

- **Database annotation**
- **Feature extraction**
  - Context
    - SIFT
    - hog+hof
  - Actor bounding-box
    - SIFT
    - hog+gof
- **Train classifier**
  - Standard SVM
    - Pros: Straightforward, Not complex
    - Cons: Not optimal in combining different types of features, temporal info is lost, 1 vs. all would be crappy (we cannot simply label "all" by -1)
  - MKL
    - Pros: Interpretable weights (appearance/motion context/actor)
    - Cons: temporal info is lost, 1 vs. all would be crappy (we cannot simply label "all" by -1), naïve implementation takes time, MKL code (by Francis Bach) should be tried, temporal info is lost
  - Spatio-temporal
    - Standard HMM
      - Pros: Already familiar, runs fast, exploits temporal info
      - Cons: Need to read some literature to see how other people use HMM in similar problems, modifications to the code is needed (convert to Linux + memory/time efficient)
    - SVM-HMM
      - Pros: Exploits temporal info, SVM is there
      - Cons: Should read the paper more carefully, seems not applicable
    - CRF
      - Pros: Seems to be perfect!
      - Cons: Needs preparation (implementation/getting familiar with toolbox + studying literature)
- **How to present/evaluate**
  - Baseline
    - Holistic representation of SIFT+hog+hof features
      - For the whole video (already done)
      - Limited to action temporal scope (using annotation information)
  - Our method
    - Actor/context decoupled
    - We can also add random tracks of bounding boxes and simulate the effect of false positives of the underlying person detector (in case one wants to automate the whole process)