Automatic Classification of Artist Visual Aesthetics: Linking Fashion and Genre

Andrew Kam

Department of Music Technology
Schulich School of Music
McGill University
Montréal, Canada

December 2019

A thesis submitted to McGill University in partial fulfillment of the requirements for the degree of Master of Arts.

2019/12/05

Abstract

Many musical preferences are strongly based on the visual aesthetics of artists. They are constructed through a combination of a music artist's geographical location, age, ethnicity, fashion, obscurity, promotional photos, videos, and many other items. We are constantly exposed to both images and music videos of artists through online sources such as music publications, social networks, and media streaming services. As such, images are a large and essential source of consuming visual aesthetics of music artists.

A novel study on artist similarity based on visual aesthetics was conducted using images. Specifically, promotional photos of artists were used, as they commonly provide an accurate depiction of the artist's branding and personality. Using a compiled list of artists taken from current music popularity charts, promotional photos of artists across four genres were retrieved from an online image source. The first stage of image analysis involved using neural networks specifically trained for object detection. The promotional photos were analyzed using two detection models in order to retrieve both the clothing garments portrayed by the artists, as well as the non-fashion objects that appear in the images. A second stage of machine learning was then applied to this new dataset. Common classifiers were trained on the extracted clothing and object text labels, and then used to make genre predictions on the unseen promotional photos.

It was found that the fashion items portrayed in the images acted as reasonable features in the genre classification task, predicting the correct genre with an accuracy significantly above chance. The object labels increased the classification precision, suggesting that the inclusion of items beyond clothing aids in this genre classification experiment. By visually clustering the images using a dimension reduction technique, it was possible to observe similar clothing items and objects that defined each genre. This provided insight into the visual stereotypes and fashion trends that are affiliated with each genre.

Résumé

Plusieurs préférences musicales sont fortement fondées sur l'esthétique visuelle des artistes. Elles sont construites à l'aide d'une combinaison de la localisation géographique de l'artiste musical, ainsi que de son âge, son ethnicité, son style vestimentaire, son exposition médiatique, ses photos promotionnelles, ses vidéos et d'autres éléments. Nous sommes constamment exposés aux images et vidéoclips d'artistes à travers des sources en ligne telles les publications musicales, réseaux sociaux et services de diffusion media. Ainsi, les images sont une grande source essentielle de consommation d'esthétique visuelle liée aux artistes musicaux.

Une nouvelle étude traitant de la similarité des artistes selon leur esthétique visuelle été conduite à l'aide d'images. Précisément, les photos promotionnelles d'artistes ont été utilisés, car elles indiquent, la plupart du temps, une représentation précise de l'image de marque et de la personnalité de l'artiste. En utilisant une liste compilée d'artistes à l'aide de classements récents de popularité musicale, les photos promotionnelles de ceux-ci à travers quatre genres ont été obtenus d'une banque de données en ligne. La première étape d'analyse d'image a impliqué l'utilisation de réseaux de neurones spécifiquement entraînés pour la détection d'objets. Les photos promotionnelles ont été analysées à l'aide de deux modèles de détection pour récupérer à la fois les vêtements représentés par les artistes, ainsi que les autres objets non-reliés au style vestimentaire apparaissant dans l'image. Une deuxième étape d'apprentissage machine a ensuite été appliquée sur ce nouvel ensemble de données. Des classificateurs communs ont été entraînés sur les étiquettes des vêtements et les objets. Ils ont été utilisés pour prédire le genre sur les photos promotionnelles non-vues.

Il a été découvert que les items reliés au style vestimentaire dépeint dans les images agissent comme une caractéristique raisonnable dans la tâche de classification de genre, réussissant à prédire correctement celui-ci avec une précision significativement plus élevée que le hasard. Les étiquettes des objets ont augmenté la précision de la classification, suggérant

que l'inclusion d'items autres que les vêtements bénéficie à cette expérience de classification de genre. En regroupant visuellement les images à l'aide d'une technique de réduction de dimension, il a été possible d'observer des articles vestimentaires et objets similaires qui définissent chaque genre. Cela a permis de mieux comprendre les stéréotypes visuels et les tendances de la mode qui sont affiliées à ces derniers.

Acknowledgements

I would like to thank my advisor, Ichiro Fujinaga. His guidance, direction, mentorship, and feedback during the writing process of this thesis was crucial for its inception and completion. Under his supervision, working in the Distributed Digital Music Archives and Libraries Lab also allowed me to gain the knowledge required to complete this study. The enthusiasm for music information retrieval in the lab inspired me throughout my time in the music technology program. A special thanks to Gabriel Vigliensoni for his advice with this thesis, and with the various projects in the lab.

I would also like to thank my family, who supported me in my move to Montréal to pursue my passion in music, and throughout my studies at McGill. I would like to thank my friends in Calgary, Vancouver, and Montréal for their encouragement in my academic activities.

Author Contributions

Under the supervision of Professor Ichiro Fujinaga, the author of this thesis, Andrew Kam, was responsible for devising the idea of this study, designing the experiment, gathering the datasets, executing the procedure, and analyzing the results. All chapters were written by the author.

Contents

List of Figures

List of Tables

List of Acronyms

AI	Artificial Intelligence
API	Application Program Interface
BAIR	Berkeley AI Research
CNN	Convolutional Neural Network
GMM	Gaussian Mixture Model
GPU	Graphics Processing Unit
ILSVRC	ImageNet Large Scale Visual Recognition Challenge
JSON	JavaScript Object Notation
k-NN	K-Nearest Neighbor
MFCC	Mel-Frequency Cepstral Coefficients
MIR	Music Information Retrieval
NB	Naive Bayes
PCA	Principal Component Analysis
SGD	Stochastic Gradient Descent
SVM	Support Vector Machine
t-SNE	t-distributed Stochastic Neighbor Embedding
TF-IDF	Term Frequency-Inverse Document Frequency
URL	Universal Resource Locator

1 Introduction

The construct of fashion, while constantly evolving, has changed from a certain lifestyle observed in fifteenth-century high society to a specific way of crafting clothes in the sixteenth century (Strähle 2017). It has a history of transporting social meaning, such as class, gender, or religion. However, Strähle (2017) explains that it has recently developed to become more of an expression of a certain lifestyle than a tool to identify classes. Fashion theory is centered on clothing fashions, with perspectives on its evolution and cyclical nature. Though many products have life cycles and evolve over time, the cycles of fashion are distinct and prominent. By definition, fashions are temporary cyclical phenomena adopted by consumers for a particular time and situation (Sproles 1981). As such, fashion can be defined as an expression of contemporary taste.

Both fashion and music have a strong social impact, with the act of dressing and our interaction with music providing identity-building features, especially for young generations (Calefato 2001). The codependency between the music and fashion industries has shaped much of contemporary consumer culture, with fashion and style being core to the intertextual taste-sharing between the two industries. Miller (2011) explains that this relationship is a natural consequence of a consumer culture where the musician has become a powerful signifier of contemporary desires. Fashion has also provided an avenue for fans to express their affiliation with a musician, building a group identity among fans with similar musical tastes.

Na and Agnhage (2013) found that consumers with similar taste in music felt connected and developed similar aesthetic preferences. Furthermore, the researchers found that certain musical genres induced stronger correlations between fashion styles and music than others.

Miller (2011) outlines a historical moment when the musician's role carried expectations of a particular lifestyle. The bohemian lifestyle, driven by class struggles in the working-class against the bourgeoisie, increased the prominence of a particular identity and way of living as the defining factor of artistic status. This notion of authenticity has become core to the musician's identity, and key to the cultural understanding of the musician (Miller 2011). In popular music, the roots and authenticity of particular styles are used to champion the superiority of certain genres, such as folk, blues, and country, over the artificial character of commercial and mainstream popular music (Strinati 2004). Thus, the concept of originality and authenticity are deployed as marketing strategies to appeal to particular segments of the audience in popular music, making the musician a fruitful marketing tool with reach beyond the music industry itself. This invites opportunities for the music and fashion industries to intersect, each benefiting from the idea of the bohemian lifestyle of the musician.

In popular music culture, many of our musical preferences are strongly based on visual aesthetics, and as a result, it is difficult for one to mentally separate the image of an artist from their audio output (Negus 1992). One is constantly exposed to both images and music videos of artists through online sources such as music publications, social networks, and media-streaming platforms. As such, images are a large and essential source of consuming the visual aesthetics of music artists. In contemporary culture, the promotional image has become almost as important, if not more so, than the object it promotes (Miller 2011). Placing a musician in promotional images, outfitted in a fashion style such as one associated with the bohemian lifestyle, is a means of escape for the consumer.

The extent in which fashion is promoted and consumed varies between music genres.

When a genre gains popularity, so does the corresponding fashion style that is linked with the particular style of music (Na and Agnhage 2013). Compared with other genres, dance music videos has historically contained the most fashion-oriented imagery, involving clothing, jewelry, and hairstyles (Englis, Solomon, and Olofsson 1993). In contrast, classic rock has generally ranked low with imagery concerning fashion. As well, visual stereotypes, ingrained in us through the media, allow us to identify such genres without listening. In the task of music genre classification, taking into account other modalities, such as visual imagery, is vital to the advancement of the domain (Liem et al. 2011). Promotional photos of artists provide an excellent display of fashion, and as such are appropriate sources of imagery for classification.

This study attempts to bridge these visual indicators with the musical genre they promote. If one can look at a promotional photo of an artist and immediately designate the associated genre of music, it is likely that a fusion of various visual features generates this implication, and it is also possible that similarities exist within a single genre at a feature level. The extraction and analysis of visual features, such as objects and clothing items, may provide insight into the indicators that are important determining factors in dividing genre boundaries. With recent suggestions of boundary blurring between genres in popular music (Van Venrooij 2009; Silver, Lee, and Childress 2016), it would be of particular interest to observe transformations in the fashion choices commonly associated with a single genre.

The focus of this thesis will be to examine these visual features ingrained into promotional photos of artists. While all the objects contained in the images will be taken into consideration, the majority of this study will concentrate on the clothing items that are put on display by the artists. By dissecting the distribution of clothing items across multiple genres, we will be able to gather information about the differences in fashion trends between the musical styles. This will be completed by extracting clothing labels, such as the styles and

classes of the garments, from online photographs by using reputable deep learning techniques specialized in analyzing visual imagery. By storing this dataset of labels in a database, they will then be filtered and manipulated to gather knowledge about the current fashion trends among music artists. These labels will then act as inputs into a second stage of machine learning, where well-known classification algorithms will be applied in order to predict the music genre. The results of this stage will allow us to conclude if the portrayed fashion in promotional photos is substantial enough to make accurate predictions of music genre, thus allowing us to analyze the impact of the promotional photo as a representation of the artist and their music, as well as study the strength of visual imagery for each specific genre.

1.1 Thesis Structure

The following outlines the content of the upcoming chapters in this thesis:

Chapter 2 provides a background of genre classification, and a history of the studies that have been completed in the field. It also dives into the advancement of multimodal strategies in music information retrieval (MIR), and how features outside of audio have been used to aid in the task of genre classification. Finally, the chapter provides a brief overview of clothing detection in images.

