Classification of YouTube Food Videos

Zeqian Li, and Yu Liu
Worcester Polytechnic Institute
{zli14, yliu25}@cs.wpi.edu

***Abstract*—ABSTRACT**

I. Introduction

Video is a popular resource for people to grasp information from internet. Meanwhile, people also are getting used to share their lives, skills, experiences, etc. online by uploading the videos they recorded to platforms like YouTube or some social media, especially food videos, which is a hot topic in video sharing market. Thus, automatically and accurately classifying those food related videos is important for video platforms to arrange reasonably. Then customers can use proper video classifications to efficiently find what they need.

There are many different schemes and models work on video classification. The problem is that the food related videos are different from other type of video in characters. Our work implements and compares three different methods to classify food videos by countries as, specifically, Chinese food, Japanese food and American food. Those videos including cooking recipes, travelling experience, culture introduction or food blog organizers, all of which contain large amount unrelated elements. Further, some food videos are very close to each other. For example, cooking teaching videos are commonly with plenty frames of raw ingredients. While Japan is famous of raw food. So the result will be of low accuracy if we classify videos based on independent frames. So it is important for video classification to specify the inner impact between frames. Otherwise, some unrelated frames my dominant and negatively impact on the result. With our work, we can know which model is most suitable for food related videos classification.

II. Related Work

CNN is commonly used in image recognition while rarely used in video classification. The main reason was CNN needs large dataset to train the model. Compared with the largest dataset like CCV (9317 videos and 101 classes) and UCF-101 [1.22](13320 videos and 101 classes), recent YouTube 8M video solved this problem totally. As we only classify videos into three classes, we grasp very small portion of data from it.

Frame extraction is an important step for video classification. Extracting Local frames in a small time period is a method used to analyze motions of a video[1.25]. But it was proved that this method is not better than nave approach of classifying individual frames as the frames are usually from small video clips. Furthermore, extracting frames in a sparse manner helps the system to establish potential connection of different frames[1.12, 1.8]. Another popular method is adding weights to different frames according to the significance of the frames in a video[1.13]. While this method can help system focus on more essential frames, we propose video-level training to solve this problem. Our method is supposed to largely reduce unrelated elements and also build strong connection between frames.

Aggregating features of frames is the last step for video classification. This step utilize the inner connection among frames to train a more accurate result. As locally extracted frames was approved not helpful, some previous work used recurrent network. However, standard recurrent networks introduces troubles learning over a long period as the problem of vanishing and exploding gradients[2.3]. Rather, Long Short Term Memory (LSTM) [2.11] records the inner states using memory cells, which works better on long-range relationships. AlexNet[1, 1.15] has built their architecture with CNN and LSTM. Our work modified some features of AlexNet to a model more suitable for our food video classification program.