Action Recognition

Hai Vu

Nội dung

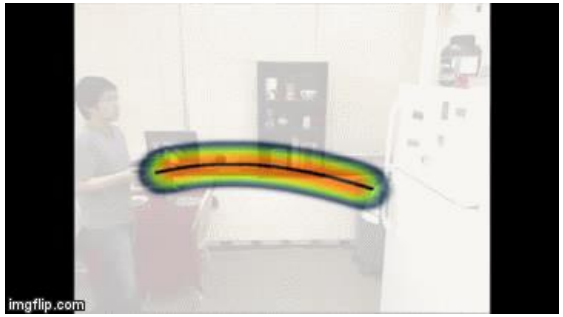
- Giới thiệu bài toán phân tích và nhận dạng hoạt động của người
- Phương pháp phân tích và nhận dạng hoạt động của người
  - Sử dụng đặc trưng không gian-thời gian (Spatio-temporal)
  - Sử dụng đặc trưng về khung xương (skeleton)
  - Một số bộ CSDL nhận dạng hoạt động của người
- Ứng dụng: bài toán phát hiện hành vi bất thường (fighting)

Giới thiệu

- Nhận dạng hoạt động của người có ứng dụng trong rất nhiều lĩnh vực

Game, giải trí, sức khỏe

Phát hiện sự kiện bất thường (ngã) trong bệnh viện



Robotics

Trong mạng camera giám sát

Giới thiệu

- Có rất nhiều hướng tiếp cận

Giới thiệu

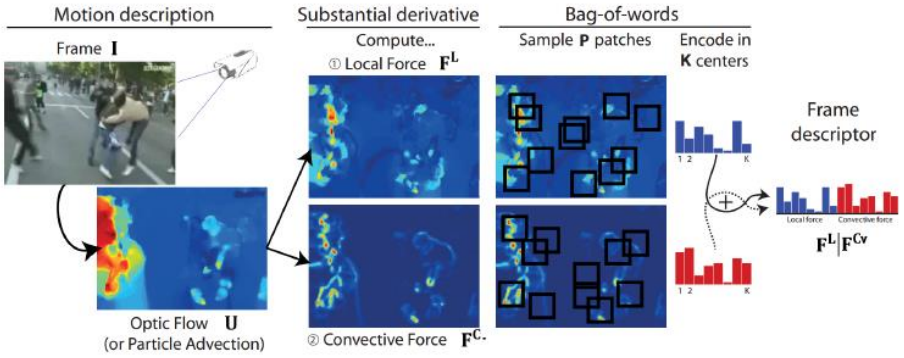
- Có rất nhiều bài toán con cần giải quyết
  - Theo bám (Tracking)
  - Phân đoạn hoạt động từ video liên tục (action spotting)
  - Nhận dạng hoạt động của nhóm người hay 1 người
  - Định nghĩa các loại hoạt động (bình thường vs. bất thường)

Giới thiệu

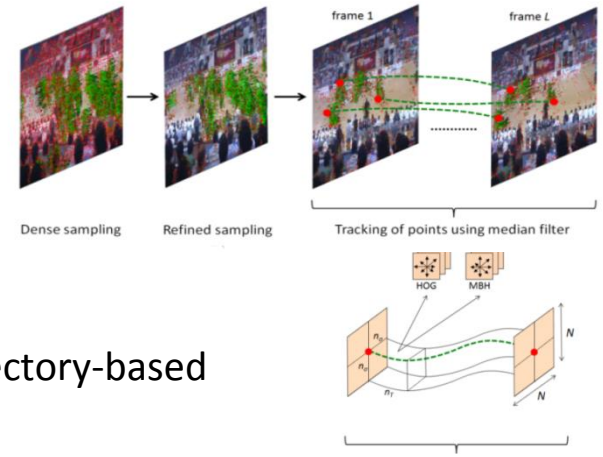
- Một số khó khăn
  - Nếu chỉ dựa vào đặc trưng trực quan (visual) → không bất biến
  - Nếu chỉ dựa vào đặc trưng khung xương → nhiễu, giá trị không ổn định theo thời gian
  - Hoạt động đa dạng, môi trường đa dạng, các góc nhìn/quan sát khác nhau
  - Dữ liệu xử lý lớn (video >> image)
  - Nhầm lẫn khi định nghĩa hoạt động

Một số phương pháp

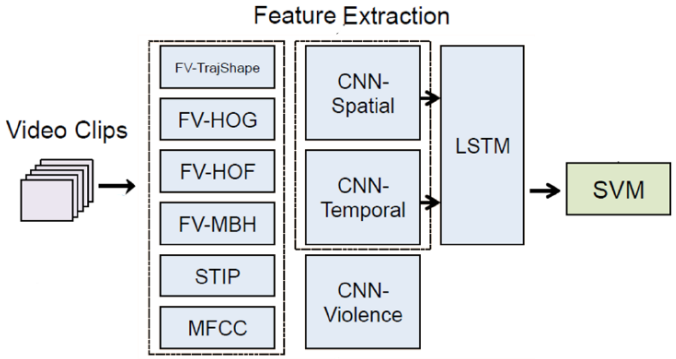
Spatial descriptions



Spatio-temporal descriptions



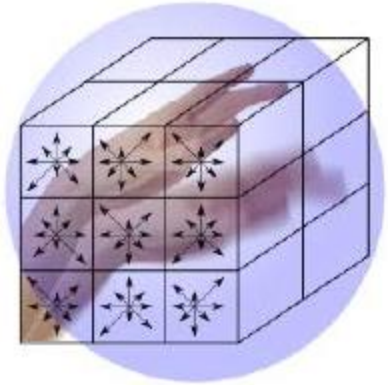
Trajectory-based



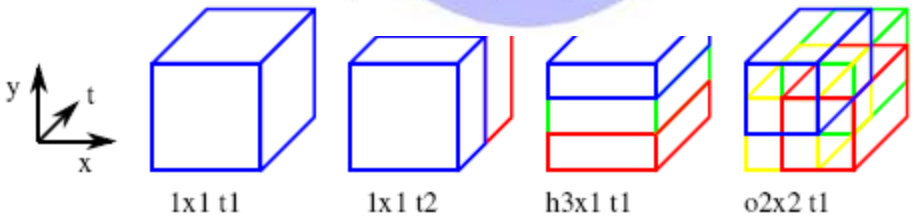
Deep learning-based techniques

Giới thiệu đặc trưng STIP

- Do Ivan Laptev (2005) phát triển



- Kết hợp các đặc trưng không gian và thời gian



STIP Detector (giống Harris Corner 3-D)

- Looks for distinctive neighborhood in the video
  - High image variation in space and time
  - Describe it using distribution of gradient and optical flow

$$H = \begin{pmatrix} L_{xx} & L_{xy} & L_{xt} \\ L_{xy} & L_{yy} & L_{yt} \\ L_{xt} & L_{yt} & L_{tt} \end{pmatrix}$$

Where L_{ij} is $\frac{\partial L}{\partial i \partial j}$

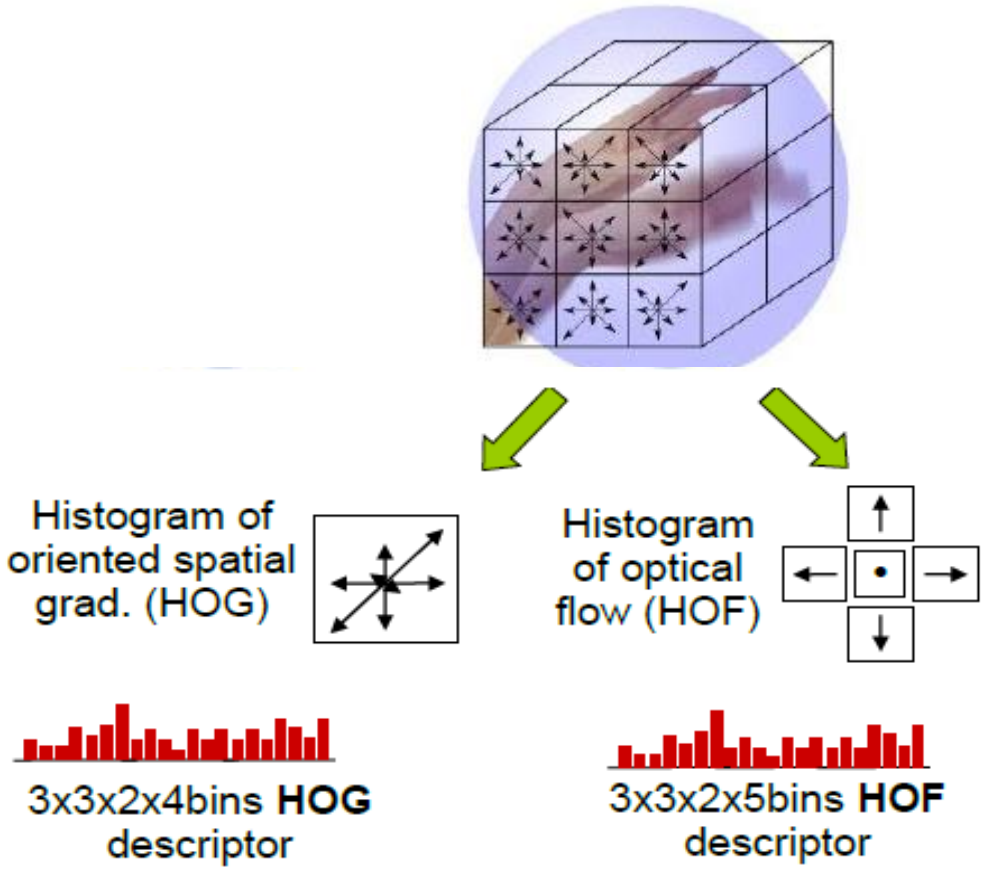
Any (x, y, t) location in the video is STIP if

$$\det(H) + \alpha * trace^3(H) > TH$$

I. Laptev. On Space-Time Interest Points. IJCV, 2005

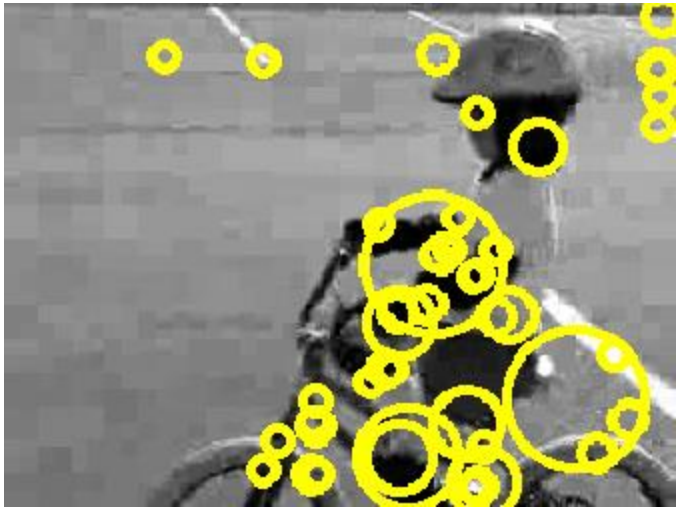
Algorithm Details: STIP Descriptor

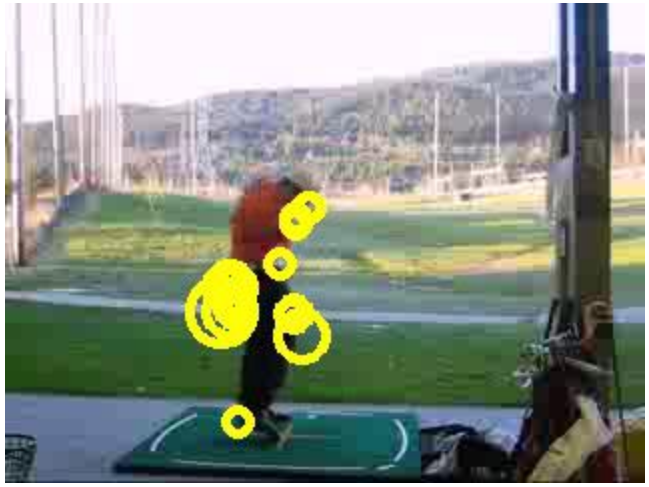
- Small spatio-temporal neighborhood extracted
- Divided into 3x3x2 tiles

