Computer Vision and Image Understanding 118 (2014) 2–15

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Automatic extraction of relevant video shots of specific actions exploiting Web data

Do Hang Nga *, Keiji Yanai

Department of Informatics, The University of Electro-Communications, Tokyo 1-5-1 Chogugaoka, Chofu-shi, Tokyo 182-8585, Japan

ARTICLE INFO

Article history:
Received 17 September 2012
Accepted 29 March 2013
Available online 6 September 2013

Keywords:
Web video
Web image
Tag relevance
VisualRank
Spatio-temporal feature
Pose estimation

ABSTRACT

Video sharing websites have recently become a tremendous video source, which is easily accessible without any costs. This has encouraged researchers in the action recognition field to construct action database exploiting Web sources. However Web sources are generally too noisy to be used directly as a recognition database. Thus building action database from Web sources has required extensive human efforts on manual selection of video parts related to specified actions. In this paper, we introduce a novel method to automatically extract video shots related to given action keywords from Web videos according to their metadata and visual features. First, we select relevant videos among tagged Web videos based on the relevance between their tags and the given keyword. After segmenting selected videos into shots, we rank these shots exploiting their visual features in order to obtain shots of interest as top ranked shots. Especially, we propose to adopt Web images and human pose matching method in shot ranking step and show that this application helps to boost more relevant shots to the top. This unsupervised method of ours only requires the provision of action keywords such as "surf wave" or "bake bread" at the beginn ing. We have made large-scale experiments on various kinds of human actions as well as non-human actions and obtained promising results.

1. Introduction

The explosion of the Internet as well as the rising need of sharing information between people on the Internet has made it a huge and unstoppably growing data source. People upload to the Internet every kind of data including images, music and videos. YouTube is one of the most popular video sharing websites which allow people to easily upload their own videos and access to those of others. By using Web API like YouTube API, we can obtain a large number of videos of various topics from Web sources without any difficulties. Especially, many of the topics are related to human and human actions; therefore, there is an increasing tendency for action recognition researchers to construct human action database exploiting Web videos.

When users upload their videos, they usually attach to the videos keywords called as "tags"- useful metadata for video retrieval. However, in general, tags are annotated to the whole video sequence, not to specific scenes. Therefore, it cannot be determined which tag corresponds to which part of the video. For example, some videos tagged "eat" might include not only the eating scene but also such other scenes as entering restaurants, ordering foods, or drinking something (see Fig. 1). People who want to search for eating scenes have to manually skip the scenes of no interest while carefully watching the whole video. This manual search makes Web videos based database construction for specific actions become a very troublesome and time-consuming task.

In this paper, we propose a new method to automatically extract from tagged Web videos relevant video shots of specific actions using metadata as well as visual context of these videos. Note that video shots here refer to small fragments of a video obtained by separating it at each point of a scene or camera change. Our unsupervised method requires only the provision of action keywords at the beginning. As for keywords, we mainly focus on words related to human actions. Our list of human action keywords contains sport activities such as "serve volleyball" or "row dumbbell" as well as activities of daily living like "shave mustache" and "tie shoelace". The list also includes some music related activities like "play trumpet" and "dance flamenco" or emotion related activities like "slap face" and "cry" as the consequence of "being angry" and "being sad" respectively. Moreover, we also tried several non-human actions such as "flowers bloom" or "leaves fall". We want to demonstrate that our proposed system can be applied to extract relevant video shots of various types of actions from the Web.

If video shots corresponding to any "action verb" can be acquired automatically from unconstrained videos like Web videos,

* Corresponding author.
E-mail addresses: dohang@mm.cs.uec.ac.jp (D.H. Nga), yanai@cs.uec.ac.jp (K. Yanai).

1077-3142/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.cviu.2013.03.009

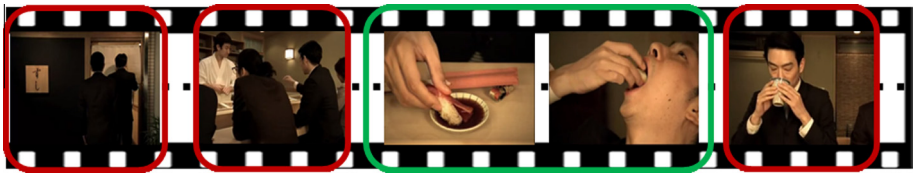

Fig. 1. A video obtained by searching with "eat sushi" keyword on YouTube. It contains scenes of interest (scenes with green bounding box) describing "eat sushi" action as well as irrelevant scenes (scenes with red bounding box) describing actions of entering restaurant, ordering sushi and drinking tea (respectively from the left to the right). Researchers who need only training data for "eat sushi" action must watch the whole video carefully to find its relevant scenes. Extracting relevant scenes in an unsupervised way is the purpose of this paper. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

