# **Scan Without a Glance:**Towards Content-Free Crowd-Sourced Mobile Video Retrieval System

**Cihang Liu**, Lan Zhang, Kebin Liu, Yunhao Liu

**September 2, 2015**

Tsinghua University

GreenOrbs

# Outline

- Background
- Preliminary
- Overview
- Techniques
- Evaluation
- Conclusion

# Background

**the Boston Bombings, April 15, 2013**

# Background


**crowd-sourced videos captured by the passengers**

# Preliminary

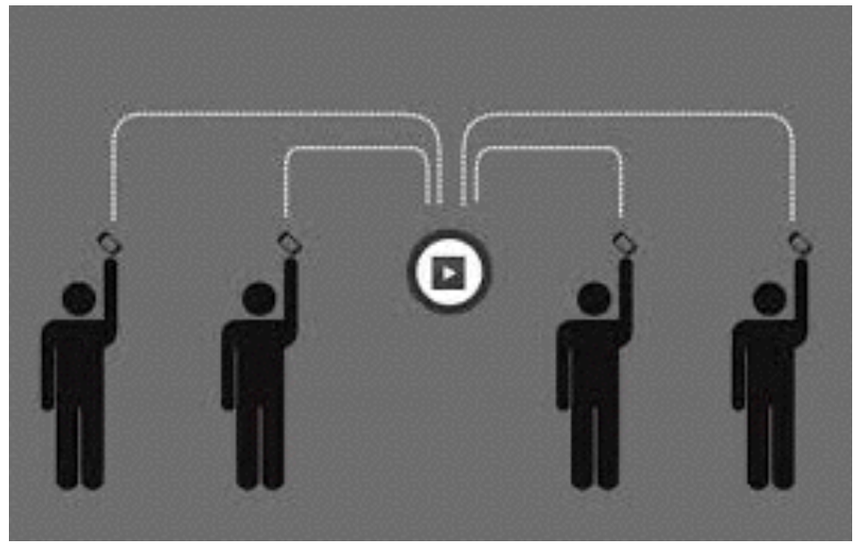
- Massive videos and video sources
  - by 2015:1.91 billion smartphone users
- Uploading videos is traffic-exhausting
  - centralized video retrieval strategy is no more applicable
- Content-based algorithms are too heavy for smart phones
  - decentralized video retrieval strategy also failed

# Preliminary

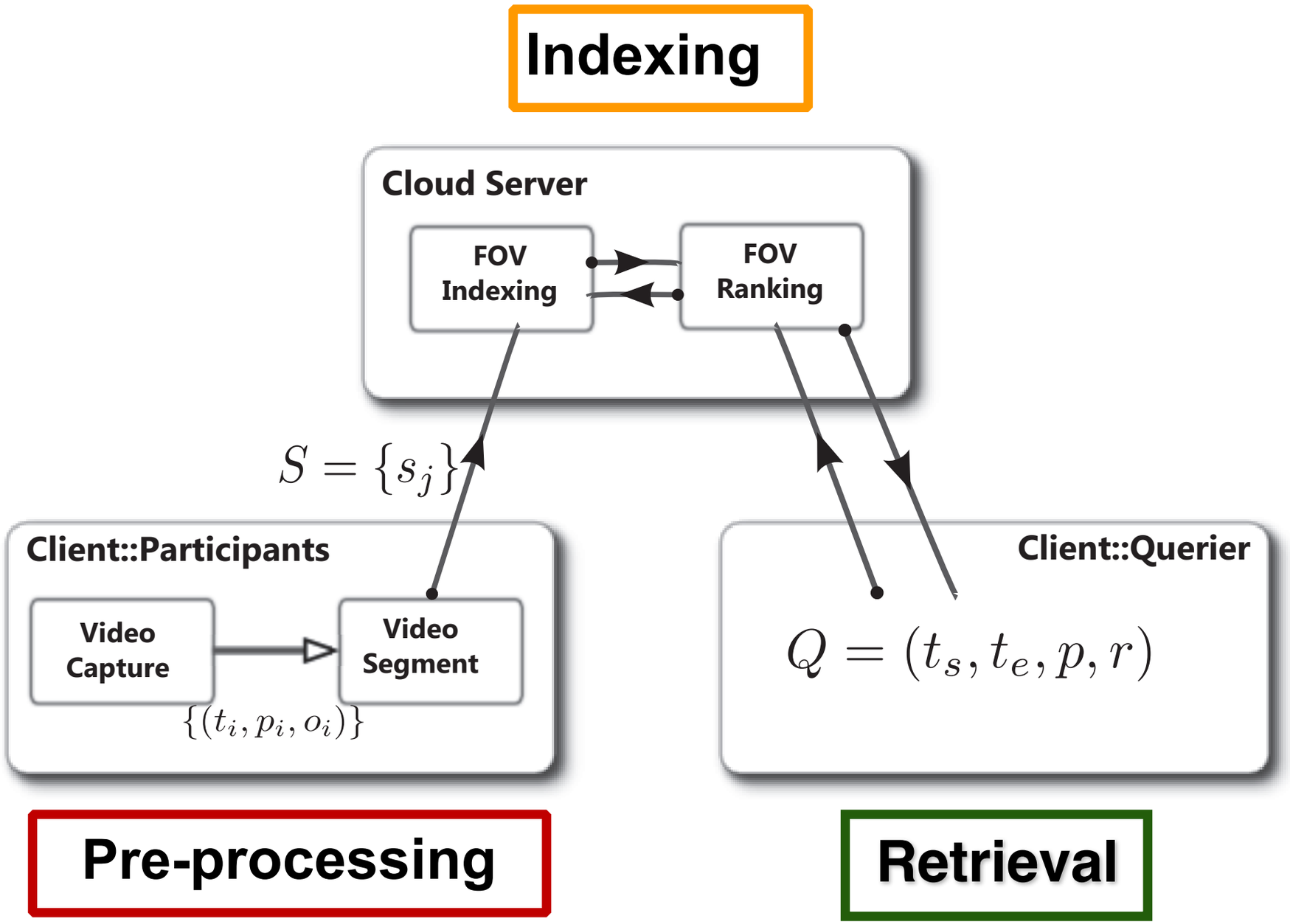
- the character of an ideal mobile video descriptor
  - rationality
  - efficiency
  - lightweight
- the indexing scheme
  - scalability

# Scan Without a Glance

- crowd-sourced mobile video retrieval system
  - content-**free** video descriptor
  - efficient **dynamic** index structure
  - **centralized** retrieval strategy
- enable the era of **crowd surveillance**

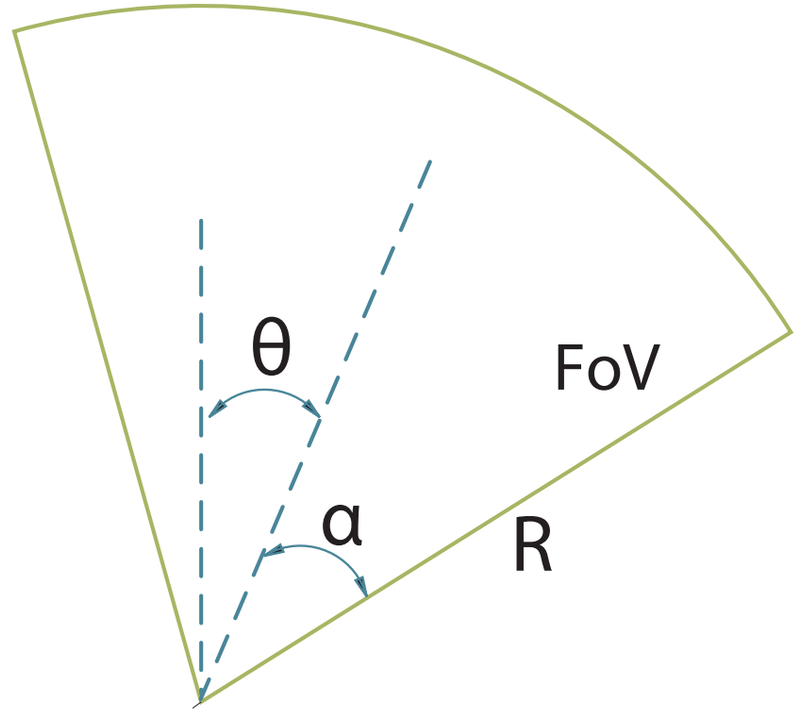


# Overview


Indexing
Cloud Server
FOV Indexing
FOV Ranking
$S = \{s_j\}$
Client::Participants
Video Capture
Video Segment
$\{(t_i, p_i, o_i)\}$
Client::Querier
$Q = (t_s, t_e, p, r)$
Pre-processing
Retrieval

# FoV: Field of View

- For each frame
  - ***t***: timestamp
  - ***p***: position of the camera
  - ***θ***: azimuth angle(orientation relative to North)
  - R: viewable radius
  - α: half of the opening angle of the camera

# Pre-processing

- Video Segmentation
- Content Free Descriptor Extraction

# Before segmentation?

- How to distinguish from different FoVs using content-free information?

