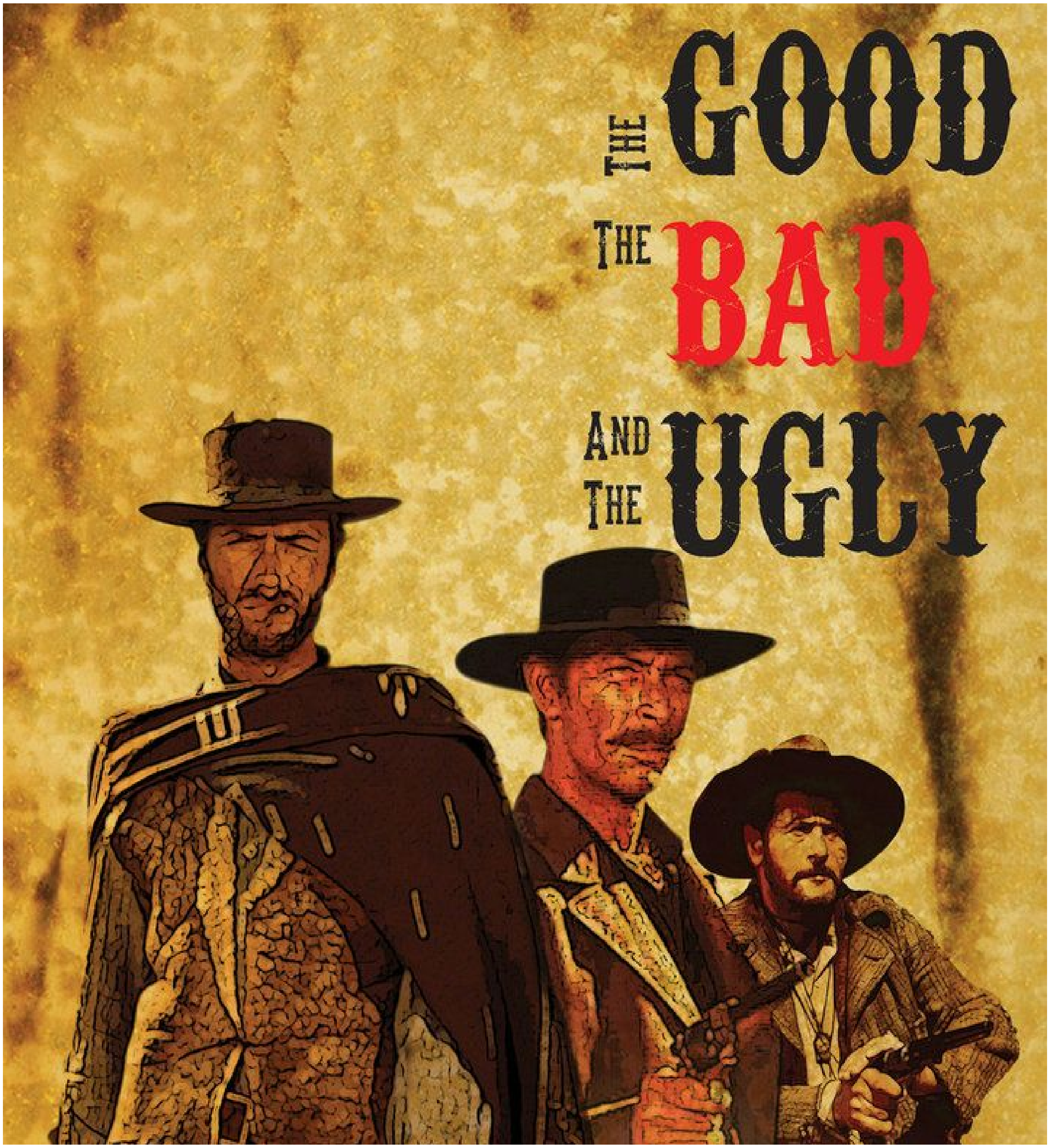
General Assembly Data Science, October-December 2015

Predicting Genre from Features of Film Posters

Job Lindsen



Introduction

Film posters occupy a space between art and advertising. Although many good film posters are designed beautifully, they always have a clear commercial purpose - to promote a film. To do so, the target audience must be able to understand quickly what the movie is about. One of the most important unique selling points of many commercial movies is a cast of well-known movie stars. Logically, they are usually given a prominent place on the film poster. In addition, the movie poster usually conveys other aspects of the film by means of its graphic design.

