# Automatic Movie Posters Classification into Genres

Marina Ivasic-Kos, Miran Pobar, and Ivo Ipsic

Department of Informatics,
University of Rijeka,
Rijeka, Croatia
{marinai,mpobar,ivoi}@uniri.hr

**Abstract.** A person can quickly grasp the movie genre (drama, comedy, cartoons, etc.) from a poster, regardless of short observation time, clutter and variety of details. Bearing this in mind, it can be assumed that simple properties of a movie poster should play a significant role in automated detection of movie genres. Therefore, visual features based on colors and structural cues are extracted from poster images and used for poster classification into genres.

A single movie may belong to more than one genre (class), so the poster classification is a multi-label classification task. To solve the multi-label problem, three different types of classification methods were applied and described in this paper. These are: ML-kNN, RAKEL and Naïve Bayes. ML-kNN and RAKEL methods are directly used on multi-label data. For the Naïve Bayes the task is transformed into multiple single-label classifications. Obtained results are evaluated and compared on a poster dataset using different feature subsets. The dataset contains 6000 posters advertising films classified into 18 genres.

The paper gives insights into the properties of the discussed multi-label classification methods and their ability to determine movie genres from posters using low-level visual features.

**Keywords:** multi-label classification, data transformation method, movie poster.

## 1 Introduction

One of the goals of a poster is to convey information about a movie (genre, etc.) to potential moviegoers without them paying a lot of attention. With just a cursory glance at a poster while driving along or looking shortly while passing by, a person can grasp the movie genre (drama, comedy, cartoons, etc.) from variety of perceptual and semantic information on the poster. Taking this phenomenon [1] into account one can suppose that relevant information for determining the genre could be contained in global low-level features such as dominant color, spatial structure, color histogram, texture, etc.

Keeping this in mind our goal was to develop a method that would automatically determine the movie genres using mostly global low-level features of movie posters.

We used data from the TMDB [2] and realized that the problem we are dealing with is a multi-label problem since most of the movies belong to more than one genre.

© Springer International Publishing Switzerland 2015

A. Madevska Bogdanova and D. Gjorgjevikj, *ICT Innovations 2014*,
Advances in Intelligent Systems and Computing 311, DOI: 10.1007/978-3-319-09879-1_32

For example, “Delivery Man” belongs to Comedy, "The Wolf of Wall Street“ belongs to Crime, Drama and Comedy genres and „The LEGO Movie“ belongs to Adventure, Fantasy, Animation, Comedy, Action and Family genres. The problem is even more complex as the number of possible genres is large and there is no limit to the number of genres a film can be classified into.

The issue of classifying a film into genres from their supporting promotional material (trailers) has recently attracted some attention. In the paper [3], low-level features are extracted from movie trailers and used to classify 100 movies into 4 genres (drama, action, comedy, horror). In [4] GIST, CENTRIST and W-CENTRIST scene features are obtained from a collection of temporally-ordered static key frames. These feature representations are used as visual vocabulary for genre classification and their discriminate ability is tested on 1239 movie trailers.

In [5] the same visual features were used as in [3]. Movies were classified into three genres (action, drama, and thriller) which were selected because of their frequency among movies that were played in Taiwan from 2004 to 2006. Some additional genres were grouped together and presented as those three (e.g. drama included comedy and romance while thriller included horror).

All these approaches [3-5] consider only a single genre per movie in order to reduce the problem to the single-label classification case and apply the classic methods for single-label classification.

However, many different approaches have lately been developed to solve multi-label classification problems. These methods were primarily focused on text classification (news, web pages, e-mails etc.), but lately there are more and more domains in which they are applied, such as functional genomics classification (gene and protein function), music and song categorization into moods and genres [6], scene classification [7], video annotations, poster classification [8], etc. Comparison of methods for multi-label learning is given in [9].

In our approach, we treat the poster classification into movie genres as a multi-label classification task.

In Section 2, two methods for multi-label problem adaptation are explained. Both methods were applied to the poster classification problem, in an experiment as detailed in Section 3. The obtained results are compared and presented in Section 4. The paper ends with a conclusion and directions for future work.

