Automatic Action Video Dataset Construction from Web using Density-based Cluster Anaysis and Outlier Detection

Nga Hang Do[1,a)] Keiji Yanai[1,b)]

1. Introduction

High quality datasets play important roles in computer vision and pattern recognition tasks. Constructing high quality datasets using noisy data such as Web data without extensive human effort of manual annotation has received increasing attention of researchers in this field recently [2], [6], [10]. In this paper, we introduce a fully automatic approach to construct a large-scale action dataset from noisy Web video search results. Previous work which also aimed to obtain data for specific action concept from noisy data with minimal manual annotation effort [4], [8] generally require additional information provided together with videos such as movie script [8] or metadata (tags) [4]. In this work, we propose an approach which exploits only visual features of videos retrieved from Web and does not require any additional material.

Our idea is based on combining cluster structure analysis and density-based outlier detection. For a specific action concept, first, we download its Web top search videos and segment them into video shots. We then organize these shots into subsets using density-based hierarchy clustering. Clusters are sets of density-connected shots. For each set, we rank its shots by their outlier degrees which are determined as their isolatedness with respect to their surroundings. Finally, we collect top ranked shots as training data for the action concept. Our work is inspired by [2] which uses density analysis of Web images for automatic image dataset construction.

2. Approach

In this work, we present an approach which autonomously extracts from noisy Web videos relevant video shots for given action concepts. Our approach consists of three steps: shot collection, shot clustering and shot ranking. See Fig. 1 for the illustration of our proposed framework. In the followings, we explain in detail each step.

2.1 Shot Collection

We first prepare keywords for given action concepts. The concepts can be defined in any form: either "verb" (such as "dive") or "verb+non-verb" (such as "throw+hammer", "cut+in+kitchen") or "non-verb" (such as "pole vault"). In case verb included in the keyword, we search for its videos in both forms: "verb" and "verb-ing" (such as "diving", "throwing+hammer"). We filter out videos belonging to "entertainment", "music", "movies", "film" and "games" categories during searching since these categories generally contain extremely long videos. Top search results are downloaded and segmented into video shots using color histogram. Each shot represents one single scene. For each concept, we download around 100-200 videos and obtain 600-2000 video shots on average.

Fig. 1 Framework of our method which consists of three steps: shot collection, shot clustering and shot ranking.

2.2 Shot Clustering

With shots obtaind after above step, we group related shots into clusters before shot ranking and selection. This step helps deal with concept diversity. With web data retrieved for a given concept, there will also be common characteristics shared among subsets of data. Therefore, we use hierarchy clustering which allows different clusters to share the same instances. We adopt OPTICS ("Ordering Points to Identitfy the Clustering Structure") [1] to find clusters. The hierarchical structure of the clusters can be obtained based on the density of the data distributed around their points. We introduce here some important definitions to briefly explain the clustering method.

Let p be an object from a dataset D, k be a positive integer and d be a distance metric, then:

Definition 1: $k-\mathrm{dist}(p)$, the k-distance of p, is defined as the distance $d(p,o)$ between p and object $o \in D$ satisfying: 1. at least k objects $q \in D$ having $d(p,q) \leq d(p,o)$, and 2. at most (k -1) objects $q \in D$ having $d(p,p) < d(p,o)$

Definition 2: $N_{k-\mathrm{dist}(p)}(p) = q|q \in D, d(p,q) \leq k-\mathrm{dist}(p)$ denotes the k-distance neighborhood of p.

Definition 3: $reach - \mathrm{dist}_k(p,o) =$

[1] The University of Electro-Communications, Tokyo, Chofu, Chofugaoka 1-5-1
[a)] dohang@mm.cs.uec.ac.jp
[b)] yanai@cs.uec.ac.jp

$maxk - \mathrm{dist}(o), d(p, o)$ represents reachability-distance of an object p with respect to object o.

The OPTICS-algorithm computes a "walk" through the data, and calculates for each object the smallest reachability-distance with respect to an object considered before it in the walk. A low reachability-distance indicates an object with a cluster, and a high reachability-distance indicates a noise object or a jump from one cluster to another cluster.

2.3 Shot Ranking

For each obtained cluster, we assign outlier factor for each shot based on outlying property relative to its surrounding space. Differently from shot clustering step, in this step surrounding space of a shot is limited within in its own cluster. In each cluster, shots are ranked according to LOF (Local Outlier Factor) as described in [3]. LOF of a point p is formally defined as follows.

$$LOF_{MinPts}(p) = \frac{\sum_{o \in N_{MinPts-dist(p)}(p)} \frac{MinPts-dist(p)}{MinPts-dist(o)}}{|N_{MinPts-dist(p)}(p)|} \quad (1)$$

LOF of an object is calculated as the average ratio of its $MinPts - dist$ and that of its neighbors within $MinPts - dist$. A large $MinPts$-dist corresponds to a sparse region since the distance to the nearest $MinPts$ neighbors is large. In the contrast, a small $MinPts$-dist means that the density is high. In each cluster, shots are ranked according to LOF. Shots with low LOF degrees are considered as relevant shots and brought to the top of the cluster.