Chapter 3 details the high-level approach that was used for the experiment. It provides a brief rundown of each stage of the experiment, and how they contributed to the end results.

Chapter 4 gives a detailed description of the experiment, including the software libraries and technologies used to execute each step, as well as their configurations. The end of the chapter includes a comprehensive breakdown of the results.

Chapter 5 discusses and interprets the results of the experiment, coming to conclusions on the success of study. A brief overview of possible future work ends the chapter.

2 Background

Although there is not a long history of musical genre classification based solely on visual imagery, the idea of automatic classification using other types of data modalities is well researched and documented. This chapter will review successful methods of automatic genre classification, recent experiments that incorporate multiple modalities to increase the performance of genre prediction, as well as techniques for image classification, specifically in the clothing domain.

2.1 Music Genre Classification

Musical genres exist as a result of humans grouping common audio characteristics and creating labels to categorize them. There are no strict boundaries that define a specific genre, but common properties such as instrumentation, rhythmic structure, and harmonic content are often shared by the music within one. As such, much of the research in musical genre recognition in the field of music information retrieval (MIR) has involved the extraction of features from audio signals. For example, an early study by Tzanetakis and Cook (2002) found that timbral texture, rhythmic content, and pitch content features were suitable for characterizing a segment of audio. These features were extracted from representative excerpts in a genre hierarchy, consisting of 20 musical genres and three speech genres, and used as

input into a variety of classifiers, including a simple Gaussian classifier, Gaussian mixture model (GMM), and K-nearest neighbor (k-NN) classifiers. With the feature set, an accuracy of 61 percent for 10 musical genres was achieved, which was comparable to the performance measured in genre classification by humans.

2.2 Multimodal Strategies

While success in the task of genre classification using audio features has been notable, other cultural and high-level features have also been taken into account (McKay and Fujinaga 2006; Whitman and Smaragdis 2002; Fell and Sporleder 2014; Sturm 2012; Neumayer and Rauber 2007). As highlighted by McKay and Fujinaga (2006), cultural information beyond the scope of musical content is of paramount importance in defining genre, and therefore should be integrated into the task of automatic genre classification. It was shown that combining features from different types of data, including symbolic, lyrical, and cultural, improved average classification accuracy (McKay et al. 2010). Using jMIR, a suite of software tools for automatic music classification, audio, symbolic, and cultural features were extracted from a dataset. The dataset consisted of audio and MIDI recordings, associated lyrics, and cultural information based on Yahoo![1] co-occurrence page counts and Last.fm[2] user tags. It was found that the cultural features were especially effective in improving classification accuracies, but lyrical features performed poorly relative to the other types, most likely due to noise in the mined lyrical transcriptions. Overall, excellent classification accuracies were obtained, with a performance of 89 percent on a 10-genre taxonomy.

With advancements in efficient computational methods, it has become possible to analyze visual information alongside audio and other associated features for annotation and

1. https://yahoo.com
2. https://www.last.fm

classification purposes. Since music takes the form of multiple modalities, a perspective on content outside the audio domain, as well as cross-domain collaboration is the key to successful MIR solutions (Liem et al. 2011). An experiment by Nanni et al. (2016) used a fusion of both acoustic and visual features for the purpose of automated musical genre recognition. As shown in the workflow diagram in Figure 2.1, the spectrogram representation of an audio signal was constructed, and the resulting image divided into sub-windows. For each sub-window, visual features were then extracted by calculating texture descriptors and bag-of-features projections. The texture descriptors included uniform local binary pattern and its Fourier histogram, as well as local phase quantization. The bag-of-features clustered localized features to create a "codebook" (Fei-Fei and Perona 2005). Compared to classification solely based on timbre-based audio features, such as mel-frequency cepstral coefficients (MFCC), it was found that genre classification performance improved when fused with the visual features.

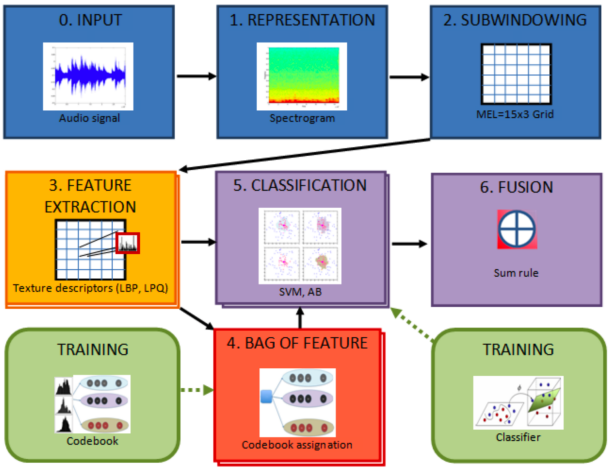


Fig. 2.1 Workflow of classification system based on visual features (Nanni et al. 2016).

Other experiments have shown that the combination of various modalities, including audio, images, and text, improves the accuracy of genre classification (Mayer and Rauber 2010).

For example, studies by Oramas et al. (2017, 2018) used a large-scale dataset consisting of cover images, text reviews, and audio tracks. With this multimodal dataset, genre prediction was performed using a deep-learning approach on the different data modalities, and also on combinations of the data types. The audio was evaluated using convolutional neural networks (CNN) to learn the features from spectrograms. A vector space model approach, an algebraic model for representing text documents as vectors of identifiers, was used to create feature vectors from the text reviews. Album covers were analyzed using deep residual networks (ResNet) for image classification, and results clustered using t-Distributed Stochastic Neighbor Embedding in Figure 2.2.

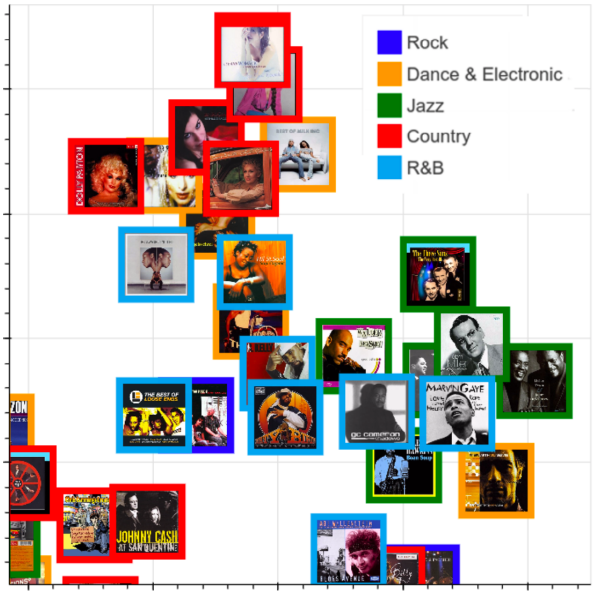


Fig. 2.2 Clusters of album covers of five of the most frequent genres using t-SNE (Oramas et al. 2017).

It was found that the text-based classification on the reviews outperformed all other modalities, while the image-based approach produced the lowest performance results. On the other hand, a multimodal approach of combining all three modalities, and using the combined feature vectors as input into a multilayer perceptron, outperformed all the single-

type approaches.

Album covers, which are designed to convey a message consistent with the music and image of an artist, have been used in various ways in MIR. Studies on the correlation between music audio and album cover art have been conducted to link the two different media types (Brochu, De Freitas, and Bao 2003; Mayer 2011). Training a set of self-organizing maps, a type of artificial neural network that produces a low-dimensional, discretized representation of the input space, with audio features such as rhythm patterns, histograms, and statistical spectrum descriptors, Mayer (2011) demonstrated that musical similarity is reflected in album covers. Album covers from the classical genre frequently used photos of people against a basic, white background, as shown in Figure 2.3.

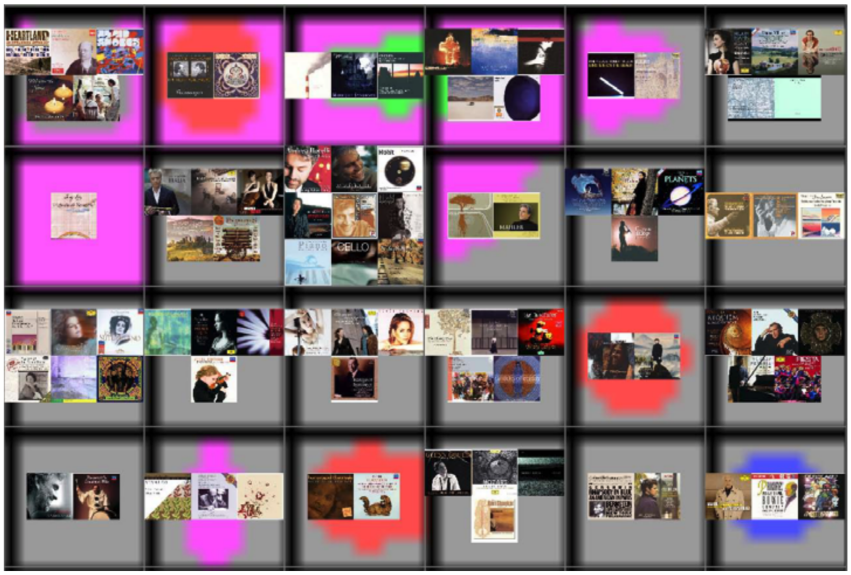


Fig. 2.3 Album covers in a cluster of classical music, using statistical spectrum descriptors as audio features (Mayer 2011).

Jazz and country albums followed similar trends, with portraits of the artists in the covers, but commonly against darker background colors. On the other hand, covers from the gothic and alternative rock genres did not contain people as frequently as the previously

discussed genres, and also appeared heavily altered and artificial. When organizing the map with image features from the album covers, such as color histograms and color names, it was found that no region of the map contained a continuous area of similar music. It was concluded that more powerful image-feature descriptors were necessary to make this inverse correlation.

Advances in extracting such image-feature descriptors form much of the work in the field of content-based image retrieval (Deselaers, Keysers, and Ney 2008). An image annotation system that calculated artist similarity and predicted genre based on promotional photos was found to perform successfully by Libeks and Turnbull (2010). The system combined seven different color and texture features from images using joint equal contribution, a method to combine and calculate distances using different descriptors (Makadia, Pavlovic, and Kumar 2008). This resulted in image-to-image distances in the range of 0 to 1, where 0 denoted identical images, and 1 indicated the most dissimilar pair of images. Using promotional photos from Last.fm, the system propagated genre labels from artist to artist, finding a notion of music similarity based on visual appearance. The classification of some genres, such as dance, classical, indie pop, and metal and its subgenres, performed better than others. The four most successful genre tags contained the word "metal", indicating that there was a specific visual appearance that made them easily identifiable. On the other hand, 10 of the 50 genre tags, including country, folk, and funk, did not perform better than chance, meaning the color and texture features may not have been adequate for extracting relevant information. It was suggested that a system that could detect concrete objects within the images would have been useful for the purposes of the experiment. The authors completed a similar study that included album cover artwork along with promotional photos, with comparable results (Libeks and Turnbull 2011). The studies proved that music-related images could be a source of information for semantic music annotation, and that they encode valuable information

that is useful for contextualizing music.

