Analyzing the similarities of actions based on video clustering

Vu Gia Truong Do Hang Nga Keiji Yanai

The University of Electro-Communications, Tokyo
{vugia,dohang,yanai}@mm.inf.uec.ac.jp

Abstract

In this paper, we propose a novel approach to represent a verb as a set of attribution vectors. Similarities between verbs are defined as distances between their representative attribution vectors sets. By ranking distances between these sets, several pairs of verbs which are visually similar are found from a collection of verbs.

We then use dendrogram to illustrate the similarities between verbs pairs. Each verb is drawn as a leaf node in a dendrogram. By counting number of nodes between leaf nodes and height of connected branches, similarities can be easily estimated.

1. Introduction

1.1. Background

Analyzing visual similarity between nouns and adjectives has been conducted by Kohara *et al.* [5]. But analyzing visual similarities between verbs has not been conducted so far. There are two difficulties that researchers must face to conduct such analysis. The first is "how to represent verb" and the second is "how to collect enough data for several different verbs".

To overcome the first difficulty, we propose a novel approach to represent a verb as a set of attribution vectors. These attribution vectors are obtained from clustered representative Bag-of-spatio-temporal feature vectors of verb. We evaluate the effectiveness of our method by applying it to UCF-101 data set [7]. This dataset contains 101 different verbs and 13320 video shots which are mined from internet.

To overcome the second difficulty, we used UCF-101 data set [7] contains 101 different actions. As an alternative way, we can use an automatic dataset construct system proposed by Nga *et al.* [1]. However, the results of the system still contains irrelevant video shots which should be excluded in order to use them as video shots corresponding to a specific "verb", because they are not perfect.

1.2. Research purpose

The main purpose of this research is to analyze visual similarities between verbs. It is equal to finding several pairs of verbs which are visually similar from a collection of verbs. A verb represents an action which can be defined as a group of continuous smaller sub-actions. So similarity between a pair of verbs can be evaluated as the ratio of sub-action which both of them share to the sum of sub-actions of verbs pairs.

For example, a pair of visually similar verbs are "type+keyboard" and "play+piano". The main sub-action of the verb "type+keyboard" is "hand move up and down continuously", and the main sub-action of the verb "play+piano" is also "hand move up and down continuously". The ratio of the same sub-action is high so they can be considered as similar.

In contrary, a pair of visually dissimilar verbs is "push+up" and "play+daf". The main sub-action of the verb "push+up" is "body move up and down continuously"and the main sub-action of the verb "play+daf" is "hand move around continuously". The ratio of the same sub-action is low so they can be considered as dissimilar.

Because each verb is represented by a set of attribution vectors, the ratio of overlapped sub-actions between a pair of verbs corresponds to simlarity between these representative sets.

2. Related work

Video recognition is one of common tasks in computer vision. The goal of this task is to automatically detecting actions in a given video sequence. The main approach is representing video as fixed-size vectors and using classifier to decide which class this video belongs to. According to the approach, at first feature descriptors are extracted from videos. Several types of features have been proposed. According to Jiang *et al.* [3], they can be classified into three categories as follows: 1) frame-based appearance features, 2) spatio-temporal visual features, and 3) acoustic features. Frame-based appearance features are extracted from frames

of videos separately. This approach ignores the change of actions through time. In contrary to frame-based appearance features, spatio-temporal visual features consider the change of actions through time. This capability make spatio-temporal visual features became most popular approach. Some example of spatio-temporal visual features are: HOG-3D [4], SURF-based spatio-temporal feature [6]. The third is acoustic features which are extracted from audio of video shots.

The number of local feature descriptors extracted from each video is different. Feature descriptors of each video need to be encoded into a fixed-size vector for later classification. The most popular encoding method is Bag-of-Features (BoF). BoF model learns a codebook offline by clustering a large set of descriptors with k-means and then assigns each descriptor of the video the closest entry in the codebook.

After representing video as a fixed-size vector, a pretrained classifier is used to decide which action a video belongs to. Because of simplicity and high accuracy SVM has been used broadly. From these decisions, we are unable to understand the relationships between actions themselves which are critical in video retrieval techniques.

Instead of classifying actions, our approach focuses on calculation of similarities between actions which allows us to further study relationships between verbs in which actions are described. Because of the complexity of actions, each verb which describes action needs a lot of video shots to be represented sufficiently. Also a large number of verbs is required to make this study deeper.