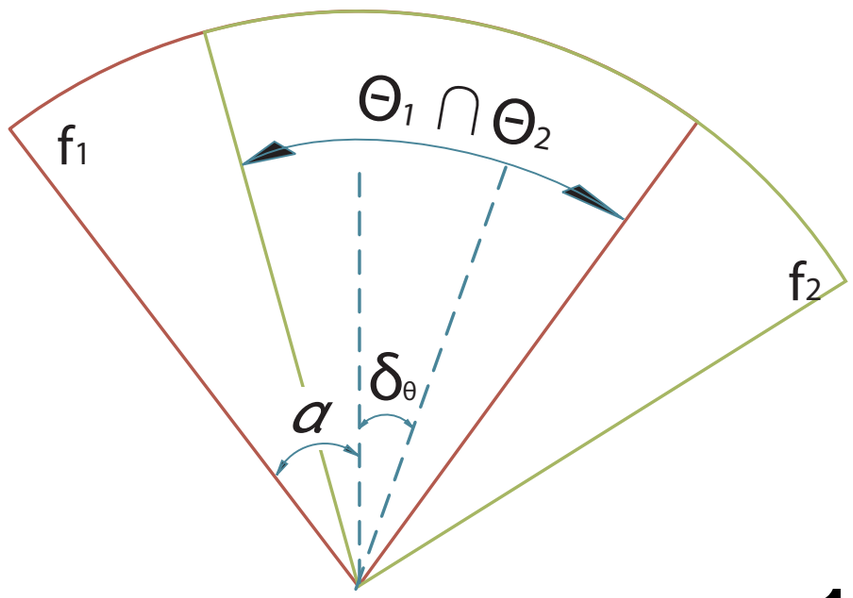
- We need a ***similarity measure***

- FoV: → Physical Movement:
  - p:position → Translation
  - θ:azimuth angle → Rotation

# (a)Rotation

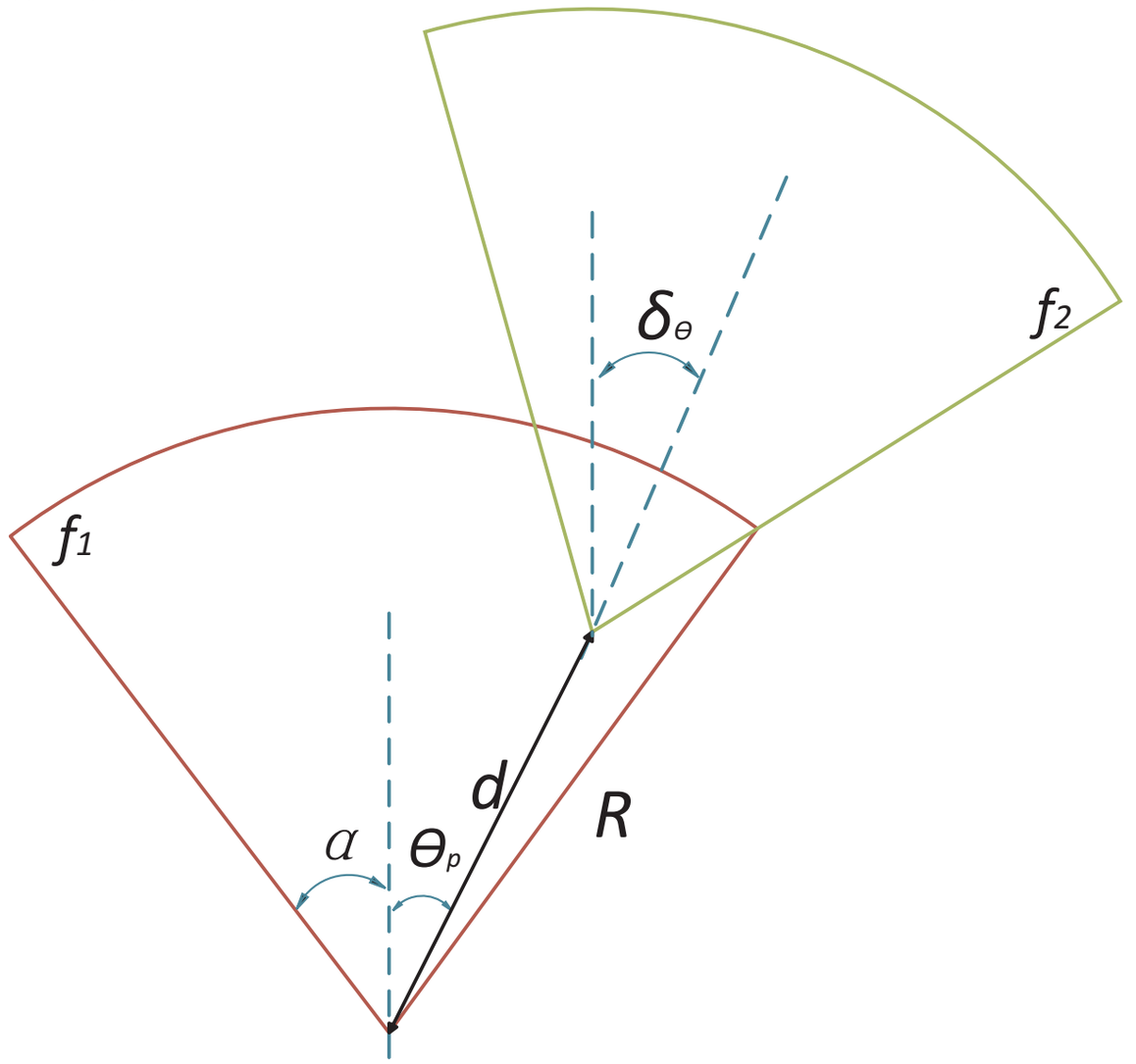
- intersection of the viewing angle

$$Sim_R(f_1, f_2) = \frac{|\Theta_1 \cap \Theta_2|}{|\Theta|} = \begin{cases} \frac{2\alpha - \delta_\theta}{2\alpha} & \delta_\theta < 2\alpha \\ 0 & otherwise \end{cases}$$



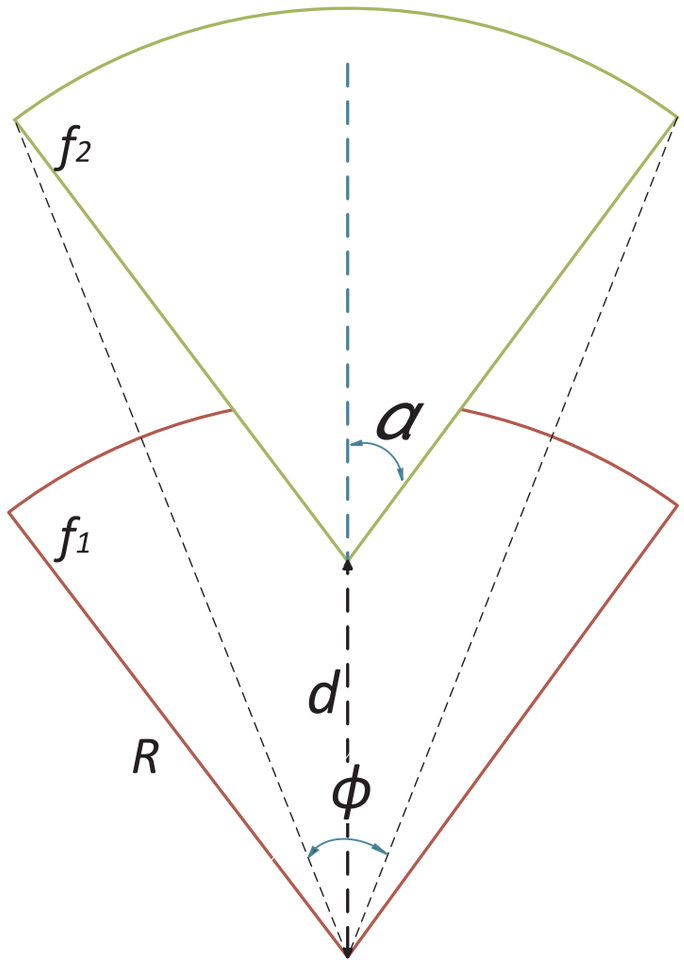
# (b)Translation

- **parallel** and **vertical** translation
- $\theta_p$ :translation direction
  - parallel:$\theta_p = \theta$
  - vertical: $|\theta_p - \theta|=90°$