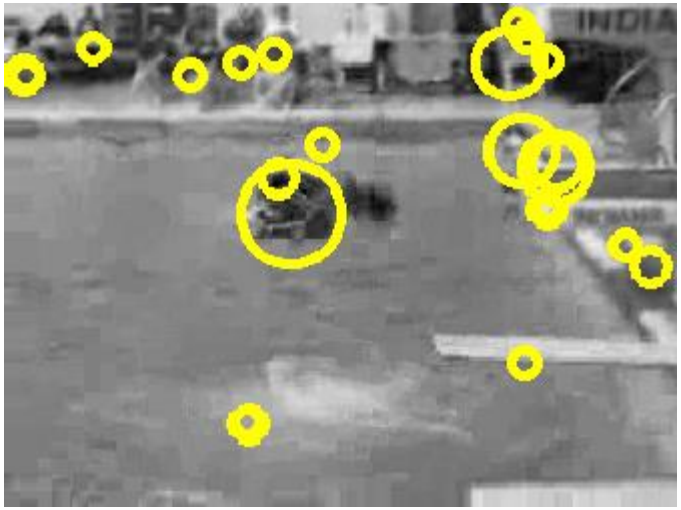


I. Laptev. On Space-Time Interest Points. IJCV, 2005

Kết quả phát hiện STIPs

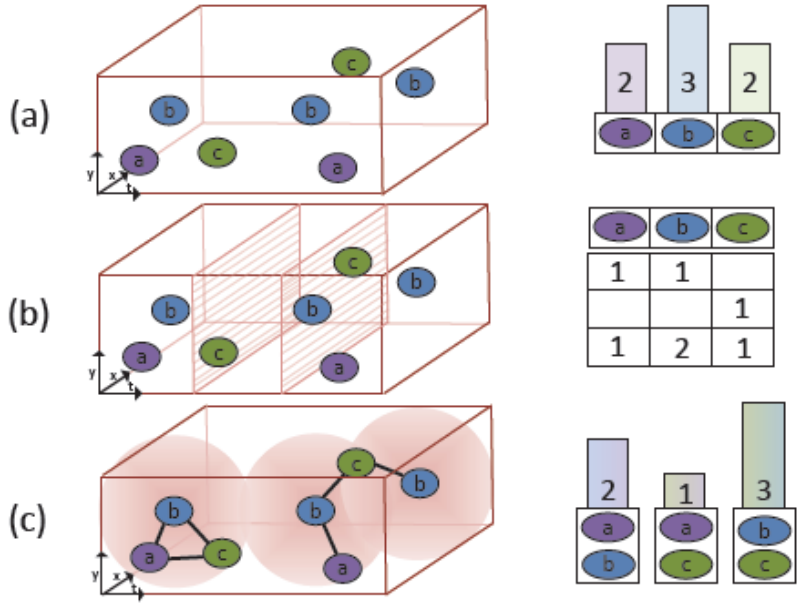






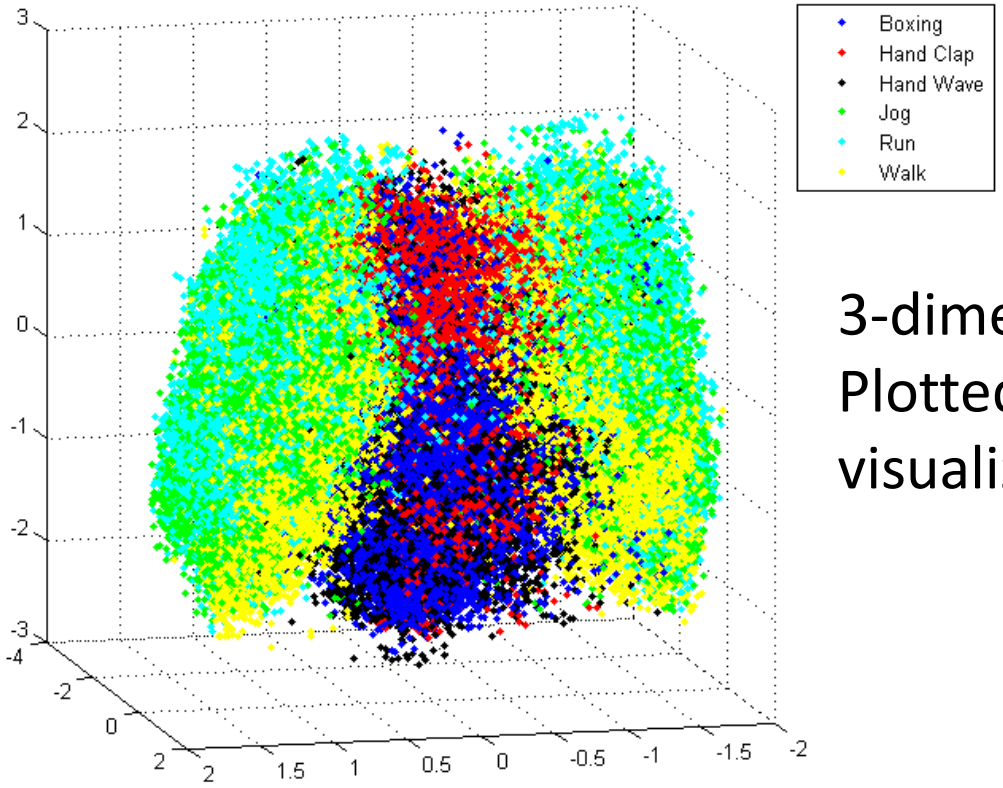
Phương pháp nhận dạng hoạt động sử dụng STIP

1. Sử dụng BoW
2. Xây dựng mạng Bayesian Networks → phản ánh trạng thái (state) của mỗi hình trạng
3. Học và so sánh Dynamic Shape sử dụng lý thuyết đa tạp (Manifolds)



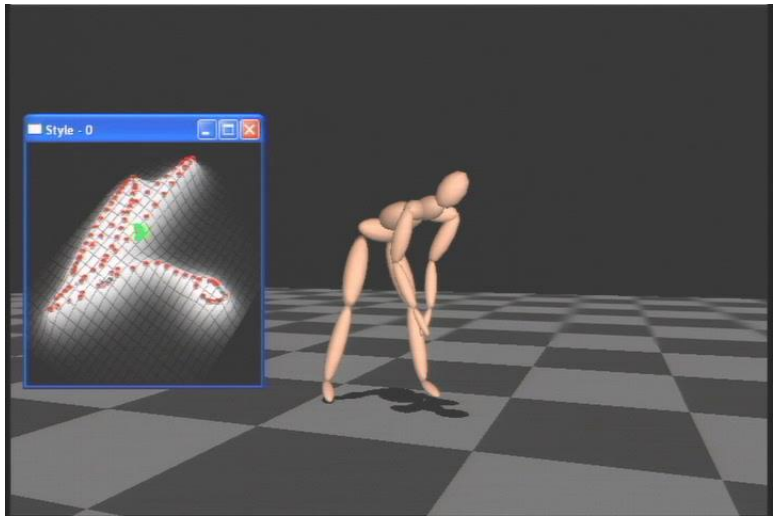
→ Làm chi tiết ở bài thực hành (sử dụng BoW)

- Ví dụ minh họa về 6 hoạt động biểu diễn bằng 3 features (sau khi dùng PCA của STIP)

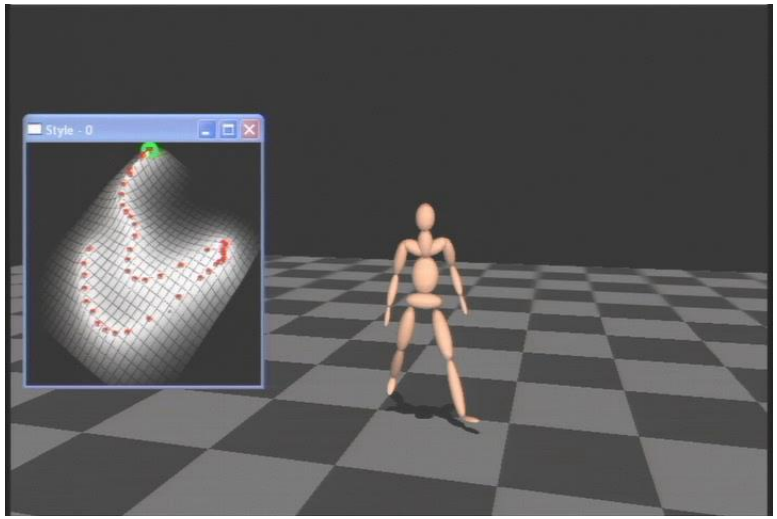


3-dimensions
Plotted for
visualization

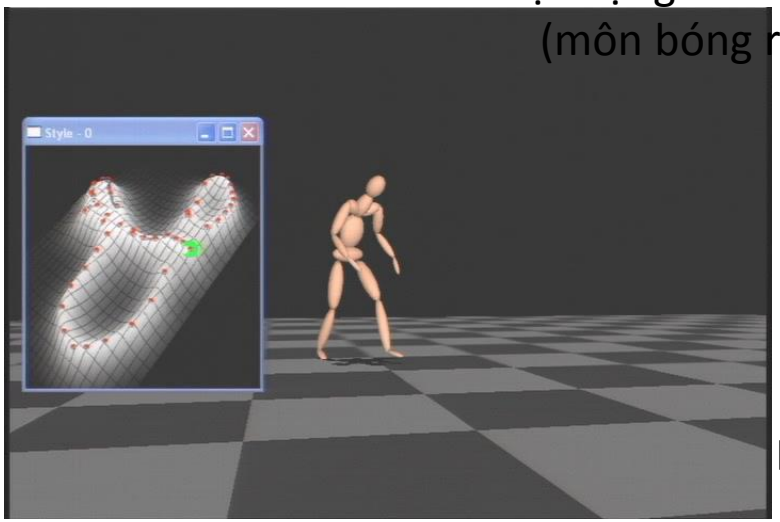
Biểu diễn bằng đa tạp (manifolds)



Hoạt động ném bóng
(môn cricket)

