Movie genre recognition

Reem Alrowaili

Seidenberg School of Computer Science

Pace University

A thesis submitted for the degree of

Master of Computer Science

December 2019

Abstract

In this thesis, we find a new way of analyzing movie trailers to classify them into the most general genre. Previous works were done using low-level feature extraction like color-histograms. Also, high-level feature extractions were done using scene categorizing and shot boundaries. However, we are trying to recognize and classify movie trailers into genres using object recognition and then calculating the frequencies of these objects in video trailers. Our approach got us 52% precision accuracy using pre-trained weights on the COCO dataset. In this thesis, we explained two methods that helped us analyze the movie trailers, object recognition, and emotion classification. Also, we used two classifiers to train on the object frequencies, neural nets, and random forest. The movie trailers dataset was collected from multiple channels on YouTube.

Table of Contents

List of Figures and Tables

1. Introduction

Recognizing movies and films by watching part of a scene of a movie is easy for humans, but could it be possible to be the same for a machine? A lot of models are constructed to recognize certain objects in an image or a video. By using labels that are placed by humans, making it possible to build near-perfect models to recognize many things in the world. How about recognizing movies and TV shows? It may be hard to construct the dataset for the movies since there are half a million movies according to the Internet Movie Database (IMDB) [1].

In this thesis, we will construct a system to recognize movie genres. There are 24 movie genres, and we want to pick the most common genres out of them to build our dataset for training. The genres that we picked to be recognized are Action, Comedy, Drama, Fantasy, and Horror. In the end, we will test the model to classify some clips and trailers to one of the genres we already specified.

This study will help show the ability of the model to distinguish common human actions to its right labeling and link these actions to the most general genres used by people to identify movies and films. It is possible to add speech recognition to get a more precise classification, but we will focus on the image and video analysis in this thesis.

To start working on such a study, we need to collect a set of movie clips and trailers and label them to the correct movie genres. These trailers will be collected from YouTube. As to what movies we start with and which genres? We could start with the most popular movies and genres that have actions in common. The collection of the dataset is an essential part of this study because all the work depends on the amount and type of data we are analyzing. Thus, we have collected approximately 900 movie trailers from YouTube and checked that it has the correct labeling of genres with IMDB. Although a movie trailer could contain multiple labeling of genres, we are going to focus on one labeling here. An example of a movie trailer that contains multiple labeling is the movie Fury. It is labeled as action, drama, and war. The action part in the movie because there is a lot of fast motion and body movement with the use of firearms. The drama part of the movie because it contains slow motion with conversations. The war part because it has the settings of war, which are tanks, the clothing of the soldiers, and firearms. All these objects and actions classify the movie into multiple genres. However, we are going to classify the movie to one labeling. Because it is easier to construct since the focus will be on one label, if we are successful in the process of recognizing the correct genre, we can expand the idea to multiple genres.

The analysis of the movies will be on a small part of the movie, which can be a trailer or a clip of the movie. The reason for this selection is that the trailer is like a summary of the whole movie. It describes the whole movie in a short clip. The trailer contains multiple scenes that are cut out of the movie. These selected scenes, in most cases, are descriptive of what kind of genre the movie is. However, some trailers give misinformation about the whole movie. For example, the scenes in the trailers may give information that the movie is an action movie. The person, in that case, is expecting an action type of movie with a lot of scenes that contain human fighting and car chasing, etc. The movie would contain two or three scenes with action type, and the rest of the movie is slow with a lot of conversations between actors.

We are not concerned if the movie trailer is giving misinformation about the movie because we want our system to recognize the information that is contained in the trailers itself, not the whole movie. Also, if the trailer is described as an action type of trailer, we will consider it as correct labeling even if the whole movie is considered drama and thriller with little action scenes in it. Because the trailer that was considered as an action, it must contain some action scenes from the movie. These action scenes would help us get a good training dataset for the action category. The same idea goes to every trained category.

In the second chapter, we will start off giving background information about the whole genre classification in the eyes of human beings in the first section of the chapter. The second section will discuss the research works that have been done to identify the genres of movies, also, the related works to the analysis of videos and human action recognition. The third chapter will discuss the methodologies used to help us analyze the movie trailers, including the structure of the whole system. The fourth chapter will present our findings from the used methodologies. The last chapter will discuss the results and limitations of our constructed system.

2. Background

In the first section, we are going to explain what a movie genre is and how each genre is identified. Next, we are going to show the different works that have been done in the video recognition field, including classifying movies into genres.

2-1. Movie Genres

A movie genre can be interpreted by narratives or characteristics or contexts of the movie. A movie can be classified into one of the genres based on these certain characteristics depicted in the movie. There are a lot of genre types; we will list some of these genres that are known and briefly explain each one.

There are difficulties in finding an obvious distinction between genres because genres may overlap or have a combination of genres. The similarities between genres can be seen, for example, between thriller and drama movies. Both have a slow-motion with lots of conversation. A combination of genres is when a movie has multiple genres that will classify it to both these genres equally, for example, a comedy and thriller movie. Genres, sometimes, are easy to recognize by a human being but may be difficult to be defined. The reason is that the features that characterize one genre are not exclusive to this specific genre. These characteristics may be seen in other genres. Thus, a genre is a combination of multiple characteristics that could be defined into it [2].

Types of genres that are recognized and used by people are Action, Adventure, Comedy, Crime, Drama, History, Horror, Musical, and Science Fiction. These genres, as explained earlier, may contain overlap characteristics when we try to define them. Also, it may contain sub-genres that specifically identifies the movie. In general, what identifies a genre is a combination of certain actions by the actors, the story, and the settings of the scene. Actions and settings of the scene are what we are focused on in our thesis. The story is hard to identify by the machines because it needs intuition to understand the story, which is not possible to do by the machines. We can do sentiment analysis on the plots of the movies to be classified into genres, but this is not the focus of this thesis. The definition of each genre in the next paragraphs is based on the film-site website by T. Dirks [3].

Action contains non-stop motion with lots of fighting scenes, car chasing, and use of firearms. Also, it could contain destructive crises, like floods, explosions, fires, etc. This genre can be in combination with other genres to identify a movie, but mainly, if the scenes contain these characteristics, the video would be recognized as an action genre.

Adventure usually contains exciting stories with new experiences, like or sometimes paired with the action genre. This genre may contain wide sceneries of jungle and desert, also, the appearance of different kinds of animals, including the use of a variety of artifacts.

Comedy is designed to amuse the audience and provoke laughter. This genre is hard to identify by the machine because the settings of the scene are similar to drama and action genres. One of the main characteristics of comedies is unusual, unexpected, and extreme human actions.

Crime is the sinister actions of criminals or mobster, particularly, bank robber’s underworld figures or ruthless hoodlums who operate outside the law. This genre is similar to the action genre because it usually contains fighting scenes and firearms usage.

Dramas are serious, plot-driven presentations portraying realistic characters, settings, life situations, and stories involving intense character development and interaction. This genre could be identified from the settings of the scene. Also, it contains slow to moderate human motions with conversation and interactions between actors.

History: medieval romps, or 'period pictures' that often cover a large expanse of time set against a vast, panoramic backdrop. The settings of the movies of this type of genre are hard to identify because the history ranges between the fifties to way older, which has completely different settings. However, these movies are considered history. That is why there are no obvious characteristics that could be identified by the machine to group all history movies.

Horrors are designed to frighten and to invoke our hidden worst fears. These types of genres contain lots of blood and dark settings. The human actions in these movies are almost similar in all horror movies. Also, it contains the use of knives and other kinds of killing tools.

Musical emphasizes full-scale scores or song and dance routines in a significant way. The human actions in these types of genres are dancing and singing, which could be identified by machines.

War has actual combat fighting on the land, sea, or in the air. These genres are similar to action genres in terms of firearm use and fast motion. It is hard to be distinguished from action movie genres.

Science fiction: Sci-fi films are often quasi-scientific, visionary and imaginative - complete with heroes, aliens, distant planets, impossible quests, improbable settings, fantastic places, great dark and shadowy villains, futuristic technology, unknown and unknowable forces, and extraordinary monsters. All these characteristics are enough to be distinguished from other genres.

The basis of these definitions is the human perceptions of the movie and how they categorize the movie into these types. Also, we mentioned some characteristics that could be used to recognize the genres in the machine perspective. It is hard to make the machine identify the category of a movie using the human perception of it because the machine cannot identify the emotional part of human interaction or the conversation that is being held between people in the movie. However, there are some ways to identify the category of a movie through the actions of people in a movie and the scenery, settings, and objects presented in the movie.