# (b_1) Parallel Translation

- $\theta_p = \delta$

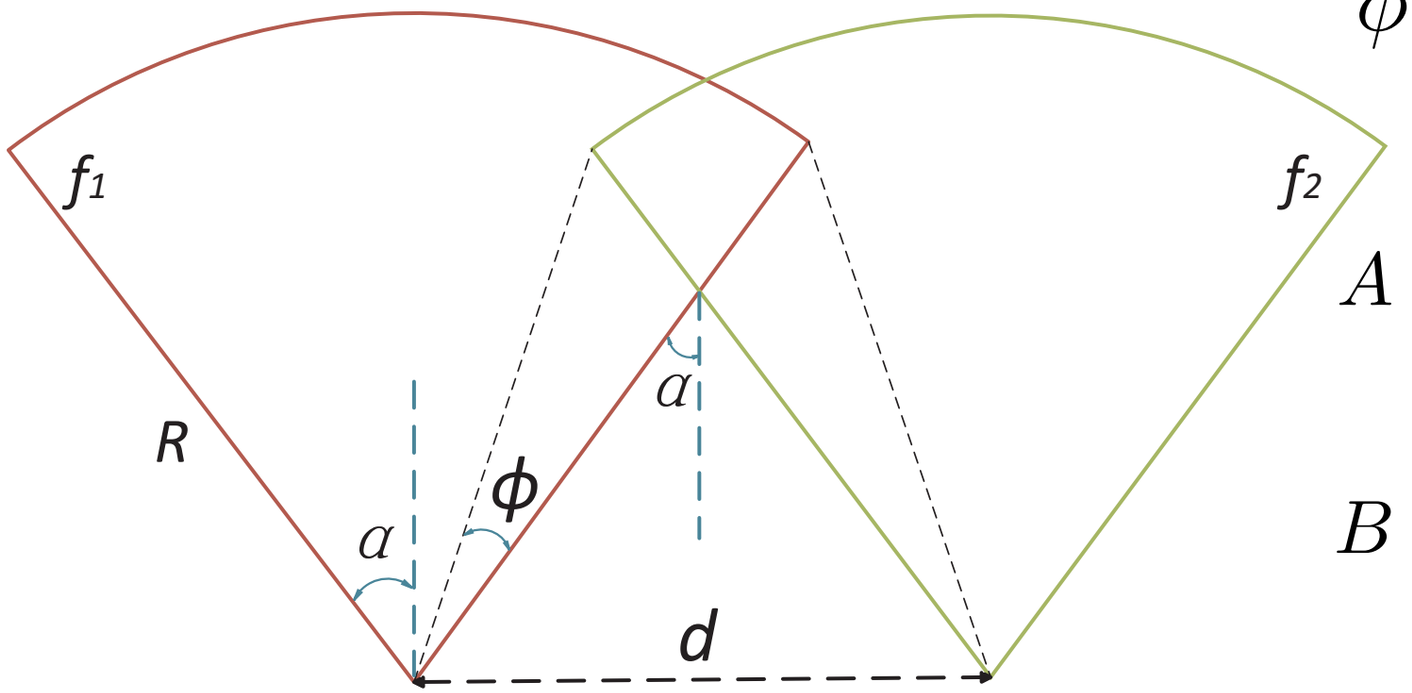



$$\phi_{\parallel} = 2\arctan\frac{R\sin\alpha}{d + R\cos\alpha}$$

$$Sim_{\parallel} = \frac{\phi_{\parallel}}{2\alpha}$$

# (b_2)Vertical Translation

- $|\theta_p - \delta| = 90°$




$$\phi_{\perp} = \arctan \frac{\begin{bmatrix} 1 & 0 \end{bmatrix} AB}{\begin{bmatrix} 0 & 1 \end{bmatrix} AB}$$

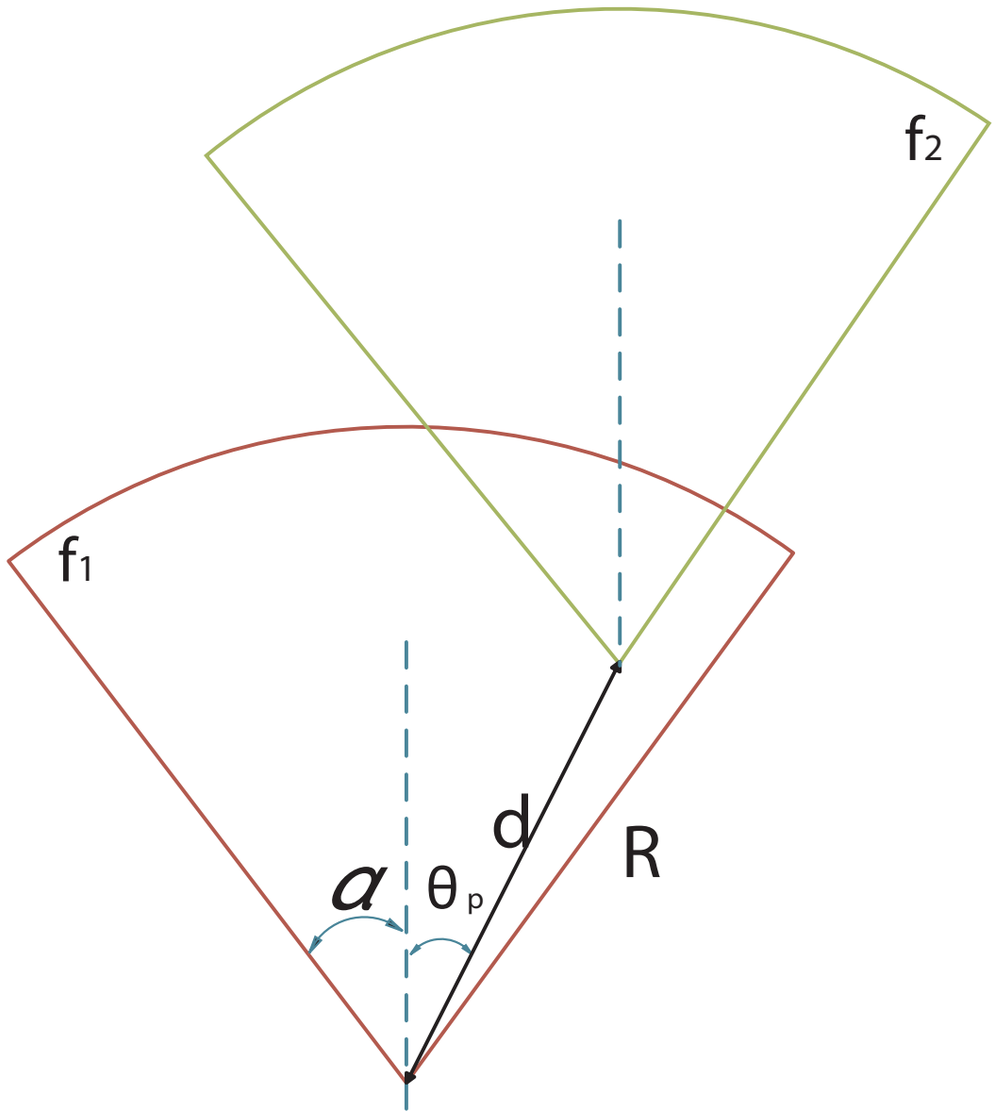
$$A = \begin{bmatrix} 2R\sin 2\alpha & -2\sin 2\alpha \\ 2R\cos 2\alpha & 1-\cos 2\alpha \end{bmatrix},$$

$$B = \begin{bmatrix} \sin\alpha \\ d \end{bmatrix}$$

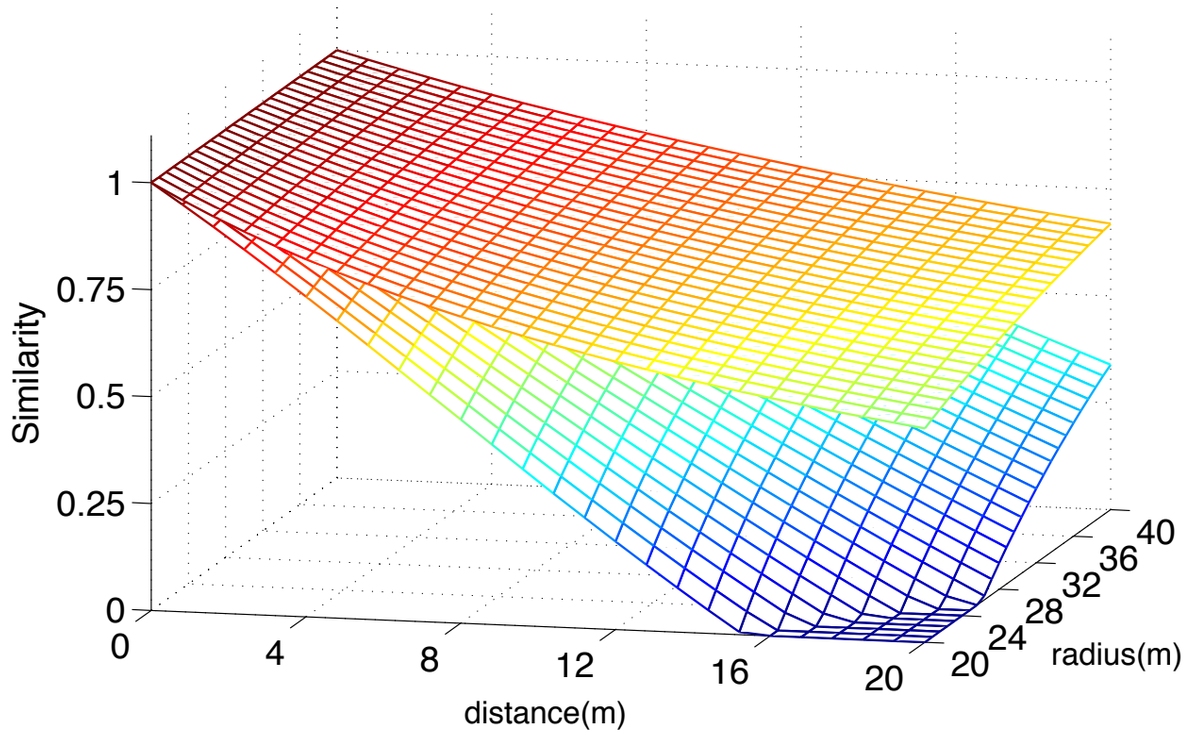
$$Sim_{\perp} = \frac{\phi_{\perp}}{2\alpha}$$

