An Object Based Graph Representation For Video Comparison

Xin Feng[1], Yuanyi Xue[2], Yao Wang[2]

1.Chongqing University of Technology, 2. NYU Tandon School of Engineering

ICIP 2017

Introduction and Motivation

➢ **How to represent the content of a video?**

•Current video representation

•SIFT, HOG, HOF, MBH -> Bag-of-words (BOW)

•Features from DNN

How does a human understand the video ?

The Proposed Framework

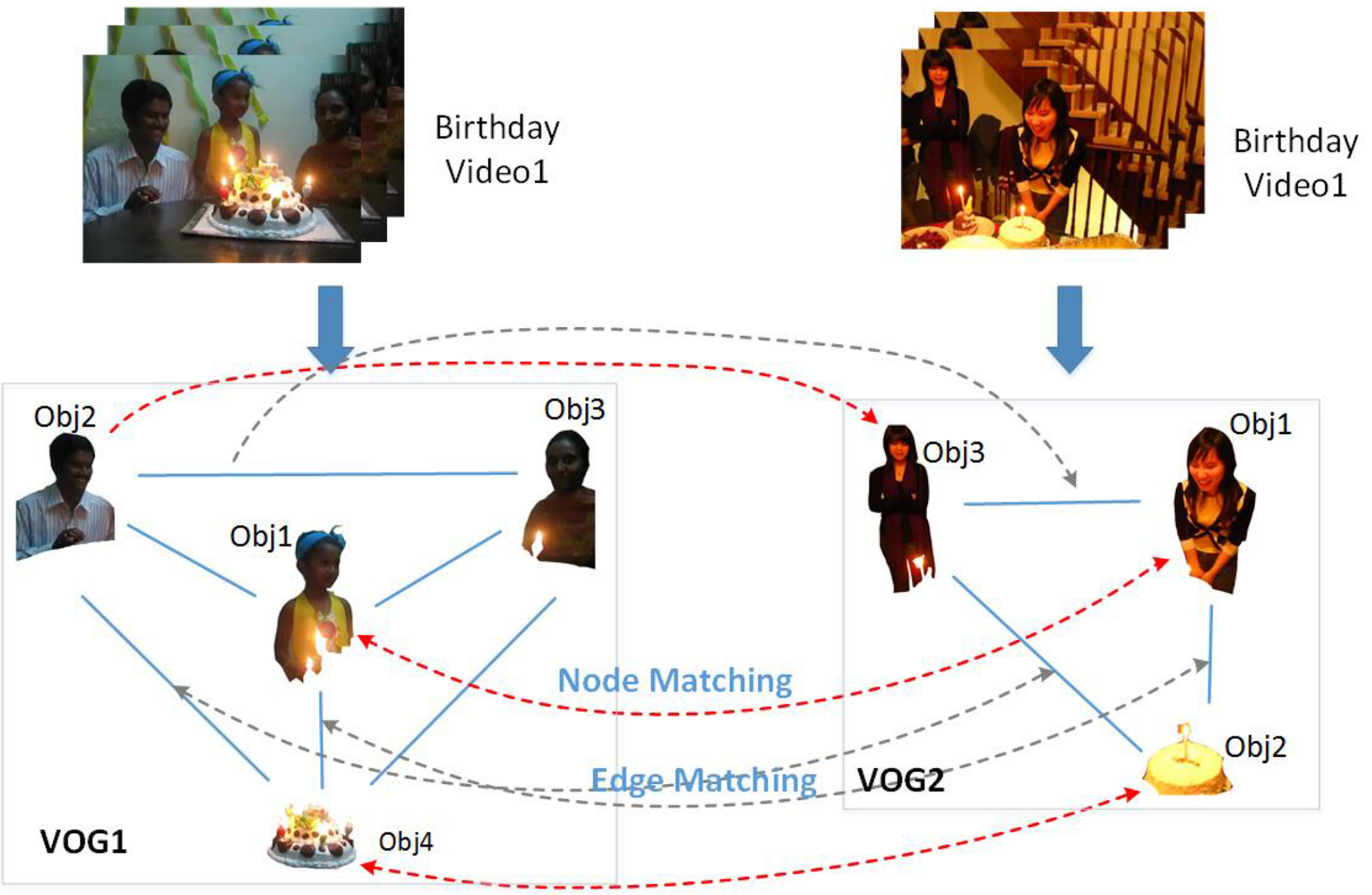
➢ Video Object Graph (VOG) : a single holistic graph representation for describing the semantic content of a video

- the detected objects
- their individual attributes
- Relationship between objects

➢ **Graph Nodes:** objects in a video, described by different spatial and temporal attributes