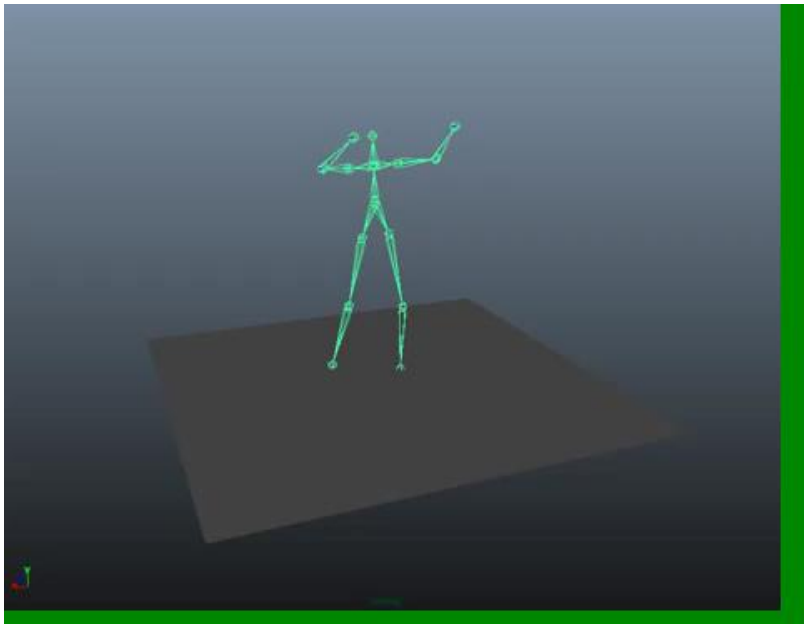


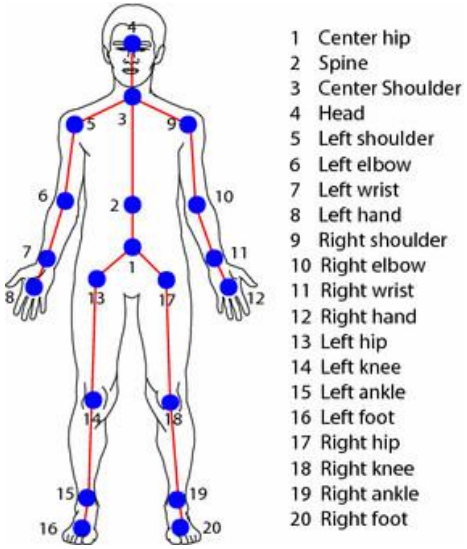
Hoạt động ném bóng
(môn bóng rổ)



Hoạt động xuất phát
(môn chạy)

Phương pháp sử dụng khung xương

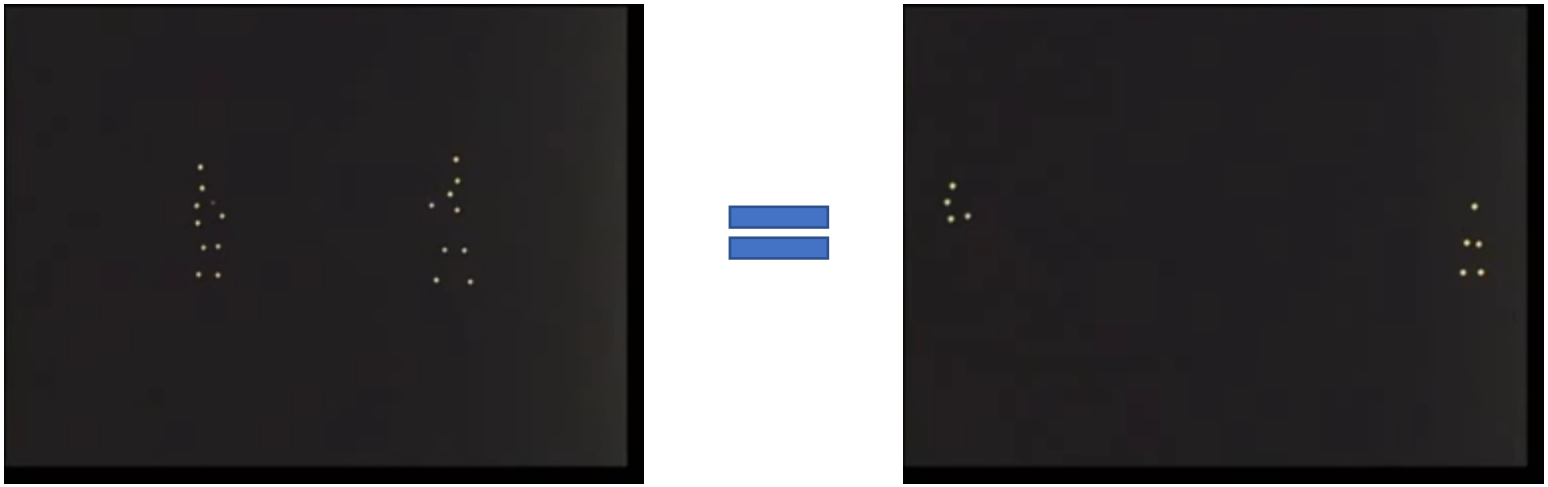




- Advantages of skeletal feature:
  - ☺Visual human appearance independence
  - ☺More discriminative
  - ☺Low computational time
  - ☺Less storage

Phương pháp sử dụng skeleton

- An experiment on motion perception is conducted by [G. Johansson 1971]



→ Specific joints engage an action.

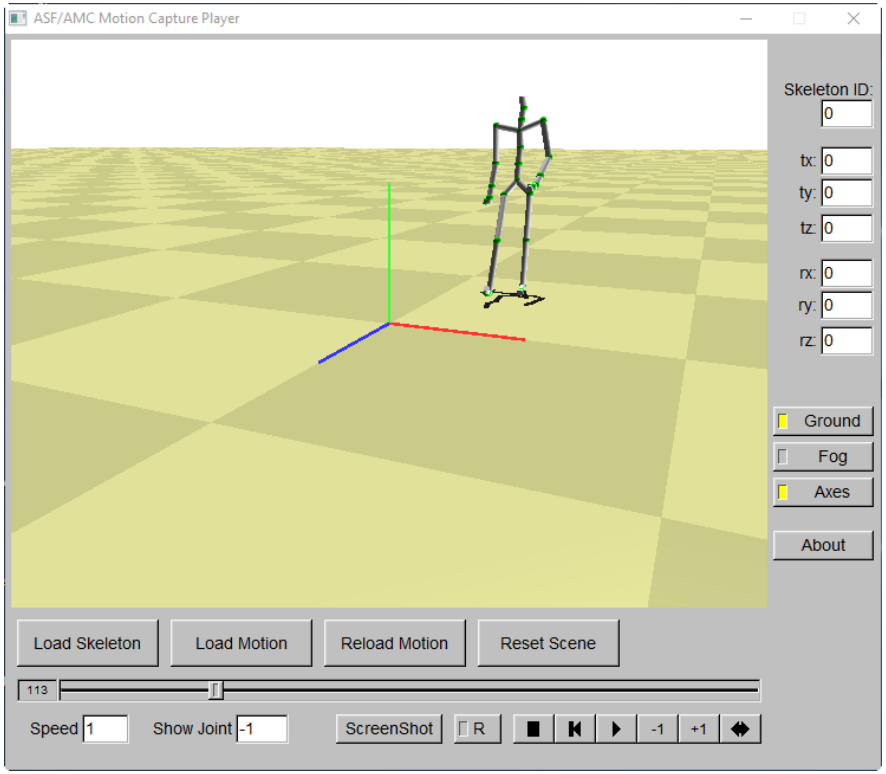
Informative Joints

- Joint Position Variance
- Joint Angle Variance
- Joint Angular Velocity

Cách trích chọn khung xương

- Sử dụng các bộ Motion capture

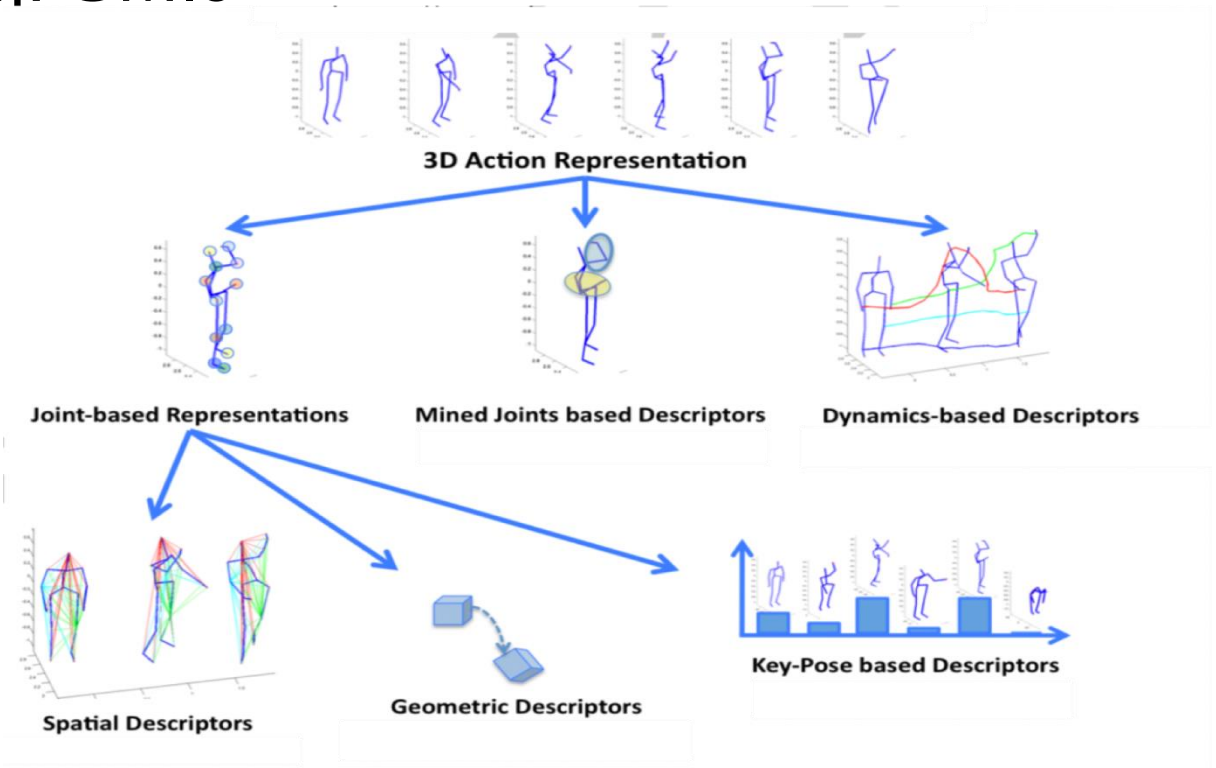
  (gồm các markers đính trên người)
- Trích chọn khớp từ dữ liệu hình ảnh (độ sâu) như: kinect
- Sử dụng các mô hình invert Kinectics



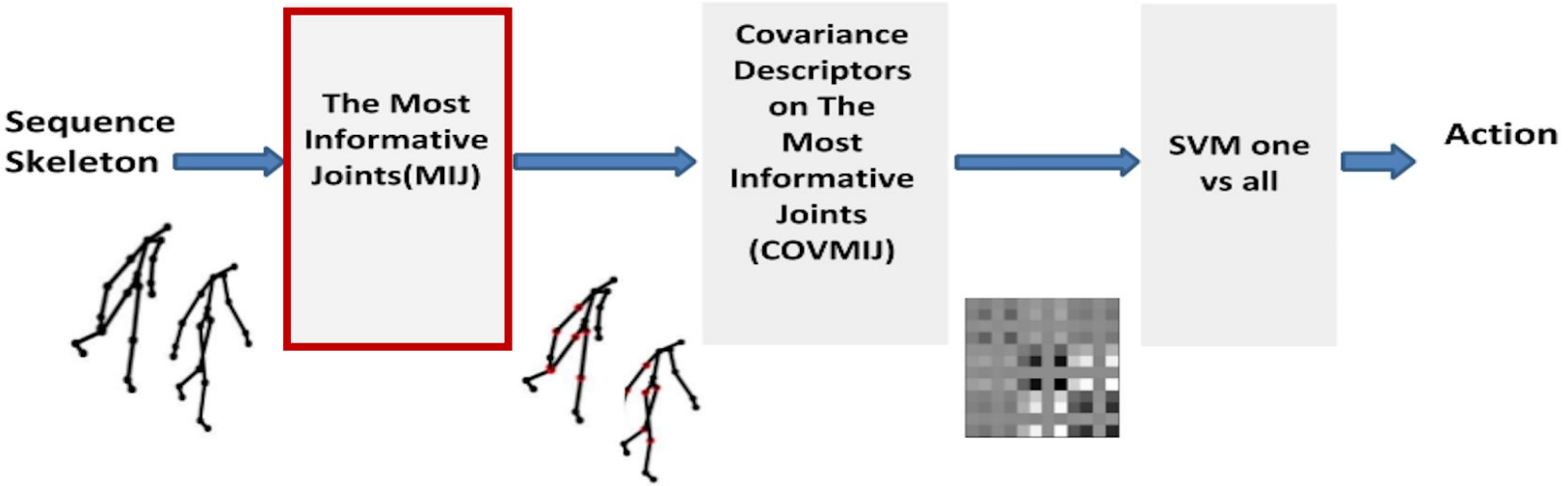
Source: http://mocap.cs.cmu.edu/tools.php

