Simultaneous Classification of Actors and Actions Within Video Sequences

Philomina Ekezie
Mercer University
xxphilominaxx@gmail.com

Nyle Siddiqui
University of Central Florida
nyle.siddiqui@ucf.edu

Abstract

Subject recognition and action detection are two significantly large domains within the world of computer vision, being especially common tasks within video processing. There do exist many video processing models that are successful at solving either one of these tasks in isolation. However, it is necessary to note that these models commonly tackle these tasks separately, such that a model that is equipped to identify a subject has no means for identifying the action, and vice versa for a model that is structured for detecting actions. Additionally, these models are trained on datasets that include a collection of either only action or subject labels. A model that can accomplish both tasks has not been proposed. Additionally, there also exists no datasets curated for such a model, which would ideally have both subject and action labels. There are instances where both the subject and the action need to be identified: a situation like this would require the novel solution posed by my project, more specifically, “simultaneous classification of who’s doing what in video sequences”. With this novel problem in mind, and the lack of works dedicated to solving it, this paper proposes a novel dataset of video sequences that include both subject and action labels for the purpose of training models to accomplish subject recognition and action detection at the same time. Also, we introduce a model that is engineered to conduct simultaneous subject and action detection and trained using the novel dataset. The results derived from training the model on this dataset show that our model is able to accomplish both subject identification and action recognition. Therefore, it is evident that our dataset is capable of training a model to perform these two big tasks and could produce better results once it is bigger and includes more data.

1. Introduction

There are many instances where, when given a video sequence, either a subject needs to be identified or an action needs to be detected. Examples from day-to-day life are plentiful: for subject detection, one instance includes autonomous driving vehicles which frequently decipher if there is a person in the rear, front, or side of the vehicle for safety purposes; another instance includes airport security cameras, which are able to identify individuals based on their face. For action detection, its applications can be found in areas such as behavior analysis, video retrieval, and human-robot interaction [Yhi Zu et. al, 2020]. Needless to say, subject recognition and action detection are two very popular domains within video processing. A less popular yet pertinent situation occurs when both of these domains are combined; that is, instances where distinguishing both subject and action in a video are less commonly addressed but are just as present in the real world. For example, in the surveillance footage for a convenience store that was recently robbed, knowing “who” is in the footage and “what” the individual is doing would prove relevant to police officials. Additionally, in sports analysis videos, knowing “who” the player is and “what” actions the player has completed is necessary for accurate play-by-play analysis.

Because existing datasets fail to contain both subject and action labels, they fail to be adequate sources of data for our novel problem. Thus, a novel dataset was needed to specifically support our project efforts. Our proposed dataset is a collection of about 3,000 RGB video samples. The video samples feature ten different subjects, specifically famous tennis players, performing one of six distinct actions in tennis.

We propose a complex model that has been strategically designed for simultaneous subject and action detection. First, its 3D-CNN backbone is used to extract spatio-temporal features from the video input. Then, these features are separately sent to a pair of transformer decoders that learn the actor and action features separately from each other. The significant aspect of this model is its use of Disentangled Representation Learning to separate the actor and subject features and learn them more efficiently.

To summarize these points, our contributions include

• A novel dataset that includes both subject and action labels to support the accomplishment of our novel problem.
• A complex model that features a 3D-CNN encoder and two transformer decoders built for the purpose of simultaneous subject and action detection.

The rest of the paper is structured as follows. Section 2 references other works that discuss methods of video action recognition and datasets similar to that of our own. Section 3 describes the statistics of our dataset, provides a walkthrough of our dataset collection process, and gives in-detail descriptions of our model architecture. Section 4 provides our experimental results and discusses the meaning of the percentages our model was able to achieve. Finally, section 5 concludes the paper by giving a summary of the main points and mentioning our future hopes for our work.

2. Related Works

The complex tasks of this project’s focus, subject detection and action detection, are two of the largest domains within video processing and understanding. In the past decade, a large amount of research has been conducted within these areas. Multiple models that aim to accomplish either actor or action reignition can be found, and there exist large datasets that support the implementation of such models. In this section, I reference related works that focus on subject and action recognition of videos, including both currently existing models and datasets.