3. Method

3.1. Overview

Our method can be summarized as three steps as follows:

1. Preparing dataset Each verb in the dataset is represented by hundreds of video shots which correspond to the action.

2. Extracting bag-of-spatio-temporal vectors

   Each video shot is represented by extracted Bag-of-Spatio-Temporal Feature (BOSTF) vector.

3. Extracting attribution vectors

   By clustering all of bag-of-spatio-temporal vectors, attribution vectors are obtained. Each verb is now considered as a set of attribution vectors.

4. Similarities evaluation

   Similarities between verbs are obtained by ranking distances for each pair of attribution vectors set.

3.2. Preparing dataset

As a dataset containing various kinds of actions, we use the UCF-101 dataset which were created by hand. The actions included in this dataset can be divided into the following five categories: 1) Human-Object Interaction, 2) Body-Motion, 3) Human-Human Interaction, 4) Playing Musical Instrument, 5) Sports. Figure 1 shows the thumbnails of all action categories in the dataset.

By using the automatic Web video shot collecting system proposed by Nga *et al.*[1], we can obtain a large-scale action shots dataset automatically. However, the dataset automatically constructed may contain irrelevant video shots which must be removed manually. The noise removal process is extremely time-consuming since the scale of the data is too large.

Figure 1. Thumbnails of all action categories in the dataset.

3.3. Extracting Bag-of-spatio-temporal vectors

A video shot can be seen as a set of frames which are continuously connected along the time axis. The change of feature point position through time contains rich information related to actions the video shot represents. These changes can be converted into spatio-temporal features. We use Noguchi *et al.*'s approach [6] to extract spatio-temporal features from video shot.

Figure 2 illustrate the detailed flow of their method. This approach can be summarized as 4 steps below.

(a) Extract visual feature

(b) Decide spatio-temporal feature points

(c) Extract motion features

(d) Convert into vector descriptors

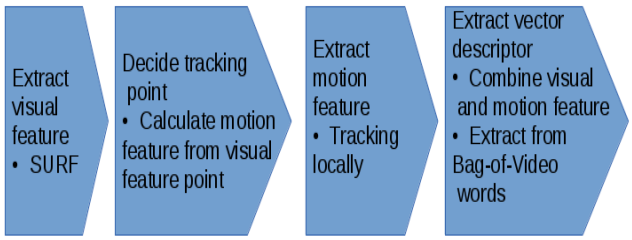

Figure 2. The processing flow of extracting spatio-temporal features

For each video shot, hundreds of spatio-temporal feature descriptors are obtained. Then using Bag-of-Features approach where the dictionary size is 5000, one video shot is converted into a 5000 length size bag-of-spatio-temporal feature (BOSTF) vector. Even though dictionary size is quite large, Bag-of-Features approach is incapable of fully describing complexity of action. Because of this reason, purely increasing size of dictionary does not come up with a better similarity evaluation result.

Actions themselves contain a lot of sub-actions. To archive better similarity evaluation result, each video shot should be split into smaller sub-actions which are simpler. We validate effectiveness of this approach by simply slitting each video shot into 2 equal parts. Then we compare the performance between using split video shots and entire video shots.

3.4. Extracting attribution vectors

As defined before, verb which describes action is composed from several continuous connected sub-actions. Similarity between a pair of verbs can be evaluated as the ratio of sub-actions which both of them share. For each pair of verbs, their similarity can be simply calculated by calculating ratio between number of BOSTF vectors which both of them share and sum of BOSTF vectors in this pair. But this approach needs a dataset which has bigger video shots count to archive acceptable similarity evaluation result.

To get a better similarity evaluation result, relationships between each video shot and verbs need to be obtained. These relationships are represented by attribution vectors which are obtained by clustering all of BOSTF vectors. In this paper, three clustering methods, K-means, pLSA, and Affinity Propagation, are used. Experiment results show that between these three clustering methods, pLSA is the most effective method.

An example of the calucation of attribution vector is as follows: video shot of verb 1 are A, B, and C, the assignment rates for verb 1, 2, 3, and 4 are shown in Table 1, then attribution vectors for video shot A, B, and C are obtained as shown in Table 2

Table 1. An example of clustering result

verb	1	2	3	4
video shot A	40%	30%	20%	10%
video shot B	30%	20%	10%	40%
video shot C	10%	40%	30%	20%

Table 2. An example of attribution vectors

video shot A= (0.4, 0.3, 0.2, 0.1)
video shot B= (0.3, 0.2, 0.1, 0.4)
video shot C= (0.1, 0.4, 0.3, 0.2)

3.5. Similarity evaluation

After extracting attribution vectors, each verb is now represented as a set which contains hundreds of attribution vectors. To calculate distance between 2 sets, instead of Euclid distance, Jensen-Shannon divergence is used.

