Human Action Recognition Using Semi-Latent Topic Models

Yang Wang and Greg Mori, 2009

SE367 Paper Presentation
- Deepak Pathak
10222
Introduction

• Human Action Recognition (What?)

• Still Images (eg: Poselets) v/s Video Sequences

• Motivation:
  Bag of words representation of image – good results in Object Recognition

Bag of Words

[Wang, Mori, 2009]
Earlier Work
(Action Recognition)

- **Motion Based:** Learning features which based on visual cues (motion + shape), optical flows
- **Temporal Dynamic Models:** Generative (e.g. HMM) and Discriminative (e.g. CRF) to model and learn features
- **Interest Point Methods:** Capture local features e.g. train SVM over the features obtained by STIP
- **Topic Models:** “Bag of Words” Paradigm. (analogous to NLP)
Bag of Words
(Analogue: NLP to VISION)

- **Word**
- **Vocabulary**
- **Topic**
- **Document**

- **CodeWord** (Each frame)
- **CodeBook** (All codewords)
- **Action Label**
- **Video Sequence**
Construction of CodeBook

Track and Stabilize person

→ Compute Optical Flow – then descriptors

→ Similarity measure between different frames

Codewords: centroid of these cluster

→ K-medoid clustering into V clusters

Affinity Matrix (among all frames of all sequences)

* Here codeword capture large scale features (containing overall temporal information of all videos in training set)
* Each video is a sequence of frames where each frame is represented by any codeword obtained above, thus video is a bag of words, removing temporal information.
Topic Models

- LDA: Generative model to learn the distribution of topics (actions) given a document (video) and distribution of topics (action) over words (codewords).
  - Dirichlet Distribution
- CTM: Similar but Logistic Distribution to properly correlate of different topics in a document.

Proposed Modification

- Semilatent LDA: Introduces supervision in LDA by making use of action labels present in training dataset.
  - Thus, better estimate the parameters of probability distribution
- Semilatent CTM: Supervised CTM
  Note: Don’t have to choose topics as they are just equal to class labels (unlike unsupervised)
Classification

• Classify each frame in the sequence:
  For each frame, given frame calculate its distribution over action labels i.e. \( p(z_i \mid W) \).
  Here, we chose \( W \) instead of just the corresponding frame so as to ensure that action label not just depend on the frame itself but video sequence as a whole.

• **SLDA**: Models/approximates this probability distribution using other distribution by minimizing KL divergence between the two.

• **SCTM**: It approximates by using coordinate ascent techniques (Variational EM-expected maximization).

• Firstly, **we can classify each frame** using distribution over action labels (take maximum) and then if video contains single action then perform majority voting.
Results
(per video classification)

- KTH Dataset:
  SLDA - 91.2%
  SCTM - 90.33%

- Weizmann Dataset:
  SLDA - 100%
  SCTM - 100%

- Hockey Dataset:
  SLDA - 87.5%
  SCTM - 76.04%

- Soccer Dataset:
  SCTM - 78.64%
  SLDA - 77.81%

- Ballet Dataset:
  SCTM - 91.36%
  SLDA - 88.66%

CTM captures correlations better than LDA, thus performs better on multiple action video datasets (i.e. soccer & ballet).
Datasets

Sample frames from our datasets

[Wang, Mori, 2009]
Conclusion

• **Proposals:**

1. A novel “Bag of words” approach for representing video sequences where each frame corresponds to a word, thus capturing large scale features.

2. Two new models: SLDA & SCTM which are basically supervised form of LDA & CTM, thus training is easy with better performance.

• **Benefit:** This paper focuses mainly on per-frame classification, thus works significantly well on datasets of video containing multiple actions.
References


Thank You

Questions ?