Skeleton-based approaches

- [Hussein2013]: Covariance Descriptor
- [Ofli2012]: SMIJ



Proposed Framework



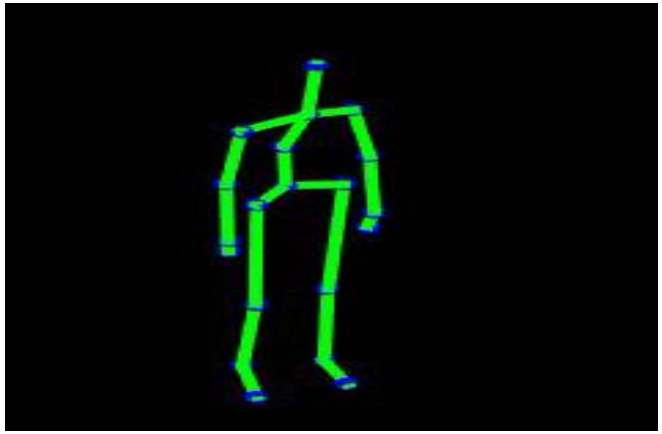
- Key idea: using Covariance Descriptors on Most Informative Joints.
- Main contributions:
  - Detection of **most informative joints**
  - Add **temporal information** into Covariance descriptors

Proposed Framework

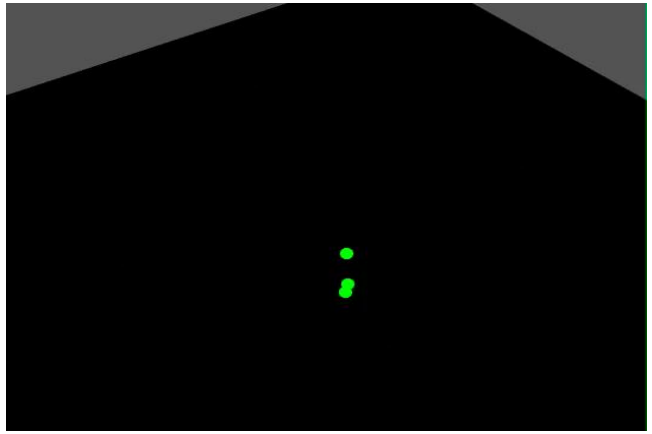
- The Most Informative Joints
  - Purpose: Identify the most informative joints (MIJ) on each action.
  - Most informative joints are defined as the joints with highest variation in 3D positions
- Process
  - Step 1 : Detecting candidates of most informative joints of each action for each people
  - Step 2 : Detecting the most informative joints of each action

Proposed Framework

- Step 1: Detecting candidates of most informative joints of each action for each people



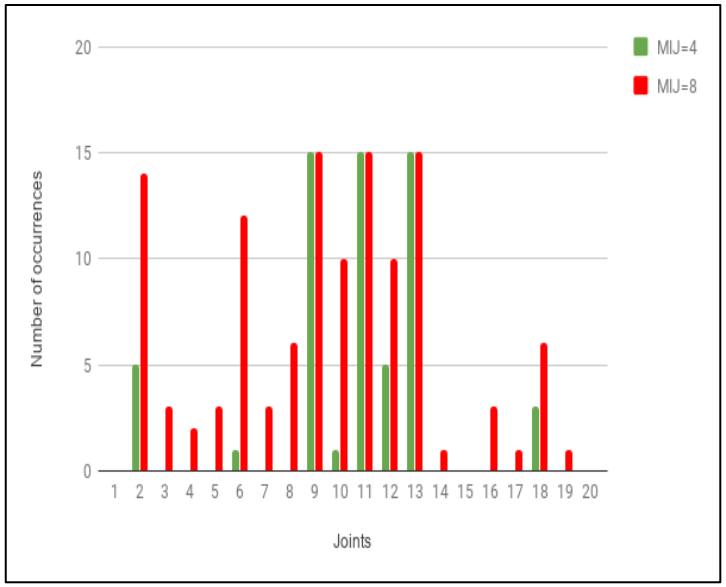
Action **Throw**



Candidates of MIJ for Action **Throw**, determined through Step 1

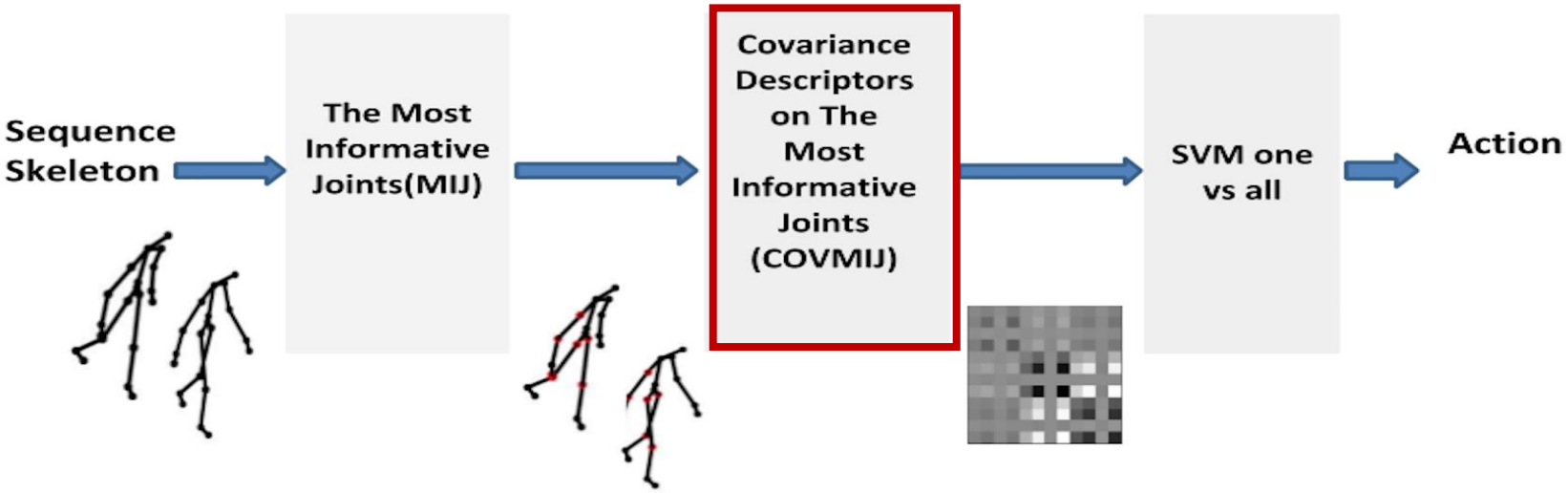
Proposed Framework

- Step 2: Detecting the most informative joints of each action
  - For each action performed by each subject other people, a set of the MIJ is determined by Step 1
  - Determining the MIJ for the to-be-considered action through voting scheme on the sets of MIJ of all peoples



Voting

Proposed Framework



- Key idea: using Covariance Descriptors on Most Informative Joints

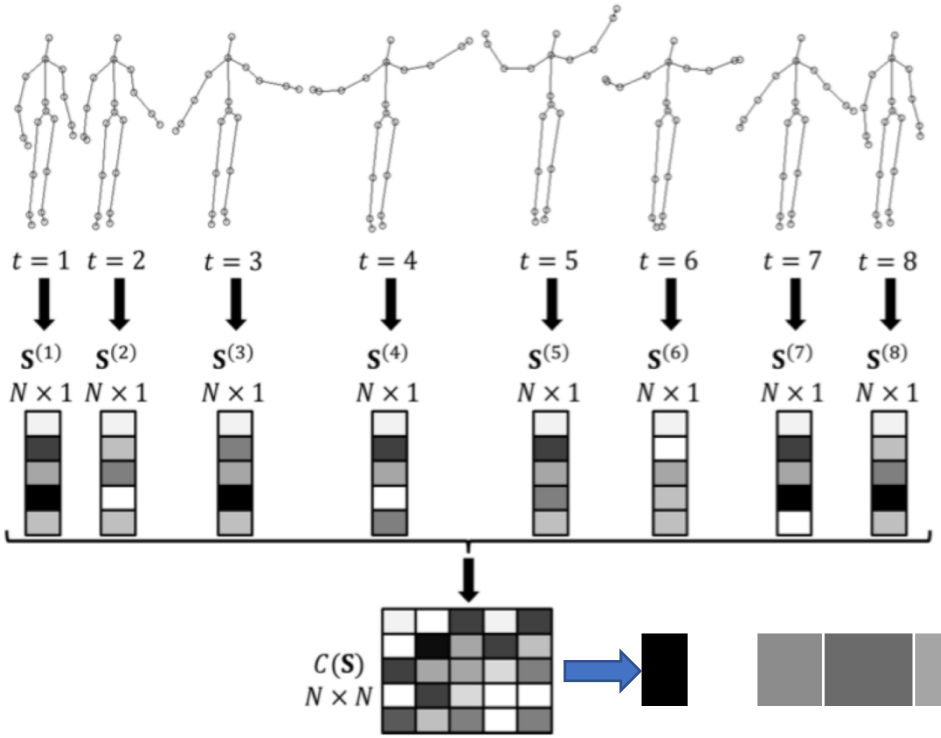
Covariance Descriptors(Cov3DJ)

- Original work by Hussein et al. in JICAI'13
- Main ideas:
  - Using Covariance Descriptor for action representation
  - Using Temporal Hierarchy for action representation at different levels
- Advantages:
  - Compact representation (a vector of 1830 dimensions for each skeletal sequence at one layer of skeleton with 20 joints)
  - Taking into account the order of motion in time
  - Efficient on MSRAction 3D, MSRC12, HMD05 datasets

This method employs information of all joints. However, the engagement of the joints in the actions is different.