We will pick some categories that have obvious content that we could set them as keys for identifying the type of movie. The genres we picked are Action, Comedy, Drama, Horror, and Fantasy. Action movies contain actions like fighting and running and objects like speeding cars and firearms. This genre can include war genres because of the characteristics' similarities. Comedy movies have smiling and laughing emotions, also, some awkward situations like falling and some exaggerated facial expressions. Drama movies have serious facial expressions with a lot of conversations. And calm situations with less motion of the human body. Horror movies contain blood and knives and weird creatures with screaming and crying facial expressions. Fantasy movies contain colorful scenes, robots, and spaceships, which is a combination of sci-fi and adventure genres.

Although more genres that will identify the movies more accurately, we picked the most general genres that could include other genres in it. For example, War movies can be included under the action genre because it has lots of human body motion and lots of shooting gun objects. Another example is the Sci-Fi and Adventure movies that can be included under the Fantasy genre because all of them contain wide sceneries of places, colorful objects, modern equipment. Combining genres that have similar features under one genre makes the recognition for the movie genre easier.

In figure 2-1 from [4], shows the relationship between genres and how each genre could include sub-genres that would identify a movie more precisely. However, most of the relationships lead to one general genre; this is because of the similarities between subgenres. Circled genres in Figure 2-1 are what we used in our thesis. As shown in figure 2-1, sci-fi and adventure genre are grouped because they are similar in plot. However, the action genre is hard to be classified in its genre because it may occur in other genres as well. For example, action can be seen in drama, comedy, sci-fi, and many other genres, as shown in figure 2-1. In our case of study, we will consider the action genre as a sperate genre. The relationship between genres shown in the figure provides an understanding of the similarities between these genres. However, these relationships are considered related because of the plot, the characters, and the setup of the story. What we are considering in our study is the setup of the story using image analysis; that is why we selected the five genres that can be easily distinguished using image analysis.

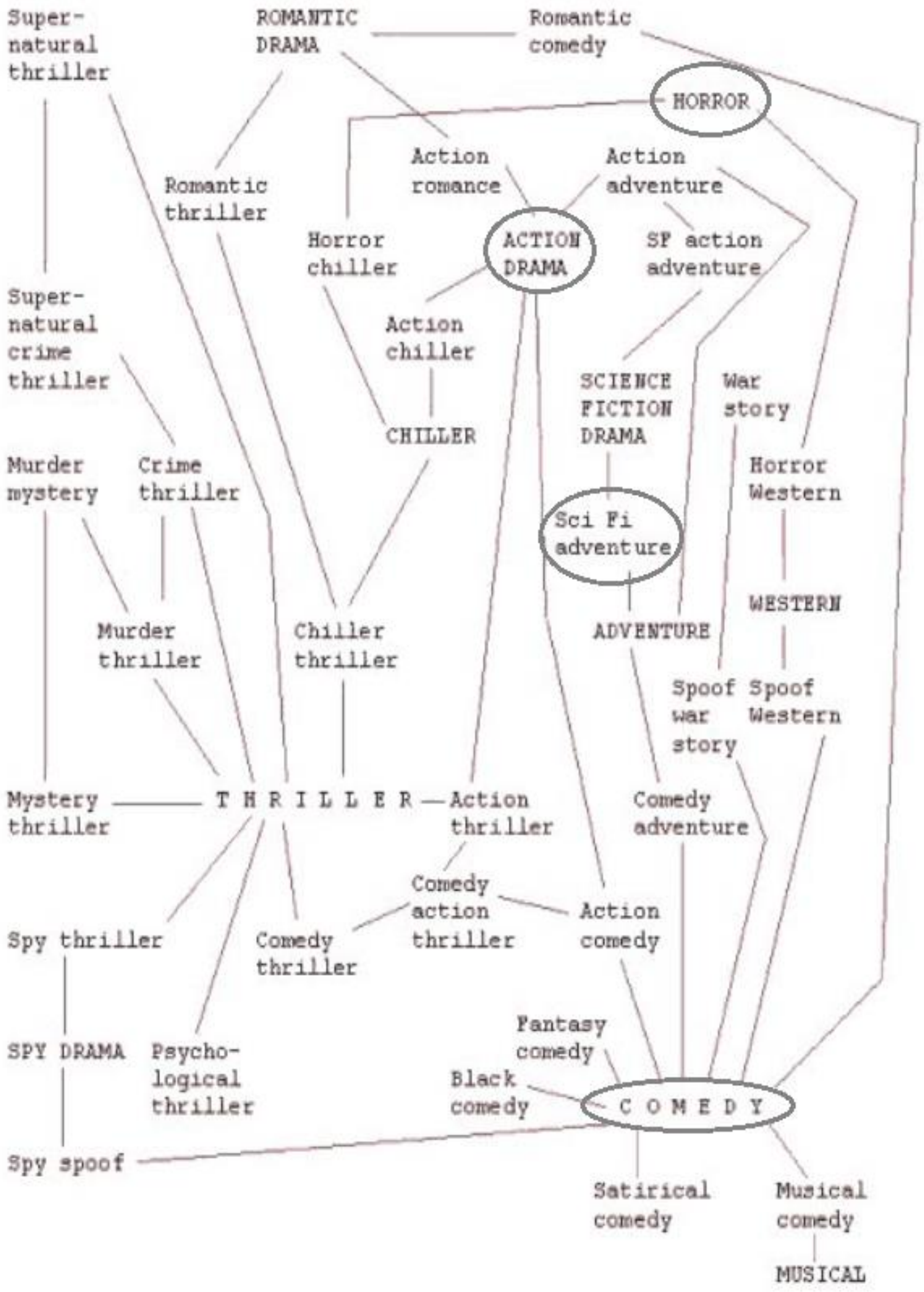

Figure 2-1: Relations between movie genres.

Because of the difficulties specifying certain human actions and human emotions, choosing most general categories is the way to help narrow down the classification. For example, romance and drama movies have almost the same features. What differentiates a romance movie from a general drama movie is the focus of the love story. The expression of love can be physically or during a conversation. Detecting the actions of people to determine if the movie contains a love story or not is possible. But it gets harder to identify romance when there is minimal to no actions in the movie. Thus, another kind of analysis that does not involve images should be done to help identify patterns. The focus of this project is to find patterns of human actions and objects that could help identify the most general genres.

2-2. Related Works

There are a lot of attempts to analyze videos and identify Human actions through a sequence of frames. One famous video dataset for human actions is UCF101 [5] they have labeled 13k clips of human actions into 101 categories. The categories grouped into Human-Object interaction, Body-motion only, Human-Human Interaction, Playing Musical instruments, and sports. This dataset helped researchers develop models that could recognize human actions. Figure 2-2 shows the number of clips in each classification.

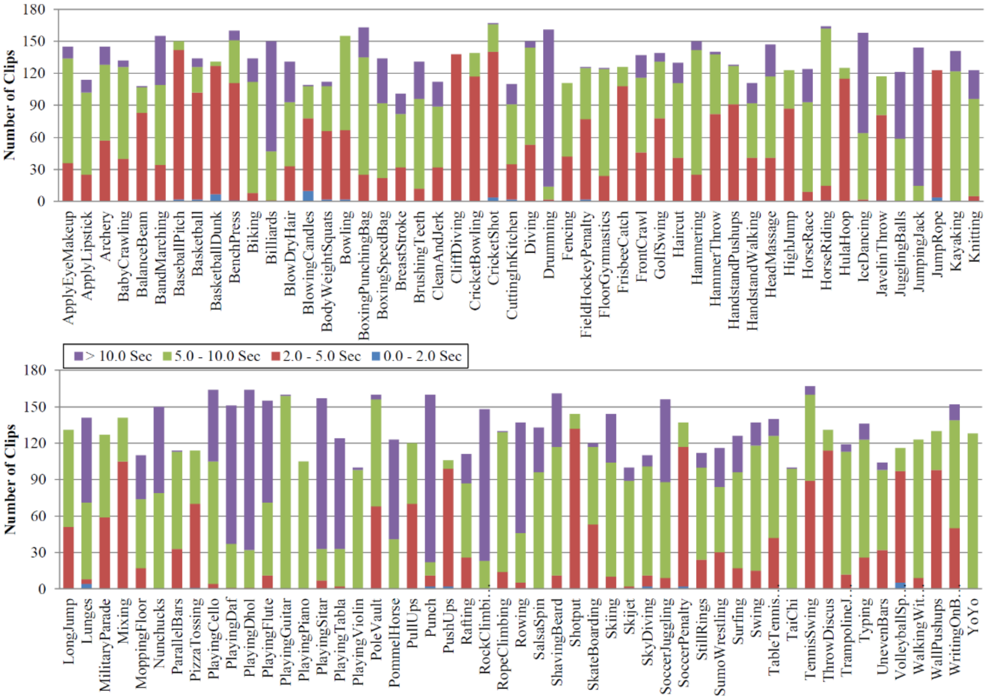

Figure 2-2: UCF dataset and number of instances for each Action.

