Movie Genre Classification based on Poster Images with Deep Neural Networks

Wei-Ta Chu
National Chung Cheng University, Taiwan
wtchu@ccu.edu.tw

Hung-Jui Guo
National Chung Cheng University, Taiwan
hguo2003@gmail.com

ABSTRACT

We propose to achieve movie genre classification based only on movie poster images. A deep neural network is constructed to jointly describe visual appearance and object information, and classify a given movie poster image into genres. Because a movie may belong to multiple genres, this is a multi-label image classification problem. To facilitate related studies, we collect a large-scale movie poster dataset, associated with various metadata. Based on this dataset, we fine-tune a pretrained convolutional neural network to extract visual representation, and adopt a state-of-the-art framework to detect objects in posters. Two types of information is then integrated by the proposed neural network. In the evaluation, we show that the proposed method yields encouraging performance, which is much better than previous works.

CCS CONCEPTS

• **Computing methodologies** → **Image representations**; *Visual content-based indexing and retrieval*; *Neural networks*;

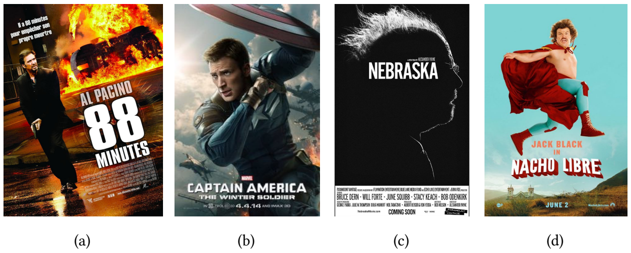
KEYWORDS

Movie genre classification, movie poster, multi-label classification, deep neural network

ACM Reference format:
Wei-Ta Chu and Hung-Jui Guo. 2017. Movie Genre Classification based on Poster Images with Deep Neural Networks. In *Proceedings of MUSA2'17, Mountain View, CA, USA, October 27, 2017,* 7 pages.
DOI: 10.1145/3132515.3132516

1 INTRODUCTION

Image attribute extraction has been widely studied in recent years, since visual attributes can boost various tasks such as image retrieval [14] and image captioning [18]. Most prior studies either focus on detecting or recognizing visual entities like object and scene, or extracting semantic concepts embedded in images. These attributes really provide widespread influence on multimedia retrieval and many computer vision applications. On the other hand, some visual attributes are implicit but can be easily perceived by human beings. Image styles [4], image aesthetics [10], and attractiveness estimation [16] are instances of such attributes.



(a) (b) (c) (d)

Figure 1: Sample movie poster images. (a) *88 Minutes* (Crime, Drama, Mystery); (b) *Caption America: The First Avenger* (Action, Adventure, Sci-Fi); (c) *Nebraska* (Adventure, Comedy, Drama); (d) *Nacho Libre* (Comedy, Family, Sport).

When studying visual attributes for different types of images, we found that movie poster image is a type of medium with very rich attributes. First, movie posters are created to attract people paying time and money to watch the corresponding movie. Information on a movie poster, therefore, should be attractive. Figure 1 shows four sample posters. In the figure caption we show movie name and the corresponding genres (in the parentheses), which are obtained from the IMDB[1] website. Most of them present the most important imagery of the corresponding movie. For example, Figure 1(a) shows a violent scene with a crashed car in fire and the actor with gun. This dramatic scene attracts people who like excitement or violence. Second, different movies target at different populations, and movie posters should concisely present genre information. For example, we can clearly perceive that Figure 1(a) and Figure 1(b) present action elements, Figure 1(c) seems mysterious, and Figure 1(d) looks funny. Third, movie posters usually present important objects like the main actor's name, like *Al Pacino* in Figure 1(a), and typical objects to demonstrate characteristics of a movie, like gun and fire in Figure 1(a) and the shield of Captain America in Figure 1(b). Overall, the first aforementioned observation is highly related to image aesthetics or attractiveness estimation; the second observation is related to genre or style classification; while the third observation is involved with object detection. Therefore, we think movie poster is a good research target with many technical challenges but was overlooked before.