Covariance Descriptor (Cov3DJ)

$$\begin{bmatrix} X_1^1 & X_1^2 & X_1^3 & X_1^T \\ Y_1^1 & Y_1^2 & \dots & Y_1^T \\ Z_1^1 & & & Z_1^T \\ & & & \dots \end{bmatrix}$$



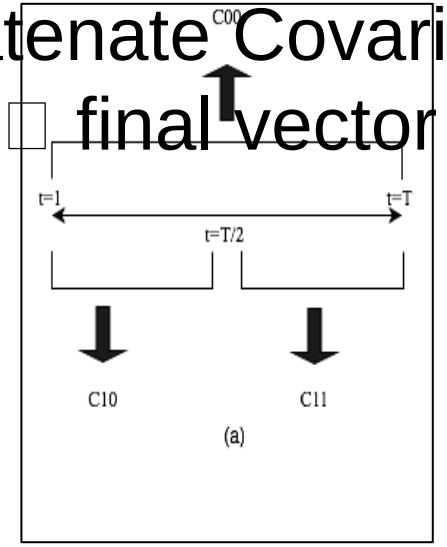
$$S = [X_1, X_2, .., X_k, Y_1, Y_2, ...Y_k, Z_1, Z_2, ...Z_k]'$$

$$C(S) = \frac{1}{T-1}\sum_{t=1}^{T}(S-\bar{S})(S-\bar{S})'$$

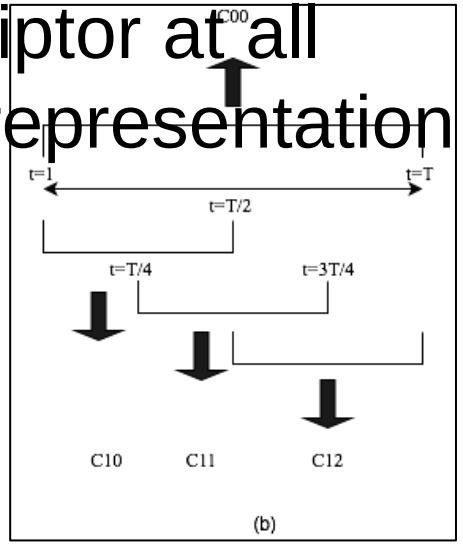
x, y, z are coordinate of k joints at time t
k is number MIJ

Temporal Hierrachy

- Compute Covariance Descriptor at different levels
- Two options: overlapping and non-overlapping
- Concatenate Covariance Descriptor at all levels final vector for action representation

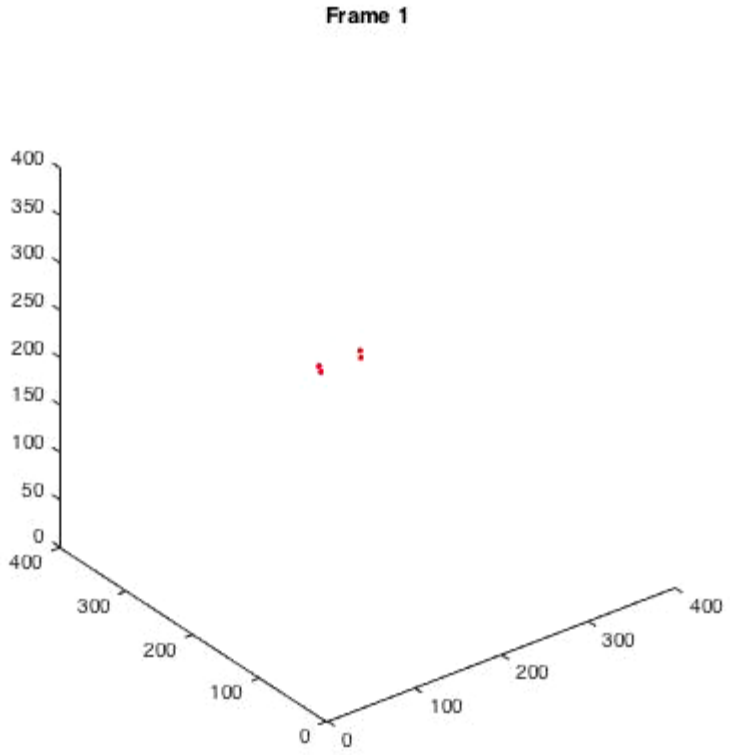


Non-overlapping



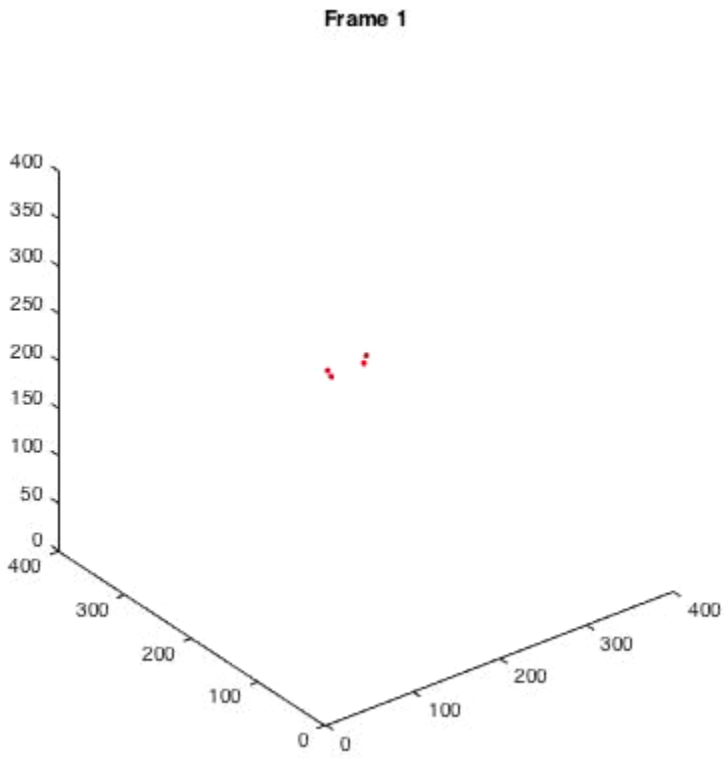
Overlapping

Most Informative Joint Detection



Hand Clap
(MSRAction 3D Dataset)

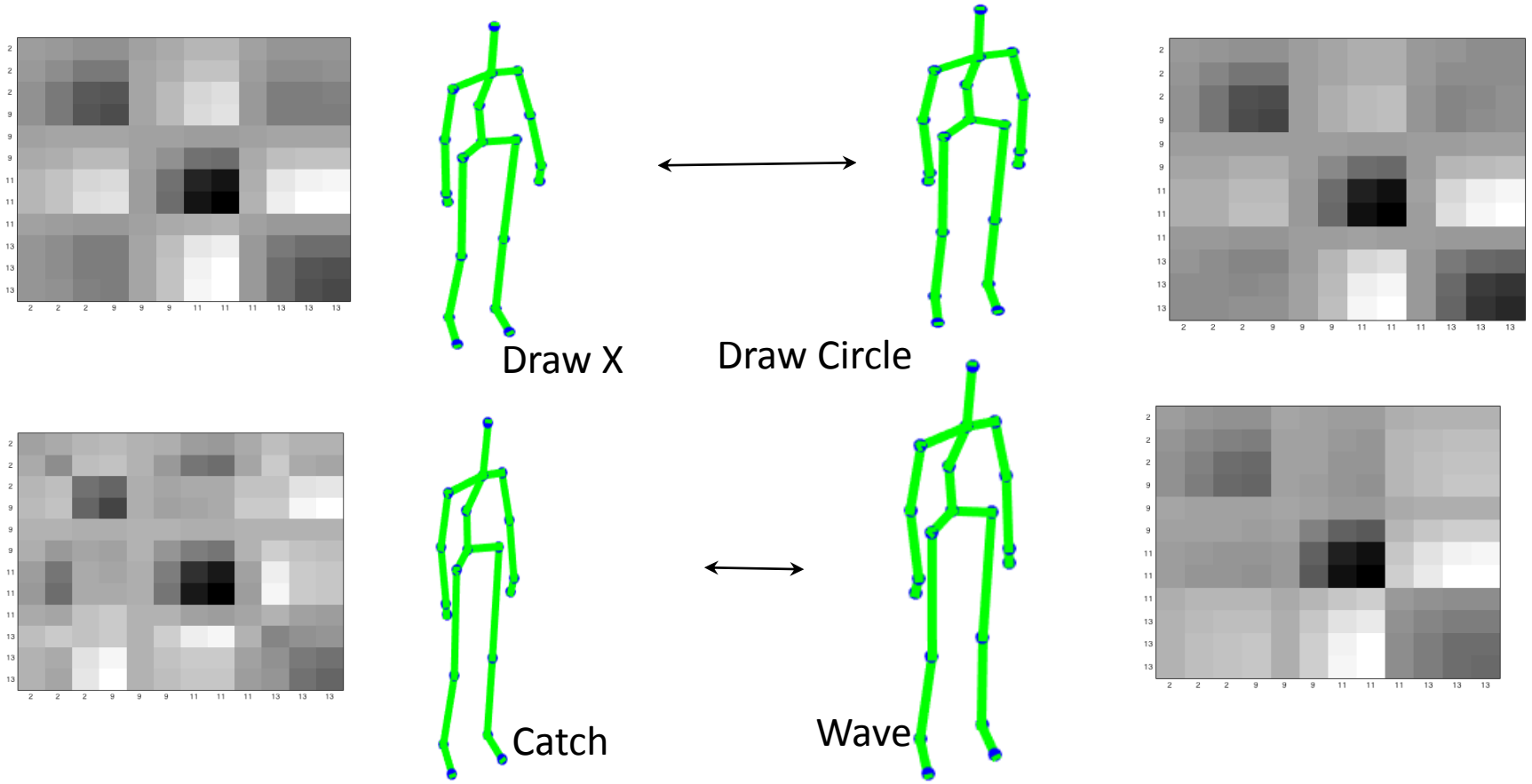
Most Informative Joint Detection



Golf Swing
(MSRAction 3D Dataset)

Hạn chế

- Hai hoạt động khác nhau nhưng lại được biểu diễn giống nhau (qua local joint)



Nhận dạng hoạt động bất thường

	Normal event	Abnormal
Airport Hall		
Bank		
ATM		

	Normal event	Abnormal
Strict area		
Hospital		
Entrance Gate		

Một số CSDL

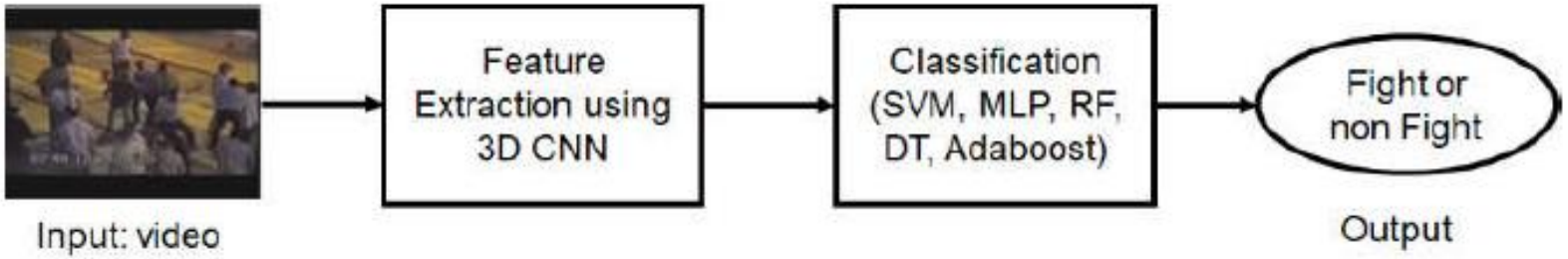
- Action Datasets: CAVIAR, BEHAVE, UT-Interaction, UCF101
- Dedicated violence datasets: Hockey, Movies, Violent Scenes, Violent Flows

Dataset	Color	Audio	Labels	BB	Sub
CAVIAR	✓	✗	✓	✓	✗
BEHAVE	✓	✗	✓	✓	✗
UT	✓	✗	✓	✓	✗
UCF101	✓	✗	✓	✗	✗
Hockey	✓	✗	✓	✗	✗
Movies	✓	✗	✓	✗	✗
VSD	✓	✓	✓	✓	✓
Violent-Flows	✓	✗	✓	✗	✗

Features of each Datasets
(Taken from [5])

Khung công việc

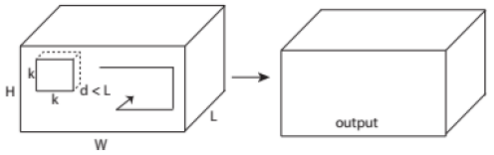
- Using short input clips from the long sequences
- Formulated as binary classification: violence vs. non-violence)



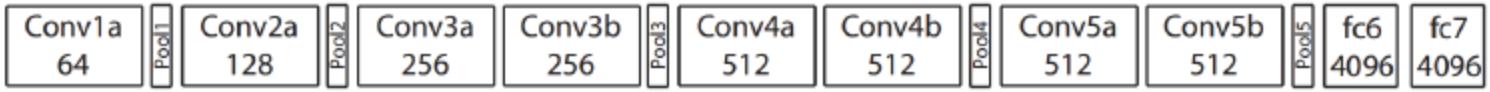
- Feature extraction:
  - Investigating **features extracted at different levels** of 3-D Convolution Neuron Network
- Classification
  - Investigating **performances of different classifiers**: SVM, MLP, RF, Adaboosts.