One paper used the UCF101 dataset to develop a two-stream model for action recognition in videos [6]. They made use of the temporal component in the videos to provide extra information for recognition. The two-stream model is a state-of-the-art approach that could be used as one of the methods to help classify trailers into movie genres. The two-stream model was used for action recognition in UCF101 [5] and HMDB-51 [7] datasets. The two streams are fed to the same CNN architecture, but the two streams are different. One stream is concerned with the spatial information, which is frames that were extracted from videos. The other stream is concerned with the optical information, which is the optical flow displacement fields between consecutive frames. The second stream is useful for temporal recognition. Both streams are inputs to the model, then the output of the model from each stream is fused with an average layer. Figure 2-2 shows the architecture of the two-stream model.

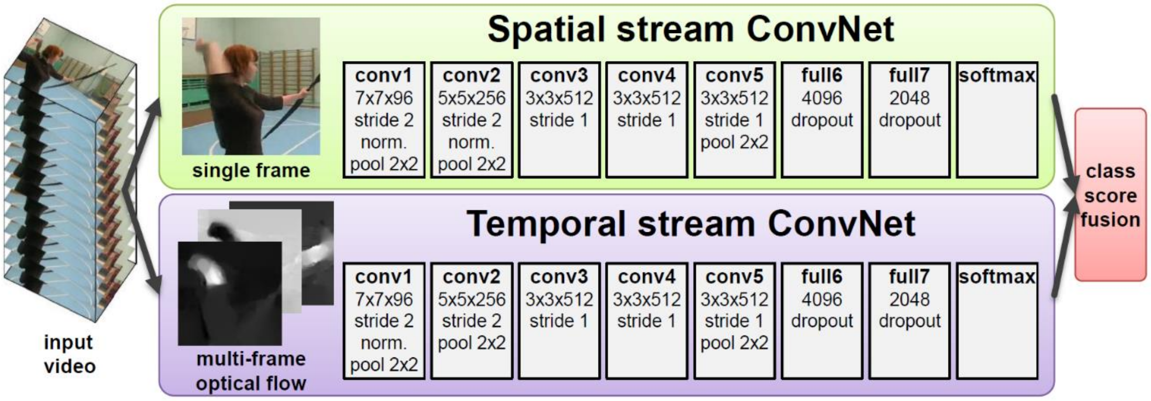

Figure 2-3: Two-stream architecture model.

An experiment on one million sports videos in [8] made use of CNN architecture to classify these videos into 487 categories. They extended the usual CNN model to get spatial and temporal information of video frames. Although it takes a long time to train on a large set of videos, they mitigated the issue by using two separate streams. One stream learns features on low-resolution frames, and the other stream is high-resolution frames, but by taking the middle part of the frame, this helped the architecture to speed up the training process for the architecture. The results of the experiment got 80% accuracy for the sports dataset. The approach was also tested on the UCF101 dataset but with 65% accuracy.

The idea of extracting human actions from movies based on script alignment and text classification was used in [9]. They got 397 action samples for eight classes of human actions. These actions were extracted using automatic annotation of human actions from 12 movies. For the classification process, they used a multi-scale approach to extract space-time features at multiple levels of Spatio-temporal scales. Characterizing motion and appearance of local features was done by computing histograms descriptors of space-time volumes in the neighborhood of detected points; these descriptors are similar to the SIFT descriptors [10]. These features are grouped in a bag-of-features and clustered using the k-means algorithm. And train a non-linear support vector machine for classifying the features into the corresponding labels. The methodology they used was tested on the KTH dataset for human action [11] and got great results of 91% accuracy.

The limitation of this approach is that human action classifications were limited to 8 categories only; this is because they used human actions that were extracted from a small set of movies. And not a lot of human actions were presented in these movies. Also, the annotations were not an efficient way of labeling all the human actions in the movies. They had misclassifications using the automatic annotation of human actions. Aligning the script of the movie with it is a good idea to automatically extract human actions with its label; however, the script does not describe all the human actions that happen in the movie.

Approaches to automatically classify movies into genres have been researched before. An approach by [12] which they used low-level features to analyze videos of movie previews. The low-level features were shot detection and average shot length, color variance, motion content, and lighting key. These four-dimensional feature spaces are combined into a joint domain representation and form a multivariant kernel. Then, a mean-shift clustering is used to analyze this feature space. They used the analysis and test on 101 movie previews with a result of 17 outliers out of the 101. They do not consider this result as the accuracy of the system, but it proves that there is a relationship between low-level features and high-level movie classes.

Another approach was presented by [13], which used high-level features of videos of movie trailers to classify them into genres. They decompose each trailer into a serious of shots using the shot boundary detection algorithm explained in [12]. Then the scene features from keyframes are analyzed using GIST [14], CENTRIST [15], and W-CENTRIST feature detectors and descriptors. W-CENTRIST is a variant of the CENTRIST descriptor that captures color information. After extracting features using the detectors and the descriptors, they used a K-mean clustering algorithm to get a bag-of-visual-words. Also, they built a

2D histogram to incorporate the temporal part of the trailer with the scene content of the trailer. Their experiment was done on 1239 movie trailers and classified the movie trailers into four general genres: action, comedy, drama, and horror. The average precision of classification using these methods got them 74% for the CENTRIST descriptor.

A more focused experiment on recognizing horror scenes from videos using a movie dataset was implemented by [16]. They used low-level features to analyze videos that are similar to [12]. The movie scene is divided into a series of shots which characterized by audio features, visual features, and emotional color features. These features are grouped into bags and fed to a support vector machine to recognize the horror movie scene. It got them 79% accuracy. This approach is limited to one category of recognizing genres, and it is similar to [12].

A recent experiment that adopted CNN to classify movies into its corresponding genres [17]. They collected a set of movie trailer videos and extracted frames and preprocessed them to fit the CNN model. Their model does not consider the temporal part of the frames. It treats each frame as a separate image. The model then is post-processed using a support vector machine using features: audio, frequency of elements in scene histogram, weighted predictions. The average precision for the genre classification is 62%.

The approaches mentioned above gave good results for movie genre classification. Most of them use low-level features to preprocess the videos and get the low-level information from them. They use these features to either do direct classification or to preprocess the videos before training them on machine learning models. However, these approaches do not use the action recognition feature, which could be useful information when analyzing movie scenes. Human emotions and actions contribute a lot in recognizing the genre of the movie, see section 2-1.

In the following chapter, we will explain the methods that will be used for classifying the movie trailers into its corresponding genres.

3. Methodology and Implementation

In this chapter, we will present a brief description of our dataset and how it has been collected, including the preprocessing methodologies used on them. Also, the methodologies used on analyzing the movie trailers and how to implement these methods on our dataset.

3-1. Dataset

The dataset used in our thesis is a set of movie trailers that are divided into five genre categories, as explained in the previous chapter. We collected 893 movie trailers from two YouTube channels that upload movie trailers and clips of movies. These trailers are pre-labeled by these channels; however, to make sure it is correctly labeled, we double-checked the accuracy of the labels from IMDB. The IMDB is considered one of the reliable resources for movie information. The exact number of videos for each genre category is shown in table 3-1.

Genre	**Number of videos**
Action	149
Comedy	150
Drama	273
Fantasy	159
Horror	162

Table 3-1: Number of videos in the dataset.

The video time length ranges between 1:30 – 2:30 (in minutes). The frame rate in the videos is 24 frames per second (fps), and the size of these frames is 1280 by 720 pixels. The total movie trailers are going to be split into training and testing sets. We are using the SciKit-learn library functions to split these data into its appropriate number for training and testing examples.

We have collected extra videos of movie trailers that are not assigned to a specific label because they are either a duplicate of videos that we already have in the dataset or clips for one scene only. These extra videos can be useful in testing the system and checking what kind of genre will be assigned to these videos.