we can build easily training database for action recognition. So far, as mentioned above, the construction of action training data has been known as exceptionally expensive work, which is totally different from building object recognition database. In fact, while object recognition categorizes up to 10,000 objects, the largest widely used action recognition dataset includes only 51 human action categories [1]. On the other hand, by applying our method, video shots associated with unlimited types of actions can be easily collected from Web sources. Although a few modest manual scanning may still be needed to use these video shots as training data for action recognition, there is no doubt that human effort can be significantly reduced in comparison to fully manual database construction. In addition, the proposed method can be applied to improve Web video searching and tagging.

Our main idea is at first, selecting relevant videos among thousands of Web videos for specified action and then, extracting the most related shots from selected videos. The video selection step is based on our assumption that videos tagged with many relevant words have high probability of being relevant videos so they should be selected. For the extraction of corresponding shots, we apply an efficient unsupervised ranking method called VisualRank [2]. We made large-scale experiments on 100 human action keywords and 12 non-human action keywords. The experimental results reflect the effectiveness of our system as we achieved high precision for many keywords. Note that here precision is considered as the percentage of relevant shots among top ranked 100 shots (Precision@100).

Furthermore, we proposed to take still Web images corresponding to given actions into account, with the intuition that the shots with more similarity to related action images have higher probability of being relevant shots, thus they should be biased in shot ranking. In fact, recent works [3–6] show that action recognition exploiting still images is possible. We collect images related to the given actions automatically via Web image search engines based only on provided keywords and measure visual resemblances between video shots and selected images. Shots with higher similarity scores will have higher chance to be ranked to the top. Note that these Web images involved processes also do not require any supervision, therefore the automaticity of the whole framework can be preserved. We verify the efficiency of introducing Web images by applying Web images exploited framework on 28 human actions and 8 non-human actions with precision achieved by original framework respectively lower than 20% and 15%. The results demonstrate that exploiting Web action images can significantly improve the performance of the original system.

The contributions of this paper can be summarized as follows: (1) a practical system of automatic construction of an action video shot database consisting of tag-based video selection from a large number of tagged videos, and visual-feature-based shot selection using VisualRank extended with novel and effective settings of non-uniform damping vector, and (2) large-scale experiments on 100 categories of human action and 12 categories of non-human action to verify the efficiency of each step of the pipeline as well as the whole system.

This manuscript is the expansion of our previous conference papers [7,8]. We made several modifications in the algorithm over the previous version and obtained significant improvements in the results. We introduce an extension to our approach that integrates pose feature into shot-image similarity measurement (Section 4), as well as new results of the improved system (Section 5).

2. Related work

In this section, we refer to some related works on action recognition, Web image mining, tag ranking and cross-media retrieval.

2.1. Action recognition

For the past five years, spatio-temporal (ST) features which describe both spatial and temporal description of movement, and their bag-of-features (BoF) representation, due to their effectiveness, have been exploited by many researches on human action recognition and content-based video analysis. By using BoF of ST features, action recognition problem can be almost regarded as the same problem as object recognition except for feature extraction process.

One of recent works which adopt this methodology is Jones et al.'s work [9]. The novelty in their work is that the results are refined based on users' relevance feedback following previous works in the image domain.

Nevertheless, the BoF model suffers from some limitations, one of which is the loss of some discriminative information in both spatial and temporal dimensions. As one of other effective models of human action recognition, a dense representation proposed by Zhen et al. [10] takes into account the motion and structure information simultaneously. In this work, high dimensional features are first extracted and then embedded into a compact and discriminative representation by discriminative locality alignment (DLA) method. On the other hand, instead of using all frames in the video sequence, Liu et al. [11] proposed to learn the most representative frames called as key frames by AdaBoost algorithm and represent action by the probabilistic distribution and temporal relationships of these frames.

The above mentioned works aimed to label to the whole content of each test video sequence one of the pre-defined categories, while our objective is to search among a large number of Web videos for only video parts which are associated with the given keywords. Moreover, while our proposed system is unsupervised, all of the above works apply supervised learning method which implements Support Vector Machines (SVMs) as the final classifiers. SVM is widely used in computer vision and machine learning in general and pattern recognition in particular.

Beside the supervised methods, there have been several attempts in unsupervised action recognition. Niebles et al. [12] categorized action videos in KTH datasets and their original ice-skating video data adopting the PLSA model. Niebles et al. [13] also proposed a method to extract human action sequences from