Journal of Visual Communication and Image Representation 55 (2018) 648–653

Contents lists available at ScienceDirect

Journal of Visual Communication and Image Representation

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

A sound-based video clipping framework toward sports scenes☆

Qunchao Mi*, Dali Xue

Regulation and Control Center, State Grid Wenzhou Power Supply Company, Wenzhou, China

ARTICLE INFO

Keywords:
Sonogram
Deep representation
Convolutional neural network

ABSTRACT

Video clipping system is very important in many intelligent applications. In order to shorten the time of video and extract the framework of the video, many methods have been proposed. But these methods just considered videos without taking sound into account. As we know, sound is also an important information for image processing. For example, many sport match videos include rich sound of audience and commentator such as NBA. In addition, human pay more attention to some video clips of interest (VCOI) such as scoring time instead of pause. So in this paper, we propose a sound-based video clipping framework toward specific sports scenes. First, we convert sound of sport videos to sonogram. For some aesthetically-pleasing images (APIM) such as slam dunk or jump shot, a set of object patches are selected using BING feature. Then, these object patches are ordered by our active object patches ranking algorithm. After that, ordered object patches and sonogram are fed into CNN respectively to obtain patch-level deep feature. In order to obtain image-level deep representation, deep feature extracted from ordered object patches are aggregated statistically into a deep representation. Finally, probabilistic model is used to select VCOI and APIM. Experiments on some NBA basketball matches have shown the effectiveness of our video clipping framework.

1. Introduction

With the development of multimedia technology, video data has exploded. How to deal with these videos effectively is still a big problem [22–27]. For example, in the field of video surveillance, many large-scale cameras are widely used in various places. Many video surveillance systems monitor the same region in different angles at the same time. Thus, how to clip these videos and extract important information is a challenge task. In addition, many videos last a long time, but human always pay attention to the most important part of them. How to find such video clips is a useful application in intelligent systems.

In sport matches, human pay more attention to some video clips of interest (VCOI), that is to say, human prefer to watch some video clips such as score time or match summary. It is because watching these video clips not only save time, but also understand the whole match. In the sports scenes, one of the important information is the sound of audience and commentator. These VCOI are always accompanied by the celebration of the audience and the commentator. These sound information is important to extract VCOI from the whole match. For many basketball matches, there are many aesthetically-pleasing images just like slam dunk, jump shot and so on. Many basketball fans prefer to using these images as their wallpaper for their smartphones. So it is useful to find these aesthetically-pleasing images automatically from the VCOI.

Biological and psychological experiments have indicated that human will allocate their gazes when they observe an image. That is to say, human will be attracted by the most salient region within an image, then their gaze will be allocated to the second salient region and so on. Obviously, it is important to incorporate human visual mechanism to find aesthetically-pleasing images.

Based on our discussions above, it is possible to design a sound-based video clipping framework toward specific sports scenes. In order to make use of the sound of the video, we first convert a set of sound of training videos to sonograms. Then, these sonograms are fed into CNN to extract deep feature. Finally, our probabilistic model is used to select video clips of interest. At the same time, in order to select aesthetically-pleasing images from these video clips of interest, some object patches are selected from a set of training images using BING feature. Then, these object patches are ordered by our active object patches ranking algorithm to form a GSP within each training image. Then, these ordered object patches are fed into CNN to extract patch-level deep features. After that, these patch-level deep features within an image are aggregated statistically into deep representation. Finally, our

☆ This paper has been recommended for acceptance by Ququ Chen.

* Corresponding author.
E-mail address: 1963561164@qq.com (Q. Mi).

https://doi.org/10.1016/j.jvcir.2018.07.008
Received 7 May 2018; Received in revised form 29 June 2018; Accepted 9 July 2018
Available online 10 July 2018

