

Words and Pictures

David Forsyth
University of Illinois
Urbana – Champaign

Abstract

It is now possible to obtain huge collections of images which carry annotations of one form or another. These annotations can take many forms. Examples include the keywords that occur in the Corel and Hemera collections; various forms of metadata – who made the artifact depicted, when it was made, etc. – that are common in museum collections; the narrative annotations that are sometimes found in museum collections; and the captions that one can collect with news images.

Such collections are interesting for two reasons: First, because visual and text information tend to be complementary, a relatively simple analysis of both the image and the text can reveal a great deal about the data item. This means that, for example, one can cluster such collections well enough to enable naive users to browse a museum's collection or browse the news in a natural way. Second, such collections can be thought of as huge but poorly supervised datasets, containing both information about the appearance of objects and various forms of world knowledge. I will demonstrate a variety of methods whereby one can build probability models linking images or image regions to their annotations. With such models, one can organize a collection in a way that makes browsing easy and quite natural. One can search for pictures using words. And, what is more important, one can attach words to pictures or even to regions. Finally, one can attach names to faces.

I will argue that there is a quite general principle here – useful supervisory information can often be extracted from quite unexpected places. I will show some of the consequences of this principle for motion tracking, face recognition and activity recognition.