Jensen-Shannon divergence formula for calculating distance between 2 distributions p and q is shown in Equation 1.

$$D_{JS}(q||p) = D_{KL}(q||(q+p)/2) + D_{KL}(p||(q+p)/2) \quad (1)$$

D_{KL} is Kullback-Leibler divergence which describes the scale of difference between distribution $f(x)$ and $g(x)$. Kullback-Leibler divergence's formula is shown in Equation 2.

$$D_{KL}(f(x)||g(x)) = \int f(x) \log \frac{f(x)}{g(x)} dx \quad (2)$$

The smaller distance between a pair of sets, the more similar their representative verbs are. We first rank calculated distances to get top 20 most similar pairs of verbs. Then we use dendrogram to illustrate similarities clearly.

Dendrogram is one type of tree diagrams often used to illustrate the arrangement of the clusters. Each cluster is drawn as leaf node of dendrogram. Leaf nodes are connected to the others by several branches which are separated by many nodes. Similarity between 2 clusters is defined as "how low is the number of nodes between them" and "how low is distance between 2 representative vectors sets which is equal to height of connected branches".

By considering each verb as a cluster, similarities between verbs can be illustrated vividly. A verb and its most similar one are drawn as a pair of leaf nodes and they have these following features: 1) There is only 1 node between them, 2) Height of connected branches is low.

4. Experiments and Results

Effectiveness of our method is evaluated through two experiments. Note that we show only results obtained by pLSA, because the results by pLSA was the best among

three clustering methods, K-means, pLSA, and Affinity Propagation.

In the first experiment, BOSTF vectors are extracted from entire video shots. In the second experiment, we split each video shot into half along the time sequence and extract BOSTF vectors from split video shots. A pair of verbs is considered as similar if their distance is smaller than 0.04 and there is only 1 node between their leaf nodes on the dendrogram.

4.1. Entire video shots

In this experiment, the top 3 most similar verb pairs are: "Mixing" and "CuttingInKitchen","ShavingBeard" and "Haircut", "HorseRiding" and "Rafting'.

"ShavingBeard" and "Haircut" 's screen shots are shown in Figure 3. Smallest distance between verb pairs is 0.0200.

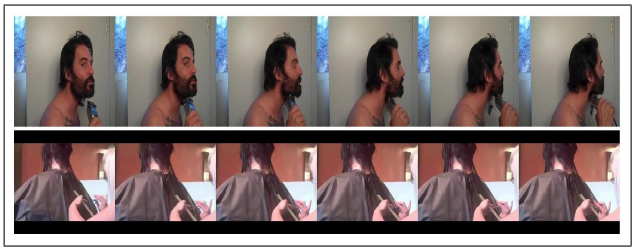

Figure 3. "ShavingBeard"and"Haircut"'s screen shot

Table 3. Top 20 most similar verb pairs

Verb A	Verb B	Verb's distance	Relative distance
Mixing	CuttingInKitchen	0.0200	1.00
Rafting	HorseRiding	0.0244	1.22
ShavingBeard	Haircut	0.0248	1.24
PizzaTossing	HeadMassage	0.0271	1.36
FrontCrawl	BreastStroke	0.0279	1.40
SoccerJuggling	Nunchucks	0.0285	1.42
ThrowDiscs	HammerThrow	0.0305	1.52
HeadMassage	BlowingCandles	0.0306	1.53
PlayingDaf	Haircut	0.0308	1.54
HorseRiding	Biking	0.0319	1.59
PlayingTabla	PlayingGuitar	0.0319	1.60
Haircut	BlowDryHair	0.0321	1.61
ShavingBeard	ApplyLipstick	0.0326	1.63
FloorGymnastics	BalanceBeam	0.0326	1.63
PlayingTabla	PlayingCello	0.0332	1.66
RopeClimbing	BabyCrawling	0.0333	1.66
Kayaking	HorseRiding	0.0333	1.67
Mixing	BoxingSpeedBag	0.0335	1.67
Haircut	BrushingTeeth	0.0336	1.68
PommelHorse	HammerThrow	0.0339	1.69

Distance between the most similar pair of verbs: "Mixing" and "CuttingInKitchen" is 0.02. As shown in Figure 4 they are connected by only one node.

4.2. Split video shots

In this experiment, the top 3 most similar verb pairs are: "Mixing" and "CuttingInKitchen","ShavingBeard" and "Haircut", "ThrowDiscs" and "HammerThrow".