3-2. Preprocessing the dataset

In the preprocessing stage, we are using OpenCV functions to read the videos and extract frames from them. The number of frames in each video will be different because of the different lengths of videos that we have in each genre. Since the size of data will be large if we extracted every frame in the video for all videos in the five genres categories, we are going to extract one frame every five frames in the video. And the same extraction will be done for all videos in all categories.

The frames that are extracted from the video are stacked in as a list of arrays. For example, if we got a video of length two minutes, which means it is 120 seconds long. Since the fps is 24, the number of frames in that video will be 120 × 24 = 2880 frames. And since we are getting one frame from every 5 frames in the 2880 frames, we will have 2880 / 5 = 576 frames for the video. These frames are resized to 512 by 512 pixels. The color of frames is kept to RGB channels. Thus, the shape of the video after going through the preprocessing will be (576, 512, 512, 3).

The preprocessing helps in reducing the size of the videos and excluding repetitive frames. The time it takes for the preprocessing for one video is approximately one minute. The total time it will take for preprocessing all videos is 14 hours. With parallel preprocessing, we reduced the time of computation to 4 hours. Due to the large size of the original frames, including the number of frames that each video contains, the computation of the preprocessing stage will take a long time to finish all the videos. After the whole selection of frames and resizing them, the frames will be stacked and saved on disk. This process also takes time to finish because of the large size of the stacked frames. The total size of the preprocessed data is approximately 400 GB.

3-3. Extracting features and clustering

Extracting the features in the videos is done by extracting the frames from the videos first because we want to deal with a single image instead of a stack of images. After getting all the frames from the videos, each one in its category, we perform the feature extractions. The process of extracting the features from these frames is by recognizing certain objects in the images using a pre-trained model. After getting the classification of each frame, we can see the top common objects and information in these frames. Getting this information will help classify the next frames.

For example, we have a set of frames that are extracted from videos in the Action category. By feeding part of the frames to a pre-trained model on ImageNet, we got torch and fire and lights. This information tells us that these objects are unique in the Action category, which we can make use of it when we want to classify frames that has similar objects.

We can see what other information is identified in these frames by checking the resulted prediction with the actual label and see if it matches the concepts of identifying a genre by people. It makes sense that the fire and torch are characteristics of an Action movie because Action movies, as explained in Chapter 2, has fighting scenes and firearms. So, we can say that the information we got from these classifications can relate to our understanding of how we, as humans, identify these categories.

Other features that can be detected by machines that are considered one of the characteristics of genres are facial emotions. The facial expressions in humans that are shown in the movies determine the type of movie genre. For example, happy faces are usually found in comedy genres. And sad faces are found in horror genres. We will see the effectiveness of using emotion recognition in identifying movie genres.

We used two models for extracting features from the frames. The first one is the object recognition model. The second one is the emotion recognition model. Both methods will be explained in the next sections. The evaluation of these models is discussed in chapter 4.

3-3-1. Extracting features using object detection model

One methodology is using an object detection model on the frames to detect multiple objects in a frame instead of classifying the whole frame into one of the classifications proposed by ImageNet. The model we are using is YOLO [18] with pre-trained weights on the COCO image dataset [19]. The model design, as proposed by the authors of the paper, is found in figure 3-1.

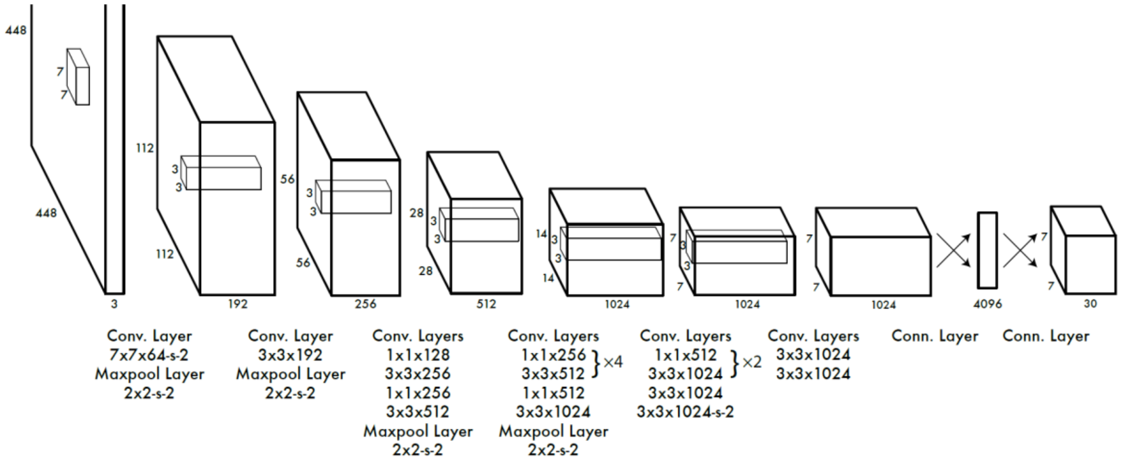

Figure 3-1: YOLO structure model.

The model is trained on annotated images of 80 objects. Their location coordinates are used as a bounding box. The accuracy of the pre-trained weights is 80%. These weights were used to detect the objects that are present in each frame of the videos.

The proposed model divides the image into an S × S grid, where S is the number of grids in each axis. If the center of an object falls into a grid cell, that grid cell is responsible for detecting that image. Each grid predicts bounding box and confidence scores for those boxes. Each bounding box consists of 5 predictions: x, y, w, h, and confidence. The model has limitations in detecting small objects that appear in groups and generalizing to object with new aspect ratios.

In figure 3-2, an example of frames taken from a video trailer and bounding boxes with their confidence percentages.

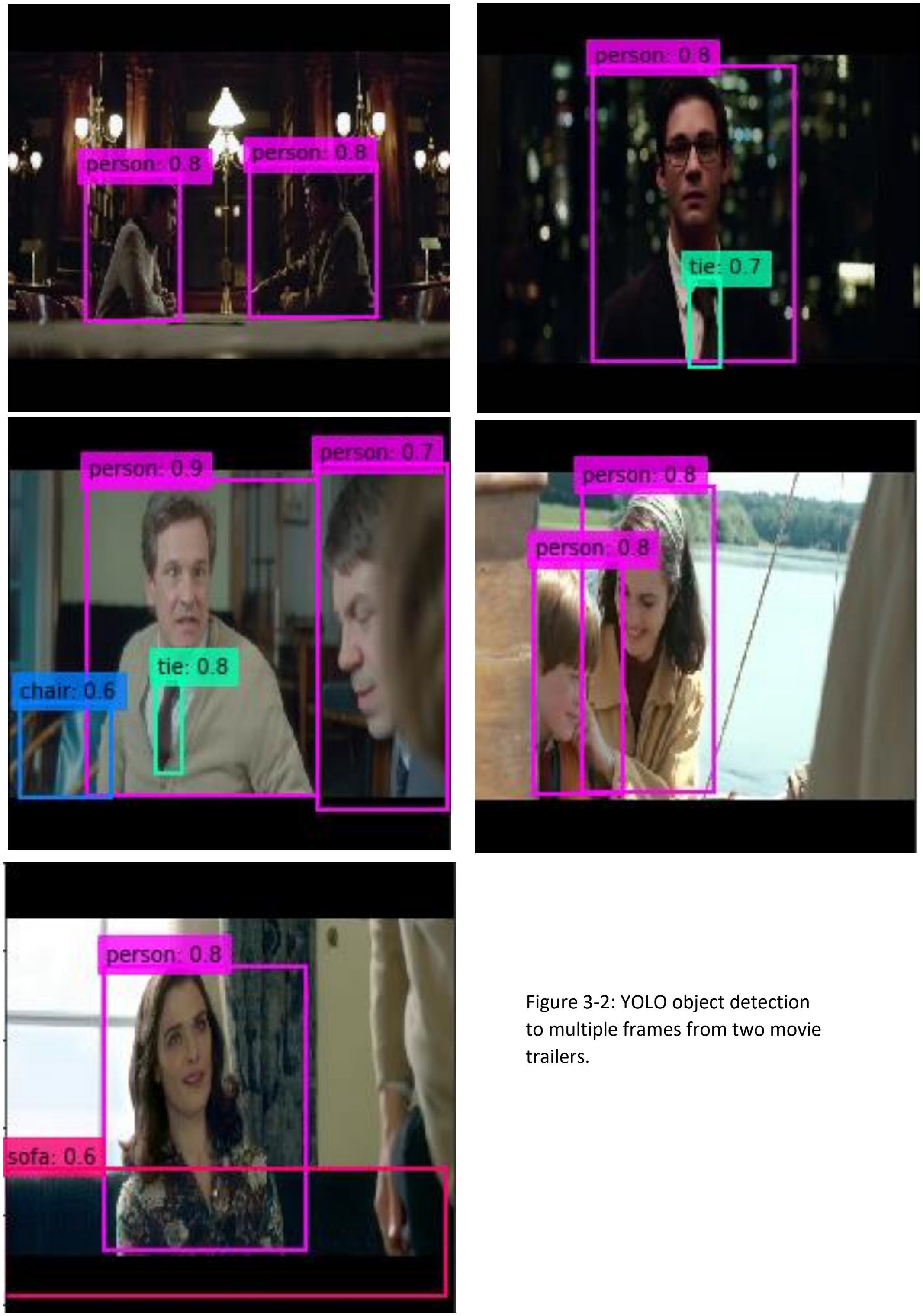

Figure 3-2: YOLO object detection to multiple frames from two movie trailers.