In this work, we propose to analyze movie poster images and classify them into movie genres based on a neural network jointly considering heterogeneous information. Motivated by the interesting work [8] that *judges a book by its cover*, i.e., determining the genre of a given book cover, we would like to investigate how likely

MUSA2'17, Mountain View, CA, USA
© 2017 ACM. 978-1-4503-5509-4/17/10...$15.00
DOI: 10.1145/3132515.3132516

[1] http://www.imdb.com

we can determine the genre of a movie poster. This work corresponds to the second aforementioned observation, and could be the fundamental module for movie content access, management, and presentation. Estimating subtle attributes from only images is actually not a totally new idea. The work in [2] shows that a person's first name can be roughly predicted based only on his/her face image. Other attributes like occupation can also be predicted to some extent based on face image [3]. The reasons for enabling such predictions are that first name is statistically related to genders, races, and when a person borns, and occupation is statistically related to genders, ages, and other body context. These inspiring works motivate us to study implicit factors related to movie genres, and encourage us to build a computational model to do classification.

Contributions of this work are summarized as follows.

- To promote and facilitate the proposed movie poster analysis, we collect a large-scale movie poster dataset from the IMDB website. This dataset mainly consists of posters of movies released from 1980 to 2015 in Hollywood, as well as the associated metadata like movie genre, names of the director and main actors, box office, and so on.
- We construct a computational model based on deep neural network to automatically classify a movie poster into genres. Note that one movie belongs to multiple genres, see Figure 1. This task is therefore a multi-label image classification problem.

Note that, although classifying a movie poster image into genres is the main target of this work, potential contributions of this study is to limited to this. With movie genre classification, we may be able to construct a movie recommender system assuming that a person likes movies of similar genres. The relationship between objects and movie genres can be discovered, and can be important clues for amateurs to design posters. Based on the experience of the model construction, we may extend this computational model to estimate other movie properties like box office.

The rest of this paper is organized as follows. Section 2 provides brief literature survey on movie genre classification. Section 3 describes the proposed model to achieve movie genre classification based on poster images. Details of the multi-label classification problem are provided. Evaluation results and discussion are provided in Section 4, and Section 5 concludes this paper.

2 RELATED WORKS

There have been a few works on movie genre classification. Rasheed and Shah proposed one of the earliest works on movie genre classification [11]. They first classified a preview (trailer) into action or non-action movie based on average shot length and motion content. Non-action movies were further classified into comedy, horror, or drama/other based on light intensity, which was inspired by cinematic principles. Action movies were further analyzed based on audio energy and fire/explosion detection. Zhou et al. [19] extracted keyframes from movie trailers. From keyframes features like GIST, CENTRIST, and W-CENTRIST were extracted, and a bag of visual word model was constructed to represent a movie. Based on such representation, similarity between movies was calculated, and a nearest neighbor classifier was used to classify a movie trailer into one of the four genres, i.e., action, comedy, drama, and horror. Recently, Simoes et al. [15] used a convolutional neural network (CNN) to classify movie trailers into genres. They proposed a movie trailer dataset and verified that features extracted by the CNN approach significantly outperforms handcrafted low-level features. Wehrmann and Barros [17] proposed a novel classification method that consists of an ultra-deep CNN with residual connections. In addition to visual information, they further extracted temporal information for movie genre classification.

The aforementioned studies are mainly working on movie trailers. On the contrary, very few works were proposed to do movie genre classification based on poster images. This may be due to that a poster image just provides limited information, which gives rise to significant challenges in feature representation as well as genre classification. The work in [7] may be the earliest published study on movie poster classification. They worked on a dataset consisting of 1,500 posters belonging to six genres, i.e., action, animation, comedy, drama, horror, and war. Low-level features like dominant colors, edge-based features, and number of faces were used to represented movie posters. A poster was classified into one or two genres based on a nearest neighbor classifier. They pointed out that one movie usually belongs to multiple genres. The same researchers later extended the evaluation dataset to 6,000 posters belonging to 18 genres [6], but similar feature representation and classification schemes were used. In our work, we try to investigate how deep learning approaches can solve the movie poster classification problem, based on a large-scale dataset associated with various metadata. We will compare with [7] and show performance of different variants of the proposed framework.