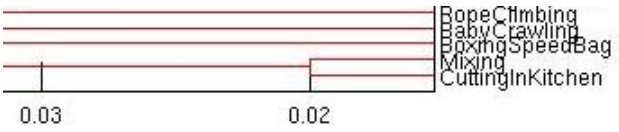

Figure 4. Part of Dendrogram

Table 4 shows top 20 most similar verb pair. Comparing to the first experiment's results, distances between verbs become smaller. Pairs of verbs which both do not change direction through time get higher rank scores. For example "ThrowDiscs" and "HammerThrow" is a pair of verbs which do not change direction frequently through time. They go from 6th rank to second rank . In contrary, "HorseRiding" and "Rafting" is a pair of verbs which do change direction frequently, They go from second rank to 14th rank.

Similar pair of verbs: "ThrowDiscs" and "HammerThrow" 's creenshot is shown as Figure 5.

Low rank score pair of verbs: "HorseRiding" and "Rafting" 's creenshot is shown as Figure 6.

Figure 5. "ThrowDiscs" and "HammerThrow"'s creenshot

Figure 6. "HorseRiding" and "Rafting"'s creenshot

Distance between pair of the verbs, "ShavingBeard" and "Haircut", is 0.00197, which is smaller than the minimum distance obtained in the first experiment. As shown in Figure 7, they are connected by only one node.

Figure 7. Part of Dendrogram

Table 4. Top 20 most similar verb pairs

Verb A	Verb B	Verb pair's 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	0.0233	1.49
PlayingTabla	PlayingGuitar	0.0234	1.50
PlayingTabla	PlayingCello	0.0238	1.53
MilitaryParade	BandMarching	0.0241	1.54
MoppingFloor	BabyCrawling	0.0245	1.56
PizzaTossing	HeadMassage	0.0245	1.56
HorseRiding	Biking	0.0246	1.57
Rafting	HorseRiding	0.0250	1.60
PlayingGuitar	ApplyLipstick	0.0253	1.61
Haircut	BlowDryHair	0.0259	1.65
RopeClimbing	BabyCrawling	0.0262	1.67
HeadMassage	BlowingCandles	0.0267	1.70
ShavingBeard	PlayingGuitar	0.0270	1.72
PlayingDaf	BlowDryHair	0.0271	1.73

5. Conclusions

In this paper, we propose a novel approach to represent verbs as sets of attribution vectors. Calculating distances between these sets allow us to evaluate similarities between verbs. Even though this approach is simple, when it is applied to 101 different verbs dataset, we obtained an acceptable similarity evaluation result. Moreover, we discovered dozens of visually similar verbs pairs.

To improve similarity evaluation result, keeping complexities of verbs is an important task. As defined before, action which is described by verb is structured from several smaller sub-actions. Representative method should separate sub-actions and use them for representing. Sub-actions of actions which do not change direction through time can be ontained by simply split video shots into half. This led to a small improvement in similarity evaluation result. But because different sub-actions have different life span, this approach is unable to extract sub-actions for all of verbs.

A promising approach for separating actions into sub-actions is the method proposed by Jain *et al.* [2]. Their method extracts representative parts from video shot of actions and uses these parts to construct BoF dictionary. By using constructed BoF dictionary, each video shot is represented by a collections of representative sub-actions. We plan to implement this method in our next work. We also plan to expand the dataset to 200 types of different verbs.

References

[1] N. H. Do and K. Yanai. Automatic construction of an action video shot database using web videos. In *Proc. of International Conference on Computer Vision (ICCV)*, 2011.

[2] A. Jain, A. Gupta, M. Rodriguez, and L. S. Davis. Representing videos using mid-level discriminative patches. In *Proc. of IEEE Computer Vision and Pattern Recognition*, pages 2571–2578, 2013.

[3] Y. G. Jiang, S. Bhattacharya, S. F. Chang, and M. Shah. High-level event recognition in unconstrained videos. *International Journal of Multimedia Information Retrieval*, 2(2):73–101, 2013.

[4] A. Kläser, A. Marszałek, and C. Schmid. A spatio-temporal descriptor based on 3d-gradients. In *Proc. of British Machine Vision Conference*, pages 995–1004, 2008.

[5] Y. Kohara and K. Yanai. Visual analysis on relations between nouns and adjectives using a large number of web images. In *Proc. of International Workshop on Modern Science and Technology (IWMST)*, 2012.

[6] 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.

[7] K. Soomro, A. R. Zamir, and M. Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.