In: J Schneider, T Strothotte, W, Marotzki, (Hg.), Proc Workshop Computational Visualistics, Media Informatics, and Virtual Communities,. Deutscher Universitäts-Verlag. Reihe Bildwissenschaft. 2003. Bd 11. S. 91-97

The representation of shape for retrieval of pictures by semantic means

Klaus D. Toennies[1], Klemens Boehm[2], Christoph Herrmann[3], Ingo Schmitt[4]

[1]AG Bildverarbeitung/Bildverstehen, ISG, Fakultät für Informatik, Otto-von-Guericke-Universität Magdeburg

[2]AG Data and Knowledge Engineering, ITI, Fakultät für Informatik, Otto-von-Guericke-Universität Magdeburg

[3]Lehrstuhl Biologische Psychologie, Fakultät für Naturwissenschaften, Otto-von-Guericke-Universität Magdeburg

[4]AG Datenbanken, ITI, Fakultät für Informatik, Otto-von-Guericke-Universität Magdeburg

Abstract

Content-based image retrieval (CBIR) lets the user search an image database by means of image features rather than keywords. Although techniques for searching the database are quite advanced, the success of a CBIR system is still limited because few methods exists that close the gap between the semantics of the request and the features into which the request is to be translated. Progress can be expected, if knowledge about rules of human vision in the search and recognition of objects and object features in pictures are used for defining an image representation. It must be possible to extract this representation and it must be capable of integratuing semantic information from user feedback during search of the image database.

Introduction

Content-based image retrieval (CBIR) is intended to relieve the user of the burden to index data base entries. Indexing of pictures has long been recognised as a tedious and error-prone task [Bess1990] which should be replaced by a search by image features. First publications appeared about 10 years ago and research interest increased ever since (see [Velt2002] for a current evaluation of 58 CBIR systems). It is a difficult task as it requires the ability of posing questions about objects or even abstract concepts to a picture of which the data is organised as an array of pixels with intensity or colour being its only feature.

The purpose of this paper is to present approaches to close the gap between semantics expressed by a request and the primitive type of features on which this request has to be mapped for carrying it out. We advocate an approach of content-based image retrieval, where complex relationships between parts of a picture are expressed in the image representation enabling a more direct mapping of requested attributes on features represented.

Content-Based Image Retrieval

Eakins et al. [Eaki1999] identify three levels of requests in their review on the state of the art in content-based image retrieval:

- Requests of the first kind (**primitive requests**) use data-based features, such as histograms, colour distributions or texture features for finding similar images. A simple request of this kind would, e.g., to search for images which have a blue region in the upper third of the picture. There is no differentiation between different semantics of this region.
- Semantics play a role in requests of the second kind (**semantic requests**). Features of the image need to be combined with a-priori-knowledge about how these features are to be interpreted in the context of the request. An example for such a request would be that for pictures containing mountains. It will not be possible to search the data base sucessfully for such images without information about how mountains are mapped into pictures.
- Requests of the third kind (**abstract requests**) search for abstract entities which are expressed in the picture but for which the picture is just an exemplified mapping or a part of it. An example for such a request would be the search for pictures that express romantic feelings.

A request in CBIR is not posed to the image itself but to a representation of features of the image. Features are often representated by a feature vector. Quantifyable attributes such as histograms of colour, grey scale or texture features are entries of the feature vector. Features may be weighted and weights may be changed during the search. The search for a picture becomes a task of identifying locations in feature space where the requested picture is to be found. Requests can be further specified by giving positive or negative feedback on pictures found.

Most CBIR systems address requests of the first kind. They are able, for instance, to search for images of a given colour and / or texture distribution. The reason for this is simple: Any system, being automatic or feedback-driven is only able to pose a request in terms of the features that are represented in or derived from the picture. Features, such as the ones mentioned above, have only a very indirect relationship to the kind of attributes of object that the user is looking for.