3-3-2. Extracting features using emotion recognition model

Since the focus in movies is people, human emotions can be a useful indicator to recognize patterns among videos of different genres. After analyzing human emotions and check the frequencies of it in the videos, we will see if this a good metric in distinguishing genres.

To detect people's emotions in video trailers, we need to detect faces that appear in the video and then classify the human emotion into one of the seven emotion categories proposed by [20]. Emotion classes are angry, disgust, fear, happy, sad, surprise, and neutral.

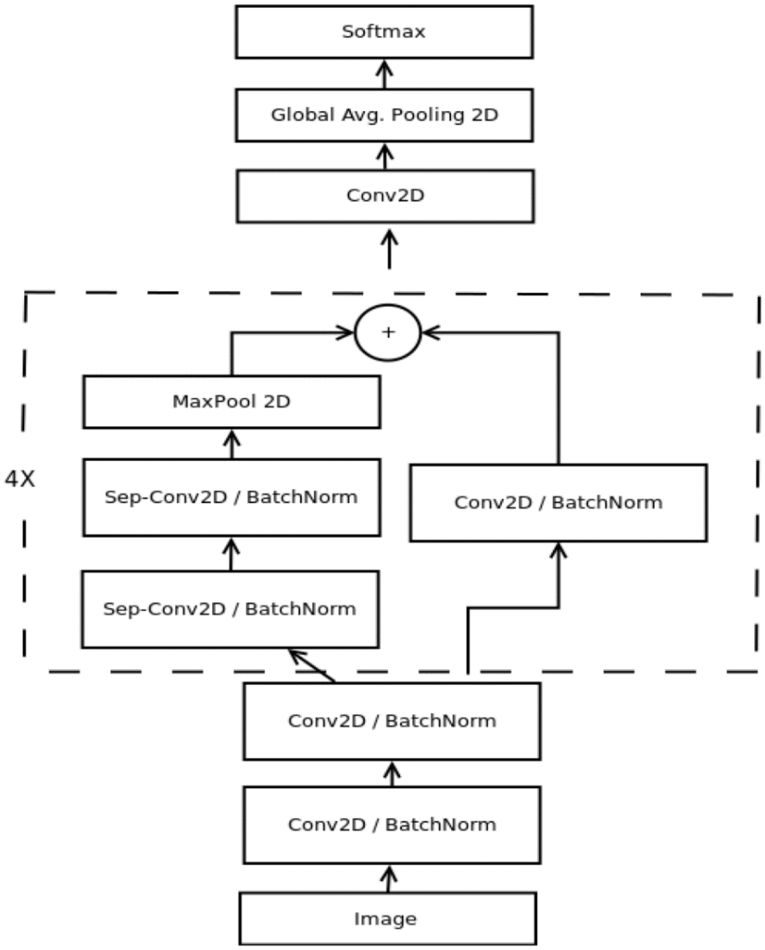

Figure 3-3: Emotion classification model structure.

In figure 3-4, we show examples of the emotion recognition model that we used on all movie trailers frames.

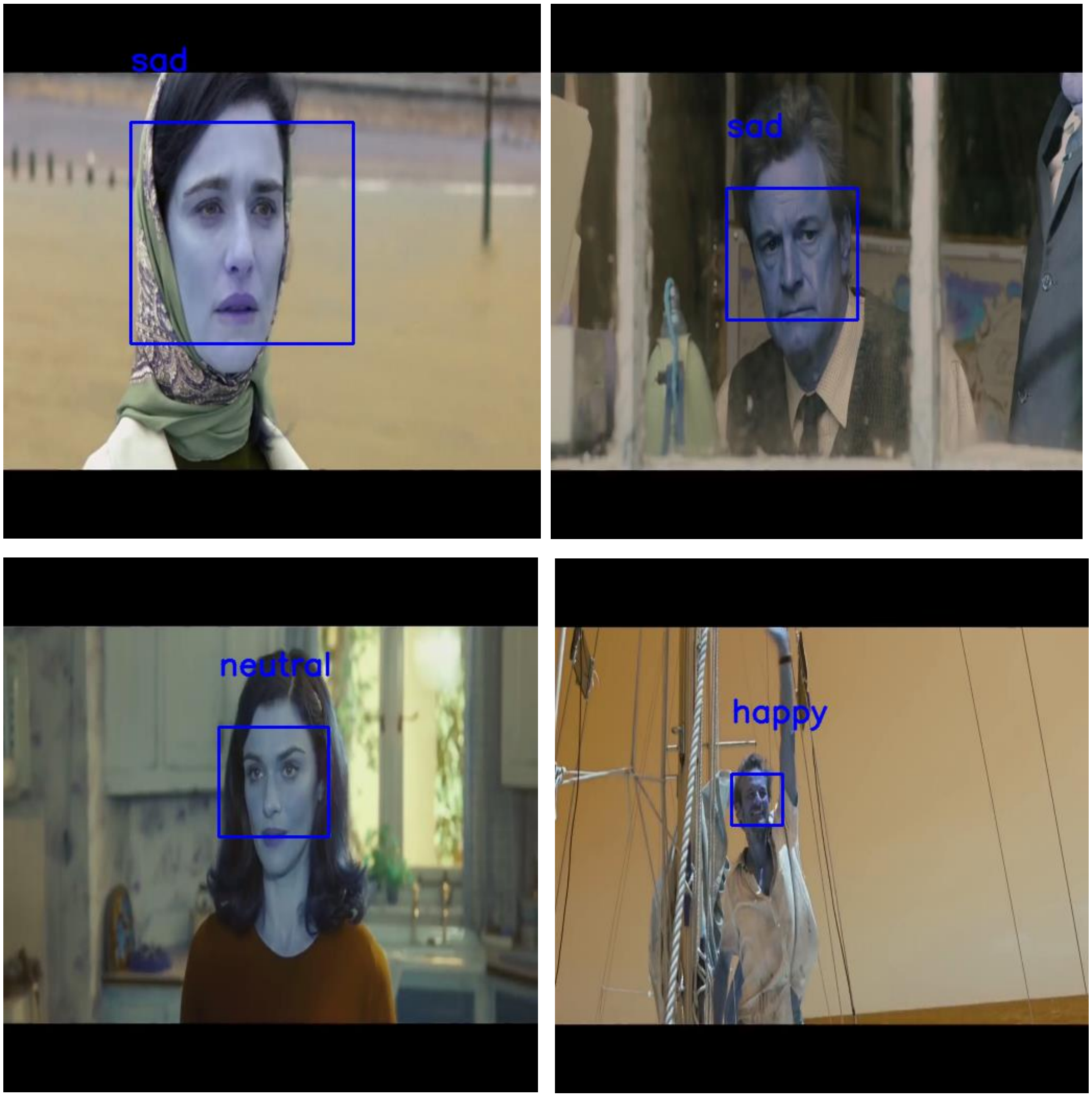

Figure 3-4: Emotion classification examples on frames from the movie trailer dataset

3-2. Structure of the overall process

After explaining all the analysis needed to construct the system for predicting the movie trailer genres. We will combine all the outputs we got from the analysis and input them to a classifier to predict the genres of a movie trailer. The structure of the process is shown in figure 3-3.

