Objective:
- Generate natural language descriptions of a video that incorporate fine-grained information extraction.
- Improve relevance of descriptions in a hybrid framework that leverages bottom-up keyword prediction semantically verified by top-down concept detection and tri-partite template graphs.

System Overview:
Translation model
Two families of multimodal (MM) correspondence
- MMLDA
- Correspondence MMLDA

Middle Level: Concepts to Language
- Concept Detectors
  - rich semantics from object, action and scene level
  - reduces visual complexity
  - similar to "visual phrases"
- deformable parts models are the base detector
- Sparse Object Stitching
  - segment video into a set of concept shots
  - Record distribution of detected concepts per shot
  - avoids need to do expensive dense detection and tracking

Low Level: Topic Model Translation
- Translation model – capture semantic correlations from low level feature codebooks to bag-of-words
- Two families of multimodal (MM) topics models (LDA):
  - Correspondence MMLDA (enforces a stronger constraint): Topic sparsity on low level features enforces stronger correspondence to text (Computationally expensive)
  - MMLDA (enforces a more diffused constraint and focuses on semantic summarization of multiple views based on observation frequencies)

High Level: Semantic Verification
- Rank nearest neighbor sentences from the training synopses by a ranking function
- \( \chi = h(x_1, x_2, x_3) \)
- \( h(x) \): at least two semantically verified similarly
- \( x \): at least one human subject
- \( x \): total number of the predicted words from the topic model
- \( R \): Run MMLDA on a vocabulary of training synopses and training concept annotations
- Semantic Verification: computing \# of concept shots
- Test Video
  - person climbing rock
  - person with microphone

Output from our system: 1. A person in on artificial rock wall 2. A person climbing a wall is on an artificial rock wall 3. Person climbs rock wall indoors 4. Young man tries to climb artificial rock wall 5. A man demonstrates how to climb a rock wall

TRECVID MER12 Dataset:
- Videos are in 5 Events
- Annotation: Human descriptions/synopses & concept bounding boxes
- Train:
  - Concept Detector: 200 Videos/Event
  - Topic Model: 120 Videos/Event (only positive instances)
  - Test:
  - 6 Videos/Event (MER12 Test Set for Recogntion)

Automatic Evaluation through ROUGE-1 Metric

YouCook Dataset:
- Cooking videos downloaded from YouTube
- Splits: Train 49 / Test 39
- Annotations:
  - Human Descriptions from MTurk
  - Object and Action Bounding Boxes
- ROUGE Benchmark Scoring

Acknowledgements:
The work was partially supported by the National Science Foundation CAREER grant (IIS-0845282), the Army Research Office (W911NF-11-1-0008), the DARPA-Multimodal Program (W911NF-09-2-0008), and the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense (DoD), Defense Threat Reduction Agency (DTRA) contract (IDN-10-1-0012). The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA, DTRA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA, DTRA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA, DTRA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA, DTRA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA, DTRA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA, DTRA, or the U.S. Government.