One of the main dimensions of design that can be used to create an atmosphere is the colour space. Some important aspects of the colour space are the choice of colours, the saturation of the colours, and the brightness of the colours. There is an intuitive mapping between the intense and dark atmosphere of, for example, Sci-Fi and Horror movies and darker colours. On the other hand, the light and funny atmosphere of a Comedy or Family movie is generally associated with saturated and bright colours.

In this paper the relationship between low-level properties, such as colour and amount spatial detail, and genre is investigated. These low-level features will be extracted from film posters, and subsequently be used to predict the film's genre. If these low-level features differ systematically between film genres, the accuracy of the prediction should be above chance level.

Hypothesis:

The design of film posters is stereotypical, so that on average the genre of a film can be predicted, with above chance level, from low-level properties of its film poster.

Methods

Data Set

The first step in constructing the data set was to identify films in each of the major film genres. IMDB.com allows you to order films based on popularity, origin, release year, seperately for each genre. After a quick survey of the number of movies in each genre, the following 7 were selected: Horror, Romance, Comedy, Sci-Fi, Thriller, Action, and Family. Furthermore, the period in which the film was released was restricted to the years 2000-2014, to limit the influence of any trends in film poster design. A web scraper written in Python using the *Beautiful Soup* library downloaded the film information for the 1000 most popular movies in each of the genres. The film poster images were downloaded in jpg-format and were all approximately 800 by 1200 pixels.

Movies can be classified in multiple genres, and therefore there was some overlap with the same titles occurring in the top 1000 list of multiple genres. These duplicates were removed, resulting in a lowest count of movies in a single genre of just over 600. To have the same number of observations in each class, I limited the number of movies in each genre to the top-600. The overall number of rows was therefore 4200.

The next step was to extract features from the image files. The Hue-Saturation-Value (HSV) representation of RGB color images allows for a relatively easy way of constructing useful features. Saturation is the intensity of a color, and the overall saturation of the image was summarized in one feature by calculating the mean saturation over all pixels. Similarly, the overall Value, i.e. the brightness of a color, was summarized in one feature as the mean Value over all pixels. Contrast was also calculated as a feature by taking the difference in brightness between the 25% brightest and 25% darkest pixels.

Hue is a rainbow-like spectrum of all unique colours ranging from red, yellow, green, cyan, blue, magenta, back to red again. The hue values were grouped in six bins and the percentage of pixels in each bin was calculated. This resulted in six colour-related features (% red-yellow pixels, % yellow-green pixels, % green-cyan pixels, % cyan-blue pixels, % blue-magenta pixels, % magenta-red pixels). The hue of pixels with very high or low brightness, i.e. almost white or black pixels, can be unstable. Therefore the percentage of (near) white and (near) black pixels was calculated first, and the remaining pixels were labeled as colour pixels. The colour-related features were taken as percentages of the number of colour pixels. The percentages of black and white pixels were also taken as separate features.

In addition to these features, luminance, calculated as L = 0.2126 Red + 0.7152 Green + 0.0722 Blue, based on the RGB representation of the image was included. Luminance takes into account the different contributions of the colours to perceived brightness. Furthermore, to quantify the amount of detail in an image the spatial frequency spectrum of the grayscale version of the images was computed. Magnitude on the log-transformed frequencies was computed to obtain a relatively symmetric distribution of frequency content, and the sum of the low spatial frequency magnitudes and the sum of the high frequency magnitudes were taken as two separate features.