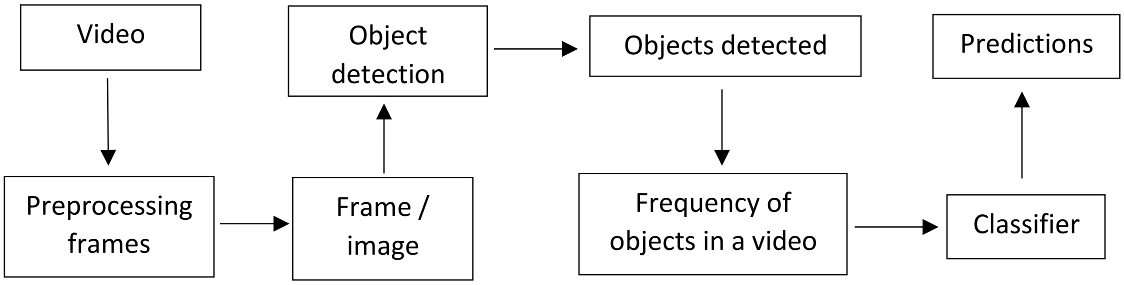

Figure 3-5: System structure and process.

In each step of the process, we tested the results if it is good enough to move to the next step. For example, in the preprocessing step, the first attempt we resized the frames to 226 by 226 and the second attempt, we resized the frames to 512 by 512. The second setup of the frames got us better object detections results than the first setup. In detail, when we resized action genre frames to 226 by 226 pixels, the YOLO model detected cakes in some of the frames; however, when we resized the frames to 512 by 512 pixels, the model detected fewer cake objects.

After running the YOLO model on all videos, we constructed a script that will collect all the objects detected by the model and built a frequency table for it. In chapter four, we will show frequency histograms that we got for each genre. The frequencies will be the features of the classifiers that we will use. We could always modify these features to include more objects or exclude some features or objects that are not significant to the classifier. The next step after getting all these frequencies, we want to input them to a classifier for finding matching patterns between videos in terms of objects. These patterns will help the classifier to identify the genre of a video if we give it its object frequencies. The classifiers that we explored were regular neural nets and random forests.

4. Results and Evaluation

In this section, I will present the results of my findings and how I evaluated them. The first part of the findings are object detection results, and the second part is the faces emotion analysis results.

4-2. Object detection results

As explained in section three, the object detection model is based on the YOLO paper [18]. After doing the preprocessing part of the videos, the frames were stacked in matrices format. The matrices are input to the YOLO model. The model weights used for the detection process were pre-trained weights on the COCO dataset [19]. The dataset has 80 classes with annotated images for objects. The YOLO model was trained and tested on the PASCAL VOC dataset, which contains 20 objects [21]. Figure 4-1 shows the objects included in the COCO dataset and how many instances of images are in each class compared to the PASCAL VOC dataset.

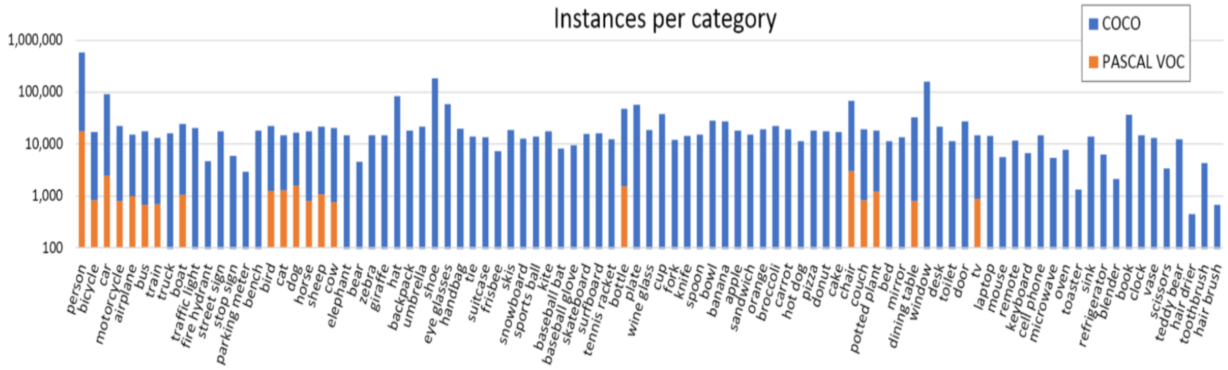

Figure 4-1: COCO dataset compared to PASCAL VOC dataset.

The results of an input of a stack of frames are translated to a frequency of objects that are detected in the frames. For example, Table 4-1 shows the detected objects with their frequencies for a video trailer in the Action category.

Table 4-1: example of objects and its frequencies in a video.

Objects	**Frequencies**
Person	982
Bicycle	14
Car	57
Airplane	27
Truck	13
Chair	26

For analysis purposes, we computed the average of the frequencies of each object for each genre. The low-frequency objects are excluded from the list of objects that differentiates a category. To exclude the low frequencies, we took the sum of the average frequencies across all genre classes. If the sum of the average frequencies is lower than the predefined threshold, then the object will be excluded. Figure 4-2 shows the frequencies for each genre.

As seen in Figure 4-2, the distribution of object frequencies is different for each genre. However, there are similarities in the distribution for the action and fantasy genre, Figure 4-2 (a, d). A similarity between the drama and comedy histogram distribution can also be seen in Figure 4-2 (b, c). The horror genre was different from all other genres. The analysis was done on all 893 video trailers, and it was done to show if there are differences between the genres. Once we found that the genres can be distinguished using the object frequencies, we can move to the next step by inputting the frequencies to a classifier. The more features we have for the classifier, the better. Features in our case of the dataset are the objects detected by the model.

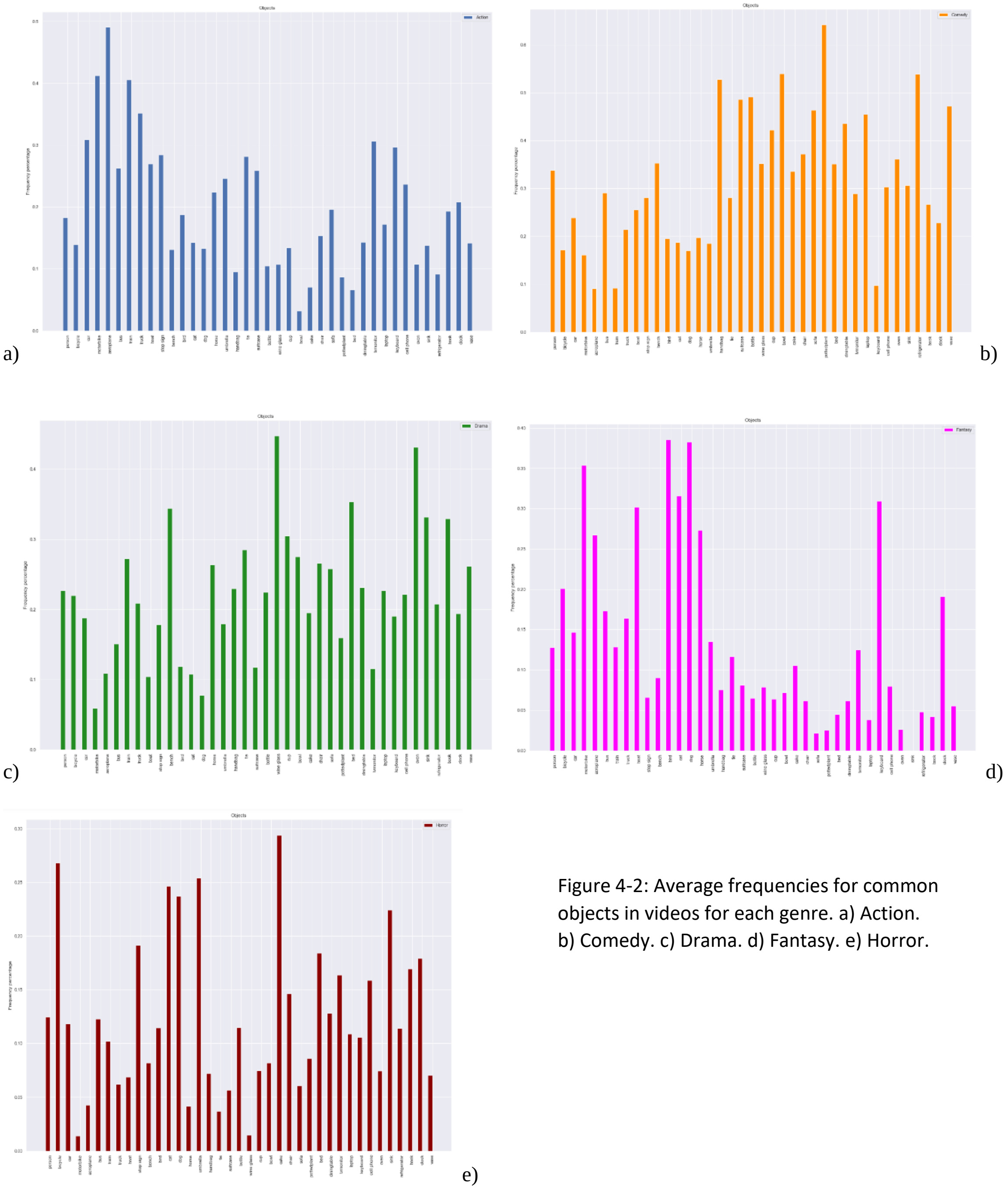

Figure 4-2: Average frequencies for common objects in videos for each genre. a) Action. b) Comedy. c) Drama. d) Fantasy. e) Horror.