## 2 Adaptation of the Multi-label Problem

The aim of our work is to develop a method that will automatically provide a list of relevant labels (movie genres) for a given, previously unseen poster, based on extracted low-level features. A movie can belong to more than one genre; therefore the task of poster classification into movie genres is a multi-label classification problem.

Multi-label classification of an example $e_j$ can be formally expressed as:

$$\exists\, e_j \in E : \varphi(e_j) = \{C_l, C_m\} \cup Z,\ Z \subseteq \cup_{i=1}^{k} C_i\,,\ \ l, m \in 1..k, l \neq m, \varphi : E \to C, \quad (1)$$

where $E$ is a set of samples, $C$ is a set of class labels and function $\varphi$ is a classifier so that exists at least one example $e_j$ that is mapped into two or more classes $C_l$ and $C_m$.

Methods most commonly used to tackle a multi-label classification problem can be divided into two different approaches [10]. These are problem transformation methods (referred as P1 in the following) and algorithm adaptation methods (referred to as P2 in the following).

In the P1 multi-label transformation approach, the multi-label classification problem is transformed into more single-label classification problems [11]. In the single-label classification problem an example $e_j$ is classified to a single class label from the set of $C$. The aim is to transform the data so that any classification method, designed for single-label classification, can be applied. We applied the P1 approach to transform the multi-label problem into single-label problems in two ways.

In the first case, binary relevance method [10] is applied, referred to as P1.1. One binary classifier is independently trained for each label. Each classifier then decides whether to assign one class label to an example or not. The overall classification result contains all class labels assigned to that instance. Therefore, each instance with multiple labels was transformed into a set of ordered pairs so that the first element of each pair is the instance and the second one is the class label. Thus, if an instance $e_j \in E$ can be classified into $Y_j = \{C_l, C_m, \dots, C_r\}$, $Y_j \subseteq C$ then that instance is replaced with $|Y_j|$ ordered pairs $(e_j, C_l), (e_j, C_m) \dots (e_j, C_r)$. For example, the movie „Non-Stop“ belongs to Action, Thriller and Mystery genres, and would be transformed into the set of ordered pairs containing individual genres: {(Non-Stop, Action), (Non-Stop, Thriller), (Non-Stop, Mystery)}.

In the second case, a label power-set method is used to create one classifier for every possible label combination (referred to as P1.2). That is, entire label set $\mathrm{Y_j} = \{\mathrm{C_l}, \mathrm{C_m}, \dots, \mathrm{C_r}\}$ is transformed into a new combined class $\mathrm{C_{l,m,\dots,r}}$ that is assigned to the example $\mathrm{e_j}$. In that way a set $\mathrm{C}$ is expanded with new combined classes into the set $\mathrm{C}'$, so that applies $\mathrm{C}' \supseteq \mathrm{C}$. Using this problem transformation method “Non-Stop” example would be transformed into the ordered pair containing one combined genre (Non-Stop, Action&Thriller&Mystery).

In both cases, a standard classification algorithm can be applied to assign a first element of the ordered pair (instance) to a second element of the pair (class label). On the other hand, algorithm adaptation methods (P2) extend specific learning algorithms in order to handle multi-label data directly.

## 3 Experiments

The experiments performed on the poster dataset obtained from the TMDB are presented below. The aim is to provide a list of relevant labels for the unknown poster using low-level features.

### 3.1 Data and Preprocessing Step

Our dataset consists of 6739 movie posters dated from 1990 onwards. We have selected 18 genres (Action, Adventure, Animation, Comedy, Crime, Disaster, Documentary, Drama, Fantasy, History, Horror, Mystery, Romance, Science Fiction, Suspense, Thriller, War and Western) and for each we picked 20 most popular movies for each year in the range. The total number of movies was smaller, because some movies were among the most popular in more than one genre.

Also, since each movie can have multiple genres, additional genre labels (e.g. Film Noir, Indie and Sport) were present in the data, so the total number of genres was 35.

The maximum number of films with a certain label is 2610, but one genre had only one instance which is not enough data to define a model for that class.

To prevent data scarcity for rare genres, we transformed the data in two ways. In the first, we joined the genres with few examples with similar genres according to our judgment, such as Neo-Noir with Crime and Road Movie with Adventure. We also joined some genres that commonly appeared together, e.g. Mystery and Crime with Thriller. After this transformation, the number of genres was reduced to 11. We refer to this data set as JG. In the second case we have simply discarded the additional genres in the data that were not among the selected 18. We refer to this data set as DG.