3 MOVIE GENRE CLASSIFICATION

3.1 Deep Neural Network

Given a set of training data $D = \{X, Y\}$, where $X = (x_1, x_2, ..., x_N)$ is the set of N poster images and $Y = (\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_N)$ is the corresponding genre information. The vector $\boldsymbol{y}_i = (y_{i,1}, y_{i,2}, ..., y_{i,M})$ is a binary vector, where $y_{i,j} = 1$ indicates that the ith poster belongs to the jth genre. Note that a movie may belong to multiple genres, i.e., $1 \leq \sum_j y_{i,j} \leq M$. Based on D, we would like to construct a computational model that outputs the probability of a given poster image x_i belonging to each movie genre. That is, the constructed model acts as a function $\mathcal{F}$ such that $\mathcal{F}(x) = (\hat{y}_{i,1}, \hat{y}_{i,2}, ..., \hat{y}_{i,M})$, where $0 \leq \hat{y}_{x,j} \leq 1$. A good model would output the estimated vector $(\hat{y}_{i,1}, \hat{y}_{i,2}, ..., \hat{y}_{i,M})$ as close to the ground truth vector $(y_{i,1}, y_{i,2}, ..., y_{i,M})$ as possible.

In this work, we construct the function $\mathcal{F}$ by jointly considering visual representation extracted from a convolutional neural network, and object information extracted by one state-of-the-art object detector [12]. Figure 2 illustrates the overall network structure. To extract effective visual representation, we construct the convolutional neural network similar to the convolutional part of AlexNet [9], as shown in the first part of Figure 2. This network consists of seven convolutional layers, where the seventh convolutional layer is followed by a batch normalization layer [5]. After normalization, feature maps are flattened as a vector to be the visual representation. The visual representation will be combined with object information in the third part of Figure 2. To reduce heterogeneity and enable feature combination, we would like to

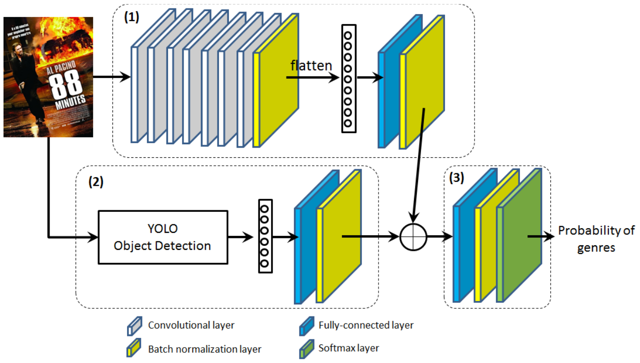


Figure 2: Structure of the proposed deep neural network.

map the visual representation and the object information into a common space. Therefore, the visual representation is embedded by a fully-connected layer with batch normalization, as shown in the end of Figure 2(1).

In addition to visual appearance, we also consider object semantics that may highly relate to movie genres. For example, in action movies, *person* and *car* may more often appear on posters; while in romantic movies, the hero and the heroine are often the main objects on posters. To explore how object information provides clues for genre classification, we detect objects by the YOLO system [12], as shown in Figure 2(2). Most state-of-the-art object detection systems first find region proposals and then classify the proposals into one of the predefined object classes. Instead, the YOLO system formulates object detection as a regression problem. An image is divided into grids, and from each grid the system predicts several bounding boxes and their associated confidence values. The confidence value conceptually represents the extent of overlap between the predicted bounding box and the ground truth. The kernel model for bounding box prediction is a convolutional neural network consisting of 24 convolutional layers followed by 2 fully-connected layers. This structure is inspired by the GooLeNet model, and provides real-time object detection and classification. Details of the YOLO system please refer to [12].

We adopted the YOLO version 2 [13] trained based on the MSCOCO dataset. Given a poster image, this system outputs bounding boxes of detected objects and their associated confidence values, e.g., $\{c_1, ..., c_K\}$ if there are K detected objects. Figure 3 shows sample results of object detection for poster images. To represent object information of a poster, we sum confidence values of objects of the same type. That is, $o_i = c_j + c_k$ if both the jth object and the kth object are object i, say a *person*. A poster image is then represented as the vector $\boldsymbol{o} = (o_1, ..., o_B)$. Note that if there are more ith objects, the value o_i tends to be larger; on the other hand, if there is no ith object in this poster, the value o_i would be zero. In our adopted YOLO system, totally 80 object classes are detected, including *person*, *car*, *dog*, and so on, and therefore the vector $\boldsymbol{o}$ is 80-dimensional ($B = 80$). As shown in Figure 2(2), the vector $\boldsymbol{o}$ is also embedded by a fully-connected layer with batch normalization.

