University at Buffalo *The State University of New York* | REACHING OTHERS

A Temporal Bayesian Model for Classifying, Detecting and Localizing Activities in Video Sequences

Manavender R. Malgireddy, Ifeoma Nwogu and Venu Govindaraju

University at Buffalo

Problem Statement

- To detect activities in videos
  - Mostly unconstrained videos (HMDB)
- Perform one shot learning
  - Learn gesture model from one sample
  - Locate and recognize gestures in a video sequence

Related Work

• Spatio temporal interest points were earlier introduced by Lapteve and Linde [4] and since then other interest-points descriptors have been proposed

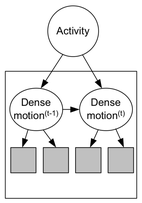
• Wang et al. [6] performed an evaluation of spatio temporal features for activity recognition and showed that dense features points perform better

• Wang et al. in [5] proposed an approach to describe videos by dense trajectories. They sample dense points from each frame and track them based on displacement information from optical flow field.

• Gaur et al. [2] proposed a model based on string representation of the video.

• Brendel et al. [1] proposed a model to represent videos by spatiotemporal graphs, where the nodes correspond to multiscale video segments and edges capture hierarchical, temporal and spatial relationships

Model description



• Standard HMM with minor variations

• Multiple observations per state

• Multiple channels

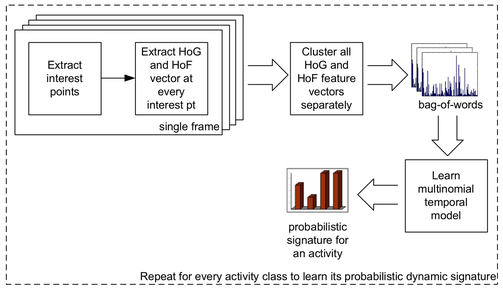
• Initial state distribution $\hat{\pi} = \sum_{r=1}^{R} \gamma_1^r(i);$

• State transition probability $\hat{a}_{ij} = \frac{\sum_{r=1}^{R}\sum_{t=1}^{T}\eta_t^r(i,j)}{\sum_{r=1}^{R}\sum_{t=1}^{T}\gamma_t^r(i)}$

• Observation densities $\hat{b}_j^d(k) = \frac{\sum_{r=1}^{R}\sum_{t=1}^{T}\gamma_t^r(j)\cdot\frac{n_t^{d,k}}{n_t^d}}{\sum_{r=1}^{R}\sum_{t=1}^{T}\gamma_t^r(j)}$

Proposed Approach

• Interest points
  - Dense sampling
  - Optical Flow points
• Descriptors
  - HOG, HOF

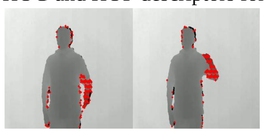


Process flow for the training cycle

• Clustering
  - Kmeans
  - Hierarchical
• Classifier
  - HMM with multiple observations

Interest points and features

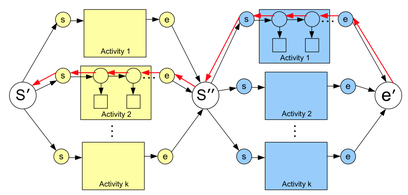
• HMDB Dataset
  - Extract dense points from each frame
  - Track each point for L frames
  - Extract HOG and HOF descriptors for each trajectory
• Chalearn Dataset
  - Extract motion points by frame differencing
  - Extract HOG and HOF descriptors for each point



• Visual words in chalearn data
  - cluster all the descriptors from training data using kmeans to form visual words
  - Use Latent Dirichilet allocation to form more meaningful visual words

• Visual words in HMDB/KTH data
  - Number of points are in the order of 30,000,000
  - Construct visual words for each activity class separately
  - Use these words as samples for constructing final visual words

Gesture spotting



Gesture spotting by computing likelihoods via Viterbi decoding

• Learn individual gesture models

• Construct a network of models with at most 5 gestures

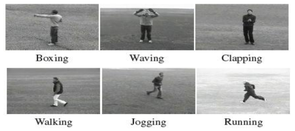
• Transition between gestures has equal probabilities

• Find most likely gestures from Viterbi decoding

Experiments and Results

• Chalearn Dataset
  - Dataset is divided into 20 batches
  - 47 videos per batch
  - First M videos are used for training
  - 3-5 gestures in each test video

• HMDB Dataset
  - 51 class
  - ~ 100 videos per class
  - 70-30 split for training and testing



• KTH Dataset
  - 4 actions
  - 25 subjects, 4 scenarios
  - 16 subjects for training

Method	Accuracy
Laptev et.al. [8]	91.8%
Yuan et.al. [9]	93.3%
Wang et.al. [5]	94.2%
Gilbert et.al. [10]	94.5%
Kovashka and Gruman et.al. [11]	94.53%
Out Method	**94.67%**

Method	Accuracy
Kuehne et.al. on 51 activities [3]	23.18%
Kuehne et.al. on 10 activities [3]	54.3%
Our results on 51 activities	**25.64%**
Our results on 10 activities (stabilized)	**66.67%**

Results for KTH (left) and HMDB (right) dataset

Method	Dataset	Edit Distance
Ours	Development	0.26336
Ours + LDA	Development	0.2409
Ours	Validation	0.26036
Ours + LDA	Validation	**0.23328**
Baseline	Validation	0.59978
Top Ranking	Validation	0.1426

Results for Chalearn dataset

Implementation details

• We obtained best results on KTH dataset with 2000 visual words and 25 states

• 1000 visual words and 25 states for HMDB dataset

• Visual words for Chalearn data depends on number of classes (10*Nclasses)

• LDA is used to reduce the number of visual words to a factor of eight (8*Nclasses)

• 10 states were found to be optimal for Chalearn data

Conclusion and Future Work

• Presented a model for activity classification and spotting that competes well with state-of-the-art systems

• Verified the model by testing it on two extreme datasets (KTH and HMDB)

• Participated in Chalearn gesture challenge and finished in top-5

References

1. W. Brendel and S. Todorovic. Learning spatiotemporal graphs of human activities. In ICCV, 2011
2. U. Gaur, Y. Zhu, B. Song, and A. Roy-Chowdhury. A "string of feature graphs" model for recognition of complex activities in natural videos. In ICCV, 2011
3. H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: a large video database for human motion recognition. In Proceedings of the International Conference on Computer Vision (ICCV), 2011
4. I. Laptev and T. Lindeberg. Space-time interest points. In ICCV, pages 432–439, 2003
5. H. Wang, A. Kl¨aser, C. Schmid, and L. Cheng-Lin. Action Recognition by Dense Trajectories. In IEEE Conference on Computer Vision & Pattern Recognition, pages 3169–3176, Jun 2011
6. H. Wang, M. M. Ullah, A. Kl¨aser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In British Machine Vision Conference, sep 2009
7. Y. Wang and G. Mori. Human Action Recognition by Semilatent Topic Models. IEEE TPattern Anal. Mach. Intell., 31:1762–1774, 2009
8. I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning Realistic Human Actions From Movies. In CVPR, pages 1–8, 2008.
9. J. Yuan, Z. Liu, and Y.Wu. Discriminative subvolume search for efficient action detection. In CVPR, 2009
10. A. Gilbert, J. Illingworth, and R. Bowden. Action recognition using mined hierarchical compound features. IEEE Trans. Pattern Anal. Mach. Intell., 33(5):883–897, 2011.
11. A. Kovashka and K. Grauman. Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. In CVPR, pages 2046–2053, 2010