Music videos, a multimedia type that has become increasingly accessible through video-streaming platforms, also plays a significant role in music marketing. The visual stereotypes we have learned to expect in promotional photos are replicated within music videos, and thus are a large source of data in MIR. Along with extracting low-level image processing features, a more high-level approach of visual concept detection can be used to analyze content in music videos.

An experiment by Schindler and Rauber (2016) used a dataset of 800 tracks of eight clearly-defined subgenres. By taking a CNN model pre-trained on the 1,000 synonym sets (synsets) of the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) (Russakovsky et al. 2015),[3] and applying it to every frame, videos were decomposed into concrete objects such as guitars, vehicles, and landscapes. The predicted probabilities of the objects were extracted and then placed into a feature vector for each track. As seen in Figure 2.4, the salient items found in the country genre were ones that are stereotypically associated with the category, such as "cowboy hat" and "pickup truck". Dance music contained items that were associated with revealing clothing, such as "brassiere" and "bikini". Videos of the metal and opera genres both displayed a relatively larger number of musical instruments, such as "electric-guitar" and "drum" in metal, and "flute" and "oboe" in the opera category.

Three classifiers were then evaluated on the feature set using stratified 10-fold cross-validation: linear support vector machines (SVM), k-nearest neighbors (k-NN), and naive Bayes (NB). Results showed high classification accuracies using the ImageNet model, with an accuracy of 74 percent for the SVM classifier. Low-level image processing features, such as global color statistics, global emotion values, colorfulness, and lightness fluctuation patterns, were also applied to the frames of the videos for classification purposes. It was found that the

3. http://image-net.org/challenges/LSVRC/2012/browse-synsets

Country	Dance	Metal	Opera	Reggae
1. cowboy hat	1. brassiere	1. spotlight	1. theater curtain	1. seashore coast
5. drumstick	3. maillot	2. electric-guitar	3. hoopskirt	2. academic gown
8. restaurant	4. lipstick	4. drumstick	5. stage	3. capuchin
9. tobacco shop	9. seashore coast	6. matchstick	11. flute	5. black stork
10. pickup truck	10. bikini	7. drum	19. harmonica	7. sunglasses
11. acoustic guitar	15. sarong	8. barn spider	21. marimba	8. orangutan
13. violin fiddle	16. perfume	10. radiator	25. oboe	9. titi monkey
16. jeep landrover	17. trunks	12. chain	26. french horn	10. lakeshore
18. tractor trailer	18. ice lolly	14. grand piano	27. panpipe	11. cliff drop
19. tow truck	19. pole	23. spider web	30. grand piano	17. elephant
21. minibus	20. bubble	24. nail	31. cello	23. steel drum
23. electric guitar	30. miniskirt	28. brassiere	48. pipe organ	24. macaw
33. thresher	42. feather boa	37. loudspeaker	55. harp	25. coonhound

Fig. 2.4 Salient ILSVRC synsets found in a selection of genres, listed in descending order according to differences in frequency from other genres (Schindler and Rauber 2016).

visual vocabulary model outperformed the low-level feature model by a large margin, which was recorded at 50 percent. The success of the high-level evaluation supported the initial hypothesis that music videos make use of easily identifiable visual concepts. A multimodal approach was also taken to investigate the effect of different modalities on classification performance. Common audio features in MIR, such as MFCC, as well as chroma and psychoacoustic music descriptors, were extracted from the same music video dataset. As the genres were well differentiated by spectral and rhythmical characteristics, accuracies over 90 percent were achieved using combinations of the audio features. Lastly, the dataset was evaluated using a combination of both the audio and video features. The multimodal approach produced results that showed noticeable improvements over the optimal combination of audio-only features, revealing that incorporating visual concepts can improve the performance of genre classification.

2.3 Clothing Detection

Clothes classification has immense potential value to the fashion industry, and as a result, extensive research has been allocated to the task. However, parsing clothing in photographs

has been a challenging problem due to the large diversity of garment items, deformation of garments due to its soft material, variations in configuration, and garment appearance and layering (Yamaguchi et al. 2012; Yamaguchi, Hadi Kiapour, and Berg 2013; Liu et al. 2016; Hadi Kiapour et al. 2015). In order to address these issues, Yamaguchi et al. (2012) used a large number of garment types (53) to explore techniques in parsing out pieces, and exploited the relationship between clothing and the underlying body pose. This relationship was evaluated in both directions: estimating clothing given estimates of pose, and estimating pose given clothing estimates. Pose estimation can be considered as an extension of work on flexible part models (Yang and Ramanan 2011), and incorporating estimates of clothing considered as an additional feature on top of that work. Using 685 labeled photographs from Chictopia,[4] a social networking site for fashion bloggers, a dataset with 56 different clothing labels was created. As seen in Figure 2.5, for each image, superpixels were first extracted for contour detection and image segmentation (Arbelaez et al. 2011). Second, pose configuration was estimated using a mixture model that captured contextual relations between body parts (Yang and Ramanan 2011). Clothing labels were then predicted for every segment, taking into account clothing appearance and location with respect to body parts. Lastly, poses were re-estimated as the calculated clothing predictions could have potentially improved the original estimations.

Taking into account pose information, the model was able to achieve 89 percent pixel accuracy, which is the percentage of pixels in the image that were correctly classified. With no pose information, the clothing parsing performance dropped to 86 percent. As well, using the estimated clothing labels led to a re-estimation probability of 87 percent. Given true clothing labels, the re-estimation probability increased to 90 percent, demonstrating the potential usefulness of incorporating clothing into pose identification.

4. http://www.chictopia.com

(a) Superpixels (b) Pose estimation (c) Predicted Clothing Parse (d) Pose re-estimation

Fig. 2.5 The clothing parsing pipeline of parsing images into superpixels, pose estimation, clothing extraction, and optional re-estimation of pose using clothing estimates (Yamaguchi et al. 2012).

2.4 Conclusion

As outlined in this chapter, a variety of approaches have been taken to increase the performance of music genre classification. Starting with a basis of audio features, other modalities have been incorporated into studies, including text, images, videos, and other cultural and high-level features. The addition of these data types have generally increased classification performance, and have become integral to the success of the task. While the goal of many of the studies was to optimize genre classification results, the intent of this thesis is to explore the fundamental relationship between fashion and music genre, a connection that was touched upon by Schindler and Rauber (2015). That study used a variety of video features, including low-level color descriptors and high-level objects found in music videos. In order to focus on the fashion portrayed by music artists, a similar approach will be taken to extract high-level objects, specifically clothing garments, from a dataset of promotional photos.

3 Methodology

The general concept of this study is to employ classification techniques on photos of artists to predict music genre. Specifically, promotional photos have been chosen as the base dataset of the experiment, as they are commonly distributed as part of press kits by the artists or their publicity management. As such, the image quality of the photos are professional, which will be consequential to the performance of the experiment, as machine vision tasks are susceptible to quality distortions (Dodge and Karam 2016). As well, the displayed visual aesthetics in promotional photos provide an accurate depiction of the artist's branding and personality. These photos may be sourced from physical or online magazines, blogs, social media, advertisements, etc. Other categories of images, such as concert and paparazzi photos, are not directly curated by the artists or their management. As a result, the image quality may be poor, and the aesthetic expression incorrectly portrayed.

There are many factors that differentiate the aesthetic expression between different images, including colorfulness, composition, texture, and statistical measurements (Khan and Vogel 2012). However, this study focuses specifically on the fashion depicted in the photographs, for motives outlined in previous chapters. Clothing garment labels will be extracted from the photos using well-known object detection models, and then recorded for analysis. These labels will then form the features of another classification stage, in which music genre will be predicted based on the detected clothing items.

3.1 Dataset

3.1.1 Image Source

The foundation of the experiment depends on a large, high-quality set of promotional photos of music artists. Online sources of promotional photos that were investigated include Last.fm,[1] Google Images,[2] Spotify,[3] and Instagram.[4] While the images residing on Spotify and Instagram are carefully maintained by the artists or their management, and Google Images attempts to find the most relevant pictures from any source, Last.fm contains images uploaded by the public. In other words, any user with an account can upload an image to an artist's profile on the website. An account is free and available to anyone who registers. Last.fm is a platform that constructs a profile of each user's musical taste by recording information about their played tracks on various platforms, including Spotify, Deezer,[5] and Tidal.[6] This data is then displayed on the user's public account, and used in a music recommendation system to suggest similar artists and tracks for the user. This process is known as "audioscrobbling". Each music artist has their own unique identifier, which is referenced whenever the audioscrobbler receives notification that one of the artist's tracks has been played. Each artist also has a public profile page with these track play counts, as well as other information, such as a biography and aforementioned promotional photos, as shown in Figure 3.1. Images are uploaded by the public, where the community must ensure that the images are of the correct artist and of acceptable quality. This is completed by an upvoting/downvoting feature available for each image. While there are no strict rules on the types

1. https://www.last.fm
2. https://images.google.com
3. https://www.spotify.com
4. https://www.instagram.com
5. https://www.deezer.com
6. https://tidal.com

of photographs that can be uploaded, the community has curated the images so that promotional photos form the basis of the collections, making Last.fm a very rich source of artist images. Using their API, promotional photos can be extracted from the Spotify platform as well, but limited to only one image per artist. The majority of popular artists have their own Instagram profiles, providing a look into their lifestyles. However, the photographs are not always of themselves, which is a requirement of this experiment. Due to the limitations of these other platforms, the promotional photos located on Last.fm will be used as the data source for this experiment.

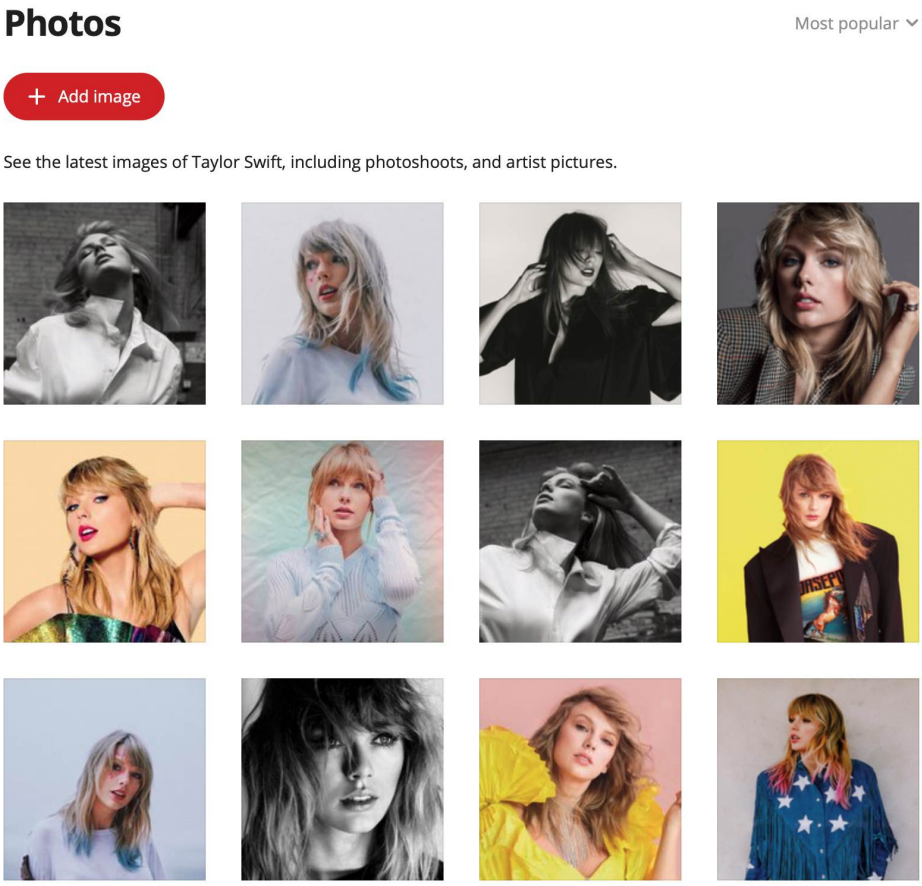


Fig. 3.1 A sample of images provided on an artist's profile on Last.fm.