The embedded visual representation and object information are concatenated as the input of the third part of Figure 2. This part consists of one convolutional layer, one batch normalization layer, and finally one softmax layer. Given a poster x_i, this network finally outputs the estimated probability vector $\hat{\boldsymbol{y}}_i = (\hat{y}_{i,1}, \hat{y}_{i,2}, ..., \hat{y}_{i,M})$, where $\hat{y}_{i,j}$ indicates the probability of the poster x_i being the jth movie genre.

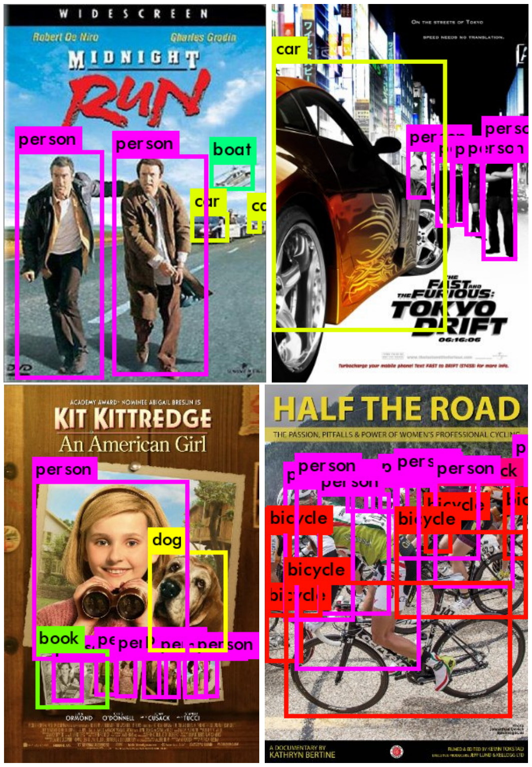


Figure 3: Sample results of object detection for poster images.

Detailed configurations of the network are shown in Table 1. The item conv3-100, for example, means that the convolution kernel is 3×3, and the number of filters is defined as 100. The activation function of each layer is ReLU, the objective function is the mean square error between estimated probability vector $\hat{\boldsymbol{y}}$ and ground truth vector $\boldsymbol{y}$, and the optimization algorithm is SGD with the learning rate 0.1. The training process was conducted in 20 epochs, with mini-batch size 128.

3.2 Multi-label Classification

The output of the network shown in Figure 2 is an M-dimensional probability vector, where each dimension indicates how likely a given poster belongs to a movie genre. As shown in Figure 1, a movie usually belongs to multiple genres. In addition, the number of genres a movie belongs to would be varied. One intuitive way to solve this multi-label classification problem is thresholding each dimension. If the value of the ith dimension of $\hat{\boldsymbol{y}}$ is larger than a threshold, we say the given poster belongs to the ith movie genre. However, how to define the threshold for each dimension (different dimensions may be associated with different thresholds) obviously is the main problem.

Table 1: Detailed configurations of the proposed network.

<table>
<tr><td colspan="9">input (100 × 100 poster images)</td></tr>
<tr><td>Part 1</td><td>conv3-100
maxpooling</td><td>conv3-32</td><td>conv3-64</td><td>conv3-128</td><td>conv3-256</td><td>conv3-128</td><td>conv3-80
batch norm.
maxpooling
flatten</td><td>fully-connected (80 nodes)
batch norm.</td></tr>
<tr><td>Part 2</td><td colspan="8">YOLO object detection
fully-connected (80 nodes)
batch norm.</td></tr>
<tr><td>Part 3</td><td colspan="8">fully-connected (256 nodes)
batch norm.
softmax</td></tr>
</table>

In this work, we adopt a grid search scheme based on Matthews correlation coefficients to determine the (nearly) best thresholds for each dimension. Details of the best threshold finding algorithm is shown in Algorithm 1. Basically, the idea is adjusting the threshold for each dimension separately. Every time when we adjust the threshold for a dimension, we binarize the predicted probability vector with the adjusted threshold, and calculate the Matthews correlation coefficient (the *MC* function shown in line 8) between the thresholded vector and the ground truth vector. Note that the notation $\boldsymbol{y}[1:i]$ denotes the vector with values of the first i dimensions are from the vector $\boldsymbol{y}$, and values of the remaining $M-i$ dimensions are set as 0. That is, when we try to find the best threshold for the ith dimension, all dimensions from 1 to $i-1$ are jointly considered. For the ith dimension, the Matthews correlation coefficients for all thresholds are stored in the vector ρ, from which we find the index of threshold that causes the largest correlation (line 13). The best threshold for the ith dimension is then determined by the corresponding value (line 14). After checking all dimensions, the set of thresholds $T = \{t_1, ..., t_M\}$ that yields the largest correlation is found.