A mapping of semantics of, e.g. a mountain into primitive features such as a colour distribution will necessarily be fuzzy. However, features of that kind are easy to generate and request by relevance feedback enables the user to identify arbitrary locations in feature space. Weights for the features may be produced, which optimally adapt the request to extracted features. Feedback then provides a-priori-knowledge, which is necessary for a request of the second or third kind. However, in general the relation between semantic features (attributes of a mountain) and primitive features (distribution of colours and/or textures) is ambiguous. Primitive feature values may have multiple meanings (something blue may be sky or water) and an object may be pictured in multiple ways. This is known in the CBIR community as the semantic gap. It can be assumed that this ambiguity increase with the distance between semantic and primitve features. Thus, we can never be sure that we have found all the pictures that show the desired object because not all the different primitive feature value combinations may have been found, which represent images containing this object. Furthermore, images containing different objects with a similar set of feature values may not always be classifyable as being different based on the features.

Shape and Visual Perception

Features for retrieval of images are often too simple for describing the semantics of a request by the representation itself. Retrieval systems rely on the assumption that a combination of such features will describe higher semantical concepts even though the features do not. While this may be true in some cases, one can easily construct a case where it is either not possible to combine desired features such as local colour or texture distributions for uniquely representig a specific meaning or it is possible but small pertubations in the image features (e.g. a shading) cause them no longer to represent the desired object while the picture still depicts the same object.

Ideally, the expressive power of a feature-based representation of image content should be directly related to the meaning of a request but this would make it dependent on the image content. An appropriate representation of features of mountains in images, for instance, would be derived from our knowledge about how mountains are mapped into pictures. It may be of little use to deduce perceptual similarity images depicting different objects (e.g., different types of apples). Therefore, we would like to have an object-idependent representation serving as a container for attributes in an image that are deemed to be relevant for describing similarity. User input then teaches such a representation the terms of perceptual similarity in a specific case. For the example above, this means that a possibly complex similarity criterion for mountains is developed through user feedback on retrieved images. However, such relevance feedback can teach similarity only in terms of a pre-defined similarity measure on the representation of image features.

Current picture descriptions often use just the above-mentioned simple features grouped into a vector on which a (weighted) norm is defined for similarity. Of all the different primitive features, shape is used the least in content-based image retrieval. Shape can be characterised fairly easy (shape is the outline of a structure) but it is difficult to represent shape in such a way that perceptually similar shapes are close to each other based on some distance metric in feature space.

On the other hand, shape has long been recognised as a an important feature for describing and differentiating objects in pictures. Humans perceive only certain details in pictures which automatically receive our attention [Para1998]. Furthermore, features which are remembered are influenced by the context of a picture [Ande1995]. Even at an age of about 8 months, elektrophysiological responses can be measured in infants when they perceive "good" shapes, ie. shapes that follow the Gestalt laws [Csib2000]. If pictures shall be retrieved in a data base based on image content, it is important that the user may use criteria which are perceived and remembered in pictures. A representation needs to be extracted from the picture which is not only capable to represent shape, colour, texture etc. but also the extent as to which such features and carriers of such features follow knowledge about perception and perceptual similarity as it was found that it is important for human processing of pictorial information that parts of the pictures are combined to a coherent object representation [Herr2001].

The representation of shape

If shape is going to be a container for knowledge taught by user feedback then it should try to incorporate as much as possible of the relation between shape and perceptual similarity. Two-dimensional and three-dimensional shape representations in image analysis have been used for quite some time for classification purposes and CBIR may be viewed as a classification task. In the following, we will consider only two-dimensional shape representations but most conclusions of it may be extended to 3-d representations.

Shape representations should be in some way deformable in order to enable them to adapt generic shapes to given shapes in the image. Various deformable shape models have been developed in recent years and have been used for segmentation, motion tracking, reconstruction and comparison between shapes. These models can be broadly classified into three classes:

- Statistical models that use a-priori knowledge about how the shape varies to reconstruct that shape.
- Dynamical models that fit the shape to the data using built-in smoothness constraints to maintain an optimal solution.
- Structural models, which extract structural features from shapes to compare and classify them.