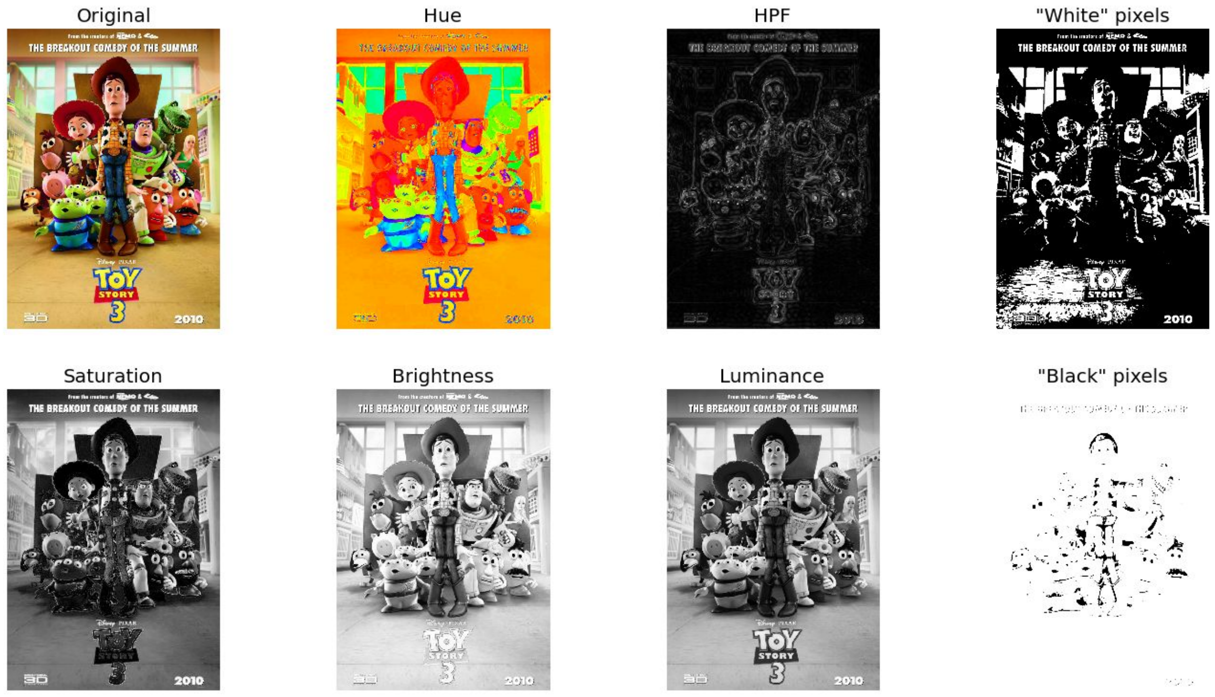


Figure 1, Illustration of some features extracted using the Toy Story 3 poster. HPF = high-pass frequency filter.

The overall number of features was 14. Figure 1 illustrates some of the features described above using the film poster of Toy Story 3 as an example. The three sub-plots Hue, Saturation, and Brightness are graphical representations of the layers in the HSV representation of the image, and together contain all the information in the original. The High-Pass Filtered (HPF) subplot shows the grayscale version of the image with all low spatial frequencies removed. The subplots on the right illustrate which pixels were classified as (near) white and (near) black.

Analysis of the features

Since the features were all measured on quite different scales, they were normalized first. Boxplots of the features, separately for the different genres (see Figure 1), showed some clear differences between the genres in the features related to the the brightness of the image. Film posters in the genres horror, action, and thriller were darker than posters from the other genres. However, by just eyeballing the boxplots it became evident that 4 features (%-white, %-black, mean brightness, and mean luminance were strongly correlated.