3.1.2 Artist Selection

One of the goals of this experiment is to investigate the current fashion trends within genres. In order to analyze the most recent trends, an updated list of the most popular artists must be compiled for each genre. However, the quantification of popularity is especially difficult with the recent emergence of online music streaming. Each streaming platform, such as Spotify and Apple Music,[7] has its own calculated popularity charts based on play counts. Last.fm also aggregates the top play counts for each genre tag. Unfortunately, these platforms do not take into account more traditional figures, such as physical sales. Billboard's genre-based charts[8] take sales, streaming, and radio airplay into account. Therefore, for this experiment, the following Billboard charts will be used to compile artist lists for three genres: Top Country Albums, Top R&B/Hip-Hop Albums, and Top Rock Albums.

Country, R&B/hip-hop, and rock were chosen as genres for this experiment, as they promote contrasts in recognizable visual imagery, and combine for 63 percent of popular music consumption in the U.S.[9] Due to the broad categorization of Western pop music and its overlap with other genres, the subset of K-pop will also be added as a category. K-pop is popular music originating in South Korea, commonly characterized by strong audiovisual elements in its marketing. The Gaon Music Chart[10] tabulates the weekly popularity of songs in South Korea, with the aim to create a national chart similar to Billboard. There are two primary charts: Gaon Album Chart and Gaon Digital Chart. Since music releases for K-pop artists are heavily focused on singles rather than the more traditional album cycles, the Gaon Digital Chart will be used for the experiment. This chart provides singles rankings based on the aggregate of downloads and streaming. With this addition, a total of four genres will be

7. https://www.apple.com/apple-music
8. https://www.billboard.com/charts
9. https://www.nielsen.com/us/en/insights/reports/2018/us-music-mid-year-report-2018.html
10. http://www.gaonchart.co.kr

used for the experiment (see Table 3.1).

Genre	Publication	Chart
Country	Billboard	Top Country Albums
R&B/Hip-Hop	Billboard	Top R&B/Hip-Hop Albums
Rock	Billboard	Top Rock Albums
K-Pop	Gaon	Digital Chart

Table 3.1 The selected genres and their chart sources.

3.2 Analysis

3.2.1 Image Classification

Classification using Convolutional Neural Networks (CNNs) has proven to be a very successful machine-learning technique in the field of computer vision research (Krizhevsky, Sutskever, and Hinton 2012). Such deep neural networks are especially remarkable in visual concept detection, making it possible to identify objects within an image. Access to a fashion-oriented, pre-trained CNN model is crucial for this study, as the procedure and performance requirements necessary to train a model from the ground up is very time intensive and resource heavy. While a more advanced and heavily-trained clothing detection model could be used for this task, such as the one described in Section 2.3 and carried out by Yamaguchi et al. (2012), a goal of the experiment is to mimic flipping quickly through artist photos for genre identification, and a simpler, pre-trained clothing model will be suitable for this objective. As well, the speed performance gained in using a lower processor-intensive image classification model will allow the analysis of more images. A more accurate clothing detection model could be considered as an option in future research.

The ImageNet Large Scale Visual Recognition Challenge (ILSVRC) is an annual image classification task, focused on object identification (Russakovsky et al. 2015). The challenge

provides a collection of 1,000 synsets to be used as a visual object vocabulary, with each synset consisting of one or more English synonyms. GoogLeNet (Inception v1), the winner of the ILSVRC 2014 challenge, is a CNN with 27 layers (Szegedy et al. 2015). The middle of the network contains an inception layer, consisting of a 1x1 convolutional layer, 3x3 convolutional layer, and 5x5 convolutional layer, filtered and concatenated together for input into the next stage. A 1x1 convolutional layer for dimension reduction and a parallel 3x3 max pooling layer complete the inception module, while global average pooling is used near the end of the network. With its high accuracy in object detection, GoogLeNet would be able to identify the visual concepts within promotional photos. As well, using a model specifically trained on clothing, the various fashion items could be extracted from the images. As such, a GoogLeNet model pre-trained on clothing classes will be used to process the promotional photos retrieved from Last.fm.

A common issue with CNNs is that increasing the network depth within the architecture results in networks that are very difficult to train due to the vanishing gradient problem (Hochreiter 1998). As a network goes deeper, its performance gets saturated or even starts to degrade. Residual networks (ResNets) are artificial neural networks that escape this issue by skipping over one or more layers. As such, a big advantage of ResNets is while increasing network depth, they avoid the negative impact of the vanishing gradient problem and performance is unaffected. This architecture was introduced in 2015 by Microsoft, and won ILSVRC 2015 with an error rate of 3.6% (He et al. 2016).

ResNet-50 is a residual network model with a 50-layer architecture. While more layers may reduce the error rate, it also decreases the speed performance of a model. As such, a model of ResNet-50, pre-trained on all 1,000 classes of the ImageNet ILSVRC (Appendix B), will be included in the experiment. This will be added to catch any objects that fall outside of the classes in the pre-trained fashion model, and to discover if such additions

could increase the prediction performance.

DeepDetect,[11] a server and high-level API for running machine-learning services, was used to run the GoogLeNet and ResNet-50 image classification models for this study. It is an open-source project, free to download and install, and implements support for supervised and unsupervised deep learning for images, text, and other types of data. A server such as DeepDetect is very useful for machine-learning tasks, as the environment setup for deep-learning frameworks can often become difficult and time consuming due to library dependencies, GPU requirements, and issues with certain operating systems. These frameworks are beneficial to deep neural-network research as they abstract the low-level details of machine-learning algorithms, allowing users to concentrate on the overall logic of the applications. They also aid in pre-processing data and building and training models. By encapsulating these frameworks within a server and available through an API, DeepDetect provides an easy way to access and manage such models. Caffe[12] is a deep-learning framework supported by DeepDetect. Developed by Berkeley AI Research, it provides easy configuration and fast performance for research experiments. It also encourages a standard distribution format for its models, and provides access to trained models in its Model Zoo[13] community. Due to this accessible collection of pre-trained models, Caffe was used for this experiment. DeepDetect provides free access to some of these specialized, custom pre-trained models for various categories.[14] Fashion, gender, sports, cars, and buildings are examples of some of the custom image models available for use. The pre-trained fashion model consists of 304 classes, listed in Appendix A. It was trained using the GoogLeNet CNN architecture, and completed with the following details:

11. https://www.deepdetect.com
12. http://caffe.berkeleyvision.org
13. http://caffe.berkeleyvision.org/model_zoo.html
14. https://www.deepdetect.com/applications/model/

- Training/test set split of 90/10
- Dataset of images with dimensions 224 x 224 pixels
- Trained with 300,000 steps
- Learning rate with step control
- Stochastic gradient descent (SGD) as optimizer
- Data augmentation with mirroring

As well, DeepDetect provides access to a ResNet-50 model, pre-trained on all 1,000 classes of the ImageNet ILSVRC (Appendix B). This was used to catch the objects that fell outside the classes of the fashion model, and to observe if doing so would increase the prediction performance.

3.2.2 Genre Classification

Sentiment analysis in machine learning uses computational tools to extract, quantify, and categorize affective states, emotional tones, and attitudes in pieces of text. The core task in a binary problem is to classify the polarity of text in a document, positive or negative, or categorize the text into one of many emotional states for a multi-class problem. The approach taken for genre classification in this study mirrors a common approach used in sentiment analysis (Yang et al. 2007). A bag-of-words model treats the text in a document as a multiset of words, disregarding the order but retaining the multiplicity. In such a vector representation, the number of items in a vector representing a document corresponds to the number of words in the vocabulary (Pang, Lee, and Vaithyanathan 2002). Words in a document are scored, and the values placed in their corresponding locations in the representation. After the processing of promotional photographs using the deep neural network,

each image will be left with a collection of clothing item labels. By treating each image in the experiment as a document, and the predicted clothing items from the CNN model output described in Section 3.2.1 as the document text, a bag-of-words representation can be constructed, with the classes of fashion items forming the vocabulary of the model. This new dataset will then be used as input to a second stage of classification. Naive Bayes (NB) and support vector machines (SVM) are common classifiers used in text sentiment analysis (Joachims 1998; McCallum and Nigam 1998). Both of these classifiers will be trained on the bag-of-words representation of clothing items. These trained models will then be used to make genre predictions on unseen promotional photos.

This two-stage classification framework provides the ability for genre prediction, and will reveal the clothing items that influence the results. This will ensure control over the features being used for prediction, as well as yield insight into the fashion associated with each genre.

4 Experiment Implementation, Results, and Evaluation

This chapter details the implementation and results of the genre-classification framework outlined in Chapter 3. The experiment is outlined in detail, including the steps in dataset construction, object detection, and genre classification. The construction of the image dataset is outlined, detailing the process used to retrieve the required promotional photos using the curated list of artists for each genre. As described in Section 3.2.1, both clothing items and non-clothing objects were to be considered as features for extraction from the photos, and the image evaluation procedure used to detect both the clothing items and objects is then described. An analysis of the detected clothing and object labels is completed, breaking down the most frequently found items for each genre. Selected classifiers were then used to process the labels, and the procedures taken to tune their hyperparameters are outlined. Afterwards, the performance of the classifiers are dissected in order to reveal the impacts of clothing and objects in this genre classification experiment. The validity of each step relied on the results of the previous cumulative steps. Therefore, the outcome of each step was thoroughly measured and examined for any errors.

4.1 Implementation

The implementation process for this experiment was completed in the Python 3.7 programming language. Each step described in the subsequent sections was modularized to run independently, and is accessible on GitHub.[1]

4.1.1 Dataset

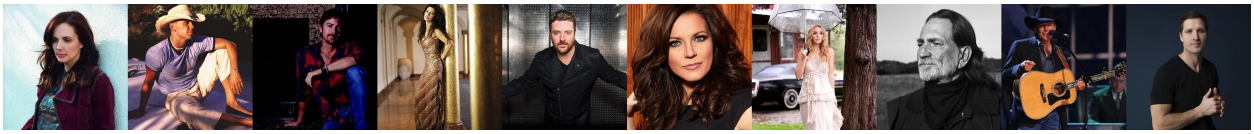
Combining the image source, Last.fm, with the artist lists generated by Billboard and Gaon was the first step in the creation of the experiment's dataset. Using the `billboard.py` Python library,[2] artist names from the Top Country Albums, Top R&B/Hip-Hop Albums, and Top Rock Albums from the first week of each month from February 2018 to February 2019 were retrieved. All three charts consist of 50 artists each. The same process was used for the Gaon Digital Chart, using the `mochart` Python library[3] to extract the weekly charts from the same time period.