The next part of these results is to use them for training the classifier and see if it can predict correctly. Two classifiers were used, neural nets, and random forests. The dataset was divided into 669 trailers for training and 224 trailers for testing.

The result of the neural nets was 44% average precision. As shown in figure 4-3 and table 4-2, the precision percentage for the Fantasy category was 60 percent, and Action comes next with 51 percent accuracy predictions. Action and comedy categories were not high enough compared to other genres.

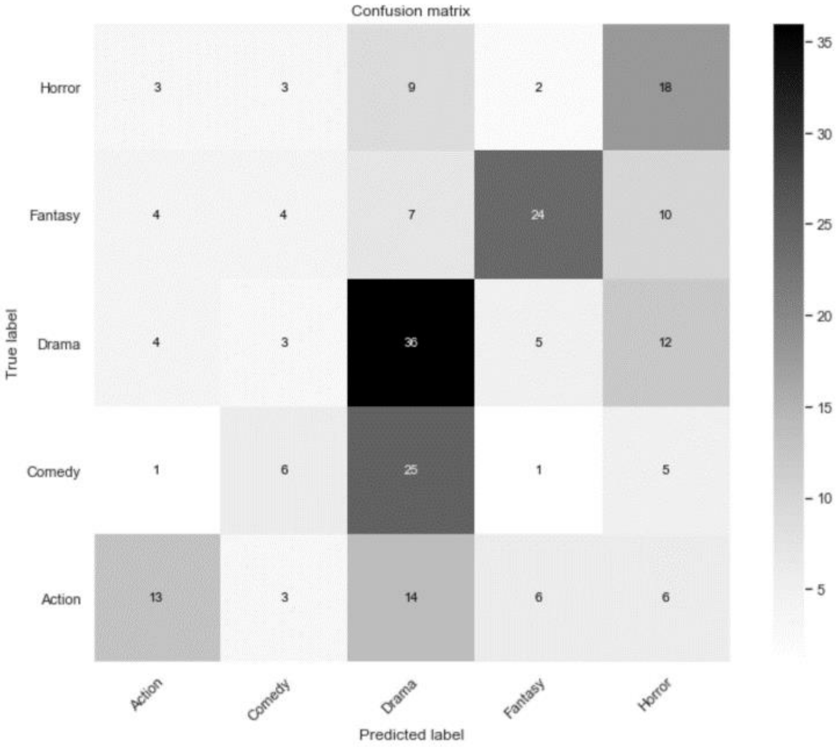

Figure 4-3: Confusion matrix for neural nets classifier.

Genre	Precision	Recall
Action	52%	31%
Comedy	32%	16%
Drama	40%	60%
Fantasy	63%	49%
Horror	35%	51%

Table 4-2: precision and recall for neural nets classifier.

The next classifier that used was a random forest classifier. The results were slightly better than the neural net classifier with an average precision of 52%. The confusion matrix of the random forest classifier is shown in figure 4-4. Also, the precision and recall percentages are shown in table 4-3. We got higher precision percentages in the drama, fantasy, and comedy. However, the horror genre got a lower percentage, and the action genre did not change.

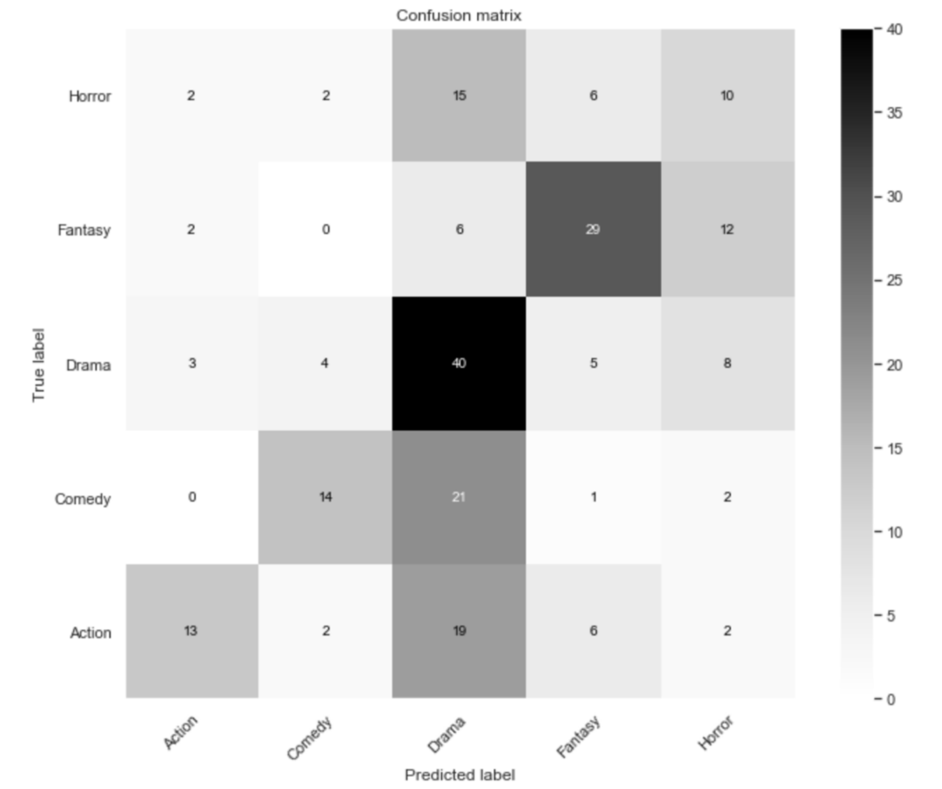

Figure 4-4: Confusion matrix for random forest classifier.

Genre	Precision	Recall
Action	65%	31%
Comedy	64%	37%
Drama	40%	67%
Fantasy	62%	59%
Horror	29%	29%

Table 4-3: precision and recall for random forest classifier.

High precision indicates that the classifier correctly identifies the relevant instances as true and false positives. High recall indicates that the classifier identifies all relevant instances. What we see in the precision and recall percentages of the random forest and neural net classifiers is that the random forest classifier got improved when it was trained on the training dataset; also it improved the overall accuracy of the predictions. The YOLO model using pre-trained weights on the COCO dataset got us this result of object frequencies. Although the overall accuracy is not high, the system can identify a couple of genres with 60% accuracy.

The results for the emotion recognition model was not helpful in this case. Because of the low accuracy of the used model, it could not identify most of the facial expressions accurately or recognized facial expressions to non-human objects. We tried to train the classifiers on the frequencies we got for the seven-facial expression, but we got very low accuracy compared to the object frequencies. We believe that low accuracy was due to the wrong predictions of facial emotion recognition. The wrong results were seen often in the predictions of the facial emotion recognition model but not in the object detection model. One of the reasons is the low accuracy of the emotion prediction for the model with 66% compared to 80% accuracy for the object detection model. The other reason, the frames may need to be bigger to correctly predict the emotion 66% percent of the time. We believe that the size of the frame is a possible reason for not predicting correctly because we face this issue in the object detection model that we discussed in section 3-2. We were limited to resizing the frames to 512 by 512 because of the huge amount of disk size that it will take if we resized the frames to a higher dimension.

In figure 4-5, we show an example of facial recognition with wrong predictions of facial emotion expression. Figure 4-5 (a, b) we got a lot of misclassified emotion for the same frame. The frames are not the same, but because it is a sequence of frames, you may have the same objects with slightly different positions of them in the next frame. Figure 4-5 (c), we have a frame of a crowd of people with no facial detection but classified into an emotion.