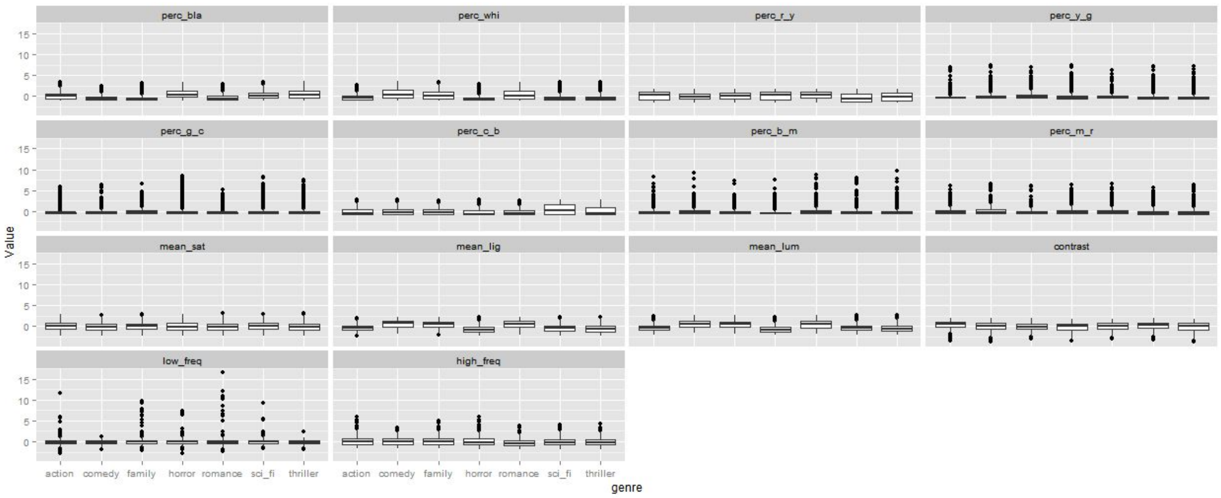


Figure 2, Boxplots of Normalized Features. Perc = Percentage, bla = black, whi = white, r = red, y = yellow, g = green, c = cyan, b = blue, m = magenta, sat = saturation, lig = lightness, lum = luminance, freq = frequency.

A PCA was applied to the features to investigate this further, and to potentially remove redundant information from the feature space by excluding some principal components. The PCA confirmed that there was strong collinearity between four features. As can be seen in Figure 4, the cumulative sum of the proportion variance explained by each component shows that the last three components carry only minimal unique information. Based on this, only the 11 first principal components were used in the modelling stage instead of the 14 original features. The amount of variance explained by these 11 components was 99.5%.

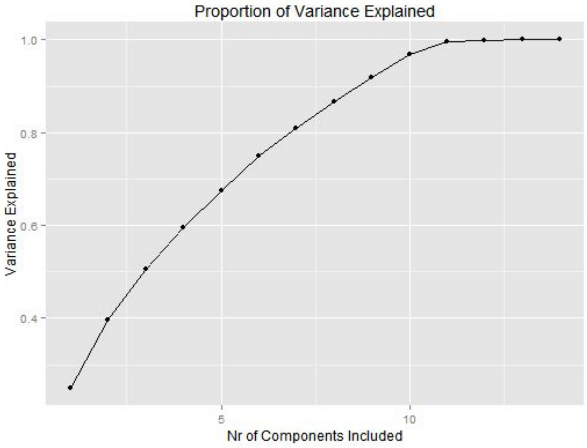


Figure 3, Cumulative Proportion of Variance Explained. The plot shows that 3 factors can be omitted with only a minimal loss in total explained variance.

Model Selection and Prediction

6 Different models were used to classify the genres: Decision Trees, Naive Bayes, ADABoost, K-Nearest-Neighbours, Random Forest, and a Support Vector Classifier. All selection and predicting of the models was done using Python and the *scikit-learn* package. To avoid overfitting the data, a 10-fold cross-validation procedure was followed. This procedure divides the data set in 10 different subsets, trains the model on 9 subsets combined and tests against the remaining subset, repeating this 10 times so that each subset is used once as the test set. For each of the 10 cross-validation steps an accuracy score was calculated, and the overall prediction accuracy of a model was taken as the average over the 10 steps. Hyperparameter optimization was performed by manually tweaking the parameters until the accuracy was maximal. Figure 4 summarizes and compares the classification accuracy of the 6 models after optimization.