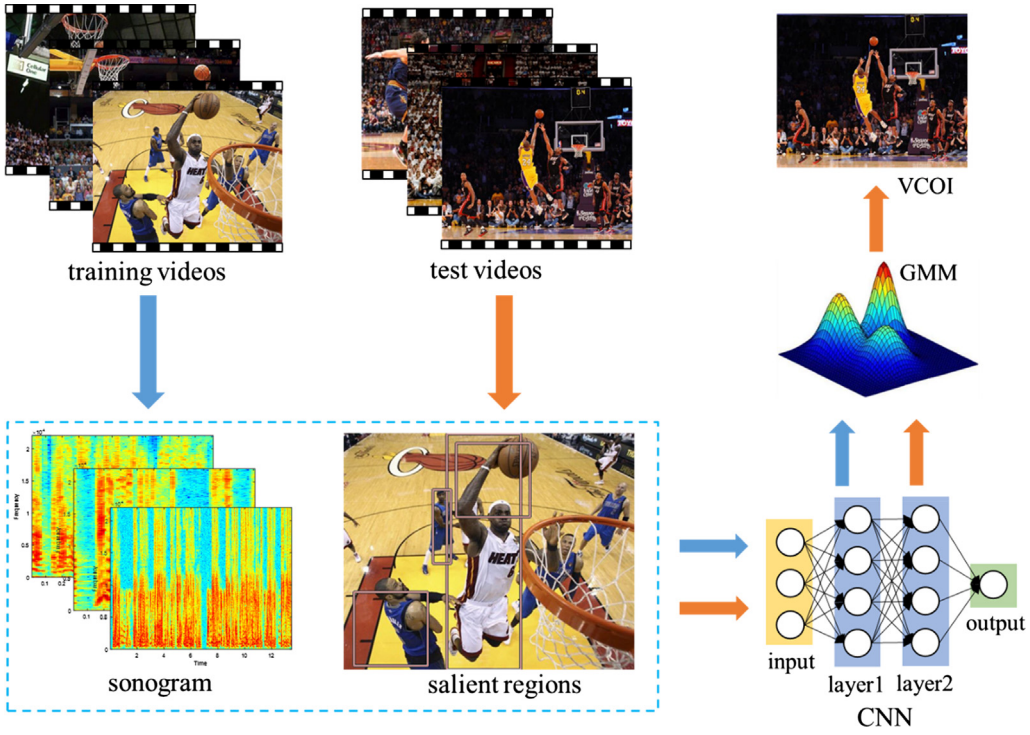

Fig. 1. The pipeline of our sound-based video clipping framework.

probabilistic model is used to select aesthetically-pleasing images from video clips of interest. Our pipeline is summarized in Fig. 1.

2. Related work

Video clipping or photo cropping is very important in intelligent systems [2,15–17]. Photo cropping can preserve areas of interest to human beings and reduce storage space. Removing the unnecessary area of the image does not affect human aesthetic experience. The video clip can remove the redundant frame part of the video and retain the part of the video that people are interested in. Zhang et al. [3] proposed a video summarization framework, the framework aims to fuse multiple videos collected from different cameras in different angles into a comprehensive one. This framework is very meaningful in video surveillance. In [4], a novel algorithm to summarize one-shot landmark videos was proposed. This model can accurately evaluate the quality of the video summary. Hazem EI-Alfy et al. [8] proposed a framework for cropping surveillance videos. The information of the video was modeled by whether the image changed at each pixel. M. Rehan et al. [9] proposed a technique for frame cropping of Moving picture experts group video.

In our work, we incorporate human gaze shifting path (GSP). Many works have shown that it is effective to take GSP into account in the field of multimedia. Studies have shown that when humans observe an image, they will proceed in a certain order. This sequence includes human perception of images. That is to say, human gaze shifting paths (GSP) potentially includes the characteristics of an image. In our work, GSP can be used to find aesthetically-pleasing images from video clips of interest. In [5], GSP is used for image/video retargeting. In their work, each aesthetically-pleasing photo can be represented by a GSP which can preserve the aesthetically-pleasing effect within an image. Zhang et al. [6] proposed a novel framework that evaluate media quality combining GSP. In this framework, the key algorithm is a locality-preserved sparse encoding algorithm that can discover GSP within an image. This framework is a big contribution in the field of multimedia. Sun [7] proposed a novel scene categorization using GSP kernel. GSP can discover a set of salient regions within an image, and then deep representation can be learned by an aggregated-based CNN. Finally, these deep representations can be converted into an image kernel, the image kernel is fed into SVM for scene categorization.

3. Our proposed method

As mentioned above, humans may pay more attention to some particular scenes toward sports games, e.g., humans prefer watching some of the goals instead of pause when they watch basketball games such as NBA. However, extracting these particular scenes manually is a tedious task. Thus, we propose our summarization framework toward such particular sports scenes.

3.1. Human gaze shifting path modeling

Human will allocate their gazes when observe an image or a scene. First, they will be attracted by the most salient region, and then the second salient region, and so on. So, it is meaningful to combine such human visual mechanism in the image processing.

In order to make use of the voice of the commentator in a sport match, we convert sound to the sonogram. The sonogram contains rich sound information. For example, when the goal is scored, there will be cheers and commentator's celebrations. Humans pay more attention to these scenes because these scenes are exciting and aesthetically-pleasing. So we are concerned about choosing such scenes.

3.1.1. Bing-based object patches selection

In the sports scene, human pay more attention to some object patches such as basketball players, basketball hoop and basketball. In order to select these object patches that may draw human attention, we leverage BING feature proposed by Cheng et al. [1]. BING feature is used for training a generic objectness measure. This generic objectness measure can find most of the foreground semantic objects within an image in a fast speed. Thus, the computation will be reduced a lot with these object proposals.

In our work, we leverage BING feature to find aesthetically-pleasing moment such as some exciting time of slam dunk. First, we select s set