# (b)translation

- arbitrary translation direction $\theta_p$

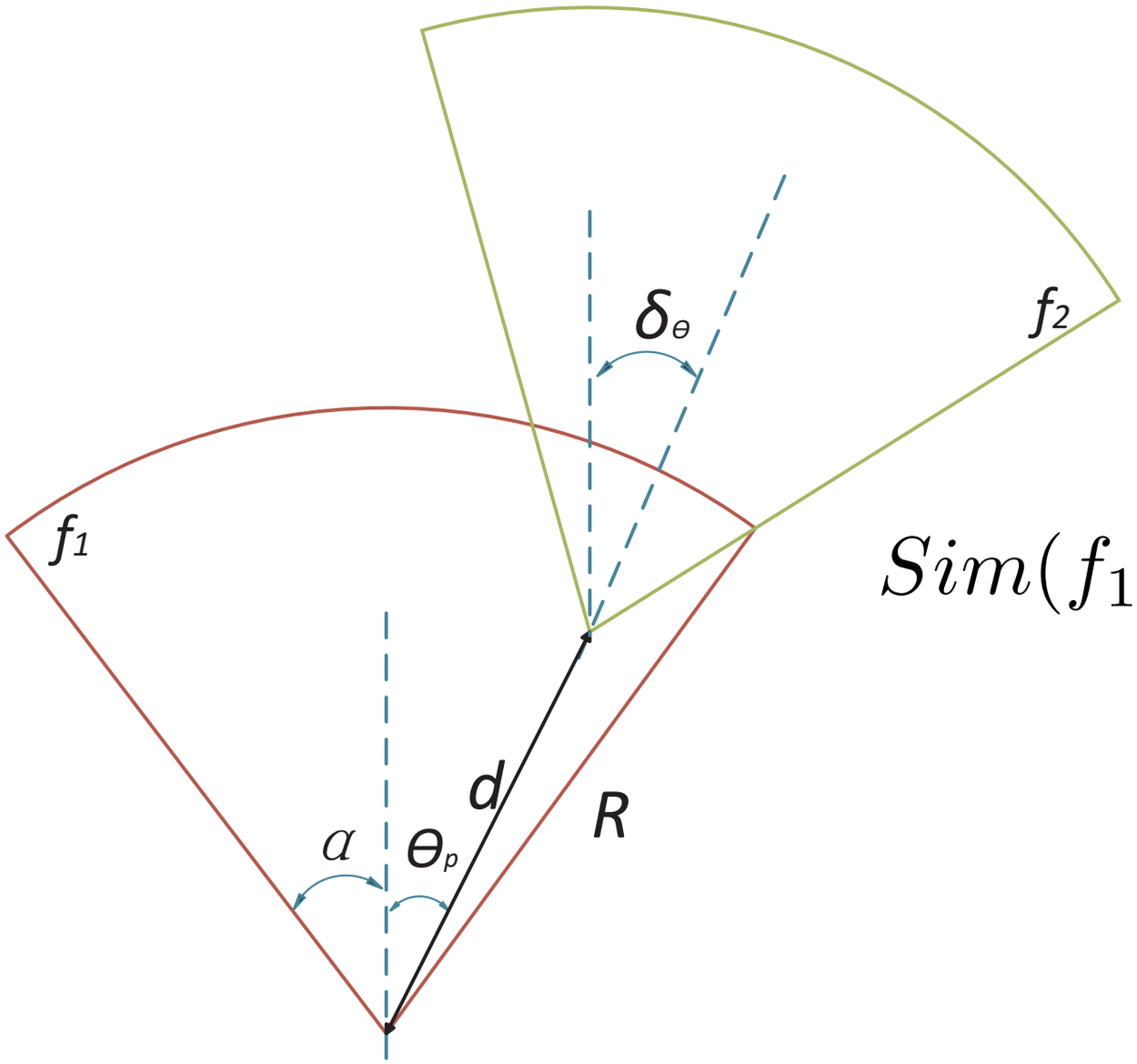



$$Sim_T(f_1, f_2) = (1 - \frac{\theta_p}{90})Sim_{\parallel} + \frac{\theta_p}{90}Sim_{\perp}$$



# (c)Reality

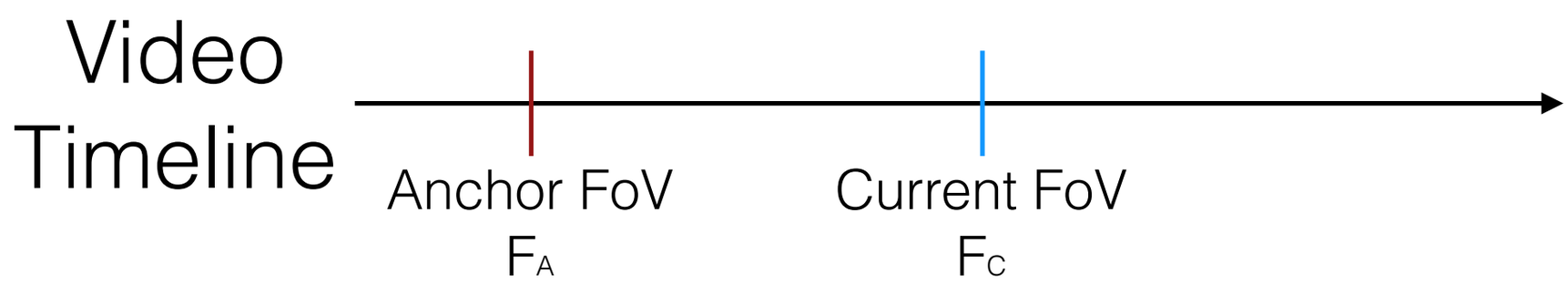
- the combination of rotation and translation




$$Sim(f_1, f_2) = Sim_R \times Sim_T$$

# Realtime Video Segmentation

- segment thresh *T*
- realtime segment point detection while recording
- if a segment point is found,
- $S_i = \{ f \mid f \in [F_A, F_C) \}$ ,
- $[F_A.t, F_C.t)$: the time interval of the segment
- update the anchor FoV: $F_A = F_C$



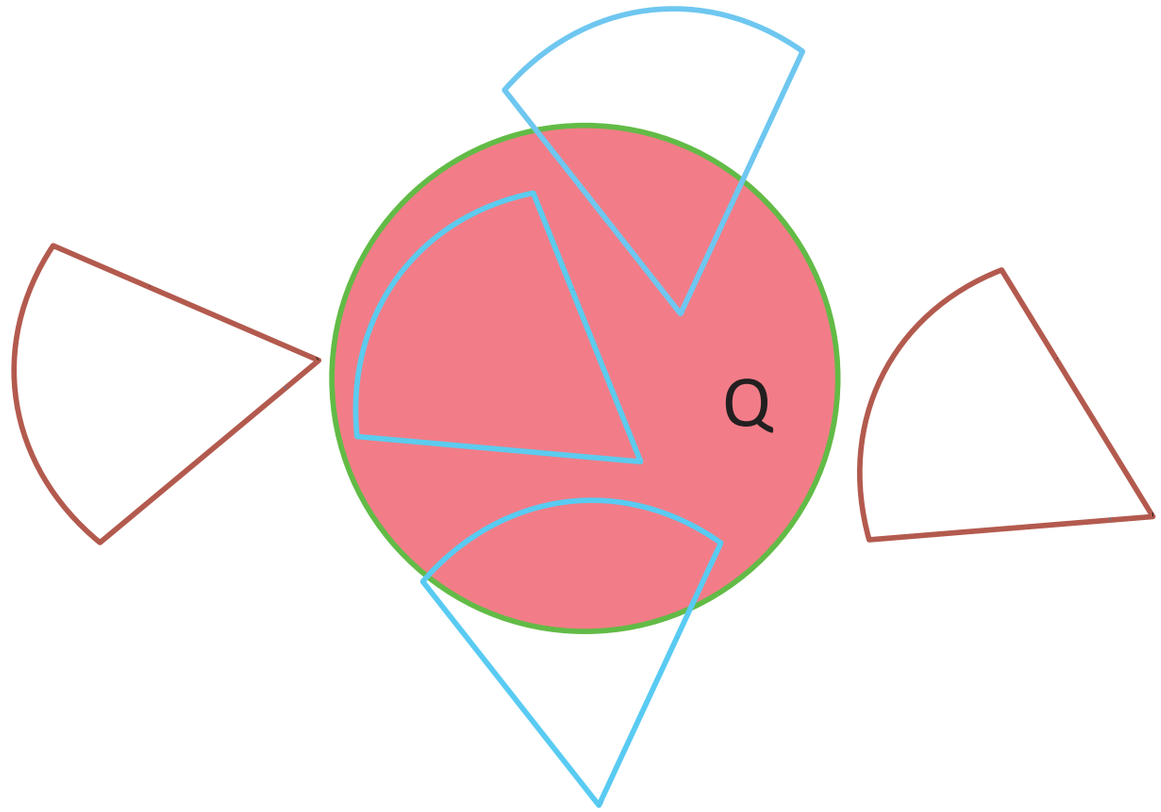
# Content Free Descriptor Extraction

- the representative FoV of each video segment $S_i$:

$$\begin{cases} \bar{p}_i = \dfrac{\sum p}{|S_i|} \\ \bar{\theta}_i = \dfrac{\sum \theta}{|S_i|} \end{cases}$$