The artist lists for each genre were then used to retrieve the promotional photos. Last.fm exposes an API,[4] with functionality to search for an artist's profile information based on their name. This was used to obtain each artist's profile page URL.

Although Last.fm exposes an API call that returns one image URL for a specified artist, multiple photos per artist were desired in order to increase the size of the dataset. Therefore, rather than using the Last.fm API for the image retrieval process, a script was developed to retrieve the images from each artist's Last.fm profile page. Each profile's photo section contains 40 promotional photos per page, in thumbnail format. By "web scraping" these thumbnail URL locations, up to 40 images were downloaded for each artist in the genre lists.

1. https://github.com/andrewkam/visual-aesthetics-classification
2. https://github.com/guoguo12/billboard-charts
3. https://github.com/hyunchel/mochart
4. https://www.last.fm/api

Figure 4.1 displays sample images from each genre.



(a) Country

(b) R&B/Hip-Hop

(c) Rock

(d) K-Pop

Fig. 4.1 Random images from each of the four genres in the dataset.

Modifying the URLs also provided the ability to specify one of a few preset thumbnail dimensions, and image sizes of 300 x 300 pixels were retrieved. These dimensions were requested as they were similar to the object detection requirements, described later in Section 4.1.2. While popular artists consistently had 40 images available for retrieval, more obscure artists did not, and the maximum number of images were extracted in these cases. All images were retrieved and saved locally in directories according to their genre. Table 4.1 displays the number of images fetched for each genre.

An item to note is that in four cases, artists were shared between two genres. For example, an artist may appear in both the Top Country Albums and Top Rock Albums charts, and

Genre	Number of Artists	Number of Images
Country	87	2092
R&B/Hip-Hop	69	2238
Rock	60	2208
K-Pop	142	2163

Table 4.1 The number of images downloaded and stored for each genre.

therefore produces duplicate image instances with different labels. Artists cannot always be compartmentalized to a single genre in practice, and such overlap was left in place for the experiment, as it is a realistic reflection of the nature of genre. This resulted in 125 duplicate images in the dataset.

4.1.2 Image Classification

Image Evaluation

Using DeepDetect, the image-prediction framework was created as a local web service. The service was configured to resize input images to 224 x 224 pixels, as the GoogLeNet fashion model was trained with photos of these dimensions. Since the images retrieved from Last.fm had dimensions of 300 x 300 pixels, reducing each dimension by the same percentage would not affect the functionality of the image classification.

As the number of downloaded images for each genre was just over 2,000, this number was used to cap the image counts. For each genre, the image list was randomly shuffled, and the first 2,000 photos used to form an equally distributed dataset of a total of 8,000 entries. This removes any issues that arise from class imbalance, such as class bias.

All 2,000 images in each genre were then sent to the prediction framework. For each image, the top-10 class predictions using the fashion model were requested, which resulted in a list of 10 labels from Appendix A, along with their prediction scores. The scores represent the probabilities of the items being correct. The output of each image input provided the

following information:

- Filename
- Chart name
- 10 clothing items with their corresponding prediction scores

This output information, formatted in JavaScript Object Notation (JSON), was then sent to Elasticsearch,[5] an open-source data collection and search engine. Installed as a local server, Elasticsearch is able to ingest and store schema-free JSON documents, then make them accessible and searchable using Kibana,[6] an open-source visualization tool. This provided the ability to both retain all clothing predictions as well as their corresponding genre labels, and then aggregate them into trends and charts for analysis. The following is an example image from the country genre (Fig. 4.2) and the associated JSON object containing its clothing prediction results:

Fig. 4.2 The image from the country genre associated with the JSON object outlined below.

5. https://www.elastic.co
6. https://www.elastic.co/products/kibana

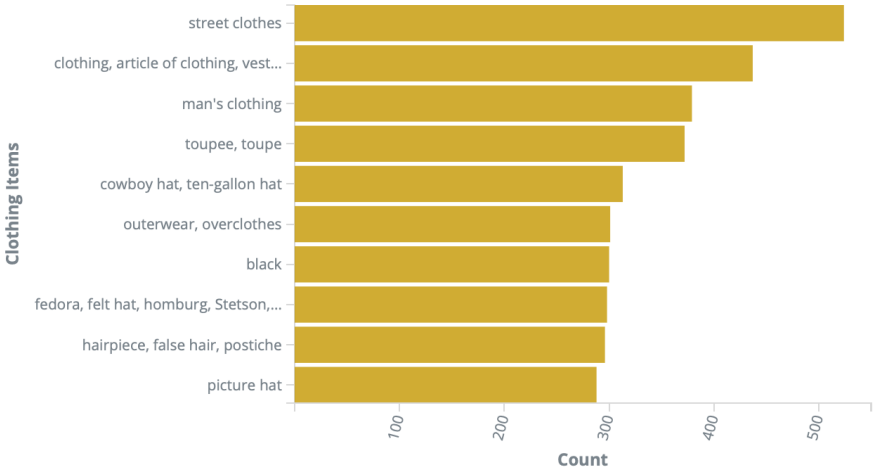
```
{
  "_index": "clothing-10",
  "_type": "img",
  "_id": "jB2DPGkBoCeb5ef4sEVD",
  "_version": 1,
  "_score": 0,
  "_source": {
    "uri": "/opt/models/images/country-albums/kenny_chesney/kenny_chesney-01.jpeg",
    "chart": "country-albums",
    "categories": [
      {
        "category": "cowboy hat, ten-gallon hat",
        "score": 0.897843
      },
      {
        "category": "boater, leghorn, Panama, Panama hat, sailor, skimmer, straw hat",
        "score": 0.0975906
      },
      {
        "category": "sombrero",
        "score": 0.00319119
      },
      {
        "category": "picture hat",
        "score": 0.00121877
      },
      {
        "category": "sunhat, sun hat",
        "score": 0.000117153
      },
      {
        "category": "millinery, woman&apos;s hat",
        "score": 0.0000298629
      },
      {
        "category": "hat, chapeau, lid",
        "score": 0.00000526341
      },
      {
        "category": "pith hat, pith helmet, sun helmet, topee, topi",
        "score": 0.00000205467
      },
      {
        "category": "fedora, felt hat, homburg, Stetson, trilby",
        "score": 0.00000166247
      },
      {
        "category": "dress hat, high hat, opera hat, silk hat, stovepipe, top hat, topper,
            beaver",
        "score": 7.39661e-7
      }
    ]
  }
}
```


As seen in the above JSON object, the analysis of the image from the “country-albums” chart resulted in 10 clothing predictions, each labeled with a “category” and a “score”. The top scoring prediction was “cowboy hat, ten-gallon hat”, with a likelihood of 89 percent. However, the likelihood of the second-highest category, “boater, leghorn, Panama, Panama

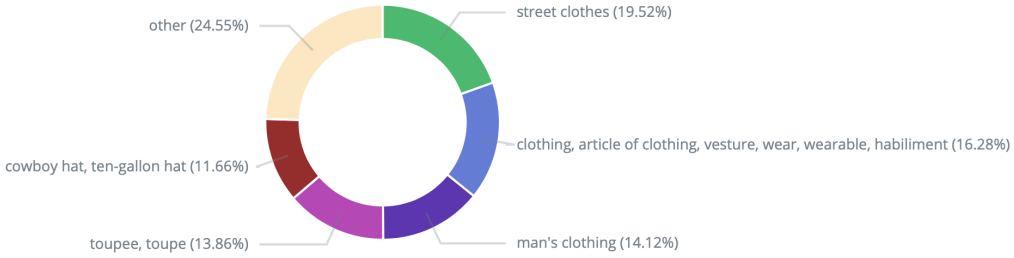
hat, sailor, skimmer, straw hat", dropped to 10 percent, and the remaining predictions all below one percent.

Clothing Analysis

Using Kibana, the clothing predictions were aggregated by count per category, for each genre. Looking at Fig. 4.3, it is clear that various types of headwear were prevalent in the most-detected clothing items in the country genre. Many of the genre's historically clichéd fashion items, such as "cowboy hat" and "Stetson hat", were among the most found within the images.



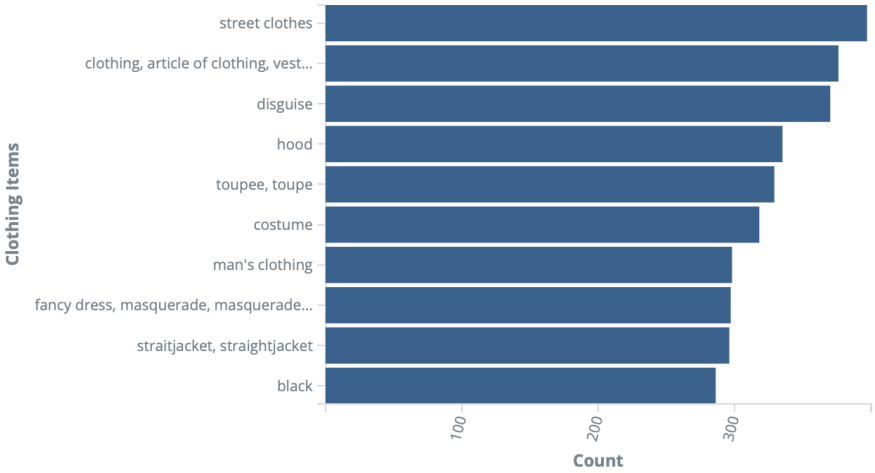
(a) The top 10 clothing items.



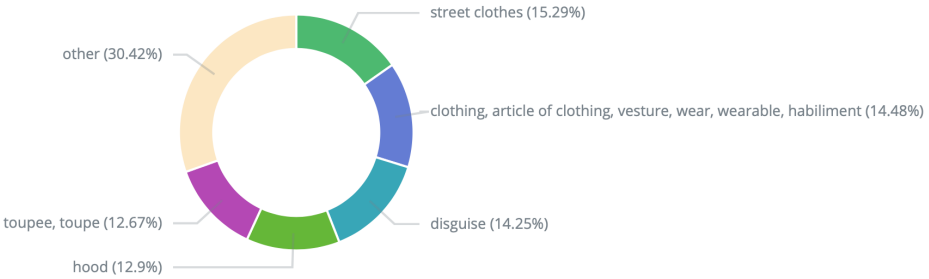
(b) The percentage occupied by the top items.

Fig. 4.3 The composition of clothing items in the country genre.

Figure 4.4 suggests that within the R&B/hip-hop genre, there was a trend in more elaborate fashion, including items such as “disguise”, “costume”, and “masquerade”, which are descriptors not commonly associated with everyday fashion. It can also be observed that the top-five items contributed to less than 75 percent of the total composition, which is not the case with the other genres in this experiment.