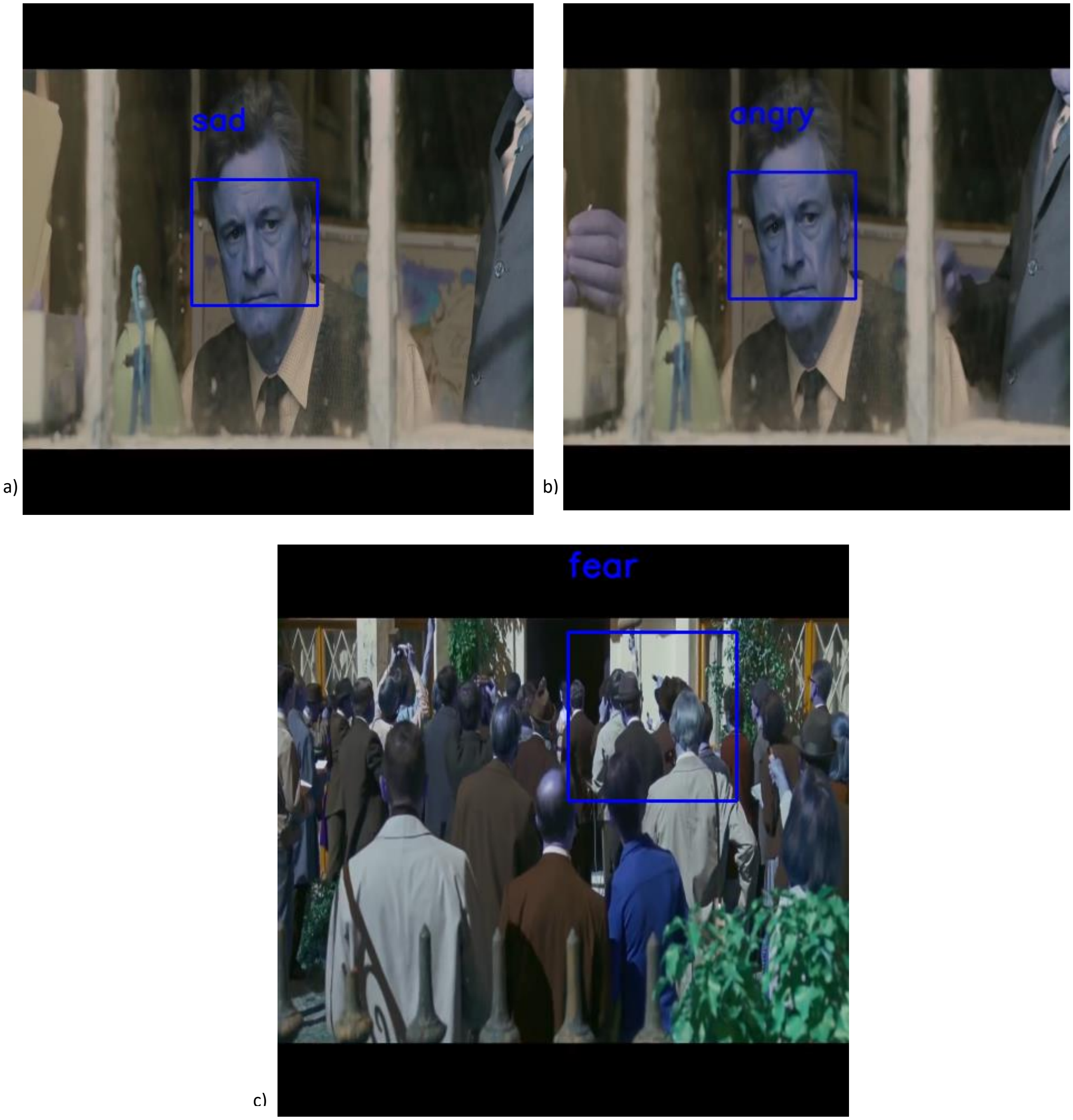

Figure 4-5: examples of misclassification of facial emotions.

5. Conclusion

The process of classifying a movie trailer is long using the proposed approach, but it gave deferential results compared to other approaches used for genre classification. Although the results were not significant, the idea of getting the trailers classified based on the frequency of objects is worth exploring and researching. We got good results for drama, fantasy, and comedies, but the object detection model needs to be improved for the action and horror genres. Although emotion recognition did not give good results, we still believe that it could be significant features to the overall flow of the system if we improved the prediction of the model.

One of the improvements that could help us get better results is increasing our dataset to more than 900 videos. We got higher accuracy in drama classification because we had more training set for this category compared to the other categories. Thus, having more data in our set will help increase the accuracy of the predictions. Also, the more feature we include, the better the classifier enhances its weights. We trained the classifier on 49 features; in other words, 49 object frequencies, but did not achieve more than 30% accuracy overall. However, when we included all the 80 features, the classifier gave us the results that we presented in chapter 4.

For further research, the YOLO model will be trained on custom objects. We will be specific on objects that distinguish the genres from each other — for example, training the model on firearms and other types of vehicles to differentiate action and fantasy movie genres from the rest. Human action recognition is another method that could help increase the accuracy of genre classification. Recognizing if a person is running or fighting or walking, etc. Analyzing the frequency of these actions appearing in multiple videos in a genre is a good metric for evaluating patterns in a genre.

In the end, our approach results were not the best in the field. However, it got us good results with pre-trained models, and we believe that we could achieve more if we implemented enhancements to the models used. That is why we can say that the approach itself is worth exploring and developing.

References

[1] IMDb, “IMDb Statistics,” 2015. [Online]. Available: http://www.imdb.com/stats.

[2] D. Chandler, “An Introduction to Genre Theory,” *http://www.aber.ac.uk/media/Documents/intgenre/chandler_genre_theory.pdf*, pp. 1–15, 2004.

[3] T. Dirks, “Main Film Genres.” [Online]. Available: https://www.filmsite.org/genres.html.

[4] D. Chandler, “An Introduction to Genre Theory,” *http://www.aber.ac.uk/media/Documents/intgenre/chandler_genre_theory.pdf*, no. January 1997, pp. 1–15, 2004.

[5] K. Soomro, A. R. Zamir, and M. Shah, “UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild,” no. November, 2012.

[6] K. Simonyan and A. Zisserman, “Two-stream convolutional networks for action recognition in videos,” *Adv. Neural Inf. Process. Syst.*, vol. 1, no. January, pp. 568–576, 2014.

[7] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, “HMDB: A large video database for human motion recognition,” *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 2556–2563, 2011.

[8] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and F. F. Li, “Large-scale video classification with convolutional neural networks,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 1725–1732, 2014.

[9] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld, “Learning realistic human actions from movies,” *26th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR*, 2008.

[10] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[11] C. Schuldt, L. Barbara, and S.- Stockholm, “Recognizing Human Actions : A Local SVM

Approach ∗ Dept . of Numerical Analysis and Computer Science,” *Pattern Recognition, 2004. ICPR 2004. Proc. 17th Int. Conf.*, vol. 3, pp. 32–36, 2004.

[12] Z. Rasheed, Y. Sheikh, and M. Shah, “On the use of computable features for film classification,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 1, pp. 52–63, 2005.

[13] H. Zhou, T. Hermans, A. V. Karandikar, and J. M. Rehg, “Movie genre classification via scene categorization,” *MM’10 - Proc. ACM Multimed. 2010 Int. Conf.*, pp. 747–750, 2010.

[14] A. Oliva and A. Torralba, “Modeling the shape of the scene: A holistic representation of the spatial envelope,” *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, 2001.

[15] J. Wu and J. R. W. A. I, “Place instance and category recognition using spatial PACT,” *Cvpr*, 2008.

[16] J. Wang, B. Li, W. Hu, and O. Wu, “Horror video scene recognition via multiple-instance learning,” *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, no. January, pp. 1325–1328, 2011.

[17] G. S. Simões, J. Wehrmann, R. C. Barros, and D. D. Ruiz, “Movie genre classification with Convolutional Neural Networks,” *Proc. Int. Jt. Conf. Neural Networks*, vol. 2016-Octob, no. April 2018, pp. 259–266, 2016.

[18] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-Decem, pp. 779–788, 2016.

[19] T. Y. Lin *et al.*, “Microsoft COCO: Common objects in context,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 8693 LNCS, no. PART 5, pp. 740–755, 2014.

[20] O. Arriaga, M. Valdenegro-Toro, and P. G. Plöger, “Real-time convolutional neural networks for

emotion and gender classification," *ESANN 2019 - Proceedings, 27th Eur. Symp. Artif. Neural Networks, Comput. Intell. Mach. Learn.*, pp. 221–226, 2019.

[21] A. Everingham, M. and Van~Gool, L. and Williams, C. K. I. and Winn, J. and Zisserman, "The PASCAL Visual Object Classes Challenge 2012 VOC2012 Results," 2012. [Online]. Available: http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.