Given a test poster image, the framework shown in Figure 2 outputs the probabilities $\{\hat{y}_1, \hat{y}_2, ..., \hat{y}_M\}$ of this poster being a specific genre. If the estimated probability $\hat{y}_i$ is larger than the corresponding threshold t_i, we say that this poster belongs to the ith genre.

ALGORITHM 1: Find Best Threshold

```
Input: Predicted probability vector $\hat{\boldsymbol{y}} = (\hat{y}_1, \ldots, \hat{y}_M)$, truth probability
       vector $\boldsymbol{y} = (y_1, \ldots, y_M)$, threshold upper bound $u$, and threshold
       stride $s$,
Output: The set of best thresholds $T = \{t_1, \ldots, t_M\}$
1  Initialize the set of threshold $T = \{t_1 = 0, t_2 = 0, \ldots, t_M = 0\}$
2  for $i = 1$ to $M$ do
3      $j = 0$
4      Initialize an empty vector $\boldsymbol{\rho}$
5      Initialize an empty vector $\boldsymbol{\theta}$
6      for $j < u$ do
7          Initialize an $M$-dim binary vector
             $\boldsymbol{b} = (b_1 = 0, b_2 = 0, \ldots, b_M = 0)$
8          if $\hat{y}_i > j$, $b_i = 1$. Otherwise, $b_i = 0$
9          $\boldsymbol{\rho} = \boldsymbol{\rho}.append(MC(\boldsymbol{y}[1:i], \boldsymbol{b}))$
10         $\boldsymbol{\theta} = \boldsymbol{\theta}.append(j)$
11         $j = j + s$
12     end
13     $k^* = \arg\max_k \boldsymbol{\rho} = (\rho_1, \ldots, \rho_k, \ldots)$
14     $t_i = \boldsymbol{\theta}[k^*]$
15 end
16 Output a set of threshold $T = \{t_1, \ldots, t_M\}$.
```

4 EVALUATION

4.1 Databases

To evaluate the proposed study, we collect a movie poster image dataset from the IMDB website. We crawled one poster per Hollywood movie released from 1980 to 2015, and collected 8,191 poster images in total. Resolutions of these poster images range from 89 × 132 to 300 × 581. The genre information associated with each movie is also collected. There are totally 23 different genres. Figure 4(a) shows the distribution of the numbers of posters in different movie genres. Note that the total number of posters shown in this figure is much larger than 8,191, because one movie usually belongs to multiple genres. From Figure 4(a), we see that the numbers of movies in different genres are quite imbalanced.

To diminish the influence of imbalance in training, we conduct different extents of augmentation for different genres. The genres *Drama* and *Comedy* are not augmented, because they already have a large number of posters. For the genres *Action*, *Romance*, *Crime*, *Adventure*, *Thriller*, and *Documentary*, we randomly crop two 150 × 150 regions from the original images. Therefore, together with the original images, the volume of data for these genres is increased twice. Similarly, the volume of data in genres *Horror*, *Biography*, *Family*, *Fantasy*, and *Sci-Fi* is increased seven times; the volume of data in genres *Short*, *Music*, *Animation*, *History*, and *Sport* is increased fifteen times; and the volume of data in genres *War*, *Musical*[2], and *Western* is increased thirty times. After data augmentation, the total number of distinct images is 16,997. By summing the number of images in all genres, the total number is 36,295.

In addition to genre information, we also collect other metadata for future study, including Box Office, Rated, Awards, Director, Writer, imdbVotes, imdbRating, Actors, and Metascore. For example, Figure 4(b) shows the distribution of the number of posters with

[2] A *Music* movie presents content related to music, like biography of a musician or a story of a band; while in an *Musical* movie, actors largely sing songs to present narrative.

different ratings. Ratings in the IMDB website range from 0 to 10. As can be seen, most movies get ratings ranging from 5 to 8, i.e., a nearly normal distribution can be seen. Figure 4(c) shows the distribution of the number of movies in different box office ranges. As we expect, most movies are not sold well. The box office of around 80% of the movies is less than 50 millions US dollars. Note that the values of box office are adjusted for inflation according to statistics provided in [1].