Decision Trees:

The Decision Tree classifier performed optimally using the 'Gini' purity measure, maximum tree depth of 6 and a minimum number of samples per leaf of 3. The mean accuracy over the 10 cross-validations was 26.5%, with a standard deviation of 2.9.

Naive Bayes:

The Gaussian Naive Bayes classifier was applied without specifying any further parameters and gave a mean accuracy of 27.4%, and standard deviation of 2.3.

ADA Boost:

The ADABoost classifier was used using the best fitting Decision Tree classifier specified above as its base classifier, and gave optimal accuracy values the number of estimators set to 200 and the learning rate to 0.1. The mean accuracy of the model was 28.7%, with a standard deviation of 2.0.

K-Nearest-Neighbours:

The K-Nearest Neighbours model performed best with the number of neighbours set to 100 and using uniform weights, meaning that all points in each neighborhood are weighted equally. The average accuracy obtained with this model was 29.5%, with a standard deviation of 3.3.

Random Forest:

The Random Forest classifier performed best with the number of estimators set to 50, and the other parameters set to the values of the best performing Decision Tree classifier tree specified above. The average accuracy obtained with this model was 30.5%, with a standard deviation of 3.0.

Support Vector Classifier:

The Support Vector Classifier performed best with a Gaussian kernel function, a kernel coefficient gamma of 0.001, and the penalty parameter C of the error term set to 750. The average accuracy obtained with this model was 30.9%, with a standard deviation of 3.1.

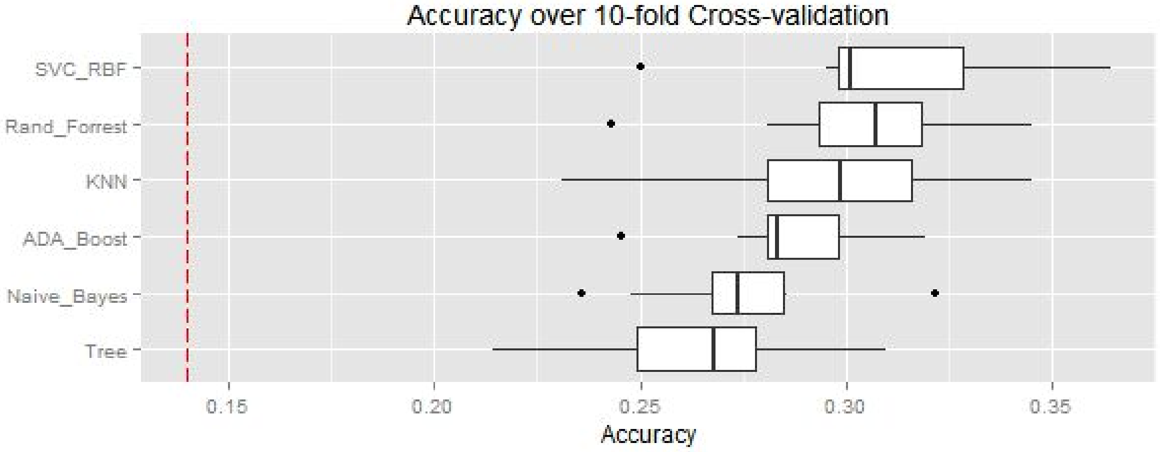


Figure 4, Boxplots of the Results of the 10-fold Cross-Validation, separately for the six models tested. The vertical red line indicates chance level (1/7 = .14). SVC_RBF = Support Vector Classifier.

Interpretation of the Classification Results

To be able to evaluate the classification performance of the best performing model, the Support Vector classifier, the predicted categories for each poster were investigated. In the 10-fold cross-validation procedure every poster is part of the test set once, and combining all these test-set predictions allows for the calculation of the Precision, Recall, and F1-score (see Table 1), as well as the confusion matrix (see Table 2).

