Abstract

The huge amount of video collections on the internet increased the demand on intelligent mining and analyses tools. This motivated the work on video understanding applications, such as semantic video annotation, indexing and retrieval. Work in this area aims to fill the "semantic gap", which is the difference between low-level visual features and human's perception. Many approaches tried to establish a textual semantic representation of the visual data in order to tackle this issue. For achieving this aim, these approaches either build a domain specific "Ontology", or utilize existing commonsense knowledge-bases. Commonsense is the information and facts that are expected to be commonly known by ordinary people, including real-life events and popular facts. Other approaches build knowledge-bases to tackle one level of semantic such as objects, events or concepts. The high demand of these applications emphasizes the need for generic knowledge on multiple semantic levels, which is our inspiration in this work.

In this work, a commonsense knowledge-base that forms the link between the visual world and its semantic textual representation is introduced. It provides knowledge at the objects, events and scenes level, besides higher-levels of semantics, which are temporal events, scenarios and psycholinguistic annotation. This is achieved by utilizing knowledge and strong functionalities of three of the largest knowledge-bases, WordNet, ConceptNet and Linguistic Inquiry Words Count (LIWC), trying to fulfill special requests of this area. This has been achieved by, first, merging different predicates' types from these different knowledge-bases. Then, eliminating non-visually-related information. Finally, fusing the resulted nodes into one unified structure. This knowledge-base is named “Semantic Visual Net” (VisualNet). Quantitative analysis and an automatic video annotation experiment show the effectiveness and comprehensiveness of the proposed enhancements.

![VisualNet building framework](image1.png)

![VisualNet structure](image2.png)
A lot of taxonomies exist in visual literature regarding the semantic levels identification. According to the most used, levels of understanding are:

- **Low-level**: where visual features can be decided directly, for example searching for "red image".
- **Object and action level**: where the properties of objects and actions can be learned as one or more visual features such as color, texture and motion.
- **Event level**: where the objects/actions information is fused with environment data to find the semantic event based on previous cognitive knowledge.
- **Scenario (or story)**: is the cognitive output about a number of events combined to achieve a purpose. Moving from the previous level to this higher level is almost non-dependant on visual features directly, but it rather depends on knowledge of merging the events.
- **Psychological category**: it contains a whole scene abstraction based on information ranging across different levels. For example, *body* and *sport* depend on object recognition whereas *sad* or *happy* may relate to gesture recognition. In contrast, *achievement* is much more based on the context.

The proposed knowledge-base provides real-life commonsense knowledge ranging from second till fifth levels of semantics. The construction engine is fully automated extracting and merging the useful knowledge for the visual domain on these different semantic levels.

**Publications:**