The resulting data distribution is more suitable for learning of classification models because sufficient examples per most genres are obtained (approximately more than 1000 posters per genre), although the width of the classes are rather uneven (std. dev is 631 in case with discarded genres and even less favorable, std. dev equals 916, in case with joined genres). The more detailed statistics of data before and after transformation is presented in the Table 1.

**Table 1.** Original and transformed data set statistics

| *Statistic* | *Original data* | *Transformed data* | |
|---|---|---|---|
| | | *joined genres (JG)* | *discarded genres (DG)* |
| No. of classes in set $C$ | 35 | 11 | 18 |
| Max examples per genre | 2610 | 3209 | 2610 |
| Min examples per genre | 1 | 279 | 64 |
| Mean examples per genre | 558 | 1497 | 982 |
| Std. dev. per genre | 641 | 916 | 631 |

Data transformed in such way was directly handled by methods proposed in approach P2. For the P1 approach an additional pre-processing step was needed to enable the use of single label classification methods. In the case of the P1.1 approach the subset of multi-label data is used more than once, in fact as many times as there are labels into which the poster is classified. In the case of P1.2 approach, for each multi-label set a new combined class was created, so the set $C'$ is built. However, with this approach a significant problem of sparse classes appears. The relation among the number of class labels in the original sets and the number of class labels contained in the set $C'$ is presented in the Table 2 for the cases of original and transformed data sets.

**Table 2.** Number of classes in original and transformed data sets when a label power-set method is used

| ***Statistic*** | ***Original data*** | ***Transformed data*** | |
|---|---|---|---|
| | | ***joined genres (JG)*** | ***discarded genres (DG)*** |
| No. classes in set $C'$ | 1480 | 387 | 891 |
| Max examples per genre | 417 | 605 | 607 |
| Min examples per genre | 1 | 1 | 1 |
| Mean examples per genre | 4.6 | 17.41 | 7.5 |
| Std. dev. per genre | 14.87 | 47.88 | 26.04 |

Considering the statistics presented in the Table 2, problem of data scarcity becomes even more obvious. In all cases many new classes are formed (e.g. 387 vs. 11 in case of joined genres, 891 vs. 18 in case of discarded genres) with a small number of examples, e.g. in cases of joined genres 44.44% genres have less than 3 elements. Due to an insufficient number of examples per most genres this kind of data transformation is not used for learning the classification model. The solution could be to form new genres that will include mostly triplets of genres but such reduction of set is not applied here.

### 3.2 Features

Motivated by the way people capture relevant information about the movie with just a glance at billboards we wanted to examine if low-level features that can be easily noticed on the poster, such as dominant colors and structure, have discriminative ability in terms of genre classification.

Before the low-level features were extracted, each poster was proportionally sized to so that it is 100 pixels wide and converted to HSV color space. Then, the image color histogram was calculated on hue (H), saturation (S) and value (V) channels of the whole image. Subsequently, histogram bins with the highest values for each channel are selected. Obtained features correspond to dominant colors (referred to as DC). We have experimentally tested different numbers of dominant colors (3, 6, 8, 12, 16, 24 and 36) and have determined that in our task 12 dominant colors per channel yield the best classification results. Thus, 36-dimensional DC vectors were used (12 dominant colors per three channels).

To preserve the information about the color layout of a poster, we have computed 5 local HSV histograms from which 12 dominant colors per channel were selected. These are referred to as DC1 to DC5, with total size of 180. DC1, DC2 and DC3 were computed from 3x1 grids, DC4 was computed on the central part of the poster that would probably contain the object and DC5 on the surrounding part that would probably contain the background. The central part was of the same proportions as the whole image, but 1/4 of the diagonal size. The arrangement of image grids from which the local dominant colors were computed is given in Fig. 1.

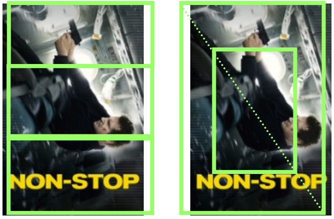

**Fig. 1.** The arrangement of image grids from which the local color histograms were computed

Additionally, we have computed the statistics and color moments (CM) for each HSV channel, such as mean, standard deviation, skew and kurtosis. The size of CM feature vector is 12.

