Recognizing Human Activities from partially observed Videos

Praneeth A S-UG201110023, Anuroop Kakkirala-UG20113022

IIT Jodhpur

Monday 3rd November, 2014

Authors

Authors

1. **Yu Cao, Yuewei Lin, Song Wang** - *Department of Computer Science and Engineering, University of South Carolina, Columbia, SC 29208, USA*
2. **Daniel Barrett, Andrei Barbu, Siddharth Narayanaswamy, Haonan Yu, Aaron Michaux, Jeffrey Mark Siskind** - *School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907, USA*
3. **Sven Dickinson** - *Department of Computer Science, University of Toronto, Toronto ON, Canada*

Abstract

Abstract

1. Propose a new method that can recognize human activities from partially observed videos in the general case.
2. Formulation of Problem:
   1. Dividing each activity into multiple ordered temporal segments
   2. Using spatiotemporal features of the training video samples in each segment as bases and applying sparse coding (SC) to derive the activity likelihood of the test video sample at each segment
   3. Finally combining the likelihood at each segment to achieve a global posterior for the activities.
3. Extend the proposed method to include more bases that correspond to a mixture of segments with different temporal lengths (MSSC), which can better represent the activities with large intra-class variations.

Past Research

Past Research

1. Simple single person human actions like walking, running
2. Recent Research - Multiple person interaction
3. Wide Approach - Train and classify the spatiotemporal features extracted from videos with different activities

Approach

1. A sequence of 2D video frames treated as a 3D XYT video volume in which interest points are located by finding local maxima in the responses of the feature detector, followed by calculating vectorized feature descriptors at each interest point
2. Bag-of-visual-words technique - Spatiotemporal features can be combined into a feature vector - describes the activity presented in the video
3. Mostly recognized after observing full video

Past Research

1. **Ryoo** - Activity Prediction.
2. Activity prediction is the maxmargin early event detectors (MMED), which try to detect the temporal location and duration of a certain activity from the video streaming.
3. *Kitani et al.* - Predict the walking path of a person in certain environments based on historical data.

Paper's Approach

Paper's Approach

1. Probabilistic formulation for human activity recognition from partially observed videos, where the posterior is maximized for the recognized activity class and the observed video frames
2. Key component in defining the posterior - likelihood that the observed video frames describe a certain class of activity
3. Bases - Set of training video samples (completely observed) of each activity class
4. Sparse coding (SC) - derive the likelihood that a certain type of activity is presented in a partially observed test video
5. Doesnot require full temporal alignment between any pair of (training or test) videos
6. Can handle Problems:
    1. Possible outliers in the training video data
    2. Limited no. of training samples
    3. Large intra-class variations

Problem Formulation

Human Activity Recognition from a Fully Observed Video

Problem Formulation

1. Fully observed video $\mathcal{O}[1:T]$ of length T. $\mathcal{O}[t]$ - frame at time t.
2. **Goal:** Classify the video $\mathcal{O}[1:T]$ into one of $\mathcal{P}$ activity classes $\mathcal{A} = \mathcal{A}_p \; p = 1, \ldots, \mathcal{P}$
3. Human actions - sequence of simple actions - contain different spatiotemporaleatures. Divide uniformly into M different segments $\mathcal{O}(t_{i-1}:t]$ where $t_i = \frac{iT}{M}$, i^{th} stage of activity $i = 1, \ldots, M$.
4. Posterior Probability($\mathcal{A}_p$ is present in video $\mathcal{O}[1:T]$) = $P(\mathcal{A}_p|\mathcal{O}[1:T])$

Human Activity Recognition from a Fully Observed Video

Problem Formulation

1. $P(\mathcal{A}_p|\mathcal{O}[1:T]) \propto \sum_{i=0}^{M} P(\mathcal{A}_p,(t_{i-1},t_i]|\mathcal{O}[1:T]) \propto \sum_{i=0}^{M} P(\mathcal{A}_p,(t_{i-1},t_i])P(\mathcal{O}[1:T]|\mathcal{A}_p,(t_{i-1},t_i])$
2. $P(\mathcal{A}_p,(t_{i-1},t_i]) =$ prior of stage i of activity $\mathcal{A}_p$
3. $P(\mathcal{O}[1:T]|\mathcal{A}_p,(t_{i-1},t_i]) =$ observation likelihood given activity class $\mathcal{A}_p$ in the i^{th} stage
4. $p^* = \arg\max_p \sum_{i=0}^{M} P(\mathcal{A}_p,(t_{i-1},t_i])P(\mathcal{O}[1:T]|\mathcal{A}_p,(t_{i-1},t_i])$

