A Light Scale Concept Ontology for Multimedia Understanding for TRECVID 2005

Milind R. Naphade, Lyndon Kennedy, John R. Kender,
Shih-Fu Chang, John R. Smith, Paul Over, Alex Hauptmann

A. Program Category
   1. *Politics: Shots about domestic or international politics
   2. *Finance/Business: Shots about finance/business/commerce
   3. *Science/Technology: Shots about science and technology
   4. Sports: Shots depicting any sport in action
   5. Entertainment: Shots depicting any entertainment segment in action
   6. Weather: Shots depicting any weather related news or bulletin
   7. *Commercial/Advertisement: Shots of advertisements, commercials

B. Setting/Scene/Site
   1. Indoor: Shots of Indoor locations
   2. Court: Shots of the interior of a court-room location
   3. Office: Shots of the interior of an Office Setting
   4. Meeting: Shots of a Meeting taking place indoors
   5. Studio Setting: Shots of the studio setting including anchors, interviews and all events that happen in a news room
   6. Outdoor: Shots of Outdoor locations
   7. Building: Shots of an exterior of a building
   8. Desert: Shots with the desert in the background
   9. Vegetation: Shots depicting natural or artificial greenery, vegetation woods, etc.
   10. Mountain: Shots depicting a mountain or mountain range with the slopes visible
   11. Road: Shots depicting a road
   12. Sky: Shots depicting sky
   13. Snow: Shots depicting snow
   14. Urban-Setting: Shots depicting an urban or suburban setting
   15. Waterscape/Waterfront: Shots depicting a waterscape or waterfront

C. People
   1. Crowd: Shots depicting a crowd
   2. Face: Shots depicting a face
   3. Person: Shots depicting a person. The face may or may not be visible

   Roles
   4. Government Leader: Shots of a person who is a governing leader e.g. president, prime-minister, chancellor of the exchequer, etc.
   5. Corporate Leader: Shots of a person who is a corporate leader e.g. CEO, CFO, Managing Director, Media Manager etc.
   6. Police/Private Security Personnel: Shots depicting law enforcement or private security agency personnel

7. Military: Shots depicting the military personnel
8. Prisoner: Shots depicting a captive person, e.g., imprisoned, behind bars, in jail, in handcuffs, etc.

D. Objects
    1. Animal (No humans): Shots depicting an animal.
    2. Computer or Television Screens: Shots depicting television or computer screens
    3. Flag-US: Shots depicting a US flag

    Vehicle
    4. Airplane: Shots of an airplane
    5. Car: Shots of a car
    6. Bus: Shots of a bus
    7. Truck: Shots of a truck
    8. Boat/Ship: Shots of a boat or ship

E. Activities

    People Activities
    1. Walking/Running: Shots depicting a person walking or running
    2. People Marching: Shots depicting many people marching as in a parade or a protest

F. Events
    1. Explosion/Fire: Shots of an explosion or a fire
    2. Natural Disaster: Shots depicting the happening or aftermath of a natural disaster such as earthquake, flood, hurricane, tornado, tsunami

G. Graphics
    1. Maps: Shots depicting regional territory graphically as a geographical or political map
    2. Charts: Shots depicting any graphics that is artificially generated such as bar graphs, line charts etc. Maps should not be included

* Features that will be available through automatic detection on the training set.

Procedure Summary

Our aim is to break down the semantic space using a small number of concepts. Since the number of concepts is supposed to be limited to around 50, it is not possible to adopt a depth first approach. One way to leverage such a small number of concepts for dividing the semantic space is to create a multi-dimensional space, where each dimension is nearly orthogonal. Each dimension can then be assigned a small number of concepts. This will achieve the effect of dividing the semantic space into a large number of hypercubes. The hope is that with N dimensions each with N_M concept along the N^{th} dimension, the space can be partitioned into $N_1*N_2*\ldots*N_M$ hypercubes.

It is therefore necessary to have each dimension as orthogonal as possible to the rest and to have concepts in each dimension that will exhaust the dimension as far as possible. This breadth first approach was then employed for concept and dimension selection.

For selecting the dimensions we studied Gans work on what makes news [1]. This led us to analyze along 7 dimensions. We then selected concepts within each dimension motivated by their utility in multiple tasks including search and detection. To validate our selection we then analyzed the lexicon with respect to the TRECVID queries from 2003 and 2004, as well as 130 queries from the BBC query log of 13000 queries.

We also evaluate the relationships between the concepts and the search terms in a real query log. The query log used is from the BBC in late 1998 and includes 13000 queries. The 26377 noun search terms from the query log are mapped to corresponding WordNet terms, retaining the hierarchical tree structure of the WordNet dictionary.

For each concept in our proposed lexicon, we conduct keyword searches through the BBC/WordNet hierarchy to find nodes which match the intended sense of the concept. From the nodes that match the sense of the concept, we choose the top one or two nodes which have the right balance of specificity and generality, as judged by human evaluators. We seek to find the largest (highest-level, most general) nodes which are specific enough to capture the meaning of the concept, but which are also not so specific as to omit other important subnodes which are also relevant to the concept. After we have determined the best nodes for each concept, we can count the number of leaves under each node as the number of noun search terms (from the 26377 total noun search terms) to which the proposed concept is related. By this analysis we came up with 57 concepts along 7 dimensions. From the 57 concepts we further pruned the lexicon down to 44 concepts taking into account the following TRECVID 2005 constraints:

1. The concepts to be annotated need to be marked up by looking at a static key-frame instead of having to play the entire video shot. This constraint is forced by the annotation tools that will be used for the annotation task. Thus all concepts

that need episode-level or story-level markup and that need the aural stream are ruled out in this lexicon

2. The concepts need to be unambiguous as far as possible. This has forced us to select representative concepts from some categories instead of categories themselves where ambiguity can creep in.
3. The concepts should be marked based on evidence from the key-frame without the need for any inference based on other context.
4. The concepts should be pre-iconographic