➢ **Graph Edge**: relative spatial relationships between object

VOG for 'Birthday' videos



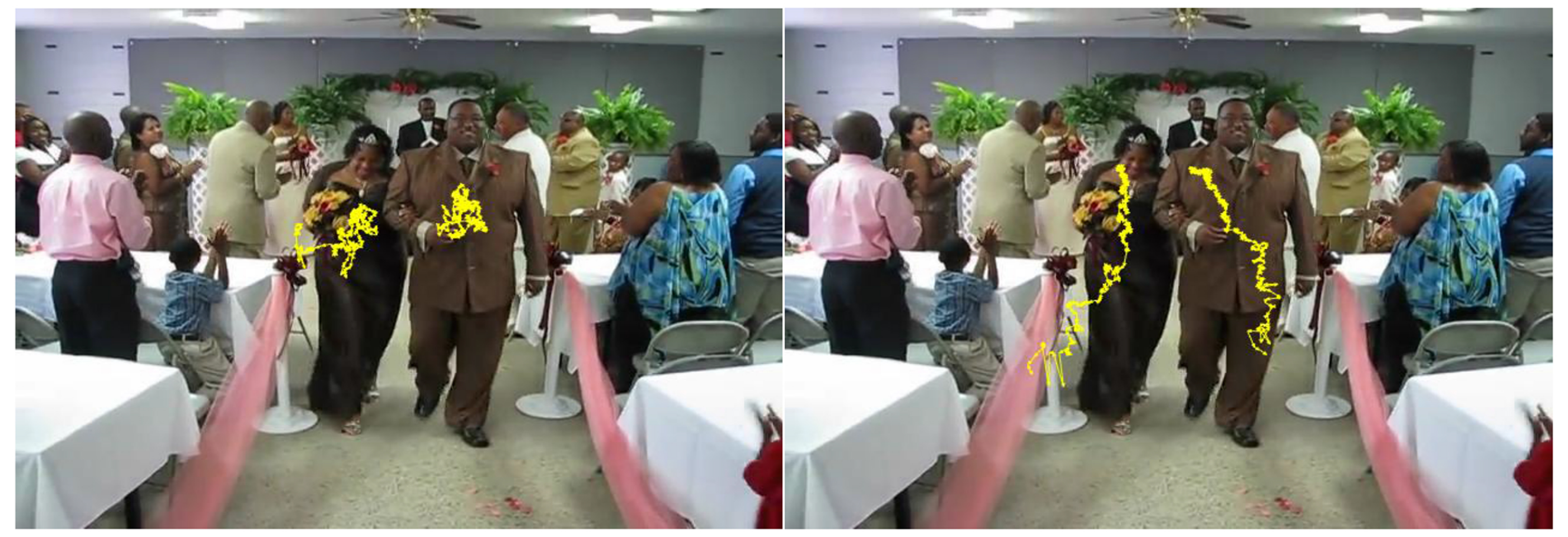
Node and Edge Attribute

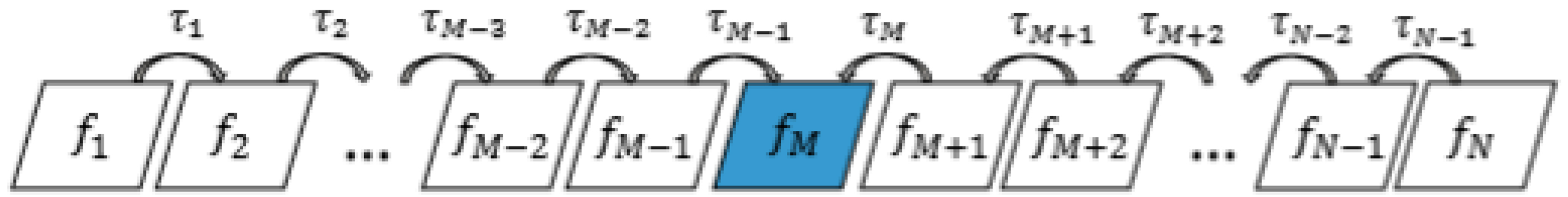
➢ **Node spatial attributes:** Shape context and RCNN feature