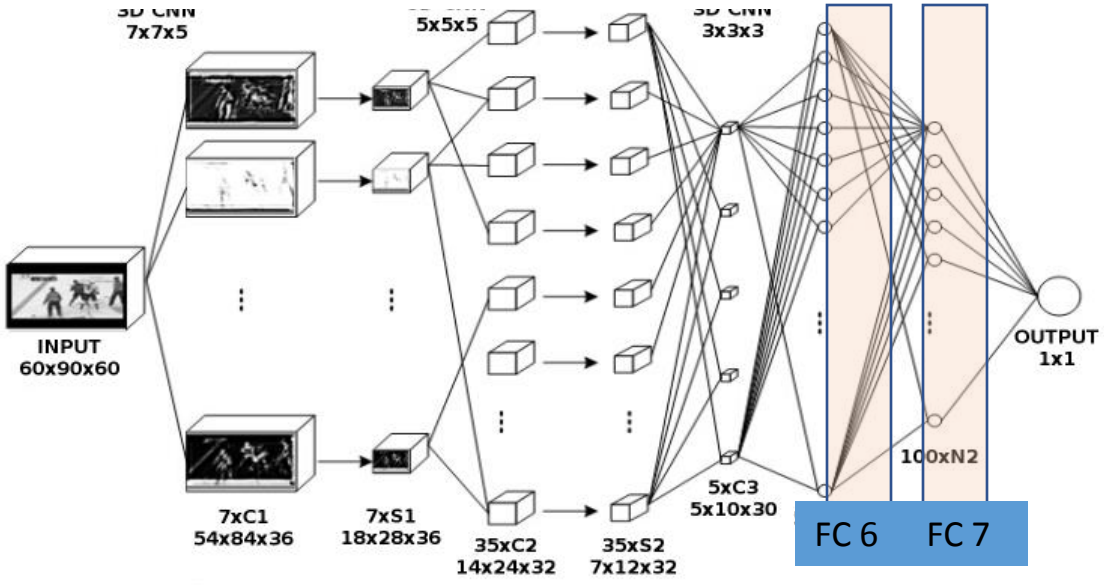
Sử dụng 3DCNN

- Proposed by [Tran15] :
  - 8 Conv. Layers by 3-D Convolution
  - 5 max pooling, 2 Fully connected layers



3-D Conv.

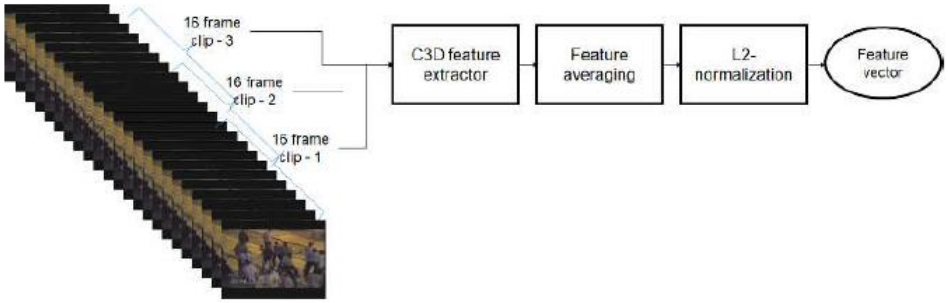




- Dimension of features extracted at FC6 and FC7: 4096
  → cover both spatial and temporal information

C3D Network training

- Utilizing a pre-trainned model of C3D on sport-1M
- For fine-tuning:
  - Five 2-sec. clips are extracted randomly
  - 16 frames clip is randomly cropped to procedure 16x128x171 volumme inputs to network for training



- Examining performances of 6 classifiers
  - K-Nearest Neighbor (K-NN), Linear and non-Linear (BRF) SVM, Decision tree, Random Forest, Neural Net Multiple Perception (MLP), Adaboost.

- Hockey Datasets

-
-
- Violence clips: fighting, pushing, beating of players
- Non-fighting clips: hockey or isolated movements

- Movies dataset

- 200 clips
- 100:100 fighting/non-fighting
- 100 Violence clips: extracted from movies or sport
- 100 non-fighting clips: extracted from public action recognition datasets

Comparative results of deep-learning based features vs. hand-designed ones (in existing works)

TABLE I. COMPARISION OF THE STUDIED METHOD WITH EXISTING ONES ON HOCKEY DATASET

Features	Classifier	Accuracy	AUC
STIP(HOG) [7]	SVM-HIK	91.7	-
STIP(HOF) [7]	SVM-HIK	88.6	-
MoSIFT [7]	SVM-HIK	90.9	-
IVF [14]	SVM	93.70	-
3D CNN [9]	Softmax	91.00	
Our studied method (Features extracted from FC6 layer of C3D)	K-NN	93.39	98.00
	SVM (Linear)	**95.50**	**99.00**
	SVM - RBF	48.30	49.00
	Decision Tree	87.70	85.00
	Random Forest	82.50	89.00
	Neural Net MLP	**95.50**	**99.00**
	Adaboost	92.50	98.00

On Hockey Dataset

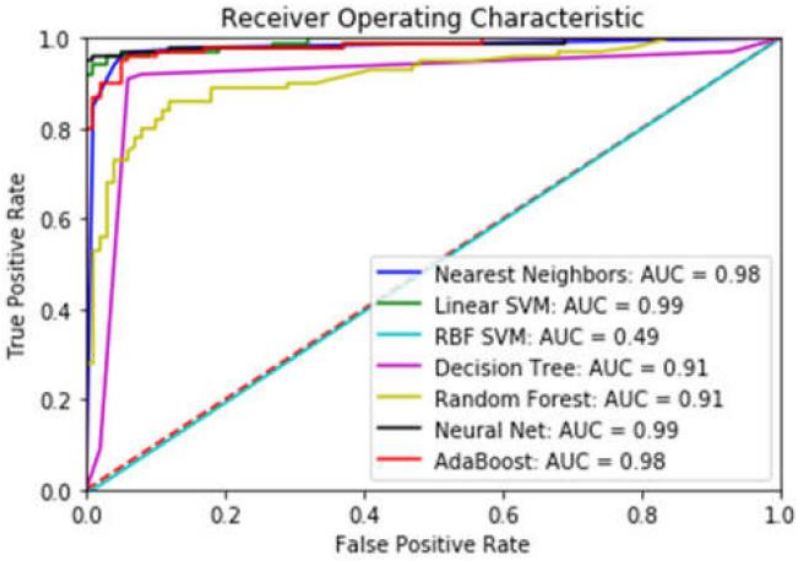
TABLE II. COMPARISION OF THE STUDIED METHOD WITH EXISTING ONES ON MOVIE DATASET

Features	Classifier	Accuracy	AUC
STIP(HOG) [7]	SVM+HIK	49.0	-
STIP(HOF) [7]	SVM+HIK	59.0	-
MoSIFT [7]	SVM+HIK	89.5	-
IVF [14]	SVM	**99.5**	-
Our studied method (Features extracted from FC7 layer of C3D)	K-NN	97.90	100.0
	SVM (Linear)	**98.94**	**100.0**
	SVM - RBF	45.35	50.00
	Decision Tree	88.01	87.00
	Random Forest	93.18	99.00
	Neural Net MLP	**98.94**	**100.0**
	Adaboost	97.39	99.00

On Movie Dataset

- The C3D-based features are efficient than hand-crafted ones: HOG, HOF or MoSIFT
- Performances are stable on both datasets
  - Robust to viewpoint and scale.
  - Depending on variation of fighting: e.g., in movies dataset is higher than Hockey datasets

Classifier performances vs. computational time



On Hockey Dataset

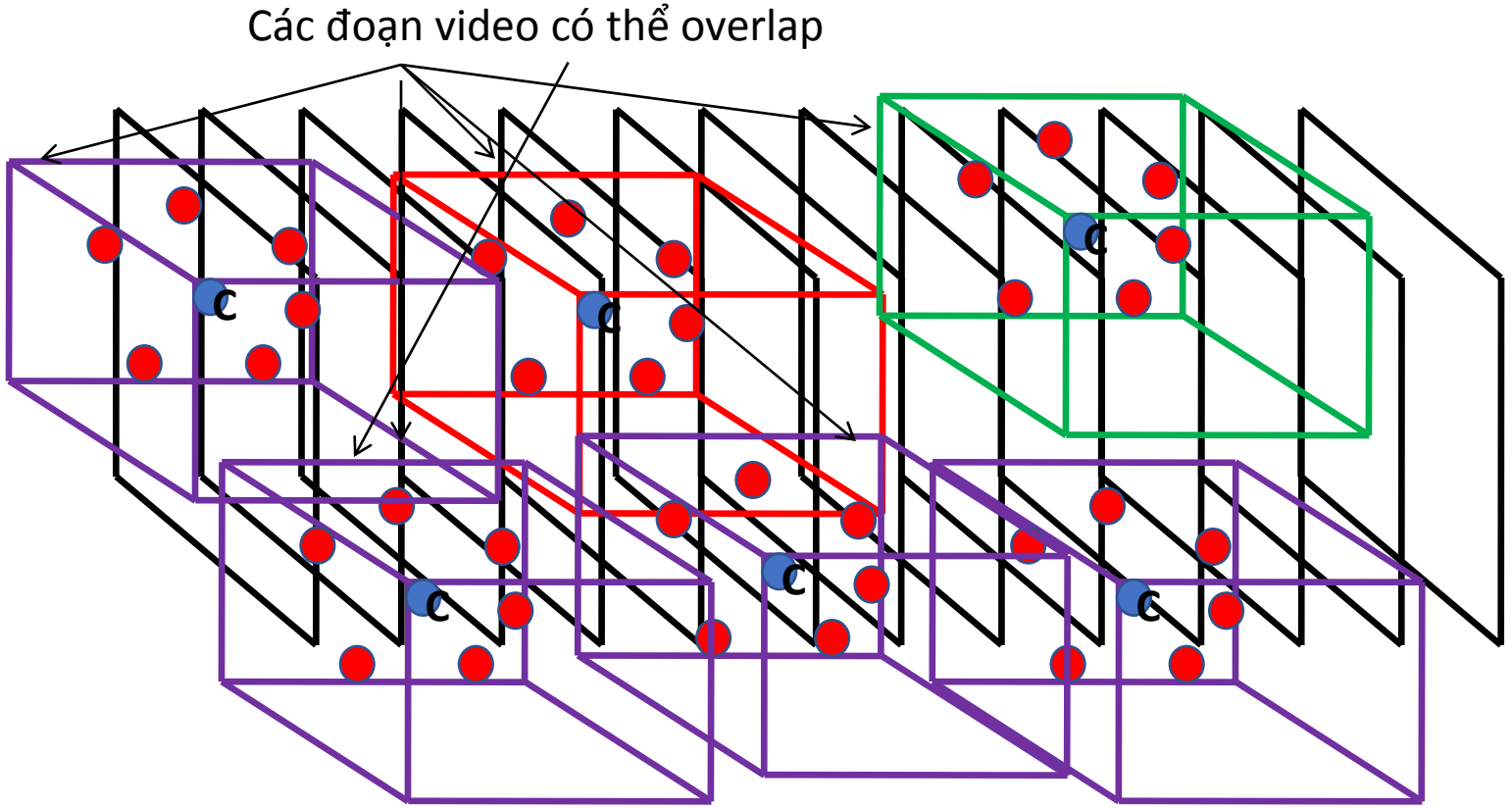
C3D –based Features extraction /videos (s)	Classification computational time (s)/video	
4.273	K-NN	4
	SVM (Linear)	0.8
	SVM-RBF	3.1
	Decision tree	0.01
	Random Forest	0.01
	Neural Net MLP	0.01
	Adaboost	0.05