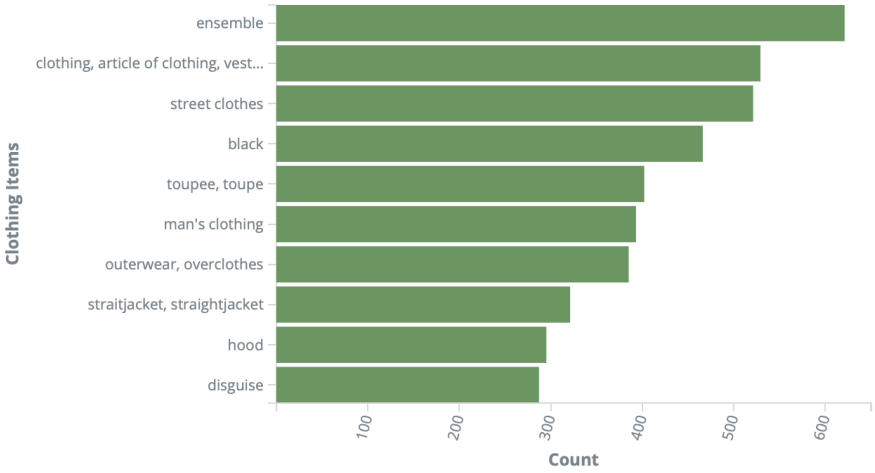
(a) The top 10 clothing items.



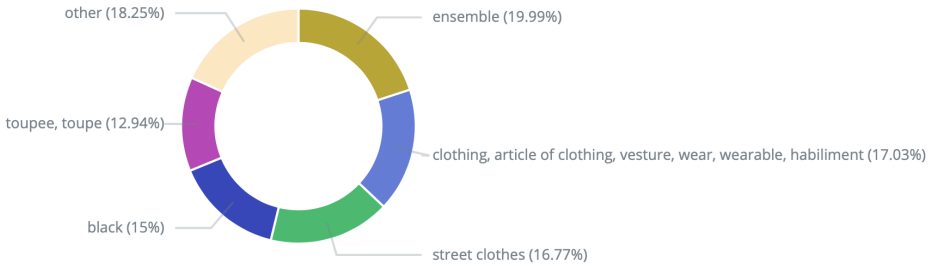
(b) The percentage occupied by the top items.

Fig. 4.4 The composition of clothing items in the R&B/hip-hop genre.

On the other hand, clothing found in the rock genre, as shown in Fig. 4.5, was more modest, containing fashionably simple items. The very generic and broad term "ensemble" dominated the top-detected clothing within the genre, found in 20 percent of the images.



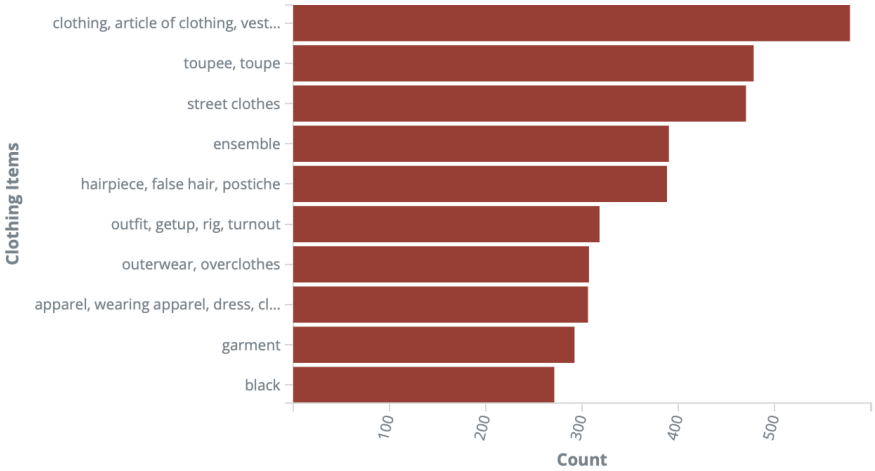
(a) The top 10 clothing items.



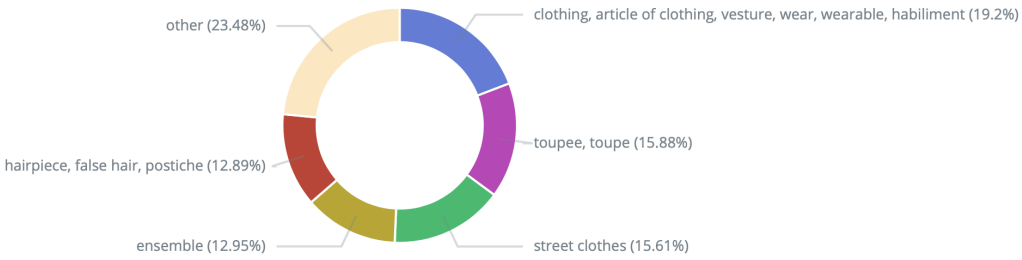
(b) The percentage occupied by the top items.

Fig. 4.5 The composition of clothing items in the rock genre.

Much like in the country genre, many of the detected items in the K-pop genre involved headwear, which can be seen in Fig. 4.6. However, hairpieces such as “toupée” and “false hair” were prominent instead of hats, such as “cowboy hat” and “Stetson hat”.



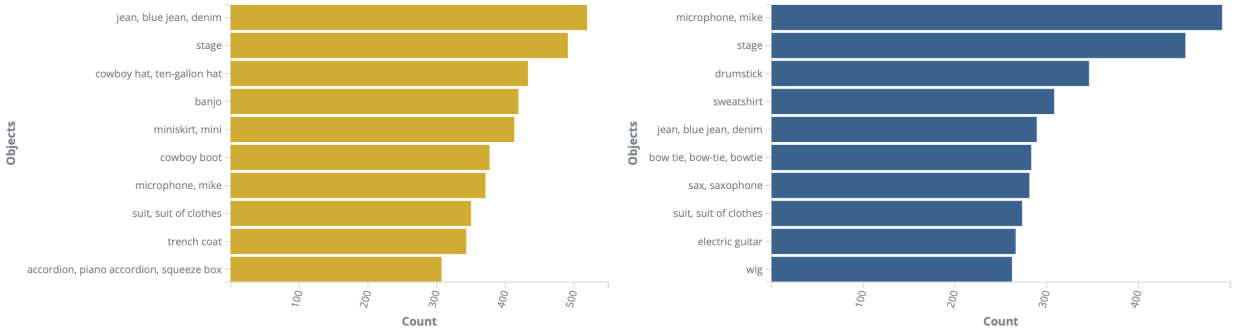
(a) The top 10 clothing items.



(b) The percentage occupied by the top items.

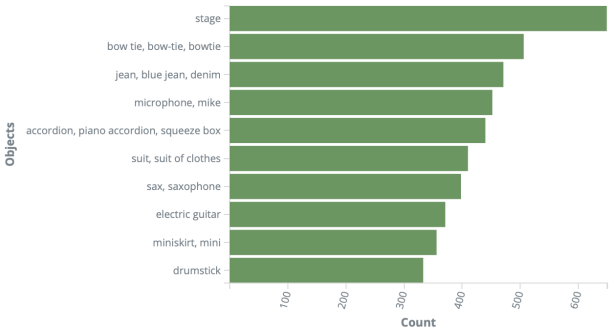
Fig. 4.6 The composition of clothing items in the K-pop genre.

The identical image-evaluation procedure was completed with the ResNet-50 model pre-trained on the 1,000 ILSVRC classes, with the results shown in Figs. 4.7a–d. As seen, non-clothing items were introduced into the top-10 detected objects. The majority of the new items consisted of musical instruments, such as "banjo", "drumstick", "saxophone", and "electric guitar". Another new item that was frequently detected in all four genres was "stage", meaning that many of the promotional photos placed the artist in a concert setting.

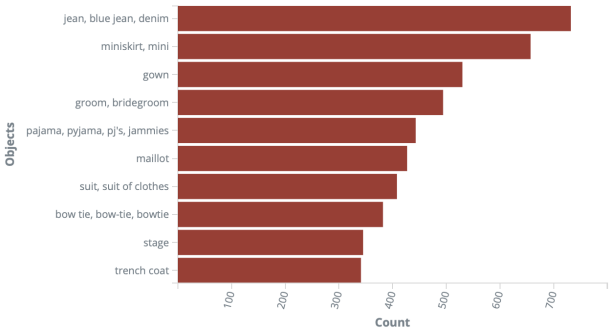


(a) Country

(b) R&B/Hip-Hop



(c) Rock



(d) K-Pop

Fig. 4.7 The top-10 objects items detected in each genre.

4.1.3 Genre Classification

The next step in the experiment was to derive music genre predictions based solely on the clothing items from the image classification procedure described in the previous section. All 10 clothing predictions for every promotional photo were stored for the 2,000 images per genre, creating a new collection of 20,000 words for each music category.

Pre-Processing

As described in Section 3.2.2, the generated clothing labels could be used to form a bag-of-words dataset, an approach commonly used in sentiment analysis. By treating each image in the experiment as a document, and the predicted clothing items as the document text, the bag-of-words representation was constructed, with the 304 classes of fashion items forming the vocabulary of the model. The scoring of each word was completed by totaling their counts within a document. Since each clothing prediction could only occur a maximum of once per image, the scores could only result in zero or one, often referred to as a binary bag-of-words. Figure 4.8 displays a sample image, with its image index and clothing item indices from the bag-of-words representation, as well as the associated synsets linked to the clothing item numbers.

As with any type of input data in a machine-learning process, text must also be cleaned for optimal performance by any classifier. Since the clothing items of the image model formed its own vocabulary, the pre-processing commonly applied to natural language, such as punctuation removal and lowercase conversation, did not apply in this scenario.

As seen in Fig. 4.3. Fig. 4.4, Fig. 4.5, and Fig. 4.6, the clothing item labeled as “clothing, article of clothing, vesture, wear, wearable, habiliment” was very prominent among all four genres, detected a minimum of 14 percent of all items, and over 19 percent in the K-pop

Image	Item	Synset
2	40	cowboy hat, ten-gallon hat
2	76	boater, leghorn, Panama, Panama hat, sailor, ...
2	93	sombrero
2	104	picture hat
2	131	sunhat, sun hat
2	175	millinery, woman's hat
2	206	hat, chapeau, lid
2	209	pith hat, pith helmet, sun helmet, topee, topi
2	246	fedora, felt hat, homburg, Stetson, trilby
2	258	dress hat, high hat, opera hat, silk hat, ...

Fig. 4.8 An example of an image and its detected synsets. The indices in the Image and Item columns were extracted from the bag-of-words representation.

genre. The label itself is also extremely broad in terms of fashion, in that it encompasses all forms of clothing. As such, this term may not be helpful in differentiating between genres, and was removed from the vocabulary. Such action parallels adding irrelevant words to a "stop word" list in text sentiment analysis.

Classification Models

As mentioned in Section 3.2.2, naive Bayes and support vector machines are common classifiers used in text sentiment analysis. There are a number of factors that make SVMs suitable for text classification: the data commonly has a very high dimensional input space, document vectors are always sparse, and most text categorization problems are linearly separable. As outlined by Joachims (1998), SVMs perform well in classification under these conditions. McCallum and Nigam (1998) explain that the large number of attributes in such classification also allows naive Bayes to excel in the task. The study also shows that the multi-variate Bernoulli model performs well with small vocabulary sizes, while the multinomial model performs better with larger sizes of vocabulary.

