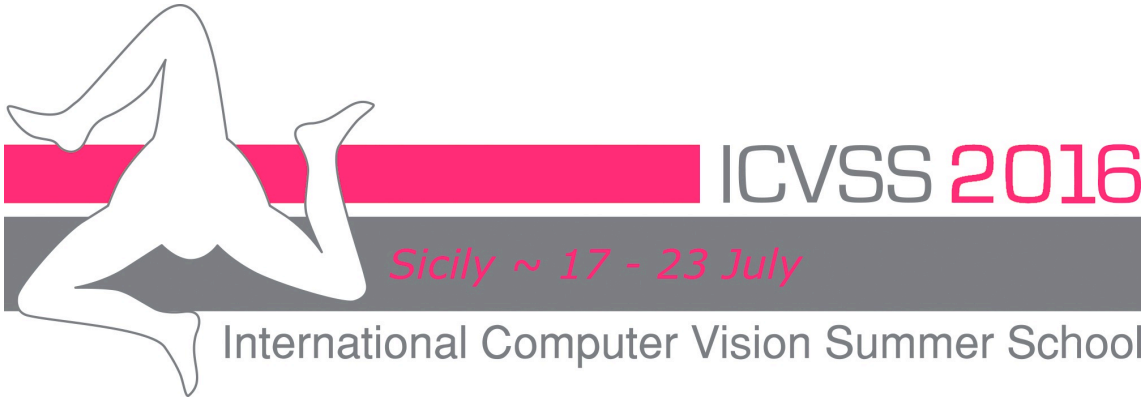

Learning to see by reading books and watching movies

Antonio Torralba
MIT, USA

Abstract

Combining images or videos with language has gotten significant attention in the last two years. Most of existing databases that combine images and text consits on captioned images providing short image descriptions, therefore, limiting what can be learnt. There is another source of very rich textual information that is rarely used in vision: books. Many books have been turned into movies and it is then possible to have the same story being described in two very different modalities. Books are a rich source of both fine-grained information, how a character, an object or a scene looks like, as well as high-level semantics, what someone is thinking, feeling and how these states evolve through a story. Grounding descriptions in books to vision would allow us to get textual explanations or stories about the visual world rather than short captions available in current datasets. In this talk I will describe work on exploiting information available in movies and books to learn to see by watching movies and reading books. I will show quantitative performance for movie/book alignment and show several qualitative examples that showcase the diversity of tasks our model can be used for.