Project: Standard Framework of Video Content Processing

Haichuan Yang

Summary

This is a proposal to launch the project of creating a framework to analyze and process the content of videos. The framework may contain data structure and algorithms. The data structure acts as the basic element to process or analysis the content of video, for example, it could be a set of image features of each frame, and the corresponding audio feature. The algorithm is defined upon this data structure, i.e., how to take use of the feature set. Both the data structure and the algorithm keep the potential to be extended. New kinds of features and processing algorithms can be easily added in this standard framework, and the standard interface will make them work together.

Intellectual Merit. The proposed research project will provide a standard way to digitally understand and describe the content of video. It is based on the thriving research on computer vision, acoustics and signal processing, and show its own way to compose these aspects together to define a set of methods which are specifically used in tasks about video content. To the academic community of video understanding and processing, the proposed framework extracts the very commonly used part from similar or different algorithms and tasks, make it easy to develop new algorithms based on it.

Broader Impacts. As the data form that takes the most percentage, videos can represent lots of thing. Almost all the aspects of our life involve videos. Video is a very efficient way to represent and send information. Nowadays, as the development of smart phone and digital camera, generating high-quality videos is quite easier. As everyone participating in this activity, the amount of videos is increasing as an incredible speed. However, the analysis tools for video do not catch this increase. For solving a new problem, people usually need to start all the things from scratch. Besides that, different with text or images, video contains much more information, both visual and audial. By decomposing the complexity of video analysis with the proposed framework, solving this problem and developing new algorithms will be much more efficient.
Project Description

The main work of this projected is to develop a set of framework to support the design of new video analysis algorithms. We anticipate this framework can unify most of the extant video analysis methods, and have good flexibility. The framework also has the implementations for feature extraction, and processing algorithms, which can be directly used for practical problems. We wish to use this framework to extract the basic component and create a scaffold to serve the researchers and engineers, to free them from the repeated and low-efficient scratch work.

Motivation

Existing video analysis methods usually face several problems: How to represent the video, what kind of feature should be used, and whether the analysis algorithm can handle the scale of the data. Take near-duplicate video retrieval [1] as an example. Near-duplicate video retrieval is a task to search the video set, and find the sample which is nearly the same to the query. The video set could be very large, e.g., IMDB has 3,446,167 titles [2]. If assume each video length is 10 minutes, and we have 12 frames per second. There are totally 24,812,402,400 images. If we directly represent video with all the image features, there will be pretty large computation in the retrieval process. On the other hand, we all know that retrieval usually require real-time, which means we do not have much time for the whole retrieval process. So it is impossible to accomplish this job with plain image representation.

From the example above, we can find that video analysis is usually a complex job. Different with image and text, video contains multimodal data: visual and audial. Compared with one dimensional audio and two dimensional image, video has three dimensions. It is the complex composition which consists of two modalities and three dimensions, so it is nearly impossible to handle it with one single algorithm. For our researchers and engineers, it usually needs many kinds of methods to cope different aspects of the problem, and it will take lots of time to handle them if always start from scratch.

Method and Plan

Actually, people already have lots of experience to decompose complex problem to be simpler ones, i.e., divide and conquer. In the history of computer science and engineering, we can take some successful cases as example.

The first example is network protocols [3]. The different network functions are assigned
to different layers, the whole function of network communication is accomplished by the cooperation of different network layers. On each layer, there may exists different kind of protocol, but they all have some similarity and can work on the base of its lower layer. The whole network model is just a layer-wise framework, and it provide basic function of communication between nodes. It allows people to implement specific protocol, and plug it into the network protocol stack. If a researcher or engineer only want to create or modify a protocol in one layer, he do not need to change anything of other layers. This framework simplifies the process of developing new network algorithm and eventually contributes to our internet world today.

Another example is the hardware architecture of modern computer. Because computer is a complex system, it need to complete lots of functions, e.g., input, output, compute, communicate, storage, etc. People divided these features and put them into different components, then gather these components, plug them in the board, then we see the computer. We must admit that the dramatic development speed of computer cannot leave the contribution made by modularization.

Taking these successful experience, we can learn something to build our video processing framework. There are several approaches can be considered in the plan of our work.

1. Compare the application background and problems in video processing and computer network, can we use layer-wise framework in our project? This depends on whether the solution of most video processing tasks have the layer-wise structure, which means the problem can be separated to several sub-problems, and each problem only depends on the solution of the previous problem. If we can get layer-wise structure in the video processing problem, we can build a layer-wise framework just as network protocol stack.

2. Try to decompose the video processing problem into disjoint modules, different module can have different function. Build a central module to gather all the modules, all the sub-modules are responsible to interact with the central module. It is not hard to see that this approach is just analogy of the computer architecture. Similarly, it is not sure that our video processing task have the modular structure, so we should analysis the most extant video analysis tasks and algorithms, to find the answer.

3. The third way is to combine the approach 1 and 2. Maybe the video processing task is not layer-wise, but it can be divided into several modules and each modules has a layer-wise structure. Or maybe it cannot be modularized directly, but it can be separated into different layers and each layer has modular structure. Video processing is a very complex task, it is definitely possible that our framework has a very complex structure.

4. Maybe neither example is suitable to our domain. In this situation, we need to create our own approach to build the framework. Firstly, we should collect the most typical and common video processing task, e.g., video retrieval, event detection, etc. Then we also should collect the classic methods for them. At last, we try to analyze the common structure or aspect of these methods, and build the framework.
Another useful tools for building our framework is the object-oriented strategy. Object-oriented strategy can hide the complex details of the implementation, and this is just what we need.

**Experiment Criterion**

The experiment criterion is nothing but to solve all the known typical video processing task by our framework. This maybe sounds incredible, but it is totally possible. It is not necessary to implement every algorithm in video processing, we only need to get the state-of-the-art level evaluation result for a certain problem. Since the number of typical problems is much smaller than the number of all the video processing algorithms, our criterion is feasible.

**Broader Impact of the Proposed Work**

Education, entertainment, commercial, sport, art, news, almost all aspects of our life involve videos. This framework will change the extant process of dealing with new problems about video analysis, and give a clear logic structure of the algorithm. Given this framework, researchers and engineers can design their new algorithm or solve their new task upon it. Maybe one can just find implemented data structure and algorithm to solve his problem, or need to plug his own implementation in the framework. It also contains completed routines to implement common jobs about video content, such as retrieval, indexing, similarity computation, and so on.
Reference

https://en.wikipedia.org/wiki/Communications_protocol