The prime example for a representation of the first class are Active Shape and Active Appearance Models developed by Cootes et. al [Coot2002], which utilize principal component analysis in order to describe variations of landmarks and textures. Another candidate is the probabilistic registration by Chen [Chen1999] that uses the per-voxel gray level and shift vector distributions to guide a better fit between a grey-level atlas representing expected image content and the data. Smoothness constraints between neighboring shift vectors improve the results. The main restriction in statistical models is that they describe the statistical variations of a fixed-structure shape but not structural differences between different shapes. If an object, for instance, consists of two parts, then this fact is not described in the representation although the shape of this object may very well be rerpesentable. If one part is missing then this may be perceived by the representation as a mere deformation which would have the same qualtity as a deformation of one of the parts (e.g., due to measurement errors) although the perceptual difference may be much greater. Without this missing part the shape may even be a different object altogether.

Examples of the second class are the front propagation methods by Malladi et. al [Mall1995], which simulate an expanding closed curve that eventually fits the shape, or the dynamic particles by Szeliski et. al [Szel1993], which simulate a system of dynamic oriented particles, which expand into the object surface guided by internal forces that maintain an even and smooth distribution between them. These deformable models are able to segment and sample objects of complex topology like blood vessels. Their restriction is that they cannot characterise the shapes either statistically or structurally. For a CBIR application this would mean that a shape may be found in the image given enough support from the data but that a similarity measure is not part of the representation. These kinds of representations are more suitable to find shapes as opposed to classify them.

Examples of the third class are the shock grammar by Siddiqi et. al [Sidd1996] or the finite element method of Pentland et al.[Pent1996]. The shock grammar defines four types of shocks, which are evolving medial axes formed from colliding propagating fronts that originate at shape boundaries. This model defines a shock grammar that restricts how the shock types can combine to form a shape. The grammar is used to eliminate invalid shock combinations. The shock graphs that describe a shape facilitate comparison between shapes. The method of Pentland et al. define a dynamic finite element model that fits the shape. The low order modal coordinates describes the object structure under its free vibration modes. A simple dot product of the modal vectors of two shapes is a strong discriminator of their structural differences. Other examples for this kind of shape representation are the super quadrics by Terzopoulos et. al [Terz1991] or shape blending by DeCarlo et. al [DeCa1998]. All models of this class are data driven in that they have no prior knowledge about the structures of the shapes they fit. This is advanageous for a shape representation in CBIR because knowledge about objects shall be incorporated through relevance feedback. However, they also cannot describe the shapes they fit statistically which makes a similarity measure difficult to install.

An intermediate solution is the active structural shape model [AlZu2002] which combines structural properties of the representations of the third class with the ability of representing and learning statistical variations of shape. However, even this representation is not able to learn structural features for classification.

A representation for shape-based CBIR

At present there exists no shape representation that suits all purposes in an CBIR application. However, the examples given above show potential to fulfill at least some of the properties. Ideally, a shape representation in CBIR should have the following features:

- A shape representation generated from an existing picture should be a combination of a classification of an object by shape and variations due to measurement or other permissable variation. The representation should facilitate separation of these different aspects. Structo-statistical representations could be a suitable solution because structure could be assumed to be mainly class-specific whereas statistical variations reflect shape changes due to permissable object variations.

- Shape features should be adaptable to specifications from user feedback such that it agrees with knowledge about perceptual similarity. By this we mean that forseeable groupings by the user according to Gestalt laws are reflected in high level features of the shape representation. Structural representation such as those of the third group may best fit this requirement.
- Shape should be separable into structural units which may be influenced by neighbouring units or underlying structural units. Separate treatment of the shape units by user request should be possible for enabling independent similarity measure for different parts of the shape description. E.g., a tree remains to be a tree if some branches are missing but it would be difficult to imagine a tree without a trunk. Thus these units need to be treated differently if trees are searched for.
- Topological and geometrical variations of a shape representation need to be separable. This requirement is probably very difficult to meet because many shapes can be thought of a geometric deformation of some base shape or a construction of different topological units. The cipher 3, for instance, can be thought of as a deformation of a line or the combination of two cups. Depending of application, either of the two representations may be the more appropriate. Which is appropriate in a given request for a shape will be needed to be learnt from user feedback.

