Chapter XII

Formal Models and Hybrid Approaches for Efficient Manual Image Annotation and Retrieval

Rong Yan
IBM T.J. Watson Research Center, USA

Apostol Natsev
IBM T.J. Watson Research Center, USA

Murray Campbell
IBM T.J. Watson Research Center, USA

ABSTRACT

Although important in practice, manual image annotation and retrieval has rarely been studied by means of formal modeling methods. In this chapter, the authors propose a set of formal models to characterize the annotation times for two commonly-used manual annotation approaches, that is, tagging and browsing. Based on the complementary properties of these models, the authors design new hybrid approaches, called frequency-based annotation and learning-based annotation, to improve the efficiency of manual image annotation as well as retrieval. Both our simulation and experimental results show that the proposed algorithms can achieve up to a 50% reduction in annotation time over baseline methods for manual image annotation, and produce significantly better annotation and retrieval results in the same amount of time.

INTRODUCTION

Recent increases in the adoption of devices for capturing digital media along with the ever-greater capacity of mass storage systems have led to an explosive amount of images and videos stored in personal
collections or shared online. To effectively manage, access and retrieve these data, a widely adopted solution is to associate the image content with semantically meaningful labels, a.k.a. image annotation (Kustanowitz & Shneiderman, 2004). Two types of image annotation approaches are available: automatic and manual. Automatic image annotation, which aims to automatically detect the visual keywords from image content, have attracted a lot of attention from researchers in the last decade (Barnard et al., 2002; Jeon et al., 2003; Li & Wang, 2006; Griffin et al., 2006; Kennedy et al., 2006). For instance, Barnard et al. (2002) treated image annotation as a machine translation problem. Jeon et al. (2003) proposed an annotation model called cross-media relevance model (CMRM), which directly computed the probability of annotations given an image. The ALIPR system (Li & Wang, 2006) used advanced statistical learning techniques to provide fully automatic and real-time annotation for digital pictures. Kennedy et al. (2006) considered using image search results to improve the annotation quality. These automatic annotation approaches have achieved notable success, especially when the keywords have frequent occurrence and strong visual similarity. However, it remains a challenge to accurately annotate other more specific and less visually similar keywords. For example, the best algorithm for the CalTech-256 benchmark (Griffin et al., 2006) reported a mean accuracy of 0.35 for 256 categories with 30 examples per category. Similarly, the best automatic annotation systems in TRECVID 2006 (Over et al., 2006) produced a mean average precision of only 0.17 on 39 concepts.

Along another direction, recent years have seen a proliferation of manual image annotation systems for managing online/personal multimedia content. Examples include PhotoStuff (Halaschek-Wiener et al., 2005) and Aria (Lieberman et al., 2001) for personal archives, Flickr.com and ESP Game (von Ahn & Dabbish, 2004) for online content. This rise of manual annotation partially stems from its high annotation quality for self-organization/retrieval purpose, and its social bookmarking functionality in online communities. Manual image annotation approaches can be categorized into two types as shown in Figure 1 (details in Section III). The most common approach is tagging, which allows users to annotate images with a chosen set of keywords (“tags”) from a vocabulary. Another approach is browsing, which requires users to sequentially browse a group of images and judge their relevance to a pre-defined keyword. Both approaches have strengths and weaknesses, and in many ways they are complementary to each other. But their successes in various scenarios have demonstrated the possibility to annotate a massive number of images by leveraging human power.

However, manual image annotation can be tedious and labor-intensive. Therefore, it is of great importance to consider using automatic techniques to speed up manual image annotation. In this work, we assume users will drive the annotation process and manually examine each image label in order to guarantee the annotation accuracy, but in addition we use automatic learning algorithms to improve the annotation efficiency by suggesting the right images, keywords and annotation interfaces to users. This is different from automatic image annotation, which aims to construct accurate visual models based on low-level visual features. But from another perspective, efficient manual annotation can bring benefits to automatic annotation algorithms, because they are typically built on manually annotated examples.

To quantitatively analyze and optimize manual annotation, it is necessary to start with some annotation time/efficiency models that formulate the annotation process. In the absence of existing studies, we propose two annotation time models to describe the process for two popular manual annotation approaches, i.e., tagging and browsing. To the best of our knowledge, this is the first attempt to quantify the manual image annotation process, and it serves as a theoretical foundation to analyze large-scale manual annotation without time-consuming user studies. Based on the time models, we further propose two hybrid annotation algorithms – frequency-based annotation and learning-based annotation – by