Recognize Human Activities from Partially Observed Videos

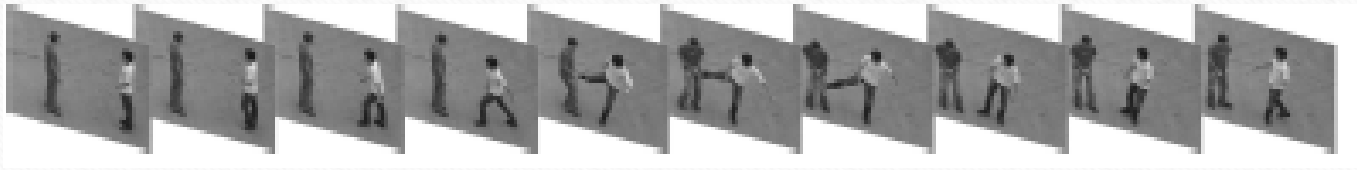
Anuroop Kakkirala Praneeth AS

Aim of the Project

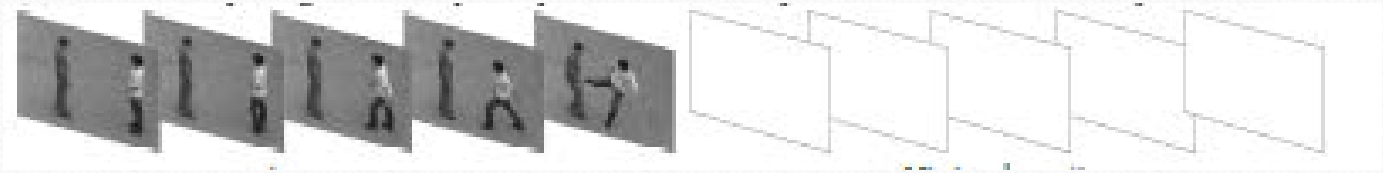
- To recognize human activities from partially observed videos and predicting the unobserved subsequence in the video.
- Two possibilities:

  1) An Unobserved subsequence is at the end of the video.

  2) It may occur at any time by yielding a temporal gap in the video



Full Observation



Missing Observation
At the end.

Gap filling

Problem Formation:
1) For a Fully Observed Video

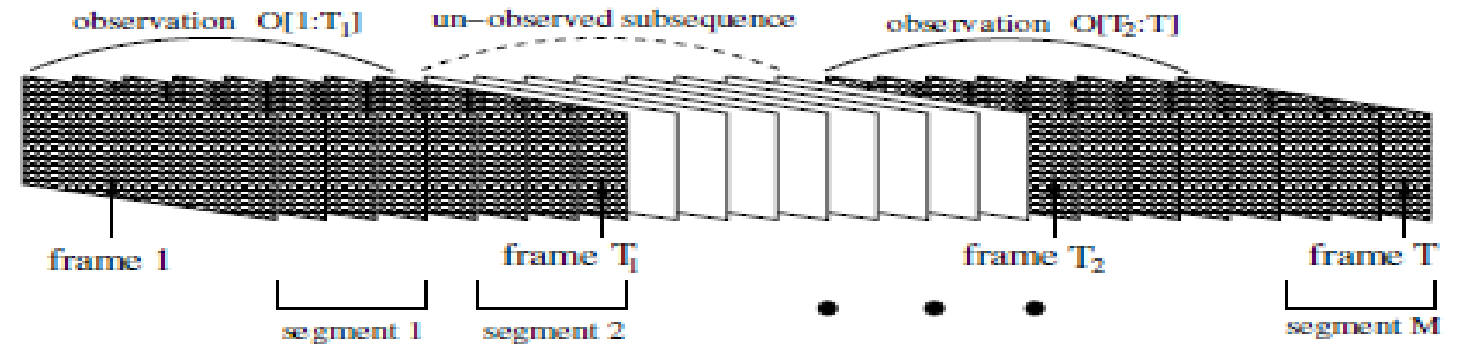
- Given a fully observed video O[1 : T] of length T, O[t] indicates the frame at time t.
- Goal is to classify the video O[1 : T] into one of P activity classes $\{A = \{Ap\}, p = 1, \ldots, P\}$.
- Human actions - sequence of simple actions - contain different spatiotemporal features. Divide uniformly into M different segments O(ti−1 : ti] where ti = i*T/M, i th stage of activity i = 1, . . . , M.

Formation continued..

- $P(Ap|O[1:T]) \propto \sum_{i|0 \text{ to } M} P(Ap,(ti-1,ti]|O[1:T])$
  $\propto \sum_{i|0 \text{ to } M} P(Ap,(ti-1,ti])P(O[1:T]|Ap,(ti-1,ti]).$
- $P(Ap,(ti-1,ti])$ = prior of stage i of activity Ap
- $P(O[1:T]|Ap,(ti-1,ti])$ = observation likelihood given activity class Ap.
- $p* = \arg\max \sum_{i|0 \text{ to } M} P(Ap,(ti-1,ti])P(O[1:T]|Ap,(ti-1,ti])$
- p* is index of recognized activity.

2) For a Partially Observed Video

- Partially observed video - O[1 : T1] ∪ [T2 : T], where frames O(T1 : T2) are missing.
- T1 is always the last frame of a segment and T2 is always the first of another segment.



Formation continued..

- Posterior probability that an activity is presented in this partially observed video
- $P(Ap|O[1:T1] \cup [T2:T]) \propto w1 \sum_{i|ti \leq T1} P(Ap,(ti-1,ti]|O[1:T1]) + w2 \sum_{i|ti-1 \geq T2} P(Ap,(ti-1,ti]|O[T2:T])$
- $w1 = T1/(T1+T-T2+1)$, $w2 = T-T2+1/(T1+T-T2+1)$
- $p* = \arg\max P(Ap|O[1:T1] \cup O[T2:T])$
- Where $P(Ap|O[1:T1] \cup O[T2:T])$ can be calculated as above.

Likelihood calculation

- Compare O[1 : T1] with the i th segment of all the training videos.
- Each segment of a video, use the bag-of-visual-words technique to organize its spatiotemporal features into a fixed-dimensional feature vector.
- $\mathbf{h}_i^n$ - feature(row) vector after applying bag-of-visual-words techniques to ith segment of the nth training video
- $\mathbf{h}_i^O$- feature for stage i in O[1 : T1]

Likelihood calculation continued..

$$\bar{\mathbf{h}}_i = \frac{1}{N} \sum_{n=1}^{N} \mathbf{h}_i^n$$

$$P(\mathcal{O}[1:T] | \mathcal{A}_p, (t_{i-1}, t_i]) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-||h_i^{\mathcal{O}} - \bar{\mathbf{h}}_i||^2}{2\sigma^2}}$$