Analyzing the similarities of actions based on video clustering

Vu Gia Truong Do Hang Nga Keiji Yanai

The University of Electro-Communications, Tokyo

Research area: computer vision

- Our lab are working on
  - Video recognition
- We are working on
  - Finding videos which have visually similar content.

Background

- Verbs
  - Express actions.
  - Represent by video shots

Running

Background

– Analyzing visual similarities between verbs has not been conducted.

Which pair is most similar ? Which pair is most dissimilar ?

– Difficulties

- How to represent verb by computable unit.
- How to collect data for several different actions.

Research purpose

- Analyzing visual similarities between actions.
  = Finding verbs pairs whose corresponding actions look similar

e.g.「play+piano」vs「type+keyboard」

Method summary

1. Extract bag-of-spatial-temporal(BOSFT)vectors
    1. Extract spatial-temporal features
    2. Encode features using Bag-of-Features approach
2. Calculate attribution vector for each verb
    1. Cluster all of BOSFT vectors
    2. Calculate distribution rate per cluster
3. Evaluate similarities
    1. Calculate distances between each verb
    2. Rank distances
    - Low distance = high similarity

1. Extract bag-of-spatial-temporal(BOSFT)vectors

1. Extract spatial-temporal features
   - Spatial-temporal feature tracks motion of special points in each frame during time.
   - For each video shot, several spatial-temporal feature vectors are extracted
2. Encode features using Bag-of-Features approach
   - Bag-of-Features counts occurrence frequency of each feature vector based on a dictionary
   - Vector represent occurrence frequency of spatial-temporal features is called Bag-of-spatial-temporal vector.

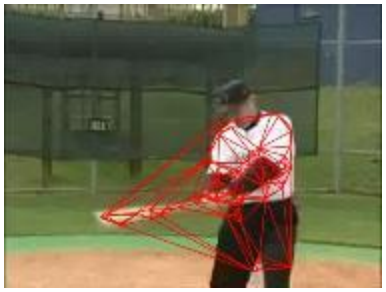
1.1 Extract spatial-temporal features

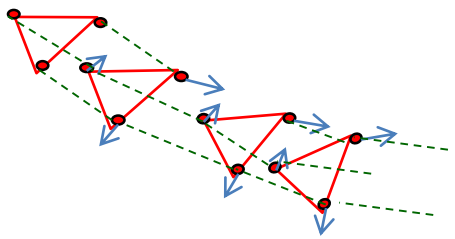
Extract visual feature

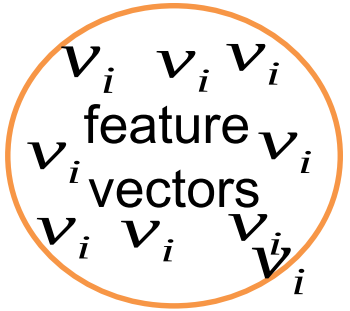
Decide tracking point

Extract motion feature

Extract vector descriptor

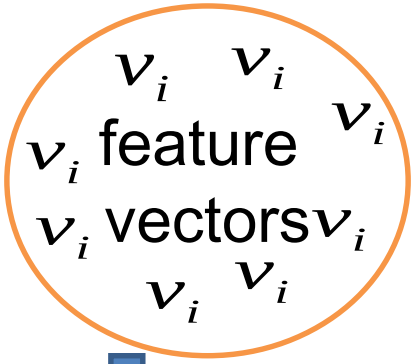




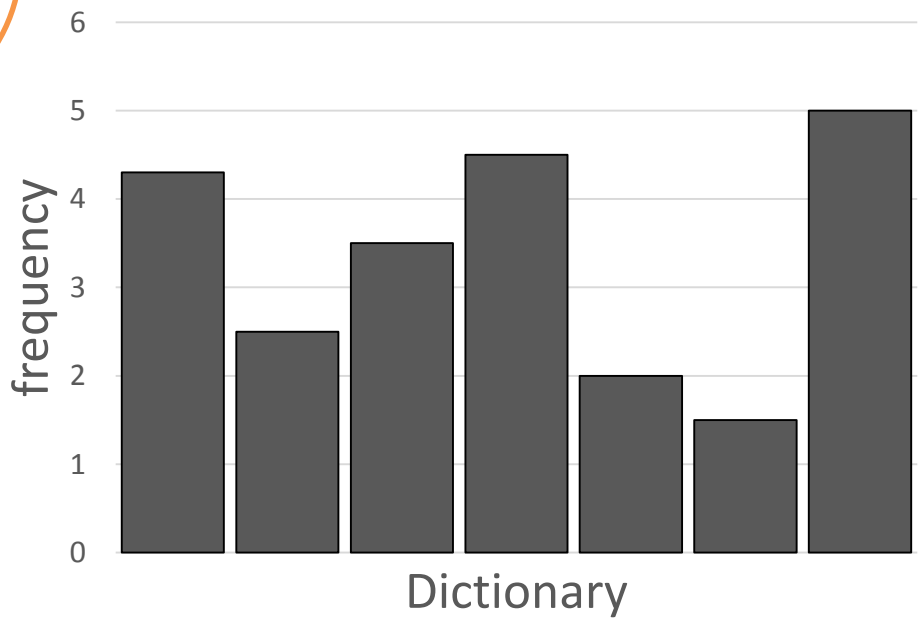


A. Noguchi and K. Yanai. A surf-based spatio-temporal feature for feature-fusion-based action recognition. In *Proc. of ECCV WS on Human Motion: Understanding, Modeling, Capture and Animation*, 2010