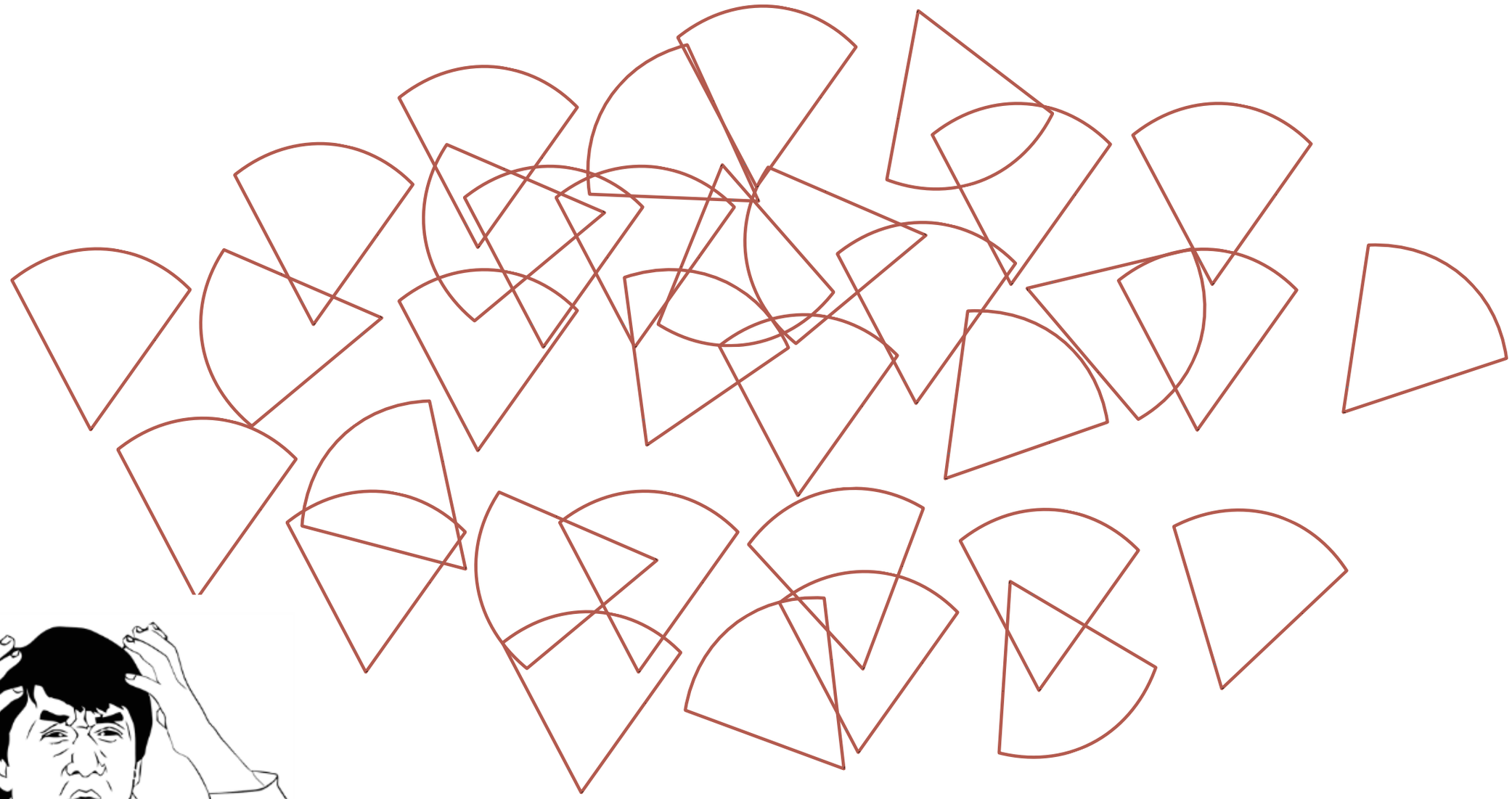
# Indexing and Retrieval

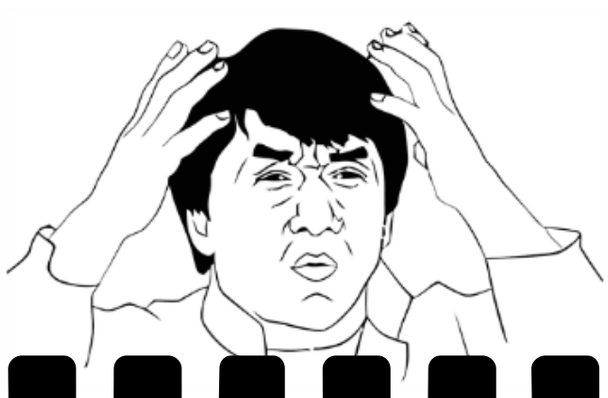
- Basic filtering strategy:
  - intersection with:
    - query range Q
    - time interval($t_s$~$t_e$)



# Indexing and Retrieval

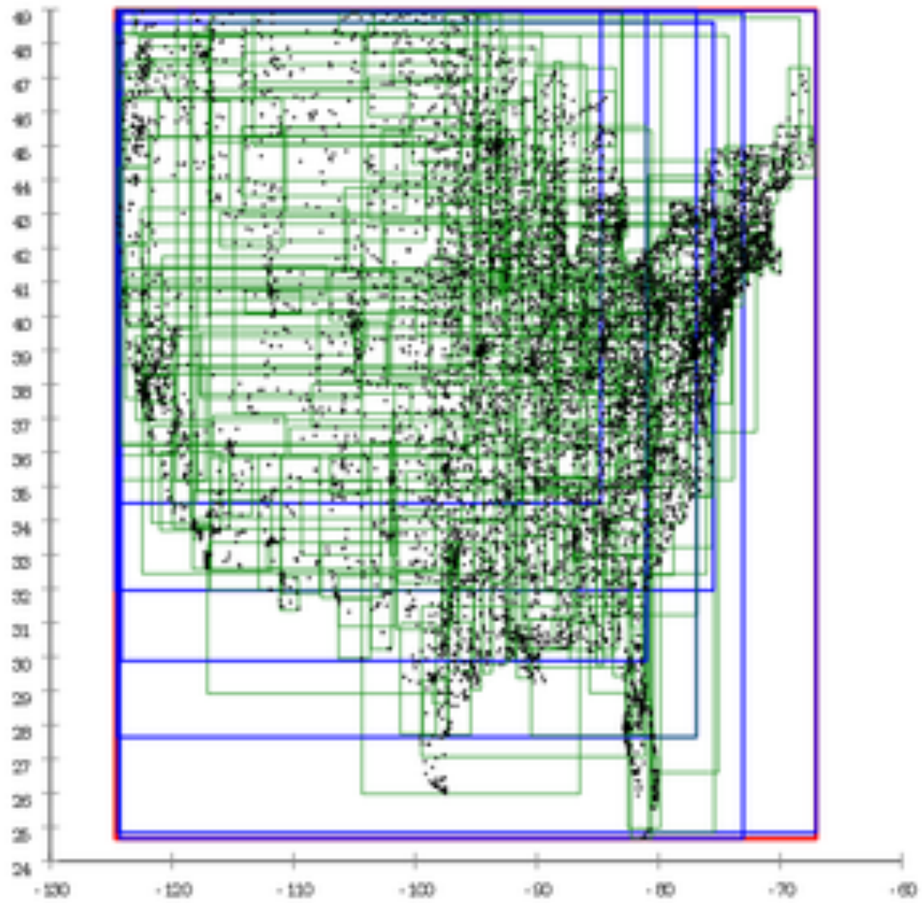
- What if there are massive video segments?