- Linear SVM and Neural Net obtain highest results
- K-NN and adaboost have comparable results
- RBF SVM is the worst

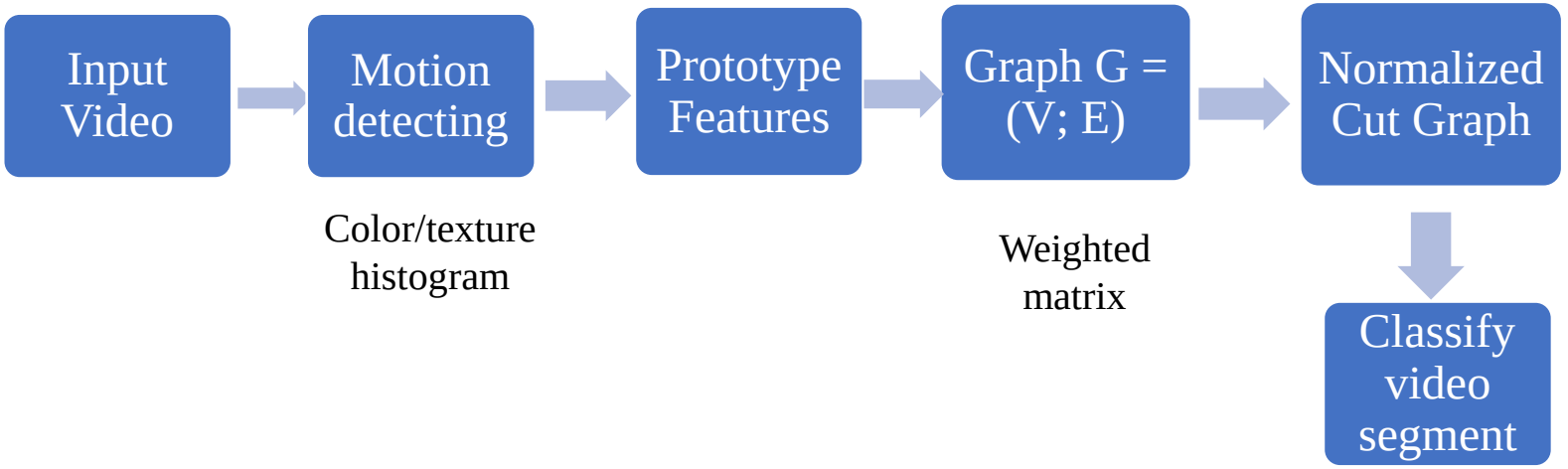
Phát hiện hoạt động bất thường từ video sử dụng học không giam sát

Hướng tiếp cận Unsupervised

- Xây dựng bộ từ điển đã định nghĩa về các hoạt động
- Tìm các cụm “hoạt động” có sự đồng xuất hiện các “từ”
- Đánh giá sự khác biệt giữa các cụm



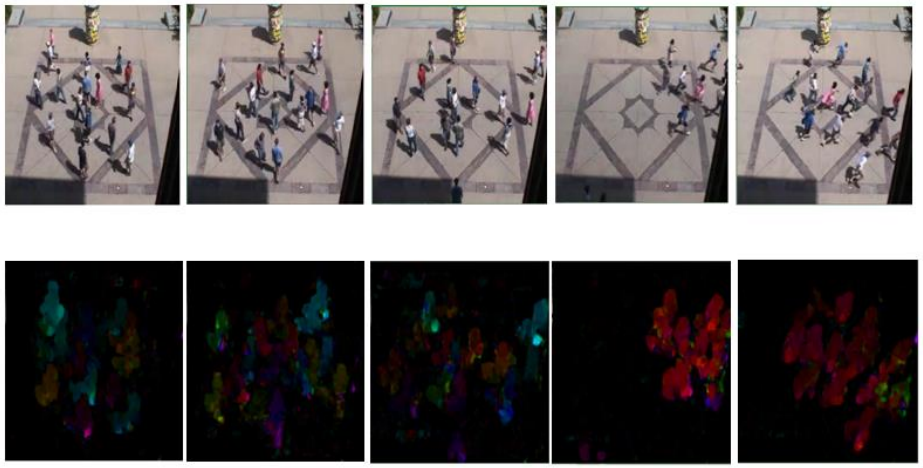
Sơ đồ tổng quát



- Nhận diện vật thể có chuyển động trong video
- Xây dựng bộ từ điển hoạt động cơ bản
- Xây dựng ma trận trọng số của biểu đồ
- Áp dụng thuật toán Normalized Cut

Trích chọn đặc trưng chuyển động

- Sử dụng dense optical flow [Gunner Farneback's algorithm]
- Sử dụng các trích chọn từ C3D

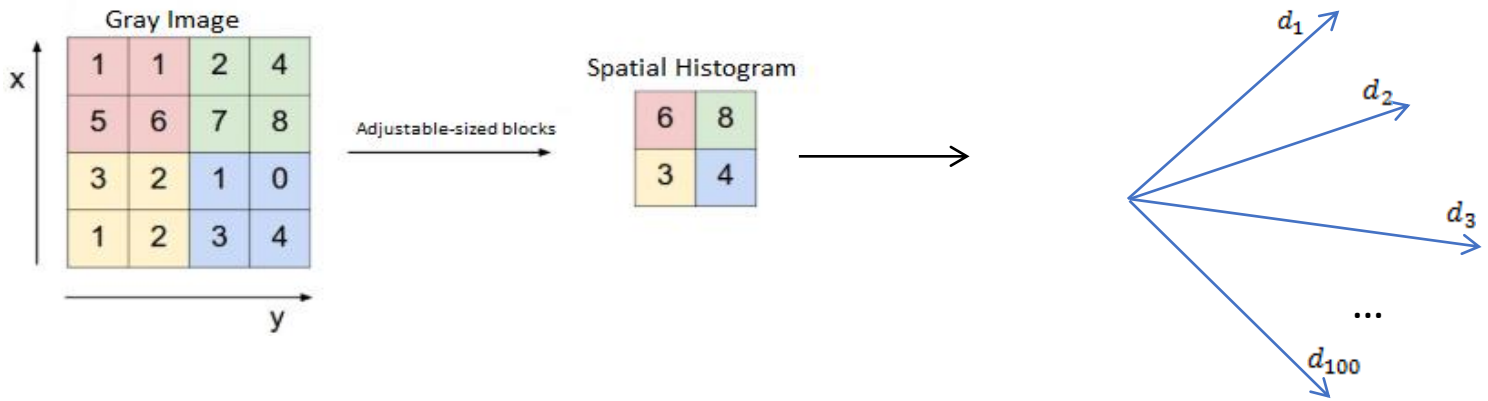


- Ngoài đặc trưng chuyển động có thể sử dụng
  - Các đặc trưng về màu sắc (color), texture
  - Sự biến đổi các đặc trưng theo thời gian

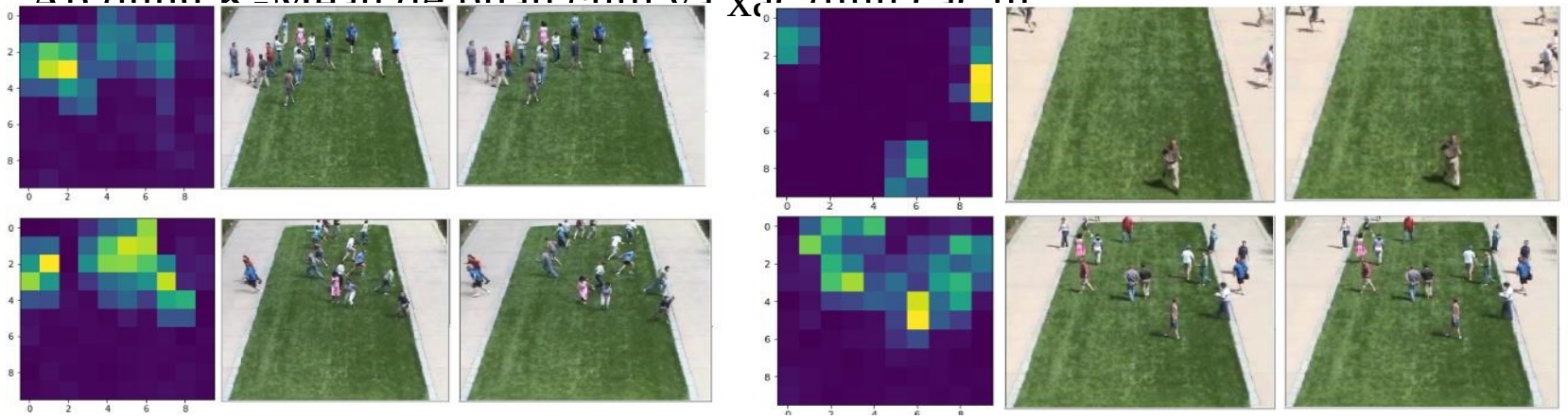
Phân tích mô hình phát hiện

❖ Xây dựng các "từ" mô tả chuyển động

- Mỗi frame ảnh được đưa về biểu đồ không thời gian có size 10x10:



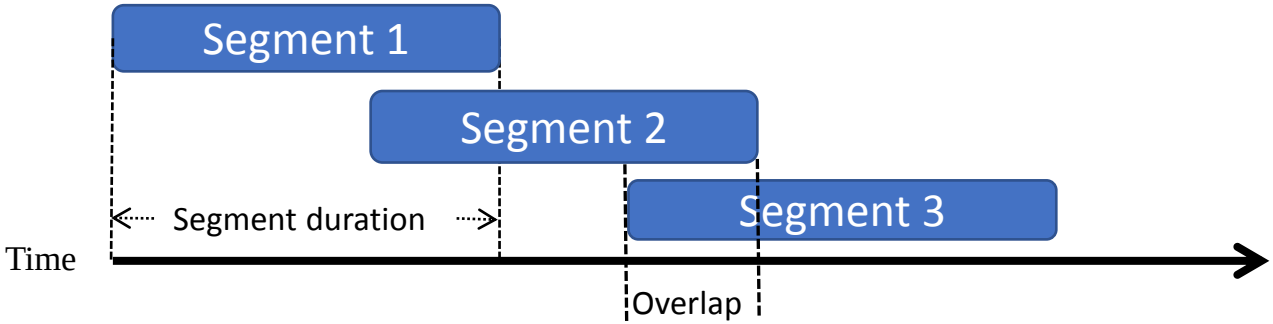
- Áp dụng K-Mean để phân cụm và xác định các từ