A shape representation with all these properties would provide for a container of user input based on his / her perception of shape. It can be expected that a request based on such a representation would be more precise and goal-oriented than one based on more primitive shape features. However, features of such complexity will no longer be representable as a simple vector of values. Structural relationships, topological and geometrical properties and a hierarchy of shape will constitute the shape description. New learning methods as well as appropriate similarity measures need to be explored.

Conclusions

Content-based image retrieval is an important aspect for any world, society or organisation relying on digital images as an important source of information because the ability to produce pictorial information exceeds the ability to retrieve this information by far. Content-based image retrieval is the attempt to free the retrieval task from dependencies on the image-gathering person as images are retrieved based on content rather than based on a pre-specified description of content. However, the relation between image content and image information representation is currently not understood satisfactorily. Representations automatically generated from an image do not possess enough expressive power to accommodate user input when he or she searches for images with a specific content. Developing, understanding and using an appropriate representation requires an interdisciplinary research effort of experts in human image interpretation, in database management and computer vision for covering all aspects of algorithmitically describing aspects of human perception of image information for the purpose of search an image database.

References

[AlZu2002] S. Al-Zubi, K.D.Toennies. Extending active shape models to incorporate a-priori knowledge about structural variability. LNCS, Vol.2449 (*Pattern Recognition, 24rd DAGM Symposium)*, Springer-Verlag, 2002, 338-344.

[Ande1995] J.R. Anderson. Cognitive psychology and its implications, New York: W.H. Freeman, 1995.

[Bess1990] H. Besser. Visual access to visual images: the UC Berkely image database project. Library Trends, Vol. 38(4), 1990, 787-798.

[Coot2001] T. Cootes, C. Taylor. Statistical Models of Appearance for Medical Image Analysis and Computer Vision. Proceedings of SPIE (Medical Imaging 2001: Image Processing), Vol. 4322, 2001, 236-248.

[Chen1999] M. Chen. 3-D Deformable Registration Using a Statistical Atlas with Applications in Medicine. Proc. MICCAI, 1999, 621-630.

[Csib2000] G. Csibra, G. Davis, M.W. Spratling & M.H. Johnson. Gamma oscillations and object processing in the infant brain, Science 290(5496), 2000, 1582-1585.

[DeCa1998] D. DeCarlo, D. Metaxas. Shape Evolution with Structural and Topological Changes using Blending. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 20(11), 1998, 1186-1205.

[Eaki1999] J.P. Eakins, M.E. Graham. Content-based image retrieval: A report to the JISC technoilogy applications programme. 1999, http://www.unn.ac.uk/iidr/report.html.

[Herr2001] C.S. Herrmann, A.D. Friederici. Object processing in the infant brain, Science, 292, 2001, p.163.

[Mall1995] R. Malladi, J. Sethian, B. Vemuri. Shape Modeling with Front Propagation: A Level Set Approach. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 17(2), 1995 158-175.

[Para1998] R. Parasuraman. The attentive brain, Boston: MIT press, 1998.

[Pent1996] A. Pentland, R. Picard, S. Sclaroff. Photobook: Tools for content –based manipulation of image databases. Intl. J Computer Vision, 18(3), 1998, 233-254.

[Sidd1996] K. Siddiqi, B. Kimia. Toward a Shock Grammar for Recognition. IEEE Conf. on Computer Vision and Pattern Recognition, 1996.

[Szel1993] R. Szeliski, D. Tonnesen, D.Terzopoulos. Modeling Surfaces of Arbitrary Topology with Dynamic particles. Proc. Computer Vision and Vision Recognition (CVPR), 1993, 82-87.

[Terz1991] D. Terzopoulos, D. Metaxas. Dyanamic 3D Models with Local and Global Deformations: Deformable Superquadrics. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 13(7), 1991, 703-714.

[Velt2002] R. Veltkamp, R. Tanase. Content-Based Image Retrieval Systems: A Survey. Tech. Report. Department of Computer Science, Utrecht University. 2000, (revised version 2002) http://give-lab.cs.uu.nl/cbirsurvey.