# Indexing and Retrieval

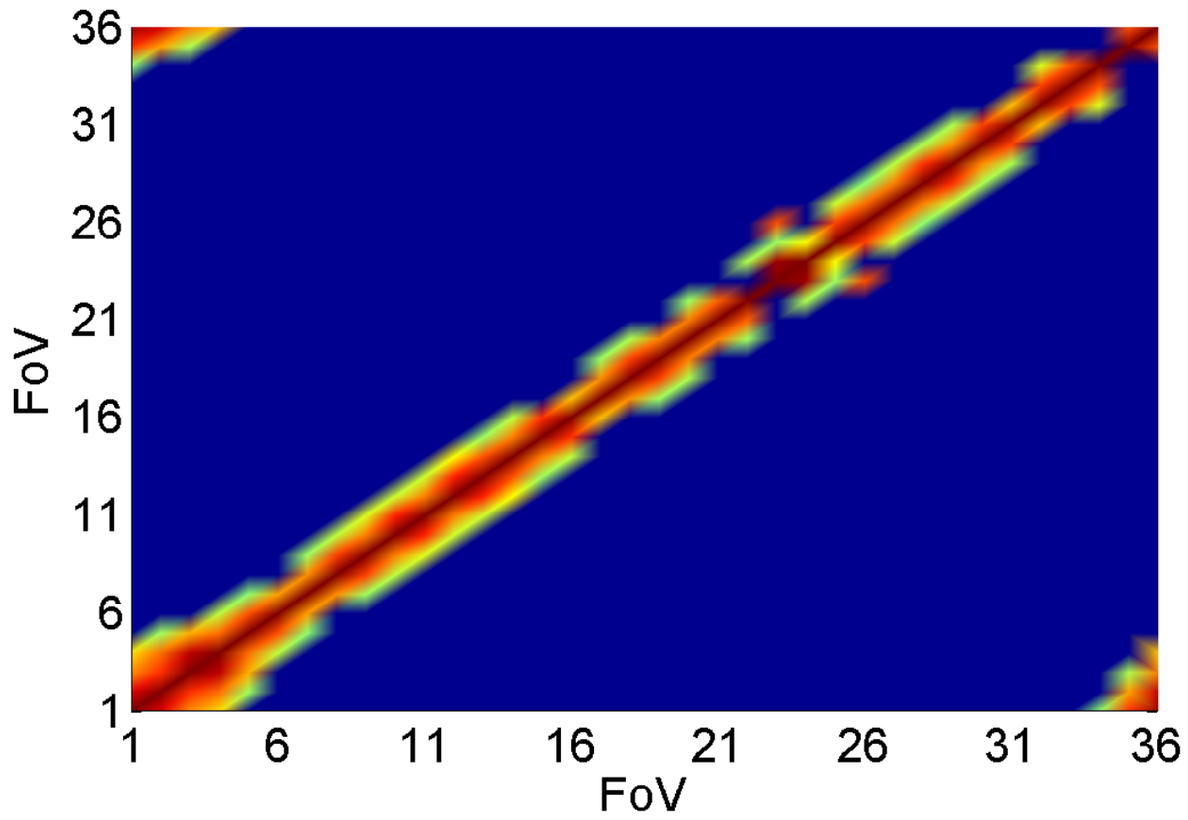
- R-Tree based indexing structure
  - index the location ***p*** and the time interval of the representative FoVs
  - **(p.latitude, p.longitude, $t_s$, $t_e$)**
  - efficiently reduce the computation burden

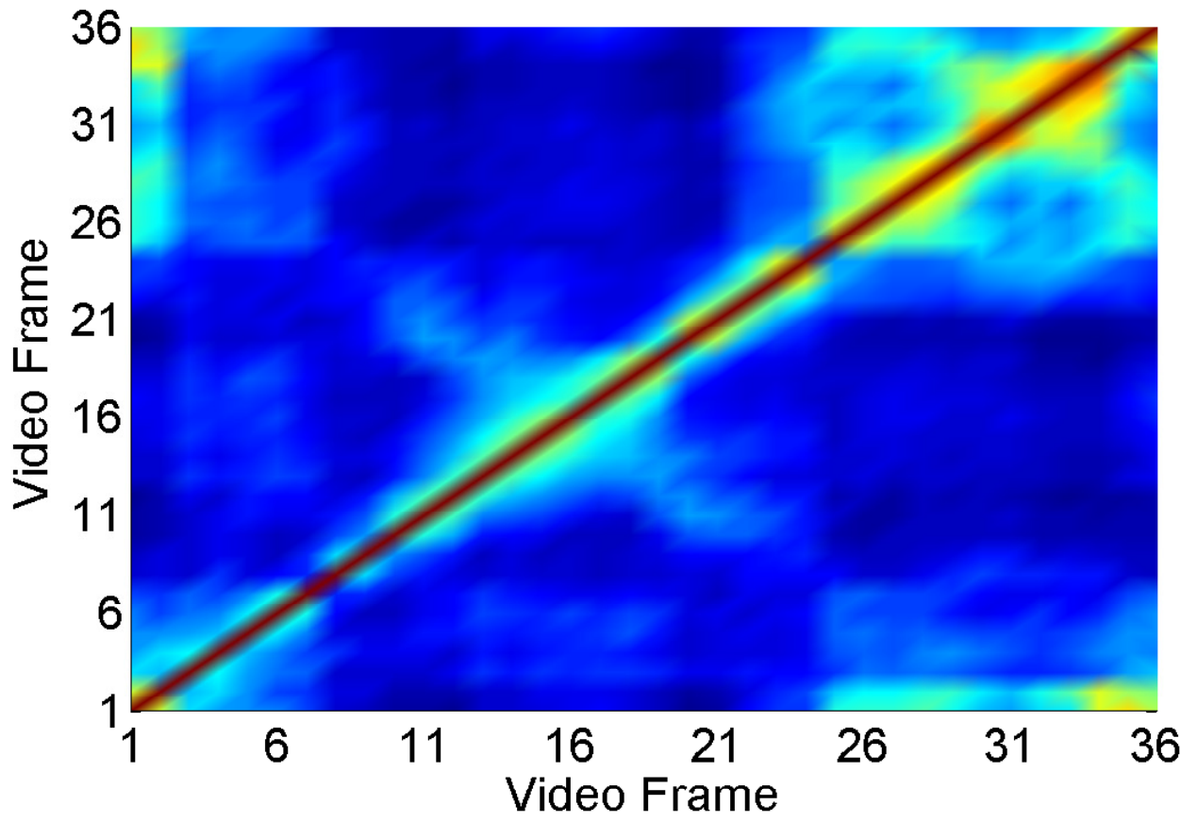


# Evaluation

- Client(smartphone):
  - HTC new One(1.7GHz quad processor and 2GB RAM)
- Server(laptop):
  - Macbook Air MD761