Phân tích mô hình phát hiện

❖ Xây dựng ma trận đồng xuất hiện :

- Phân tách video thành các phân đoạn – trượt video với tốc độ 60 frame ảnh 1 phân đoạn và độ chồng lấp giữa 2 phân đoạn là 30 frame ảnh



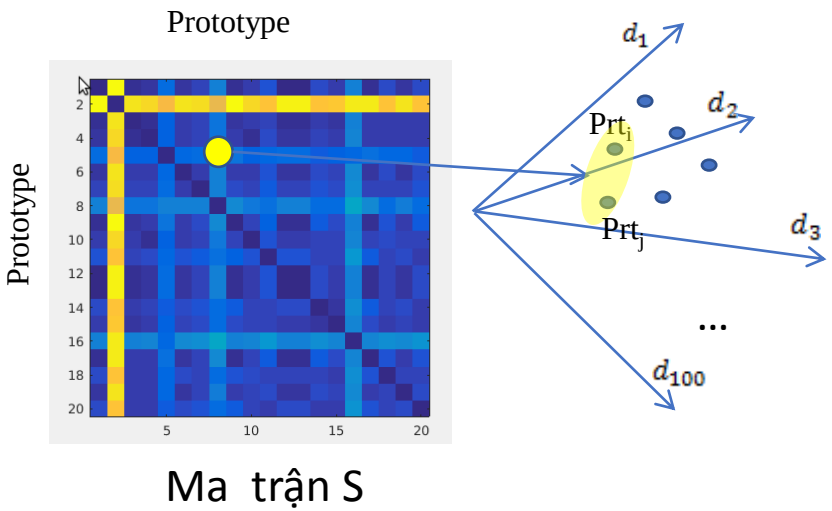
- Xây dựng ma trận đồng xuất hiện giữa phân đoạn video đặc tính cơ bản:

Bow type	Prototype 1	Prototype 2	Prototype 3	…	Prototype K
Segment 1	1	1	0	…	0
Segment 2	0	1	1	…	1
…	…	…	…	…	…
Segment N	0	1	1	0	0

Phân tích mô hình phát hiện

❖ Xây dựng ma trận trọng số:

	Seg1	Seg2	…	Prt1	Prt2	…
Seg1						
Seg2		I			C	
…						
Prt1						
Prt2		C^T			β.S	
…						



Ma trận S

- Ma trận trọng số W biểu hiện mối quan hệ giữa các nodes trong graph G

Trong đó:

I: ma trận đơn vị biểu diễn quan hệ giữa các phân đoán

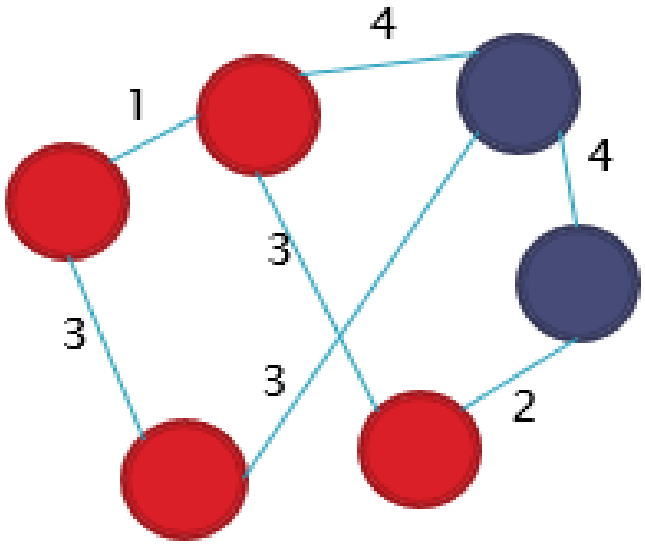
S: ma trận biểu diễn quan hệ giữa các prototype features với nhau, được đo bằng thử nghiệm chi-square

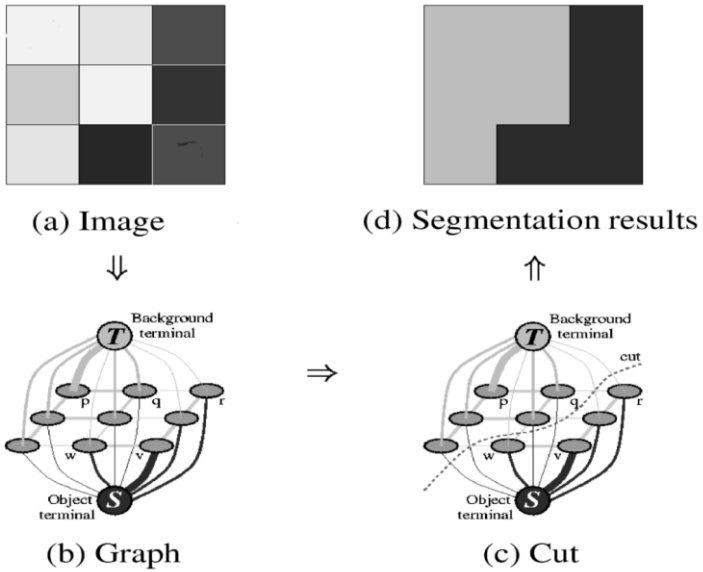
Phân tích mô hình phát hiện

❖ Thuật toán Graph-cut:

Lý thuyết xây dựng đồ thị:

"Các điểm trong không gian dữ liệu hay các pixel trong bức ảnh có thể được biểu diễn dưới dạng các nodes thuộc đồ thị $G = (V, E)$*. Giữa các nodes trong đồ thị được liên kết với nhau*

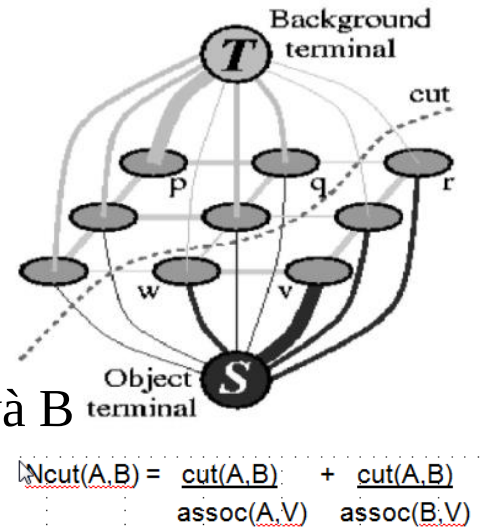




Phân tích mô hình phát hiện

❖ Thuật toán Graph-cut:

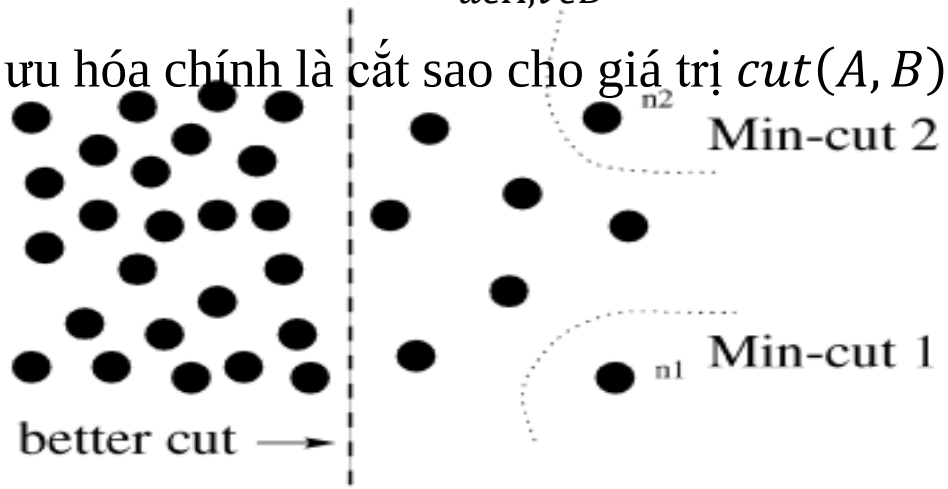
- Giả thuyết:
  - Đã xây dựng được đồ thị $G = (V, E)$
  - Muốn chia tập dữ liệu thành các tập nhỏ: A và B



- Độ không tương đồng giữa 2 tập con:

$$cut(A, B) = \sum_{u \epsilon A, v \epsilon B} w(u, v)$$

=> Bài toán tối ưu hóa chính là cắt sao cho giá trị $cut(A, B)$ là nhỏ nhất.



Kết quả phát hiện

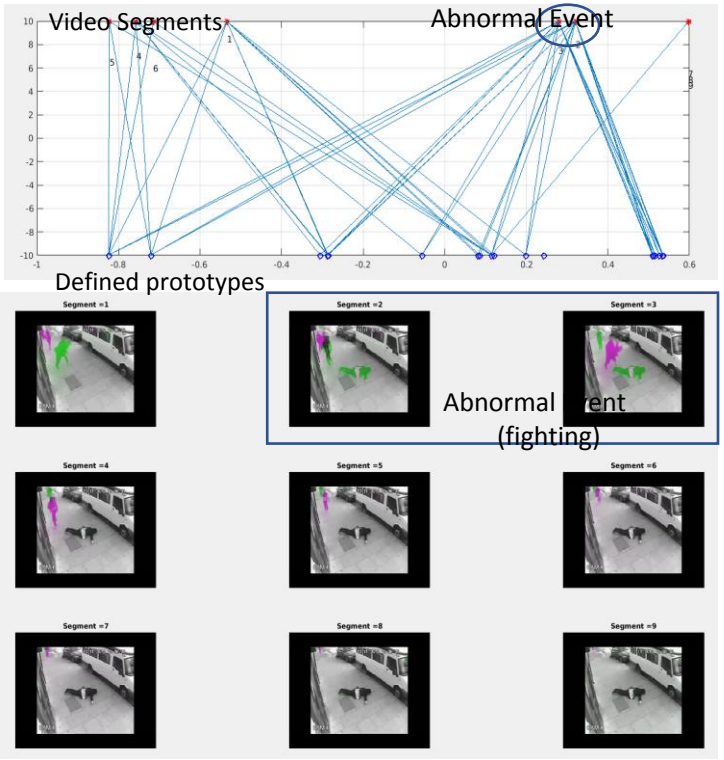
❖ Áp dụng thuật toán Normalized Cut với biểu đồ:

- $x: V \cup P \rightarrow \mathrm{R}^N$ thể hiện vị trí sắp xếp của các phân đoạn video và đặc tính cơ bản ở không gian 1 chiều

- xác định trị riêng nhỏ nhất khác 0 và vector riêng tương ứng của phương trình đặc trưng:

$$(D - W)x = \lambda Dx$$

➢ Tiếp tục sử dụng thuật toán K-Mean để xác định phân đoạn nào là khác biệt nhất



Một số kết quả

Nội dung thực hành

- Chọn video từ dataset (gồm 1 sự kiện bất thường)
- Trích chọn đặc trưng motion dense
- Xây dựng bộ từ điển và định nghĩa các “từ” (là các thành phần của chuyển động)
- Chia các đoạn video thành segment (có overlap)
- Xây dựng ma trận đồng xuất hiện (từ - segment)
- Xây dựng ma trận similar giữa các từ
- Xây dựng ma trận W (graph) thể hiện quan hệ giữa các từ và thời điểm xảy ra, với trọng số tương ứng
- Thực hiện thuật toán Normalized cut để cut các đoạn thành các cụm
- Tính khoảng cách giữa các cụm