The authors of [Mora et al., 2017] believe that finer-grain classification is better than more coarse data. In their paper, they propose a neural network architecture whose purpose is to classify videos within the THETIS dataset. The THETIS dataset, which will be further discussed shortly, is a dataset of fine-grained tennis actions including footage from 55 subjects performing 12 distinct tennis shots multiple times [Mora et al., 2017]. The main steps of their algorithm are first, feature extraction, and secondly, classification through an LSTM. In the feature extraction portion, they use Inception, which is a well-known deep CNN architecture. Every video input is shortened to the first 100 clips, and then Inception is used to make predictions about each frame of the video. In the classification stage, the input passes through a stack of three LSTMS and then passed through a softmax layer to produce the predicted layer. Our work is slightly similar to the work by [Mora et al., 2017], in the sense that we also use a CNN to extract features from our input video, though we use another common one known as R3D; however, instead of using LSTMS, we use a pair of transformer decoders. These decoders are extremely efficient in disentangling the actor and action features, something that may not be possible when using the LSTM networks.

Although the number of datasets specifically curated for video action recognition within tennis are extremely scarce, there does exist the THETIS dataset. This dataset was presented in 2013 and consists of fine-grained tennis actions including footage from 55 subjects performing 12 distinct tennis shots multiple times [Mora et al., 2017]. There are 1980 RGB videos within the dataset. These videos are shot-in-the-wild, low-definition, and include both occlusions and dynamic backgrounds. Although this dataset is specifically designed for recognition of tennis actions, a quality that is very scarce among datasets, it does have some poor characteristics that our dataset made sure to avoid. First, the videos do not include tennis balls and simply showcases the subjects performing the action. This causes a large opportunity for error in classifying certain actions, specifically the serve and the volley. As mentioned in [Mora et al., 2017], their network had trouble deciding between serves and smashes, and they mention the lack of tennis balls in the videos as a possible reason for this. Secondly, the videos being shot-in-the-wild mean that they videos are not taken from tennis courts. Being that tennis actions are usually performed on tennis courts, this can prevent the dataset from being as applicable to real-world situations as possible [Mora et al., 2017]. Unlike the THETIS dataset, our dataset features videos that showcase the subjects performing actions on actual tennis courts and includes the person, the racket, the tennis ball, the net, and all other realistic features. We also strategically chose our action classes so that no two actions were too similar in nature so as to limit any chance of confusion.

3. Methodology

When it comes to solving the problem of simultaneous subject and action detection, we found that the solution has two fundamental parts: a model and a dataset. First, a model with the architecture to achieve both tasks was crucial. A common limitation found within previous models was that they could only achieve one of the two tasks but were not built to tackle both in a merged manner. So, we introduce a model that is designed to accomplish both tasks in a unified approach. Secondly, a dataset specifically curated for the problem at hand prior to our research was nonexistent; many of them were constructed for sole subject recognition or sole action detection, and those that featured both subject and actor labels were not built specifically for this problem. Resultingly, we established and populated a novel dataset to specifically tend to simultaneous subject and action recognition.

3.1. Dataset Statistics

Our dataset is comprised of about 3,000 RGB video samples, a number that is ever-growing as the dataset is still in the process of being expanded. Each of these video samples is a short video clip of a single subject performing a single action. The dataset has ten subject classes as well as six distinct action classes. The individuals and actions that comprise the subject and action classes all stem from the sport of
tennis. The dataset contains clips that have varying camera-angles and court set-ups. Camera angles included face-on, side, and ¾ angles, capturing all sides of the individual excluding the back. The court views included overhead court views from the opposite side of the subject, courtside views from either the left or the right, and up-close, on-court views as seen in personal training videos or practice sessions. To organize the files within our dataset, we used the naming convention “S-A-x-”. For clarification, a file with the label “S1A2x4” would refer to a video that contains Subject 1 performing Action 2 for the 4th repetition.