Both of these classifiers were then trained and tested on the bag-of-words representation of clothing items with 5-fold cross-validation. Using the `naive_bayes.BernoulliNB` class in `scikit-learn`,[7] a free machine learning library for Python that was downloaded and installed, naive Bayes classification was completed. This specific classifier was chosen since each feature in the dataset was a binary value, and the classifier is designed for binary/-boolean features. This classifier has no hyperparameters for tuning. For SVM classification, `svm.LinearSVC`, the linear support vector classification class in `scikit-learn`, was used. The classifier supports both dense and sparse input, and the multiclass support is handled according to a one-vs-the-rest scheme. The following penalty and loss function hyperparameters were tuned for optimal performance for this model:

- Penalty parameter of error term
- Dual/Primal optimization problem
- Loss function
- Penalization norm

The output in Fig. 4.9 shows the hyperparameter tuning of the SVM model using primal optimization, squared hinge as the loss function, and L2 as the penalization norm. The primal optimization problem was chosen as the number of training data points was much greater than the number of dimensions, therefore making the calculation more efficient than the dual problem. With the `svm.LinearSVC` library, the loss function must be set to squared hinge alongside primal optimization, and the penalization norm set to L2 to avoid sparse coefficient vectors. With these three hyperparameters set, the value of the penalization norm,

7. https://scikit-learn.org

represented by 'C' below, was tuned to produce the highest F1 score and find the optimal configuration.

```
Best Param:  {'C': 0.01, 'dual': False, 'loss': 'squared_hinge', 'penalty': 'l2'}
0.437 (+/-0.019) for {'C': 0.0001, 'dual': False, 'loss': 'squared_hinge', 'penalty': 'l2'}
0.460 (+/-0.029) for {'C': 0.001, 'dual': False, 'loss': 'squared_hinge', 'penalty': 'l2'}
0.465 (+/-0.022) for {'C': 0.01, 'dual': False, 'loss': 'squared_hinge', 'penalty': 'l2'}
0.457 (+/-0.027) for {'C': 0.05, 'dual': False, 'loss': 'squared_hinge', 'penalty': 'l2'}
0.456 (+/-0.025) for {'C': 0.1, 'dual': False, 'loss': 'squared_hinge', 'penalty': 'l2'}
0.457 (+/-0.027) for {'C': 0.05, 'dual': False, 'loss': 'squared_hinge', 'penalty': 'l2'}
0.454 (+/-0.024) for {'C': 1, 'dual': False, 'loss': 'squared_hinge', 'penalty': 'l2'}
0.455 (+/-0.023) for {'C': 5, 'dual': False, 'loss': 'squared_hinge', 'penalty': 'l2'}
0.454 (+/-0.023) for {'C': 10, 'dual': False, 'loss': 'squared_hinge', 'penalty': 'l2'}
```

Fig. 4.9 The tuning of the penalization norm hyperparameter in the SVM model.

As seen in the output of the tuning process, a penalization norm of 0.01 produced the highest F1 score of 0.465. This hyperparameter configuration of primal optimization, squared hinge as the loss function, L2 as the penalization norm, and penalization norm of 0.01 was used to train the SVM classifier on the clothing dataset.

k-Fold cross-validation was implemented using the `model_selection.KFold` class in `scikit-learn`. This was completed in order to estimate the accuracy of the two classifiers on unseen data, as there was no explicit test dataset for evaluation. By setting k=5, 6,400 of the 8,000 images were used for training, while the remaining 1,600 images were used for testing for each of the five instances. The dataset was shuffled prior to being split into batches.

Accuracy, precision, recall, and F1 score were chosen as metrics to evaluate the classifiers, as they are adequate measures of relevance in classification tasks (Zheng 2015). Macro averages for precision, recall, and F1 score were used instead of micro averages, as the classes in the dataset were already balanced, and would therefore represent the performance overall across all sets of data.

4.2 Results

The following section presents in detail the resulting performance metrics produced by both the naive Bayes and SVM classifiers. The prediction success of each of the four genres is then evaluated and compared. A visual approach is then applied to the classification procedure in order to examine observable similarities between the images.

The identical classification procedure was applied to both the clothing dataset and the object dataset. The results of the procedure using the clothing dataset are reported first in this section, and the object dataset afterwards.

4.2.1 Clothing Items

Training the naive Bayes and SVM classifiers on the clothing dataset and then running predictions using 5-fold cross-validation both produced performance results significantly above chance (25 percent). Performance was measured using accuracy, precision, recall, and F1 score as metrics. The accuracy metric is the fraction of predictions that the model identified correctly. Precision outlines the proportion of positive identifications that were actually correct, while recall is the proportion of actual positives that were identified correctly. F1 score is the harmonic mean of precision and recall. As shown in Table 4.2, the linear SVM classifier performed better of the two, with an F1 score of 0.465.

Classifier	Accuracy	Precision	Recall	F1 Score
NB	0.449	0.456	0.446	0.452
SVM	0.465	0.466	0.465	0.465

Table 4.2 Performance of genre classification with clothing items. Reported values are the means of 5-fold cross-validation results.

It is of importance to break down the performance of the classification to discover the impact of each genre on the prediction success. In order to further investigate the behavior

of the classification procedure, a hold-out method was applied by using a train/test split of 75/25 on the clothing dataset. This hold-out process mimics an independent test set by leaving a percentage of the dataset aside for evaluation, and would also produce performance results outlining the model's predictive success for each specific genres. Since it outperformed the naive Bayes classifier using the previous cross-validation method, the SVM classifier was used to generate the 2,000 predictions (25 percent of the 8,000 images). As seen in Fig. 4.10 and Table 4.3, the recall for the K-pop and R&B/hip-hop genres, 0.507 and 0.510 respectively, were significantly higher than the others, providing evidence that it was easier for the classifier to identify these two genres out of the four.

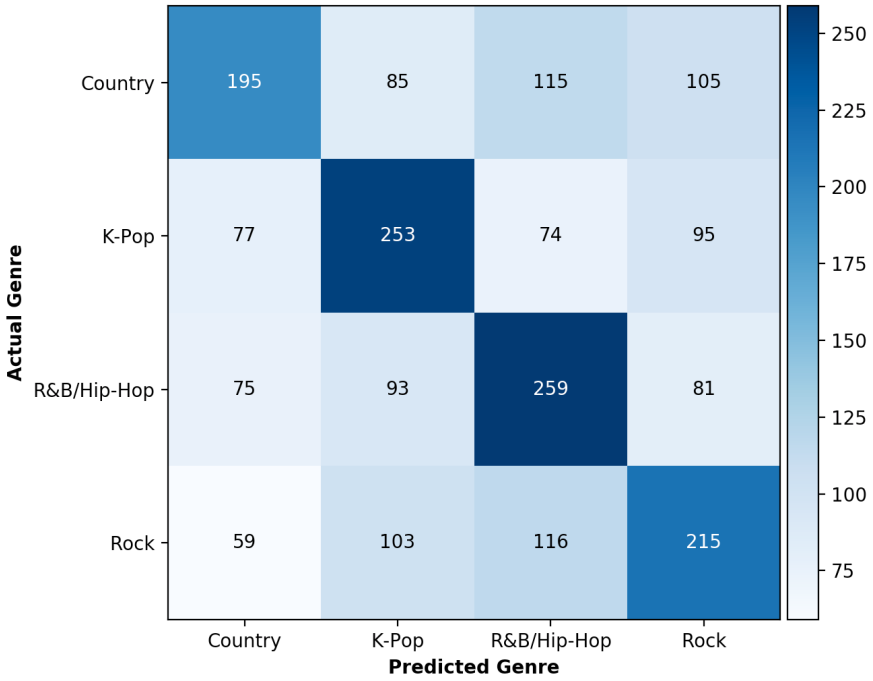


Fig. 4.10 Prediction results of genre classification with the clothing items dataset.

On the other hand, the country genre was predicted only 406 times, the least number out of all genres, but predicted most successfully when done so. This was reflected in the lowest recall value (0.390) and highest precision (0.480) among all categories. This could very

Genre	Precision	Recall	F1 Score
Country	0.480	0.390	0.430
K-Pop	0.474	0.507	0.490
R&B/Hip-Hop	0.459	0.510	0.483
Rock	0.433	0.436	0.435

Table 4.3 Class specific performance of genre classification with clothing items.

much be due to the unique fashion within the genre, such as the "cowboy hat" and "Stetson hat" items not commonly found in others. Rock was predicted with the lowest precision, which could be attributed to the relatively simple and generic clothing items found within the genre.

Performance Visualization

While the breakdown of precision and recall shown in Table 4.3 provides insight into the success of the classifier for each genre, it does not provide any answers into the reasoning behind such success. In order to visualize how specific clothing items could impact the genre classification process, t-distributed Stochastic Neighbor Embedding (t-SNE) was applied to the dataset of clothing labels. t-SNE is a technique for dimensionality reduction that is particularly well suited for the visualization of high-dimensional datasets (Maaten and Hinton 2008). Applying principal component analysis (PCA) dimension reduction using the `decomposition.PCA` class and t-SNE using the `manifold.TSNE` class from `scikit-learn`, the scatter plot shown in Fig. 4.11 was produced. As seen in the plot, a few clusters consisting mostly of a single genre were produced.

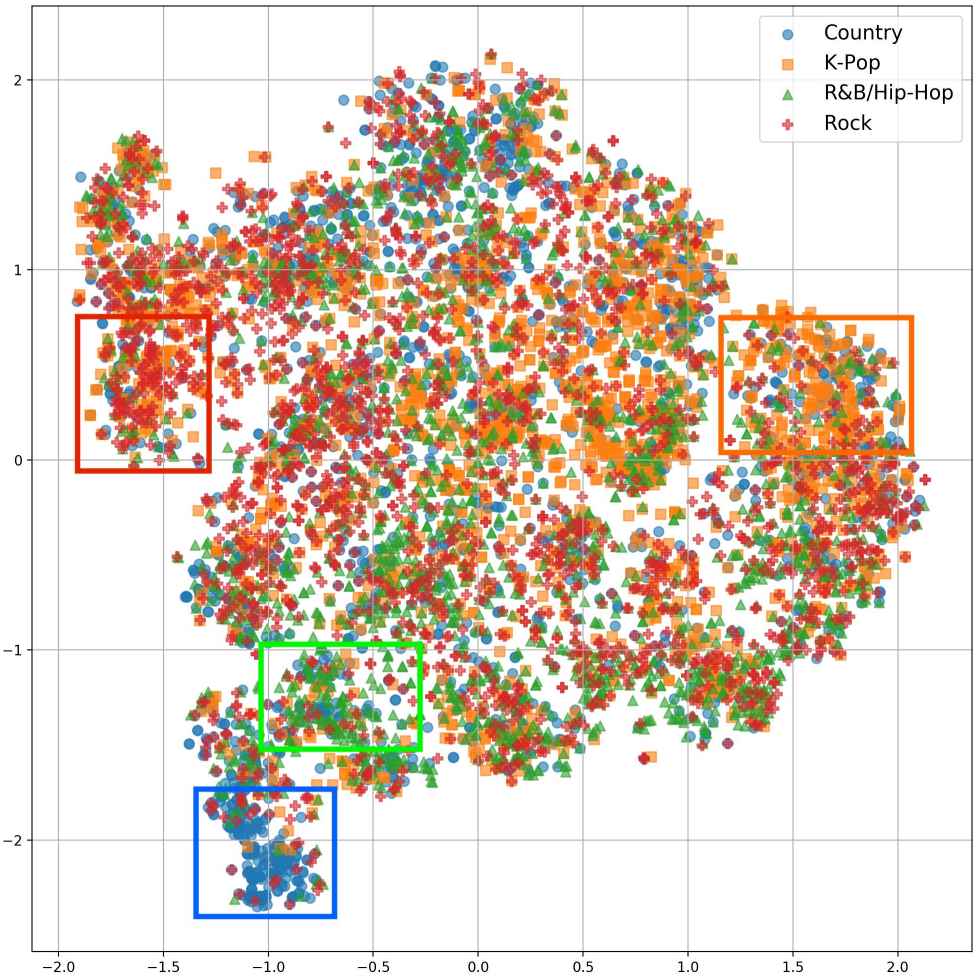


Fig. 4.11 Images are clustered based on clothing items, using t-SNE and PCA dimension reduction. The blue, green, red, and orange boxes highlight clusters of genres.