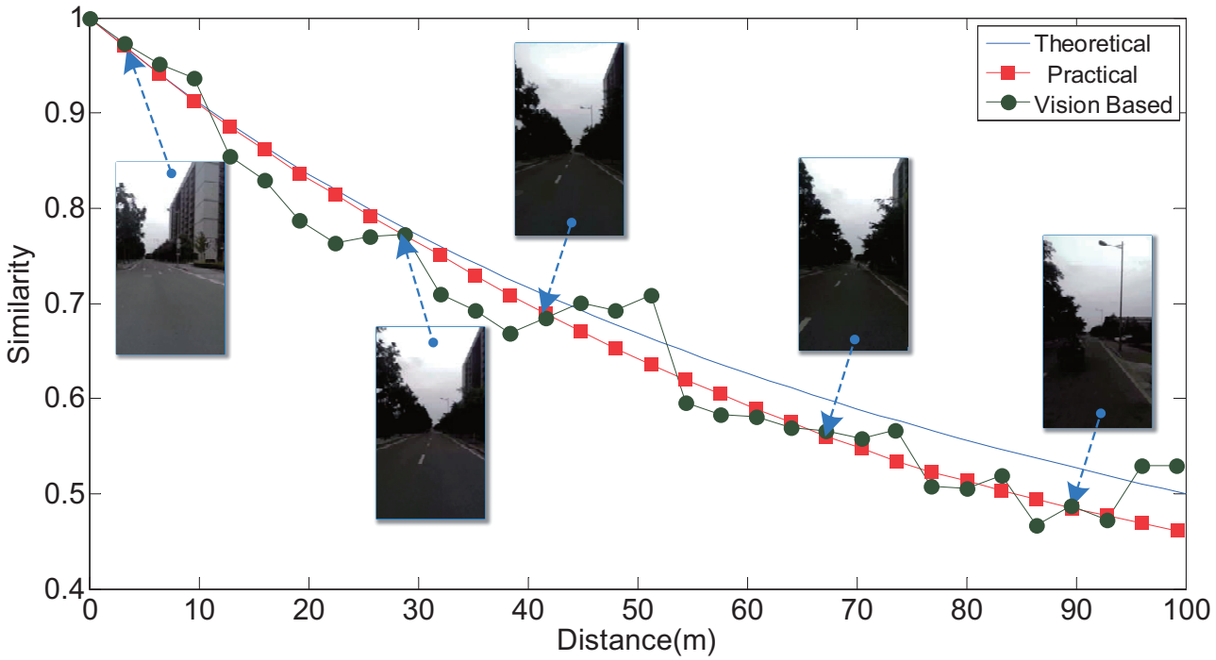
# Similarity Model: Rotation

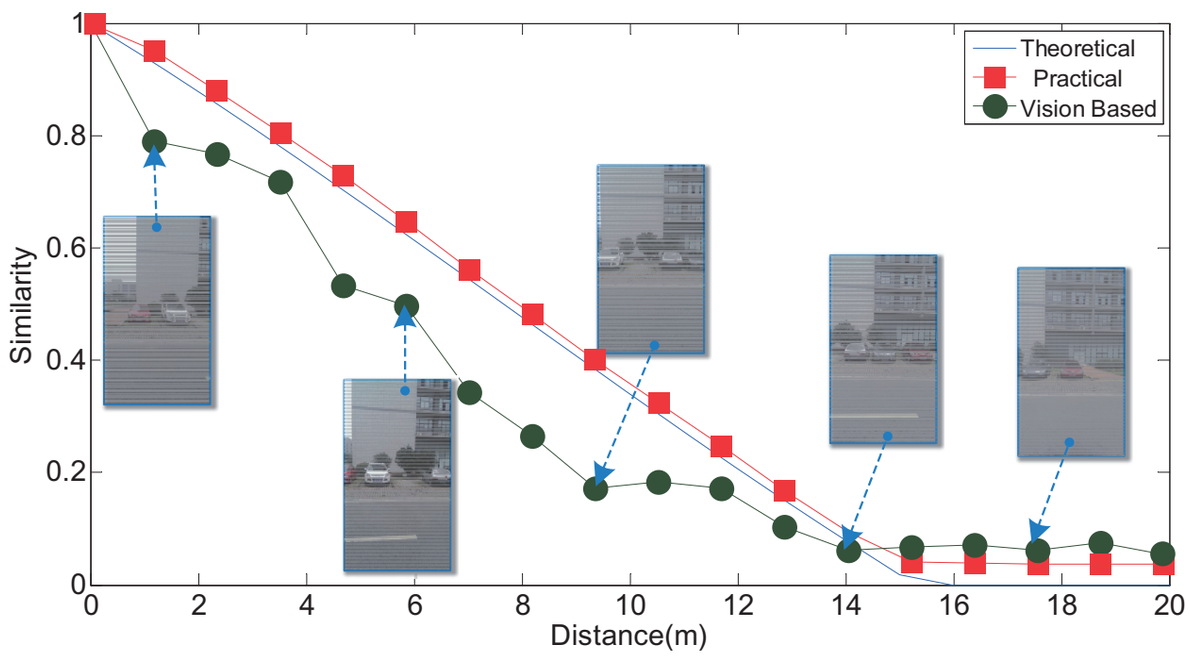



- similarity measure using the FoV based algorithm (content-free) and frame differencing based algorithm