Also, the GIST image descriptor, available at [12], was used. It is a structure-based image descriptor created for recognition of similar scenes, like mountains, streets, etc. This descriptor refers to the dominant spatial structure of the scene characterized by properties of its boundaries (e.g., the size, degree of openness, perspective) and its content (e.g., naturalness, roughness) [13]. Spatial properties are estimated using global features computed as a weighted combination of Gabor-like multi scale-oriented filters. The dimension of GIST descriptor is n x n x k where n x n is the number of samples used for encoding and k is the number of different orientation and scales of image components. GIST descriptor of each genre is implemented with 8x8 encoding samples obtained by projecting the averaged output filter frequency within 8 orientations per 8 scales. The size of GIST feature vector is 512.

### 3.3 Classification Methods

We have used three classification methods for classifying unknown posters into movie genres. Naïve Bayes classifier was used on data adapted according to the P1.1 approach described above. RAKEL [14], a kind of problem transformation method (P1) and ML-kNN [15], an algorithm adaptation method (P2), can be directly used on multi-label data. All methods are tested with both transformed data sets JG and DG (with joined or with discarded genres).

When using the Naïve Bayes (NB), binary relevance method was used and a single NB classifier was trained per each genre to distinguish that genre from all other genres. To classify an unseen poster sample all NB classifiers are applied.

RAKEL (random k-label sets) was run using the nearest neighbor (1-NN) classifier based on the Bhattacharyya distance (2) as base classifier,

$$d_B(p,p') = \sum_{i=1}^{N} \sqrt{p(i)p'(i)}, \tag{2}$$

where $p$ and $p'$ are histograms and N is the number of bins. The RAKEL subset size was 3 and the number of models was 12.

ML-kNN is a variant of the lazy learning algorithm derived from the traditional k-Nearest Neighbor (kNN) algorithm. For each unseen poster instance, its k nearest neighbors are firstly identified in the training set. Then, based on the statistical information gained from the genre label sets of these neighboring instances, maximum a posteriori (MAP) principle is utilized to determine the genre set for the unseen poster instance.

## 4 Experimental Results

From 6739 collected posters, 80% were used for training and 20% for testing. We have used the feature set that includes GIST features, dominant color (DC) features, local dominant color features (DC1 to DC5) and color moments (CM). The size of feature vector is 740. Since the size of feature vector is large in proportion to the number of posters, we have also tested the classification performance using five subsets.

We have used accuracy, precision, recall and F1 score as instance-based and label-based evaluation measures. The instance-based evaluation measures are based on the average differences of the actual and the predicted sets of genres over all posters in the test dataset. The label-based evaluation measures assess the predictive performance for each genre separately and then average the performance over all genres [10].

The Table 3 shows the label-based evaluation results obtained using NB on data transformed with joined genres (JG) and on data transformed by discarding additional genre labels beyond the selected 18 (DG), for different feature subsets. The results obtained using all features are only slightly better than with other subsets with much smaller number of features for both data transformation methods JG and DG, and accuracy was even better when using only the feature subset DC+CM.

Thus, in further experiments, we only tested the subset with the smallest number of features (DC+CM) that performed similarly and the set with all features. The idea was to keep the number of features low enough to emphasize the information that is relevant for classification and adequately train the classifier.

The results obtained with the DG data transformation are significantly lower than with JG method, which is not surprising due to much larger number of classes.

**Table 3.** Label (genre) based evaluation results for datasets JG with joined genres (11) and dataset DG with 18 genres, using NB

| *Label-based evaluation measure* | *All features* | | *GIST* | | *DC+CM* | | *DC1..DC5 + CM* | | *DC + DC1..DC5 + CM* | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **JG** | **DG** | **JG** | **DG** | **JG** | **DG** | **JG** | **DG** | **JG** | **DG** |
| Accuracy | 0.62 | 0.62 | 0.63 | 0.65 | **0.65** | **0.70** | 0.61 | 0.59 | 0.61 | 0.59 |
| Precision | **0.30** | **0.21** | 0.29 | **0.21** | 0.28 | **0.21** | 0.28 | 0.19 | 0.28 | 0.19 |
| Recall | **0.61** | **0.60** | 0.57 | 0.54 | 0.48 | 0.39 | 0.56 | 0.55 | 0.56 | 0.56 |
| F1 score | **0.38** | **0.29** | 0.37 | **0.29** | 0.34 | 0.21 | 0.36 | 0.27 | 0.36 | 0.27 |