Combining poster images with these metadata, we believe this is the first heterogeneous movie poster dataset that may facilitate not only genre classification, but also several other studies such as the relationship between actors and genres, the relationship between director and box office, and so on. This database will be released in public soon.

4.2 Performance of Movie Genre Classification

To evaluate movie genre classification based on poster images, we randomly select 75% of images from each genre for training the proposed model, and the remaining images are used for testing. The performance metric is accuracy defined as

$$Accuracy = \frac{\|\boldsymbol{y} \text{ AND } \boldsymbol{b}\|}{\|\boldsymbol{y} \text{ OR } \boldsymbol{b}\|} \times 100\%, \tag{1}$$

where $\boldsymbol{y}$ and $\boldsymbol{b}$ are truth labels and predicted labels (in binary vector form), respectively. The notation $\|\cdot\|$ denotes L1-norm of a vector.

We compare performance obtained by one baseline system and several neural network variants. Detailed settings of these methods are described as follows.

- (1) Baseline system [7]: To our best knowledge, this is one of the few prior works on movie poster classification. We implement the best setting reported in [7] to be the baseline system. They represented a poster image by low-level features including dominant colors, edge-based features, and number of faces. They first find feature centroids of images in each genre. Given a test poster image, the distance between it and centroid of each genre is calculated, and the label of each poster is determined as the genre with the nearest feature centroid.
- (2) Fully-connected neural network: We first try a simple fully-connected neural network to do poster classification. Five fully-connected layers with each consisting of 512 nodes are connected, followed by one softmax layer. Other training settings, such as activation function and mini-batch size, are the same as the proposed method mentioned in Section 3.1.
- (3) Convolutional neural network: We also try a CNN to do poster classification. Eight convolutional layers are connected, with the numbers of filters 100, 32, 64, 128, 256, 256, 512, and 512. Output of the final convolutional layer is flattened, and then fed to one fully-connected layer (512 nodes) followed by a softmax layer. Settings for training are the same as the aforementioned.
- (4) Object information only: Only the second part and the third part of Figure 2 are considered. This setting is used to demonstrate how well genre classification can be done if only object information is adopted.
- (5) The proposed structure illustrated in Figure 2.
- (6) The proposed structure without batch normalization layers: Comparing with the setting in (5), this setting is used to demonstrate the effectiveness of batch normalization.
- (7) The proposed structure with the third part replaced by SVM: The third part of Figure 2 acts as a classifier. It is interesting to investigate a classification scheme different from neural network. Therefore, we try to construct an SVM classifier based on the fused feature vectors, and evaluate performance obtained by this approach.

Table 2: Accuracy of movie genre classification based on different variants (%).

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Accuracy	7.31	14.05	14.34	15.79	18.73	17.70	15.36

Table 2 shows performance comparison of different variants. The first column shows that simple low-level features and nearest neighbor classifiers do not work well when the number of genre is increased to 23 (in [7], only a limited poster dataset consisting of 6 genres was used). Table 2(2) and Table 2(3) show that neural network-based methods provide much better performance than [7]. This confirms to the current research trend, especially when a large number of training data are available. Table 2(4) shows that the effectiveness of object information, even if the object detectors in YOLO are not specially constructed for movie posters. The proposed model combines visual appearance and object information, and the best performance can be achieved, as shown in Table 2(5). Comparing with Table 2(5) with Table 2(6), we verify the effectiveness of batch normalization [5]. Last, Table 2(7) shows that the constructed SVM classifier does not work better than the neural network classifier in our case.

Figure 5 shows sample classification results, including two success cases (top two) and two failure cases (bottom two). Figure 5(a) shows warm colors, which are highly related to romance. For Figure 5(b), the YOLO object detection module works quite well in detecting animals, which often appear in animation movies. On the other hand, both Figure 5(c) and Figure 5(d) are presented in a more abstract way. The visual appearance of Figure 5(c) looks a little bit horrible, and Figure 5(d) looks like showing an animation. The truth labels of Figure 5(c) are crime, drama, and mystery, where crime is actually related to horror. The truth labels of Figure 5(d) are short, comedy, and drama, which is quite difficult to infer if we just look at one poster image without watching the movie or the movie trailer.