1.2 Encode features using Bag-of-Features approach



Dictionary

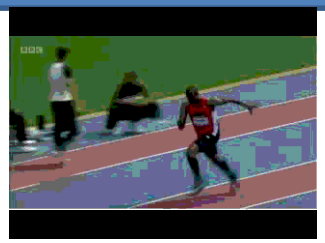


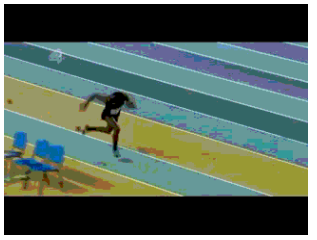
2. Calculate attribution vector

1. Cluster all of BOSFT vectors
   - Use Probabilistic latent semantic analysis (pLSA) to cluster
     - Set cluster count from 100 to 200
   - Calculate attribution vector of each video shot
2. Calculate attribution vector for each verb
   - Calculate mean value of all attribution vectors of all video shots belong to each verb

2.1 Cluster all of BOSFT vectors

- Use pLSA method to cluster all of BOSFT vectors

	Cluster i			
Video				
rate	0.006	0.0042	0.0028	0.0024

	Cluster (i +1)			
Video				
rate	0.004	0.0034	0.0017	0.0011

2.1 Cluster all of BOSFT vectors

- Attribution vector of each video shot
  - Is distribution rate per cluster
  - Has sum equal to 1

Video shot
j of verb

=

Cluster 1 : x_{j1}

Cluster 2: x_{j2}

⋮

Cluster i: x_{ji}

⋮

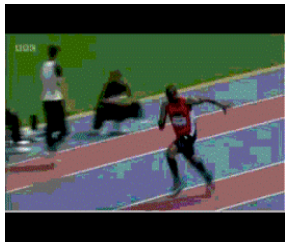
Cluster n: x_{jn}

Attribution vector

$$x_{j1}, x_{j2}, .., x_{jn}$$

2.2 Calculate distribution vector for each verb

Video shots



Attribution vector of each video shot

$$(x_{11}, x_{12}, .., x_{1n})$$

$$(x_{m1}, x_{m2}, .., x_{mn})$$

Attribution vector of verb

$$(x_1, x_2, ..., x_n)$$

$$x_i = \frac{1}{m} * \sum_{j=1}^{m} x_{ij}$$

3. Evaluate similarities

1. Calculate distances between each verb
   - Use Jesen-Shannon divergence to calculate
2. Evaluate similarities based on distances
   - Rank calculated distances
   - Find pairs of verbs which are visually similar

3.1 Calculate distances between each verb

- Jesen-Shannon divergence

$$D_{JS}(q \parallel p) = \lambda D_{KL}(q \parallel \lambda q + (1-\lambda) p) + (1-\lambda) D_{KL}(p \parallel \lambda q + (1-\lambda) p)$$

$$D_{KL}(f(x) \parallel g(x)) = \int f(x) \log \frac{f(x)}{g(x)} dx$$

3.2 Evaluate similarities

1. Find pairs of verbs which are visually similar
   - Rank calculated distances ascending
   - Find pairs of verbs which are visually similar
     - Low rank score = high similarity

Experiment

- Dataset: 13320 videos for 101 verbs

1. Evaluate visual similarities between verb pairs

2. Find verb pairs which are very similar

Experiment data

- Dataset: UCF-101
  - Contain 13320 videos for 101 verbs
  - Divide into 5 main categories
    - Human–Object Interaction
    - Body–Motion
    - Human–Human Interaction
    - Playing Musical Instruments
    - Sports

Experiment result : similar verb pairs

Verb A	Verb B	Absolute distance	Relative distance
Mixing	CuttingInKitchen	0.0157	1.00
ThrowDiscs	HammerThrow	0.0181	1.16
ShavingBeard	Haircut	0.0197	1.26
SoccerJuggling	Nunchucks	0.0205	1.31
ShavingBeard	ApplyLipstick	0.0222	1.42
Haircut	BrushingTeeth	0.0224	1.43
FrontCrawl	BreastStroke	0.0234	1.49
PlayingTabla	PlayingGuitar	0.0234	1.50
PlayingTabla	PlayingCello	0.0239	1.53
MilitaryParade	BandMarching	0.0241	1.54

Examples of visually similar verb pair

- "ThrowDiscus" and "HammerThrow"
  - Body swings around

Examples of visually similar verb pair

- “PlayingGuitar” and “PlayingTablar”
  - Hands move in small area

Examples of visually un-similar verb pair

- ”SoccerJuggling” and “NunChuck”
  - Body moves

Examples of visually un-similar verb pair

- ”Haircut” and “BrushingTeeth”
  - Hand moves

Conclusion

- Propose a novel approach to represent verbs
- Evaluate visual similarities between verbs
- Future work
  - Improve represent method to find visual similar verb pairs
  - Expand dataset
    - Add cooking related verbs
    - Add eating related verbs

Thank you for listening

Future work details

1. Improve representing method
   - Apply mid-level feature approach
     - Extract feature descriptors from volumes of video
2. Expand dataset
   - Add other verbs
     - Add cooking, eating related verbs
   - Add more sample video shots for each verb
     - Expand to 200 vide shots per verb

Contributions of our work

- Analysis similarities between verbs

- Can be applied to
  - Text-based video retrieval
  - Content-based video retrieval