Instance-based classification results obtained using NB are presented in the Table 4 with both data transformation methods, for different feature subsets. Instance-based results are lower than genre based results for all evaluation measures. Also, all feature subsets perform similarly with respect to F1 score. This suggests that most of the features are interdependent. The F1 score is a measure of classification accuracy that considers both precision and recall. It can be interpreted as a weighted average of the precision and recall. The F1 score reaches its best value at 1 and worst score at 0.

**Table 4.** Instance (movie poster) based evaluation results for datasets JG and DG, using NB

| *Instance-based evaluation measure* | *All features* | | *GIST* | | *DC+CM* | | *DC1..DC5 + CM* | | *DC + DC1..DC5 + CM* | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **JG** | **DG** | **JG** | **DG** | **JG** | **DG** | **JG** | **DG** | **JG** | **DG** |
| Accuracy | 0.55 | 0.57 | 0.56 | 0.60 | **0.61** | **0.68** | 0.57 | 0.56 | 0.57 | 0.56 |
| Precision | **0.25** | - | **0.25** | **0.17** | - | - | - | - | - | - |
| Recall | 0.47 | 0.44 | 0.45 | 0.41 | 0.44 | 0.33 | **0.49** | **0.46** | **0.49** | **0.46** |
| F1 score | 0.31 | **0.23** | 0.31 | 0.23 | **0.32** | 0.21 | **0.32** | **0.23** | **0.32** | **0.23** |

Label-based classification results with RAKEL and ML-kNN are presented in Table 5. Results with RAKEL are significantly better for all evaluation measures than with ML-kNN, but actually slightly worse than with NB. As with NB, DC+CM feature set performs similarly as all features for RAKEL.

**Table 5.** Label based evaluation results for datasets JG and DG, using RAKEL and ML-kNN

| *Label-based evaluation measure* | *RAKEL* | | | | *ML-kNN* | | | |
|---|---|---|---|---|---|---|---|---|
| | *All features* | | *DC+CM* | | *All features* | | *DC+CM* | |
| | **JG** | **DG** | **JG** | **DG** | **JG** | **DG** | **JG** | **DG** |
| Precision | 0.33 | 0.26 | 0.33 | 0.31 | 0.53 | 0.52 | 0.44 | 0.57 |
| Recall | 0.32 | 0.25 | 0.32 | 0.31 | 0.06 | 0.03 | 0.03 | 0.12 |
| F1 score | 0.33 | 0.26 | 0.32 | 0.31 | 0.1 | 0.05 | 0.05 | 0.20 |

The results for instance based evaluation, shown in Table 6, show similar relationships between classification methods, feature sets and data transformation methods as for genre based evaluation.

**Table 6.** Instance (movie poster) based evaluation results for datasets JG and DG, using RAKEL and ML-kNN

| *Instance-based evaluation measure* | *RAKEL* | | | | *ML-kNN* | | | |
|---|---|---|---|---|---|---|---|---|
| | *All features* | | *DC+CM* | | *All features* | | *DC+CM* | |
| | **JG** | **DG** | **JG** | **DG** | **JG** | **DG** | **JG** | **DG** |
| Accuracy | 0.23 | 0.19 | 0.22 | 0.24 | 0.06 | 0.07 | 0.03 | 0.15 |
| Precision | 0.33 | 0.26 | 0.32 | 0.32 | 0.13 | 0.09 | 0.07 | 0.23 |
| Recall | 0.33 | 0.25 | 0.32 | 0.32 | 0.06 | 0.07 | 0.03 | 0.16 |
| F1 score | 0.31 | 0.24 | 0.30 | 0.30 | 0.08 | 0.07 | 0.04 | 0.17 |

Overall the best results are obtained using the Naive Bayes classification algorithm with all features on data transformed with joined genres, however only slightly worse performance was observed with only a small number of color-based features DC+CM.