Human Activity Recognition from a Partially Observed Video

Problem Formulation

1. Partially observed video - $\mathcal{O}[1 : T_1] \cup [T_2 : T]$, where frames $\mathcal{O}(T_1 : T_2)$ are missing
2. **Assumption:** T_1 is always the last frame of a segment and T_2 is always the first of another segment.
3. Posterior probability that an activity is presented in this partially observed video
   $P(\mathcal{A}_p|\mathcal{O}[1 : T_1] \cup [T_2 : T]) \propto \omega_1 \sum_{i|t_i \leq T_1} P(\mathcal{A}_p, (t_{i-1}, t_i]|\mathcal{O}[1 : T_1]) + \omega_2 \sum_{i|t_{i-1} \geq T_2} P(\mathcal{A}_p, (t_{i-1}, t_i]|\mathcal{O}[T_2 : T])$
4. $\omega_1 = \frac{T_1}{T_1 + T - T_2 + 1}, \omega_2 = \frac{T - T_2 + 1}{T_1 + T - T_2 + 1}$

Human Activity Recognition from a Partially Observed Video

Problem Formulation

1. Using previous equations(Posterior Probability):
   $P(\mathcal{A}_p|\mathcal{O}[1:T_1] \cup [T_2:T]) \propto \omega_1 \sum_{i|t_i \leq T_1} P(\mathcal{A}_p,(t_{i-1},t_i])P(\mathcal{O}[1:T_1]|\mathcal{A}_p,(t_{i-1},t_i]) + \omega_2 \sum_{i|t_{i-1} \geq T_2} P(\mathcal{A}_p,(t_{i-1},t_i])P(\mathcal{O}[T_2:T]|\mathcal{A}_p,(t_{i-1},t_i])$
2. Index of recognized activity:
   $p^* = \arg\max_p P(\mathcal{A}_p|\mathcal{O}[1:T_1] \cup \mathcal{O}[T_2:T])$

Likelihood Calculation

Likelihood Calculation

1. Compare $\mathcal{O}[1:T_1]$ with the i^{th} segment of all the training videos.
2. Each segment of a video, use the bag-of-visual-words technique to organize its spatiotemporal features into a fixed-dimensional feature vector
3. $\mathbf{h}_i^n$ - feature(row) vector after applying bag-of-visual-words techniques to i^{th} segment of the n^{th} training video
4. $\mathbf{h}_i^{\mathcal{O}}$ - feature for stage i in $\mathcal{O}[1:T_1]$
5. $\bar{\mathbf{h}}_i = \frac{1}{N}\sum_{n=1}^{N}\mathbf{h}_i^n$
6. $P(\mathcal{O}[1:T]|\mathcal{A}_p,(t_{i-1},t_i]) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{\frac{-||h_i^{\mathcal{O}}-\bar{\mathbf{h}}_i||^2}{2\sigma^2}}$

Likelihood Calculation

1. Mean vector not a good feature. Problems:
   1. Mean feature may not be a representative of the true 'center'
   2. Activity label for a training video is actually incorrect
2. Feature vectors from training data as bases. Construct the bases matrix A_i using the segment-i feature vectors from N training

$$A_i = \begin{pmatrix} \mathbf{h}_i^1 \\ \mathbf{h}_i^2 \\ \dots \\ \mathbf{h}_i^N \end{pmatrix}$$

Likelihood Calculation

1. $x^* = \min_x ||\mathbf{h}_i^{\mathcal{O}} - \hat{\mathbf{h}}_i^{\mathcal{O}}||^2 + \lambda||x||_0$
2. $P(\mathcal{O}[1:T_1]|\mathcal{A}_p, (t_{i-1}, t_i]) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-||h_i^{\mathcal{O}} - A_i x^*||^2}{2\sigma^2}}$ - Sparse Coding Calculation.
3. **Benefits:** Can automatically select a proper subset of bases for approximating the test video segments

Likelihood calculation using sparse coding on a mixture of segments

Likelihood Calculation

1. Include more bases that are constructed from a mixture of segments with different temporal lengths and temporal locations in the training videos.
2. Construct 8 more bases from $(t_{i-1}, t_i]$ - $(t_{i-2}, t_{i-1}], (t_i, t_{i+1}], (t_{i-2}, t_i], (t_{i-1}, t_{i+1}], (t_{i-2}, t_i], (t_{i-1}, \frac{t_i - t_{i-1}}{2}]$, $\frac{t_i - t_{i-1}}{2}, t_i], (t_{i-1} + \frac{t_i - t_{i-1}}{4}, t_i + \frac{t_i - t_{i-1}}{4}]$

Experiments

Experiments

1. **State-Of-Art comparison Methods:** Ryoo's human activity prediction methods (both non-dynamic and dynamic versions), early event detector - MMED, C2, and Action Bank.
   Another Comparison: KNN (K Nearest Neighbor)

Evaluation on the special case: prediction

Experiments

1

Evaluation on the special case: gap-filling

Experiments

1

Evaluation on degenerate case: recognition

Experiments

1