Automatic Image Annotation based on WordNet and Hierarchical Ensembles

Wei Li and Maosong Sun

State Key Lab of Intelligent Technology and Systems
Department of Computer Science and Technology, Tsinghua University
Beijing 100084, China
wei.lee04@gmail.com, sms@mail.tsinghua.edu.cn

Abstract. Automatic image annotation concerns a process of automatically labeling image contents with a pre-defined set of keywords, which are regarded as descriptors of image high-level semantics, so as to enable semantic image retrieval via keywords. A serious problem in this task is the unsatisfactory annotation performance due to the semantic gap between the visual content and keywords. Targeting at this problem, we present a new approach that tries to incorporate lexical semantics into the image annotation process. In the phase of training, given a training set of images labeled with keywords, a basic visual vocabulary consisting of visual terms, extracted from the image to represent its content, and the associated keywords is generated at first, using K-means clustering combined with semantic constraints obtained from WordNet, then the statistical correlation between visual terms and keywords is modeled by a two-level hierarchical ensemble model composed of probabilistic SVM classifiers and a co-occurrence language model. In the phase of annotation, given an unlabeled image, the most likely associated keywords are predicted by the posterior probability of each keyword given each visual term at the first-level classifier ensemble, then the second-level language model is used to refine the annotation quality by word co-occurrence statistics derived from the annotated keywords in the training set of images. We carried out experiments on a medium-sized image collection from Corel Stock Photo CDs. The experimental results demonstrated that the annotation performance of this method outperforms some traditional annotation methods by about 7% in average precision, showing the feasibility and effectiveness of the proposed approach.

1 Introduction

With the exponential growth of multimedia information, efficient access to a large image/video databases is highly desired. To address this problem, Content-Based Visual Information Retrieval, has become a hot research topic in the domain of both computer vision and information retrieval in the last decade.

Traditionally, most of the content-based image retrieval techniques is based on the query-by-example (QBE) architecture, in which user should provide an image example firstly, the visual similarity of low-level visual features such as color, texture and shape descriptors is then computed to find the visually similar images compared to