PERGAMON

Available at
www.ElsevierComputerScience.com
POWERED BY SCIENCE DIRECT®

Pattern Recognition 38 (2005) 919–934

PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY
www.elsevier.com/locate/patcog

A Bayesian network-based framework for semantic image understanding

Jiebo Luo[a,*], Andreas E. Savakis[b], Amit Singhal[a]

[a]Research and Development Laboratories, Eastman Kodak Company, 1850 Dewey Ave., Rochester, NY 14650-1816, USA
[b]Department of Computer Engineering, Rochester Institute of Technology, 83 Lomb Memorial Dr., Rochester, NY 14623, USA

Received 9 August 2004

Abstract

Current research in content-based semantic image understanding is largely confined to exemplar-based approaches built on low-level feature extraction and classification. The ability to extract both low-level and semantic features and perform knowledge integration of different types of features is expected to raise semantic image understanding to a new level. Belief networks, or Bayesian networks (BN), have proven to be an effective knowledge representation and inference engine in artificial intelligence and expert systems research. Their effectiveness is due to the ability to explicitly integrate domain knowledge in the network structure and to reduce a joint probability distribution to conditional independence relationships. In this paper, we present a general-purpose knowledge integration framework that employs BN in integrating both low-level and semantic features. The efficacy of this framework is demonstrated via three applications involving semantic understanding of pictorial images. The first application aims at detecting main photographic subjects in an image, the second aims at selecting the most appealing image in an event, and the third aims at classifying images into indoor or outdoor scenes. With these diverse examples, we demonstrate that effective inference engines can be built within this powerful and flexible framework according to specific domain knowledge and available training data to solve inherently uncertain vision problems.

Keywords: Semantic image understanding; Low-level features; Semantic features; Bayesian networks; Domain knowledge

1. Introduction

Low-level features, such as color, texture, and shape, have been widely used in content-based image processing and analysis [1–3]. While low-level features are effective for certain tasks, such as “query by example”, they are rather limited for many multimedia applications, such as efficient browsing and organization of large collections of digital photos and videos, that require advanced content extraction and image understanding [3]. Therefore, the ability to extract semantic features in addition to low-level features and to perform fusion of such varied types of features would be very beneficial for scene interpretation.

Since a large number of semantic understanding tasks are performed on photographic images, it is important to understand some of the fundamental characteristics of photographs: (a) they have unconstrained picture content and are taken under unconstrained imaging conditions; (b) they serve the purpose of recording and communicating memories and, therefore, enable a certain degree of consensus among first- and third-party observers with respect to the intent of the photographer; (c) the feature extraction process is imperfect, due to limitations in the accuracy of feature extraction algorithms, as well as the limitations in our understanding of the problem.

[*] Corresponding author. Tel.: +1 585 722 7139; fax: +1 585 722 0160.
E-mail address: jiebo.luo@kodak.com (J. Luo).

0031-3203/$30.00 © 2005 Pattern Recognition Society. Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2004.11.001

Due to the unconstrained nature of photographic images, and the lack of fully reliable low-level features, it is advantageous to select a diverse set of features that extend beyond the standard color/texture/shape types. Semantic features are excellent candidates for providing diversity in the feature set and their use has been proposed in addition to low-level features. The challenge with such an approach is that knowledge from diverse feature sets needs to be integrated, so that specific inferences can be made. In addition, the inference engine should be capable of resolving conflicting indicators from various features, which are likely to occur due to the imperfect nature of the feature extraction algorithms. A unified framework for integrating both low-level and semantic features would be extremely valuable for image understanding, because it would allow for diversity in the feature extraction process and the incorporation of features that are different in nature.

Classic work on feature fusion can be categorized primarily into rule-based methods, voting methods, and discriminant-based methods. Rule-based methods require that the rule designer have knowledge of all possible conditions, which allows for the design of complete and efficient rules [47,4,5]. Rule-based methods can be effective in restricted environments, however, the unconstrained nature of photographs makes it difficult to effectively employ rule-based methods in general situations. Fuzzy logic-based algorithms can also be considered in this category. Voting methods can be as simple as majority voting, or they may involve more sophisticated weighting approaches, which in some cases resemble rule-based methods [6–8]. The difficulty with voting methods lies in determining the weights of different feature types. While it is convenient to assume equal weights for all features, in practice it becomes necessary to adjust the weights according to feature type, as dictated by the application on hand. Discriminant-based methods include neural networks and other equivalent classifiers, e.g., fuzzy neural networks [9–11] and support vector machines (SVM) [12,13], which treat all features as combined vectors. The difficulty with the feature vector approach is the lack of insight into how each feature influences the combined decision. Nevertheless, discriminant-based methods have received considerable attention and have proved effective in a variety of applications [12,13,16,17].

In this paper, we present a framework for semantic image understanding based on belief networks. The framework is suitable for applications where semantic understanding of pictorial images is important. Three examples are presented as case studies for using the proposed framework. The first example is main subject detection (MSD), i.e., determining the likelihood of a given *region* in an *image* being the main subject. Another example is emphasis image selection (EIS), i.e., selecting the most appealing *image* in an *event* comprising of a number of related images. The final example is scene classification of images into indoor or outdoor scenes. The proposed general framework and Belief networks are discussed in Sections 2 and 3. The applications are outlined in Section 4, followed by a benchmarking study of Bayesian networks (BN) vs. neural networks for MSD in Section 5 and conclusions in Section 6.

2. A general framework for image understanding

2.1. Review of existing image understanding frameworks

Image understanding is the process of converting “pixels to predicates”, i.e., iconic image representations to symbolic form of knowledge [21]. Image understanding is the highest (most abstract) processing level in computer vision [22], as opposed to image processing, which converts one image representation to another, for instance, converting raw pixels to an edge map.

Much of the early successes in image understanding have been made in constrained environments, e.g., automatic military target recognition [23], and document [24] and medical [25] image understanding. While image understanding in unconstrained environments is still largely an open problem [16,22], progress is being made in scene classification where the goal is to place an image into one of a set of *predefined* physical (e.g., indoor or outdoor, upright or upside-down image orientation) or semantic categories (e.g., beach, sunset).

Complete object recognition is not necessary and often not possible, especially given the current capabilities of computer vision systems. Fortunately, scenes can be classified without full knowledge of every object in the image. It may be possible, in some cases, to use low-level information, such as spatial distribution of color and texture, to classify some scene types with a high level of accuracy.

One major approach to semantic image understanding is based on the above premise. Examples or training data are collected. These exemplars are thought to fall into clusters in the feature space. They are used to train an appropriate classifier to classify novel test images. In essence, exemplar-based approaches apply pattern recognition techniques (discriminants) to vectors of low-level image features (such as color, texture, or edges) and semantic image understanding is achieved according to the similarity between novel test images and the training exemplars according to a distance metric measured in the selected feature space.

Exemplar-based systems using low-level features have demonstrated successes as well as limitations. Low-level features have the advantage of simplicity. Global or local features are calculated for each image without having to first segment the image or recognize objects in the scene, which can be as challenging as image understanding itself. For tasks such as indoor–outdoor image classification [1], scene classification [2], and image orientation detection [12,17], respectable performance has been achieved. However, even for the same tasks, higher level features or cues are clearly demanded. For instance, some natural images pose difficulty even for a human to decide the correct orientation at a low