in [13]. A news story is represented by multimodal features which include a sequence of keyframes in the visual track, and

a set of words and named entities extracted from the speech transcript in the audio track. Vector space models and language models on individual modalities are constructed to compare the similarities among stories. Furthermore, the multiple modalities are further fused to improve performance. Experiments on the TRECVID-2005 crosslingual news video corpus show that modalities and measures demonstrate variant performance for novelty and redundancy detection.

## 2. Concluding remarks

Several major problems are currently addressed by state of the art research for similarity matching in computer vision and multimedia but more efforts are needed in the following directions:

- Study of the distribution of measures for various feature spaces on large real-world sets of image. In particular, how well is the perceptive similarity order preserved by the measure when the number of images/videos grows?
- Study of ranking visual items that correspond to human perception.
- Definition of methods for the segmentation of images in homogeneous regions for various feature spaces, and definition of models of this spatial organization which could be robustly combined with the similarity of the local features.
- Detection of salient features to a type of images or objects, so that to free the user from specifying a particular set of features in query process.
- Combination of multiple visual features in image query and search.
- Developing efficient indexing schemes based on image similarity features for managing large databases. It has been shown that traditional database indexing techniques such as using R-trees fail in the context of content based image search. Therefore, ideas from statistical clustering, multi-dimensional indexing, and dimensionality reduction are extremely useful in this area.

Apart from these issues, extraction and matching of higher (semantic) level image/video attributes (such as recognition of object, human faces, and actions) are perhaps the most challenging tasks. Only when the features extracted at both these levels are combined, can similarity-based indexes be built.

In addition, to the success of the field, formalization of the whole paradigm of visual similarity is essential. Without this formalism it will be hard to develop sufficient reliable and mission critical applications that are easy to program and evaluate. Some early applications may be implemented without such a rigorous formalism, but the progress in the field will require full understanding of the basic requirements in visual similarity.

## References

[1] A. Basharat, Y. Zhai, M. Shah, Content based video matching using spatiotemporal volumes, Computer Vision and Image Understanding 110 (3) (2008) 360–377.

[2] H. Bay, A. Ess, T. Tuytelaars, L. Van Gool, Speeded-up robust features (SURF), Computer Vision and Image Understanding 110 (3) (2008) 346–359.

[3] R. Brooks, T. Arbel, D. Precup, Anytime similarity measures for faster alignment, Computer Vision and Image Understanding 110 (3) (2008) 378–389.

[4] R. Datta, D. Joshi, J. Li, J. Wang, Image retrieval: Ideas, influences, and trends of the new age, *ACM Computing Surveys*, accepted, 2008.

[5] M.F. Demirci, R.H. van Leuken, R.C. Veltkamp, Indexing through laplacian spectra, Computer Vision and Image Understanding 110 (3) (2008) 312–325.

[6] Y. Fu, Z. Li, T.S. Huang, A.K. Katsaggelos, Locally adaptive subspace and similarity metric learning for visual data clustering and retrieval, Computer Vision and Image Understanding 110 (3) (2008) 390–402.

[7] W-B. Goh, Strategies for shape matching using skeletons, Computer Vision and Image Understanding 110 (3) (2008) 326–345.

[8] P.H. Gosselin, M. Cord, S. Philipp-Foliguet, Combining visual dictionary, kernel-based similarity and learning strategy for image category retrieval, Computer Vision and Image Understanding 110 (3) (2008) 403–417.

[9] V.N. Gudivada, V. Raghavan, Design and evaluation of algorithms for image retrieval by spatial similarity, ACM Transactions on Information Systems 13 (2) (1995) 115–144.

[10] N. Sebe, M.S. Lew, D.P. Huijsmans, Toward improved ranking metrics, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (10) (2000) 1132–1141.

[11] A. Smeulders, M. Worring, S. Santini, A. Gupta, R. Jain, Content based image retrieval at the end of the early years, IEEE Transactions on Pattern Analysis and Machine Intellingence 22 (12) (2000) 1349–1380.

[12] A. Tversky, Features of similarity, Psychological Review 84 (4) (1977) 327–352.

[13] X. Wu, A.G. Hauptmann, C.-W. Ngo, Measuring novelty and redundancy with multiple modalities in cross-lingual broadcast news, Computer Vision and Image Understanding 110 (3) (2008) 418–431.

[14] J. Yu, J. Amores, N. Sebe, P. Radeva, Q. Tian, Distance learning for similarity estimation, IEEE Transactions on Pattern Analysis and Machine Intellingence 30 (3) (2008) 451–462.

Nicu Sebe
*Faculty of Science, University of Amsterdam,*
*Kruislaan 403, 1098 SJ Amsterdam, The Netherlands*
*E-mail address:* nicu@science.uva.nl

Qi Tian
*University of Texas at San Antonio,*
*San Antonio, TX 78249-1644, USA*
*E-mail address:* qitian@cs.utsa.edu

Michael S. Lew
*LIACS Media Lab, Leiden University,*
*Niels Bohrweg 1, 2333 CA, Leiden, The Netherlands*
*E-mail address:* mlew@liacs.nl

Thomas S. Huang
*Beckman Institute, University of Illinois*
*at Urbana-Champaign, 405 N. Mathews Ave.,*
*Urbana, IL 61801*
*E-mail address:* huang@ifp.uiuc.edu