3. Experiments and Results

Here we report our experiment results on 11 actions defined in UCF YouTube Action dataset [9]. Note that we do not use videos of that dataset. Our videos are collected as described in Section 2.1. As distance metric for shot clustering and shot ranking, we use Euclidean distance. As visual features, we extract Spatio-Temporal Features as proposed in [5]. Our baseline is our most related work [4]. According to this method, first videos are ranked based on usage frequencies of tags. Shots are collected from videos which have tags with high co-occurance frequencies. Next shots are ranked using VisualRank [7] which is a ranking method with a visual-feature-based similarity matrix. Shots sharing the most visual characteristics with others are ranked to the top and selected as relavant shots. Since it became hard to obtain tag information, we could not perform tag co-occurance based video ranking step as proposed in [4]. Here we use our method of shot collection and apply their idea of using VisualRank to shot ranking to compare with our proposed method of shot selection which composed of diversity based shot clustering and LOF based shot ranking. We show that our method can obtain higher precision rate for most of experienced actions and our results look more diverse than those by the baseline. Precision rate is calculated as percentage of relevant shots among top 100 shots following our baseline [4]. Precision for all actions are shown in Table 1. Some example results are shown in Figure 2.

Table 1 Results on 11 action keywords. Baseline here means method with our shot collection and VisualRank based shot ranking.

Action	Proposed	Baseline
basketball	59	67
biking	30	35
diving	25	19
golf_swing	59	52
horse_riding	49	48
soccer_juggling	76	72
swing	36	22
tennis_swing	38	37
trampoline_jumping	42	44
volleyball_spiking	36	45
walking	25	11
Average	43.2	41.1

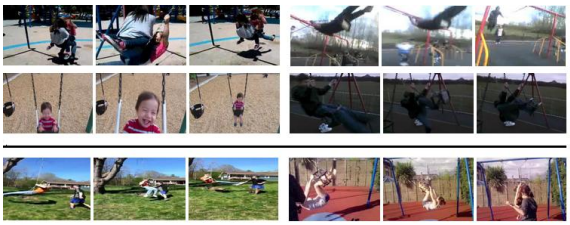

Fig. 2 Relevant shots among top 15 results of "swing" obtained by our method (top two rows) and the baseline (bottom row). As shown in this figure, our method could selection action shots with more various look (taken from various viewpoints).

4. Conclusions

In this paper, we proposed a fully automatic approach for action dataset construction with noisy Web videos. Our approach aims to solve the problem of limitation in quantity of training data for the task of action recognition. It demonstrated that concept detection in web video is feasible and offers the advantage of a fully automatic, scalable learning of human actions.

References

[1] Ankerst, M., Breunig, M. M., peter Kriegel, H. and Sander, J.: OPTICS: Ordering Points To Identify the Clustering Structure, *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 49–60 (1999).

[2] Chen, Xinlei, A. S. and Gupta, A.: Neil: Extracting visual knowledge from web data, *Proc. of IEEE International Conference on Computer Vision* (2013).

[3] Chiu, A. L.-m. and Fu., A.-C.: Enhancements on local outlier detection, *Proceedings of IEEE Database Engineering and Applications Symposium* (2003).

[4] Do, H. N. and Yanai, K.: Automatic extraction of relevant video shots of specific actions exploiting Web data, *Computer Vision and Image Understanding*, Vol. 118, No. 1, pp. 2 – 15 (2014).

[5] Do, N. H. and Yanai, K.: A Dense SURF and Triangulation Based Spatio-temporal Feature for Action Recognition, *Proc. of International Conference on Multimedia Modelling*, pp. 375–387 (2014).

[6] Golge, E. and Duygulu, P.: ConceptMap: Mining Noisy Web Data for Concept Learning, *Proc. of European Conference on Computer Vision*, Vol. 8695, pp. 439–455 (2014).

[7] Jing, Y. and Baluja, S.: VisualRank: Applying PageRank to Large-Scale Image Search, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 30, No. 11, pp. 1870–1890 (2008).

[8] Laptev, I., Marszalek, M., Schmid, C. and Rozenfeld, B.: Learning realistic human actions from movies, *Proc. of IEEE Computer Vision and Pattern Recognition* (2008).

[9] Liu, J., Luo, J. and Shah, M.: Recognizing realistic actions from videos, *Proc. of IEEE Computer Vision and Pattern Recognition*, pp. 1996–2003 (2009).

[10] Xia, Y., Cao, X., Wen, F. and Sun, J.: Well Begun Is Half Done: Generating High-Quality Seeds for Automatic Image Dataset Construction from Web, *Proc. of European Conference on Computer Vision*, Vol. 8692, pp. 387–400 (2014).