Precision is the number of correctly classified posters in a genre, divided by the total number of posters predicted for that genre, The Action movie genre had the lowest Precision score (0.23), and table 2 shows that this is mainly due to the large number of posters that is classified as an Action movie. The Family movie genre had the highest Precision (0.40), which seems to be mostly due to a large number of correct classifications for this genre.

	Precision	**Recall**	**F1-score**
Action	0.23	0.34	0.28
Comedy	0.35	0.32	0.33
Family	0.40	0.40	0.40
Horror	0.33	0.36	0.34
Romance	0.32	0.25	0.28
Sci-Fi	0.28	0.34	0.31
Thriller	0.30	0.15	0.20
Average	0.32	0.31	0.31

Table 1, Precision, Recall, and F1-score for each of the genres, and the overall Average of the SVC_RBF model.

Recall is the number of correctly classified posters in a genre, divided by the number of posters in that genre (600). The Thriller genre performs most poorly of all the genres (0.15), and Table 2 shows that this is mostly due to the low number of posters classified in that genre. The Recall score of the Romance genre (0.25) seems to suffer because of the same reason. The F1 score can be interpreted as a weighted average of the precision and recall score, and does not provide much additional insight in the current analysis.

The Confusion Matrix in Table 2 shows that for each genre the number of correctly classified poster is the column maximum by a reasonable margin. In other words, there seems to be genre-specific information in many film posters. However, the confusion matrix also reveals a pattern that can be interpreted as favouring a more simple classification into two, instead of seven, classes. For example, when looking at the column for the posters that were predicted to be of the Sci-Fi genre it is clear that most of the errors are coming from the Horror, Thriller and Action genres. At the other hand, the column for predicted Family shows most misclassifications come from the genres Comedy and Romance.

This pattern of classification into two classes, [Horror, Thriller, Sci-Fi, Action] and [Comedy, Family, Romance], could to be related to darker/brighter atmosphere associated with those two sets. Inspection of the best fitting Decision Tree model showed that the first split of the data set is made based on the lightness feature. In other words, this feature seems to hold, at early stages of classification, most information about which genre a film poster belongs too.

	PREDICTED GENRE								
		Action	**Comedy**	**Family**	**Horror**	**Romance**	**Sci-Fi**	**Thriller**	Sum
ACTUAL GENRE	**Action**	206	41	41	101	60	109	42	600
	Comedy	118	189	110	23	93	50	17	600
	Family	82	95	238	45	81	44	15	600
	Horror	129	33	22	214	25	112	65	600
	Romance	108	102	116	49	151	47	27	600
	Sci-Fi	124	42	43	100	34	206	51	600
	Thriller	119	35	27	123	34	169	93	600
	Sum	886	537	597	655	478	737	310	4200

Table 2, Confusion Matrix of the SVC/RBF model.

Conclusion

In general the analysis supported the hypothesis that the design of film posters is stereotypical, and that on average the genre of a film can be predicted, with above chance level, from low-level properties of its film poster. The 6 models tested in the analysis all performed better than chance level (0.14), with the SVC/RBF and Random Forest model giving the best performance.

With 7 genres, the SVC/RBF-model was able to predict genre correctly 31% of the time, just under one-in-three. Although this score was considerably higher than chance level, it still is some way off from making consistently making reliable decisions. This is not unexpected given the limited information about the film posters that was entered into the prediction. To increase the accuracy of the prediction additional features could be extracted from the posters, such as the presence of faces, the amount of text and the fonts used, etc.

Interestingly, a deeper investigation of the prediction results showed that misclassification between genres was mostly within the subsets [horror, action, sci-fi, thriller] and [family, comedy, romance]. Although further analysis is needed into this finding, it could tentatively be associated with the darker/brighter atmosphere associated with those two sets.

To reach the correct target audience of a new movie, the film poster is an important marketing tool. Adhering to certain genre specific design stereotypes, the atmosphere, and to lesser extent the genre, of a movie can be communicated effectively.

Take home message:

The genre of a film can be to some extend predicted from low-level properties of its film poster, but classification seems to favour a more simple 2-class separation presumably based on brightness of the poster.