3.2. Dataset Population

The technique that was used to oversee dataset collection is described as follows. The first step was choosing one subject from the ten subjects that made up the subject classes. Once chosen, a full-length tennis match featuring this subject and an opponent was found via YouTube and downloaded for further use (all URLs are accounted for in a separate document). Then, the video would be split into smaller clips where the subject performed only one action from the action class. The video samples were then loaded into our dataset annotation/storage tool, VIAME, an open-source computer vision software platform designed for projects in AI. Once these clips were in VIAME, the data could be fed to the model.

3.3. Annotations

During the beginning stages of the data collection, the video clips were annotated using our annotation tool, VIAME. Due to the fact that our video samples featured varying camera views from different positions around the court, the subject of our videos was not guaranteed to remain in the center of frame at all moments. Thus, we needed to incorporate per-frame annotations so that the subject could be located at all parts of the video. We accomplished this by using the annotation tool provided by VIAME to add bounding boxes around the individual at each frame within the video. The incorporation of these per-frame annotations were especially helpful to the model, which could better locate the subject as a result of their presence.

3.4. Model Architecture

In addition to the novel dataset, we have proposed a model that is capable of performing subject and action recognition simultaneously mainly through its disentanglement feature. The model receives a group of three videos during the input stage: the original video, a second video where the same actor completes a different action, and a third video where a different actor completes the same initial action. The number of frames along with the number channels, height, and width of each frame is taken into account. After the input portion of the architecture, the model features two main parts to be described shortly: a 3D-CNN backbone and a set of two transformer decoders for disentangled feature learning.

3D-CNN Backbone The backbone of the model is a 3D-CNN, specifically a common CNN known as I3D. The purpose of this 3D-CNN is to extract spatio-temporal features from the input videos. This backbone generates a feature map as output; a learnable positional encoding is added to this output directly after. Then, the feature map along with the added positional embedding is fed to one of the two transformer decoders, for subject and actor respectively, to undergo the process of disentanglement.

Transformer Decoders The model consists of two transformer decoders, the subject decoder for subject disentanglement and the actor decoder for action disentanglement. After the spatio-temporal features are extracted from the input video by the backbone, the features for actor and action are separately sent to these decoders. Both of the ac-
tor and action decoders consist of twenty queries, which are vectors that are initialized with random values to be learned later by the model.

4. Results

The following results were produced from the initial training of our model on our dataset when it had been newly created and only contained about 200 samples. To get some preliminary results before having a larger, more established dataset, we trained the model on this “baby dataset”. During this training of our model, seven subjects were part of the training set and three subjects were part of the testing set. We recognized two types of results: top 1 percent accuracy (for the model’s subject recognition) and test accuracy (for the action recognition).

The metric used to measure the status of our subject recognition capabilities was feature retrieval. The subjects within the training and testing sets were different and unique, so to gain meaningful results, we could not simply use test accuracy. The metric used to measure the status of our action recognition capabilities was simply test accuracy. Because all the action classes were present during both training and testing, the main consideration was how many samples the model guessed the correct action for in the test set.

The model achieved a 66 percent Top-1 accuracy when it came to recognizing the subject. This revealed that the model was having trouble locating the subject. The model achieved a 50 percent accuracy when it came to detecting the action. This percentage revealed that the model needed more examples to efficiently distinguish the actions.

![Image of a table showing Top 1% Accuracy - Subject: 66.6% and Test Action - Action: 50%]

Figure 4. The results produced by our model.

5. Conclusion and Future Works

In conclusion, subject and action detection are both incredibly prominent domains in computer vision. There has been extensive research in each area, which has led to state-of-the-art models that are either able to efficiently process the subjects in a video or determine the action(s) performed in such a video. Additionally, there exist datasets that include subject labels for subject detection and action labels for action detection. However, there is a major lack of works that merge both tasks into a single problem. Thus, we introduce simultaneous classification of subjects and actions in videos as a novel problem. Our project aims to solve this novel problem by introducing a novel model that accomplishes the tasks. We also introduce a novel dataset that is specifically curated for this novel problem. We show the results produced by our model when trained on this dataset. We hope that our work can foster further research in this novel topic and that both our dataset and model can be utilized by future works that aim to tackle the complex problem of simultaneous subject and action recognition.

6. References