## 5 Conclusion and Future Work

In this paper, automated detection of movie genres from posters was modeled as a multi-label classification task, where a single movie may belong to more than one genre. The experiment was conducted on a dataset containing 6739 movie posters, classified into one or more of 18 genres. Since some genres had too few examples to effectively train the classifier, the performance of classification was compared with the same dataset where some genre labels were merged, yielding 11 genres.

As the usual single-label classification algorithms can't directly be used to solve the multi-label problem, either the problem or the algorithms must be adapted in some way. Two different methods for problem transformation were applied and use of appropriated classifiers is described in this paper. These are: ML-kNN, Naïve Bayes and RAKEL. ML-kNN and RAKEL methods are directly used on multi-label data. For the Naïve Bayes the task is transformed into multiple single-label classifications. The features used in the classification were low-level features based on color histograms and color moments combined with the GIST descriptor. Obtained results are evaluated and compared on a poster dataset using different subsets of color and structural features.

The best result considering the F1 score was about 0.38 for the case of Naive Bayes classifier on the complete feature set and for 11 genres. Reducing the number of features from 740 to only 48 features related to the dominant colors of the HSV histogram didn't significantly impact the results, yielding the F1 score of 0.34. This suggests that few dominant colors indeed carry discriminative part of information about the movie genres, and that other tested features might be largely interdependent.

In the future work, we plan to test the dense SURF [16], other visual features used for scene representation [13] as well as features for text recognition. We also plan to test the classification on a much larger dataset and with different classification methods.

Also, a subjective test will be conducted to determine human ability to detect genres from poster images, and the results will be used for comparison with automatic detection.

## References

1. Potter, M.C.: Short-term conceptual memory for pictures. Journal of Experimental Psychology: Human Learning and Memory 2(5), 509 (1976)
2. The movie database, `http://www.themoviedb.org/`

3. Rasheed, Z., Sheikh, Y., Shah, M.: On the use of computable features for film classification. IEEE Transactions on Circuits and Systems for Video Technology 15(1), 52–64 (2005)
4. Zhou, H., Hermans, T., Karandikar, A.V., Rehg, J.M.: Movie genre classification via scene categorization. In: ACM Proceedings of the International Conference on Multimedia, pp. 747–750 (2010)
5. Huang, H.-Y., Shih, W.-S., Hsu, W.-H.: A film classifier based on low-level visual features. In: IEEE 9th Workshop on Multimedia Signal Processing, MMSP 2007, pp. 465–468 (2007)
6. Allmusic, `http://www.allmusic.com/`
7. Boutell, M.R., Luo, J., Shen, X., Brown, C.M.: Learning multi-label scene classification. Pattern Recognition 37(9), 1757–1771 (2004)
8. Ivašić-Kos, M., Pobar, M., Mikec, L.: Movie Posters Classification into Genres Based on Low-level Features. In: IEEE Proceedings of International Conference MIPRO, pp. 1148–1153 (2014)
9. Madjarov, G., Kocev, D., Gjorgjevikj, D., Džeroski, S.: An extensive experimental comparison of methods for multi-label learning. Pattern Recognition 45(9) (2012)
10. Tsoumakas, G., Katakis, I.: Multi-Label Classification: An Overview. International Journal of Data Warehousing & Mining 3(3) (2007)
11. Read, J., Pfahringer, B., Holmes, G., Frank, E.: Classifier Chains for Multi-label Classification. Machine Learning Journal 85(3) (2011)
12. GIST, `http://people.csail.mit.edu/torralba/code/spatialenvelope/`
13. Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. International Journal of Computer Vision 42(3), 145–175 (2001)
14. Tsoumakas, G., Vlahavas, I.: Random k-label sets: An ensemble method for multi-label classification. In: Machine Learning: ECML, pp. 406–417. Springer (2007)
15. Zhang, M.L., Zhou, Z.H.: ML-KNN: A lazy learning approach to multi-label learning. Pattern Recognition 40(7), 2038–2048 (2007)
16. Bay, H., Tuytelaars, T., Van Gool, L.: Surf: Speeded up robust features. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006, Part I. LNCS, vol. 3951, pp. 404–417. Springer, Heidelberg (2006)