5 CONCLUSION

We have presented a system to classify movie poster images into genres. A deep neural network is proposed to jointly consider visual appearance and object information, and a classifier is constructed to estimate the probabilities of a poster belonging to different genres. Multi-label classification is achieved by thresholding the estimate probabilities, with the thresholds adaptively determined by a grid search scheme. To evaluate the proposed system and facilitate future research, we collect a large-scale movie poster dataset consisting of

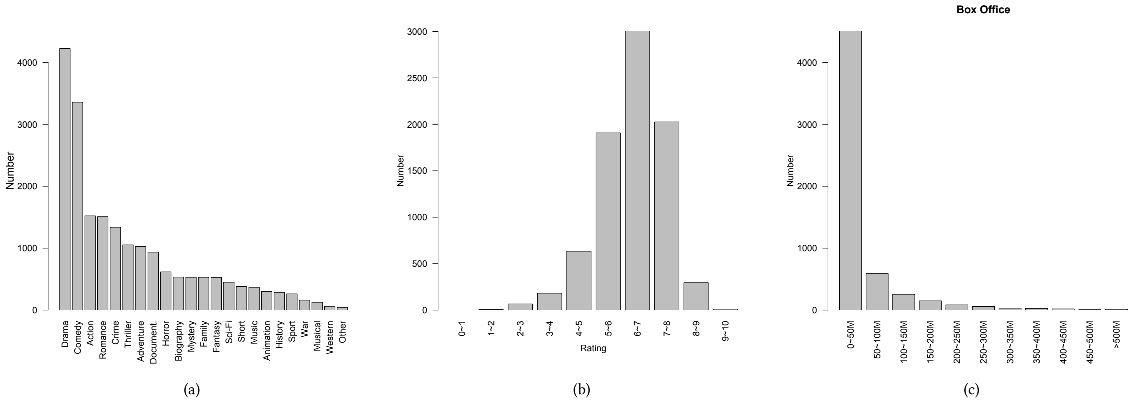
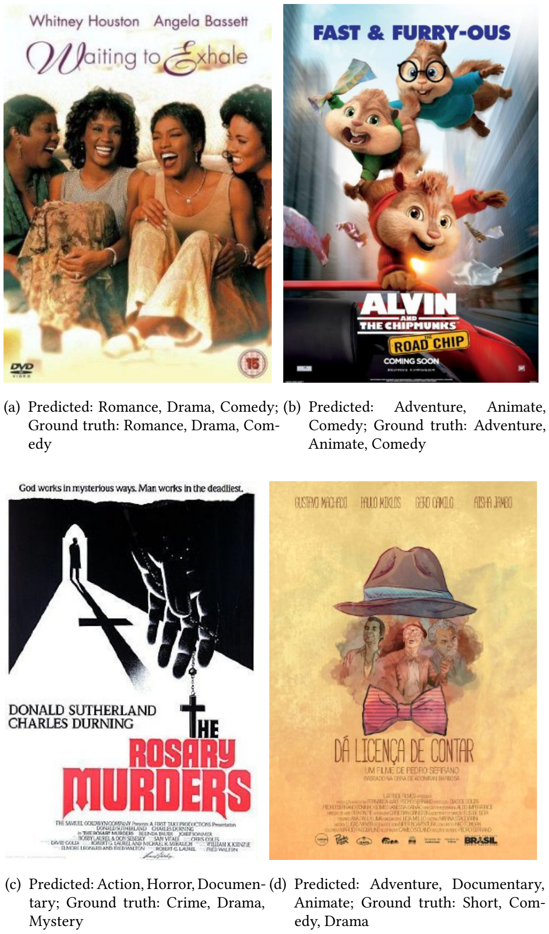


Figure 4: (a) The distribution of the numbers of posters in different movie genres. (b) The distribution of the numbers of posters with different ratings. (c) The distribution of the numbers of posters in different box office ranges.



(a) Predicted: Romance, Drama, Comedy; Ground truth: Romance, Drama, Comedy

(b) Predicted: Adventure, Animate, Comedy; Ground truth: Adventure, Animate, Comedy

(c) Predicted: Action, Horror, Documentary; Ground truth: Crime, Drama, Mystery

(d) Predicted: Adventure, Documentary, Animate; Ground truth: Short, Comedy, Drama

Figure 5: Success (top two) and failure (bottom two) classification cases.