➢ **Node temporal attribute:** Motion trajectories of objects

- ● Object tracking
- ● Correcting global motion for object trajectory

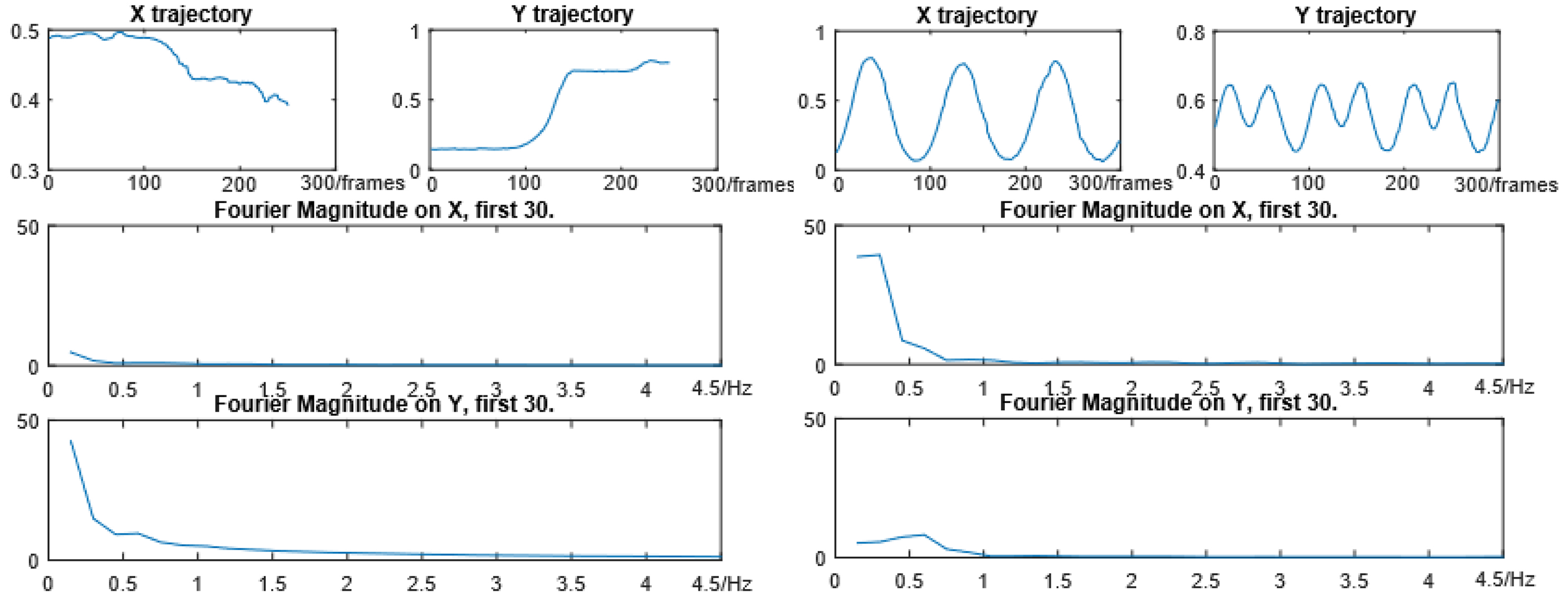
The global motion correction





- ● Two descriptors on the **Fourier spectrum** of the trajectory
  - Magnitude spectrum
  - Peak frequencies

"Sliding" "Swinging"



➢ **Edge attribute:**

- ● trajectory differences between the two connecting objects
- ● to reveal the consistencies of the motion between two objects as well as their spatial displacement

Graph Matching for Video Comparison

➢ Comparison between two videos using graph matching between two VOGs

$$\mathcal{M}(\mathbf{X}) = \sum_{i_1,i_2} x_{i_1,i_2} k^o_{i_1,i_2} + \alpha \sum_{i_1 \neq i_2, j_1 \neq j_2} x_{i_1 i_2} x_{j_1 j_2} k^e_{e(i_1,j_1),e(i_2,j_2)}$$

$$\text{s.t. } \mathbf{X} \in \{0,1\}^{n_1 \times n_2}, \mathbf{X}\mathbf{1}_{n2} \leq \mathbf{1}_{n1}, \mathbf{X}^T\mathbf{1}_{n1} \leq \mathbf{1}_{n2}$$

Datasets

➢ **User generated videos**

➢ Have well-defined objects (people, dog,etc)

➢ Columbia Consumer Video (CCV) dataset

In our study , we use videos from 4 event categories

Experiment Results

The propose VOG is verified on video comparison on the four event detection task. Compared with : **3D sift +BOW , C3D**

	BoW	C3D_*fc6*	VOG (proposed)
SimRatio	0.87	1.10	1.47
Precision	14.4%	41.7%	68.9%
Recall	33.3%	77.8%	100%
Time Cost(s)	0.59	0.06	0.61

Conclusion

The proposed VOG provides a good visual "thumbnail" representation of the video content, and is used for video retrieval, video event clustering ,video summarization, etc