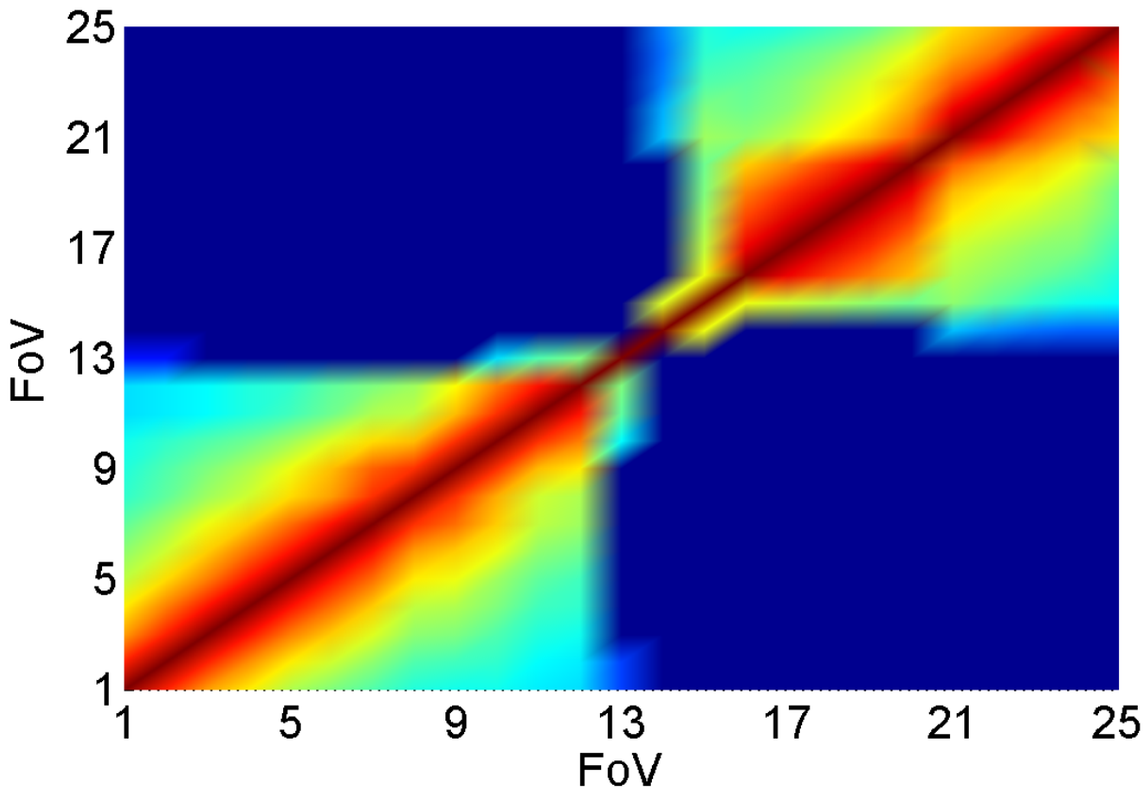
# Similarity Model: Translation

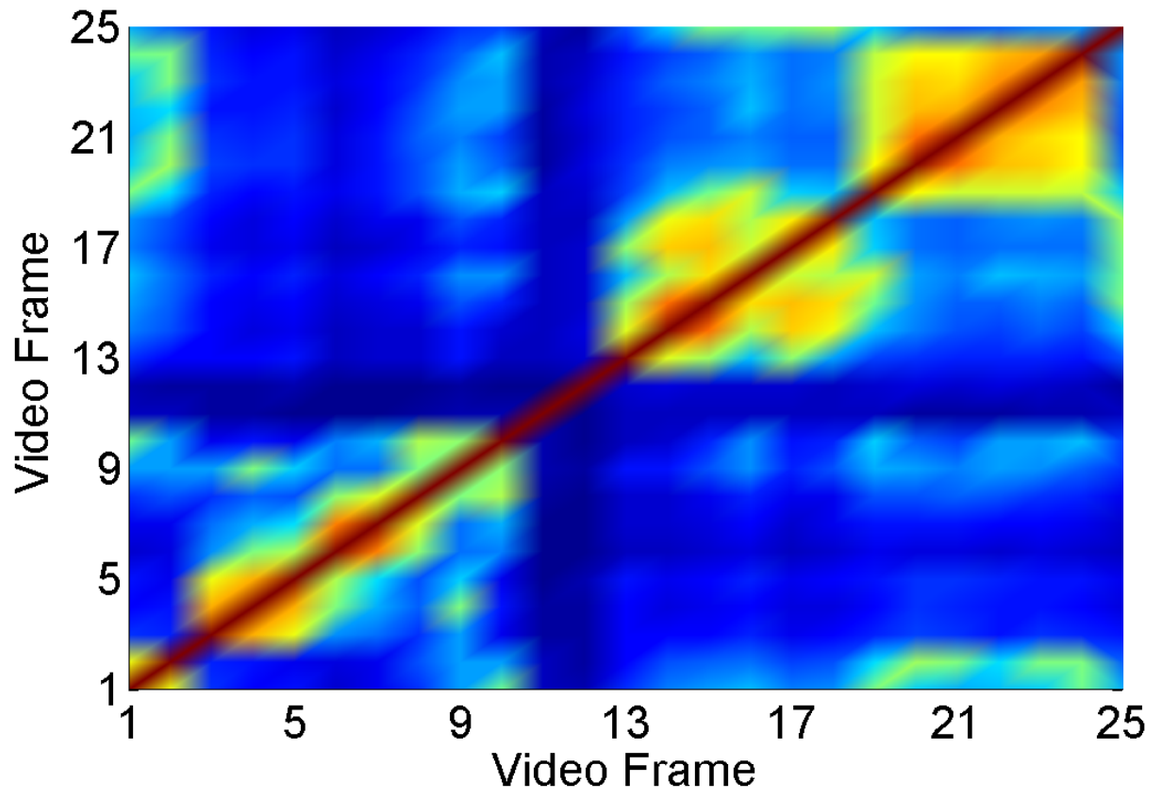



- parallel translation and vertical translation

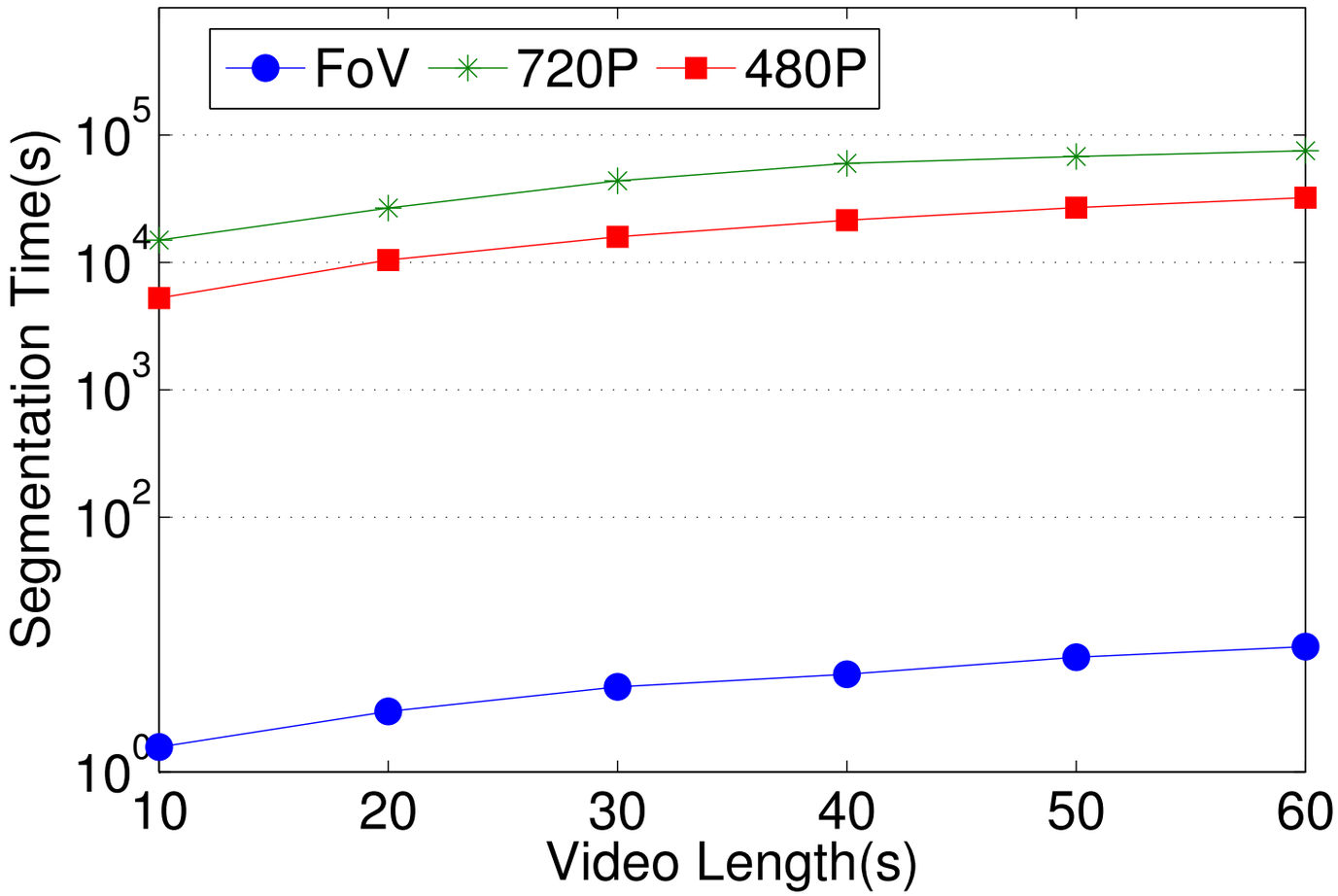
# Similarity Model: Reality



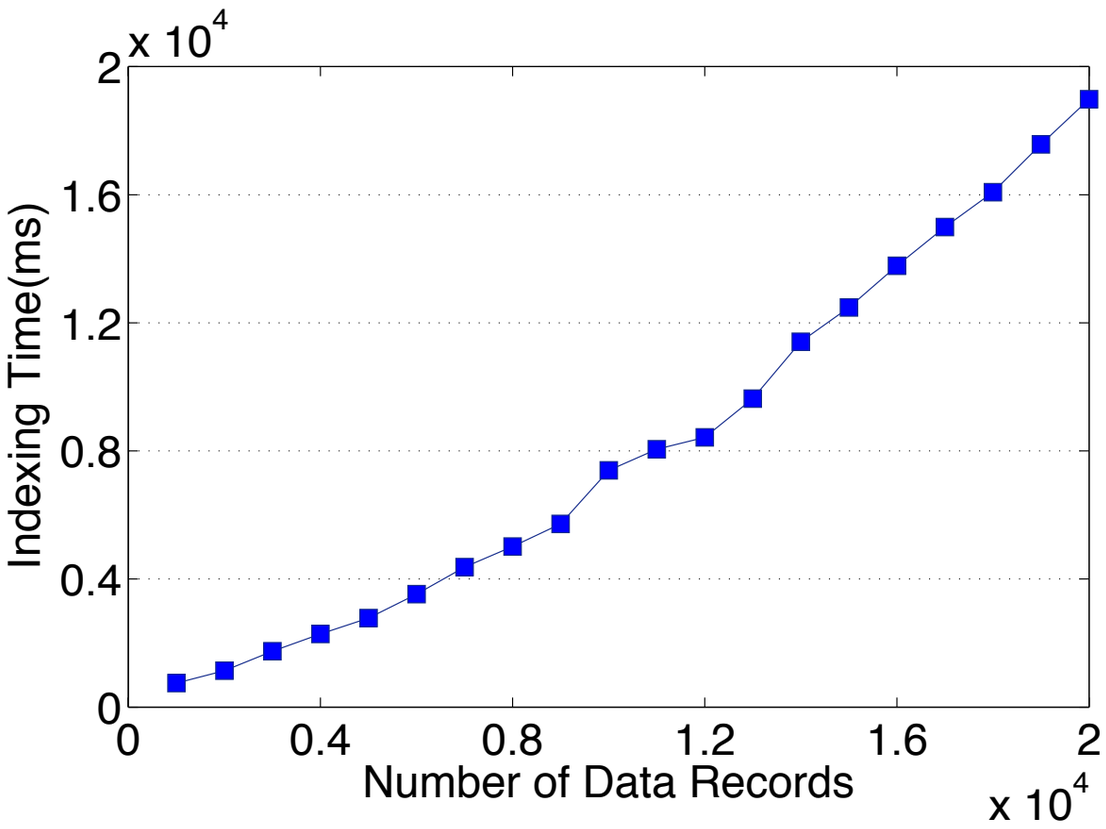

- a combination of rotation and translation

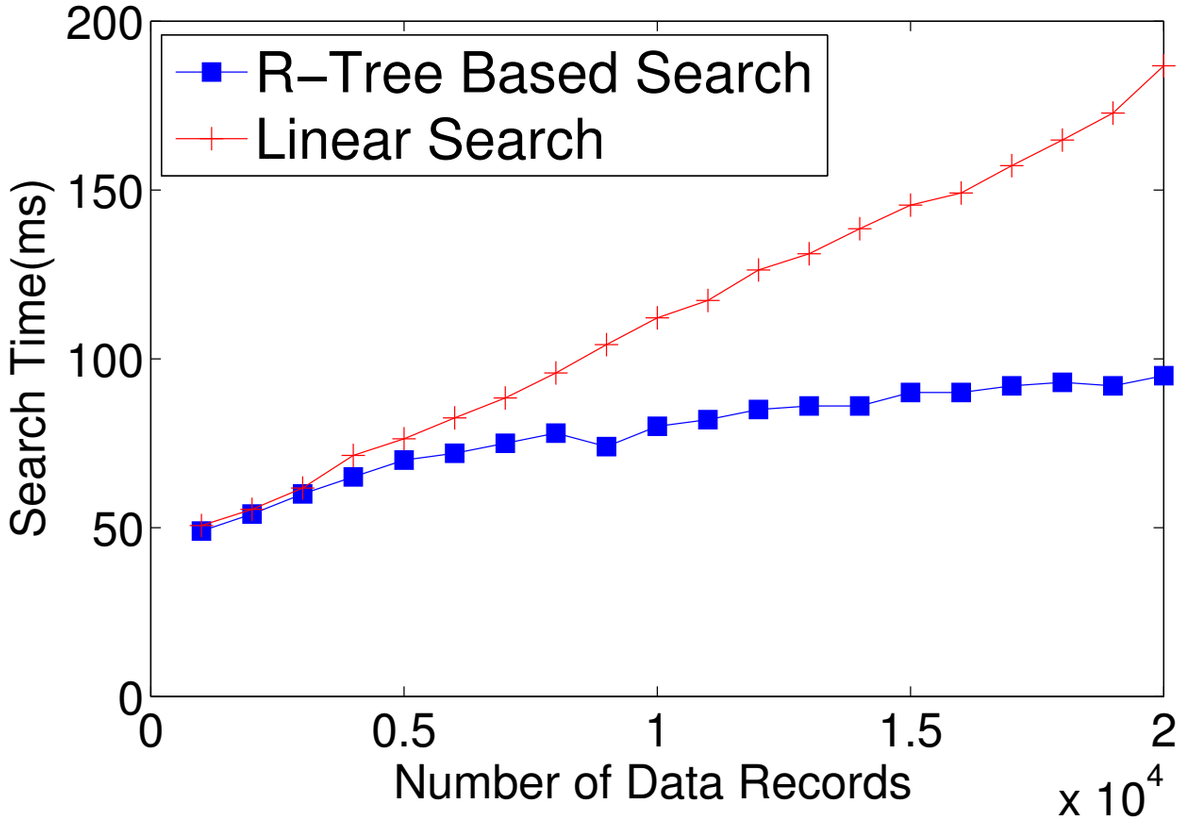
# Segmentation Efficiency




- segmentation time with **different algorithm** using videos with **different resolution**

# Indexing Efficiency




- time of index building and searching

# Conclusion

- Content-free video descriptor
  - **real-time video segmentation**
  - **minimize the web traffic**
  - **comparable performance to Content-based descriptor**
- Efficient mobile video retrieval system
  - **dynamic structure**
  - **high retrieval efficiency**

# *Thanks*

cihang@greenorbs.org