8,191 distinct movies belonging to 23 different genres. The evaluation results show that promising performance over previous works can be achieved by the proposed model. In the future, we plan to improve classification performance by incorporating more information, and utilize the rich metadata to investigate more movie properties.

ACKNOWLEDGMENTS

This work was partially supported by the Ministry of Science and Technology of Taiwan under the grant MOST 105-2628-E-194-001-MY2 and MOST 106-3114-E-002-009.

REFERENCES

[1] 2017. *Box Office Mojo.* http://www.boxofficemojo.com.

[2] Huizhong Chen, Andrew C. Gallagher, and Bernd Girod. 2013. What's in a Name? First Names as Facial Attributes. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition.*

[3] Wei-Ta Chu and Chih-Hao Chiu. 2017. Predicting Occupation from Images by Combining Face and Body Context Information. *ACM Transactions on Multimedia Computing, Communications, and Applications* 13, 1 (2017), Article No. 7.

[4] Wei-Ta Chu and Yi-Ling Wu. 2016. Deep Correlation Features for Image Style Classification. In *Proceedings of ACM International Conference on Multimedia.* 402–406.

[5] Sergey Ioffe and Christian Szegedy. 2014. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *Proceedings of International Conference on Machine Learning.*

[6] Marina Ivasic-Kos, Miran Pobar, and Luka Mikec. 2014. Automatic Movie Posters Classification into Genres. In *Proceedings of ICT Innovation.* 319–328.

[7] Marina Ivasic-Kos, Miran Pobar, and Luka Mikec. 2014. Movie Posters Classification into Genres based on Low-level Features. In *Proceedings of International Convention on Information and Communication Technology, Electronics and Microelectronics.*

[8] Brian Kenji Iwana, Syed Tahseen Raza Rizvi, Sheraz Ahmed, Andreas Dengel, and Seiichi Uchida. 2016. *Judging a Book By its Cover.* https://arxiv.org/abs/1610.09204.

[9] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Proceedings of Advances in Neural Information Processing Systems.*

[10] Xin Lu, Zhe Lin, Hailin Jin, Jianchao Yang, and James. Z. Wang. 2015. Rating Image Aesthetics Using Deep Learning. *IEEE Transactions on Multimedia* 17, 11 (2015), 2021–2034.

[11] Zeeshan Rasheed and Mubarak Shah. 2002. Movie Genre Classification by Exploiting Audio-Visual Features of Previews. In *Proceedings of International Conference on Pattern Recognition.*

[12] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You Only Look Once: Unified, Real-Time Object Detection. In *Proceedings of IEEE*

Conference on Computer Vision and Pattern Recognition.

[13] Joseph Redmon and Ali Farhadi. 2016. YOLO9000: Better, Faster, Stronger. *arXiv preprint arXiv:1612.08242* (2016).

[14] Behjat Siddiquie, Rogerio S. Feris, and Larry S. Davis. 2011. Image Ranking and Retrieval based on Multi-Attribute Queries. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*. 801–808.

[15] Gabriel S. Simoes, Jonatas Wehrmann, Rodrigo C. Barros, and Duncan D. Ruiz. 2016. Movie Genre Classification with Convolutional Neural Networks. In *Proceedings of International Joint Conference on Neural Networks*.

[16] Kazuma Takahashi, Keisuke Doman, Yasutomo Kawanishi, Takatsugu Hirayama, Ichiro Ide, Daisuke Deguchi, and Hiroshi Murase. 2016. A Study on Estimating the Attractiveness of Food Photography. In *Proceedings of IEEE International Conference on Multimedia Big Data*.

[17] Jonatas Wehrmann and Rodrigo C. Barros. 2017. Convolutions through Time for Multi-label Movie Genre Classification. In *Proceedings of the Symposium on Applied Computing*. 114–119.

[18] Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. 2016. *Boosting Image Captioning with Attributes*. https://arxiv.org/abs/1611.01646.

[19] Howard Zhou, Tucker Hermans, Asmita V. Karandikar, and James M. Rehg. 2010. Movie Genre Classification via Scene Categorization. In *Proceedings of ACM International Conference on Multimedia*. 747–750.