By replacing the plot markers with the original source images, it was possible to observe the clothing items that influenced the formation of these clusters. Figure 4.12 displays a sample of the images within the blue box located in the bottom-left corner of Fig. 4.11, which mainly consists of images from the country genre. As seen, the images contains a high prevalence of headwear, specifically cowboy hats. This reinforced the previous notion that the low recall value and high precision found in Table 4.3 was due to this unique fashion within the genre.

Fig. 4.12 Cluster of country images in t-SNE visualization (Fig. 4.11). It can be seen that the cluster is defined by the presence of cowboy hats.

Figure 4.13 below displays the cluster of R&B/hip-hop images outlined by the green box at the bottom-left corner of Fig. 4.11. While the presence of one distinct clothing item is not obvious in this group of photos, various forms of headwear can be observed. Baseball caps and bandanas can be found in many of the photos.

Fig. 4.13 Cluster of hip-hop images in t-SNE visualization (Fig. 4.11). Headwear, such as baseball caps and bandanas, are present in many of the photos.

Figure 4.14 below displays the cluster of images mainly from the rock genre, outlined by the red box on the left side of Fig. 4.11. The first obvious trait in this group of images is that the majority of them are in greyscale. Secondly, most of the artists in these images are wearing suits, many with ties or bow ties.

Fig. 4.14 Cluster of rock images in t-SNE visualization (Fig. 4.11). Suits are the prevalent clothing apparel in this cluster of photos. The majority of the photos are also in greyscale.

Figure 4.15 below displays the cluster of images mainly from the K-pop genre, outlined by the orange box on the right side of Fig. 4.11. It is noticeable that the majority of the photos are head shots of the artists, with their faces being the focal points of the images. Since the images are zoomed into faces of the artists, clothing items are either non-existent or cropped out. Most of the artists in this cluster are female, and these photos may have been grouped together based on longer hair styles.

Fig. 4.15 Cluster of K-pop images in t-SNE visualization (Fig. 4.11). The majority of the photos are head shots of the artists, with long hairstyles on display.

4.2.2 Objects

Similar to the procedure with the clothing item dataset, the naive Bayes and SVM classifiers were trained on the 1,000 ILSVRC objects outlined in Appendix B. Again, the `naive_bayes.BernoulliNB` library was used for the naive Bayes classifier, while the `svm.LinearSVC` library was used with the exact same hyperparameter tuning, as this configuration was once again found to produce the optimal performance. Running predictions using the identical cross-validation procedure on the object items produced even better performance results than with the clothing item dataset. As seen in Table 4.4, an F1 score of 0.533 was obtained with the SVM classifier, demonstrating that the correct genre could be predicted in the majority of instances.

Classifier	Accuracy	Precision	Recall	F1 Score
NB	0.505	0.517	0.505	0.499
SVM	0.535	0.535	0.535	0.533

Table 4.4 Performance of genre classification with the ILSVRC object dataset. Reported values are the means of 5-fold cross-validation results.

Similar to the clothing dataset, the performance results in Table 4.5 and confusion matrix in Fig. 4.16 was generated by training the SVM classifier on 75 percent of the objects dataset, and testing on the remaining holdout subset. As expected, there was an increase in both precision and recall for all four genres, reflected in Table 4.6. There was an extra improvement in the precision, recall, and F1 score of the K-pop genre, with all three metrics gaining a 0.1 increase at minimum.

Performance Visualization

Using PCA dimension reduction and t-SNE, the scatter plot shown in Fig. 4.17 was produced. As seen in the plot, a few clusters consisting mostly of a single genre were produced.

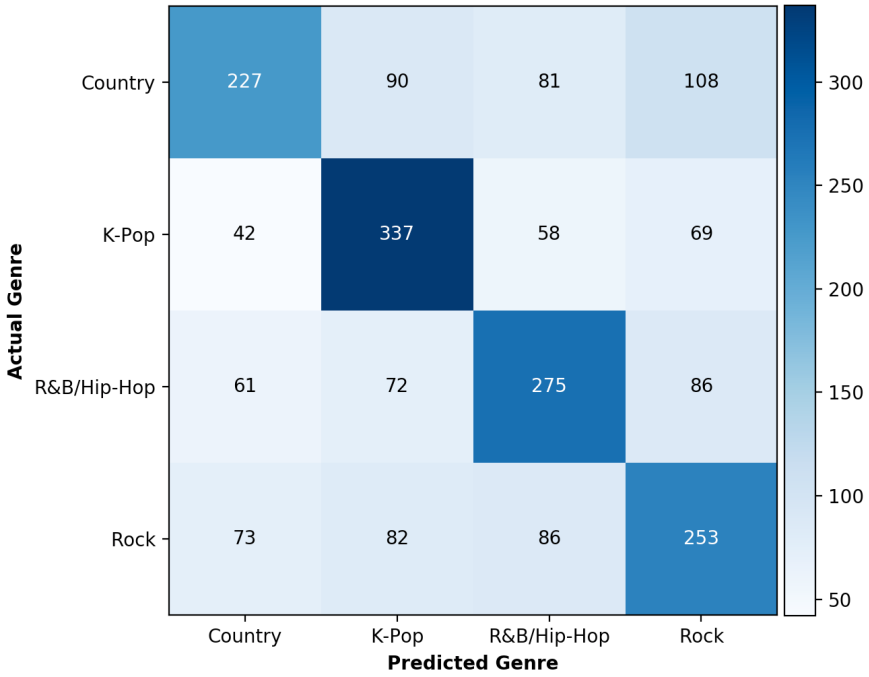


Fig. 4.16 Prediction results of genre classification with objects.

Genre	Precision	Recall	F1 Score
Country	0.563	0.449	0.499
K-Pop	0.580	0.666	0.620
R&B/Hip-Hop	0.550	0.557	0.553
Rock	0.490	0.512	0.501

Table 4.5 Class specific performance of genre classification with objects.

Genre	Precision	Recall	F1 Score
Country	+0.083	+0.059	+0.069
K-Pop	+0.106	+0.159	+0.130
R&B/Hip-Hop	+0.091	+0.047	+0.070
Rock	+0.057	+0.076	+0.066

Table 4.6 The increase in class specific performance of genre classification using the object dataset compared to the clothing dataset.

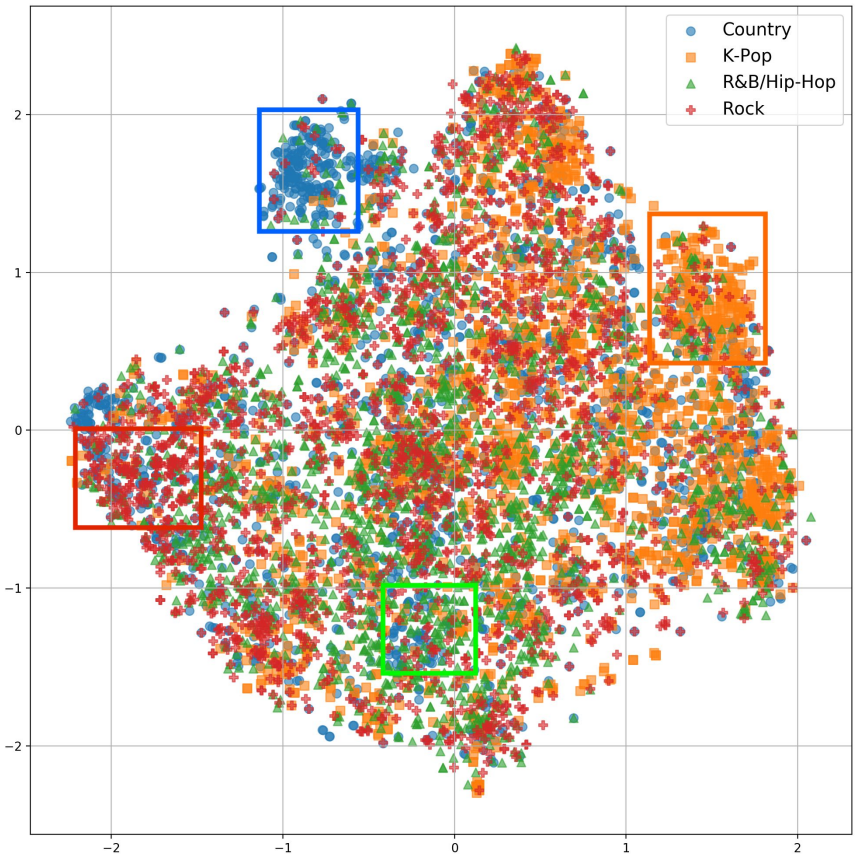


Fig. 4.17 t-SNE visualization with objects.

As with the clothing items, by replacing the plot markers with the original source images, it was possible to observe the objects that influenced the creation of these clusters. Figure 4.18 displays a sample of the images within the blue box located in the top-left corner of Fig. 4.17, which consists of images almost solely from the country genre. As seen, almost every image contains a cowboy hat, providing a clear signal for the associated country genre. Similarly to the clothing dataset, this distinct item could have been the main reasoning behind the low recall value and high precision found in Table 4.5.

Fig. 4.18 Cluster of country images in t-SNE visualization (Fig. 4.17). It can be seen that the cluster is defined by the presence of cowboy hats.

While there are no obvious clusters of R&B/hip-hop images in Fig. 4.17, the green box in the bottom of the plot contains a large percentage of images from this genre. Looking at Fig. 4.19 below, there isn't a specific clothing item that stands out among the rest. However, most of the images are of male artists, wearing relatively simple clothing, such as T-shirts and sweaters. It is a possibility that men's clothing was the commonality for this group of images.

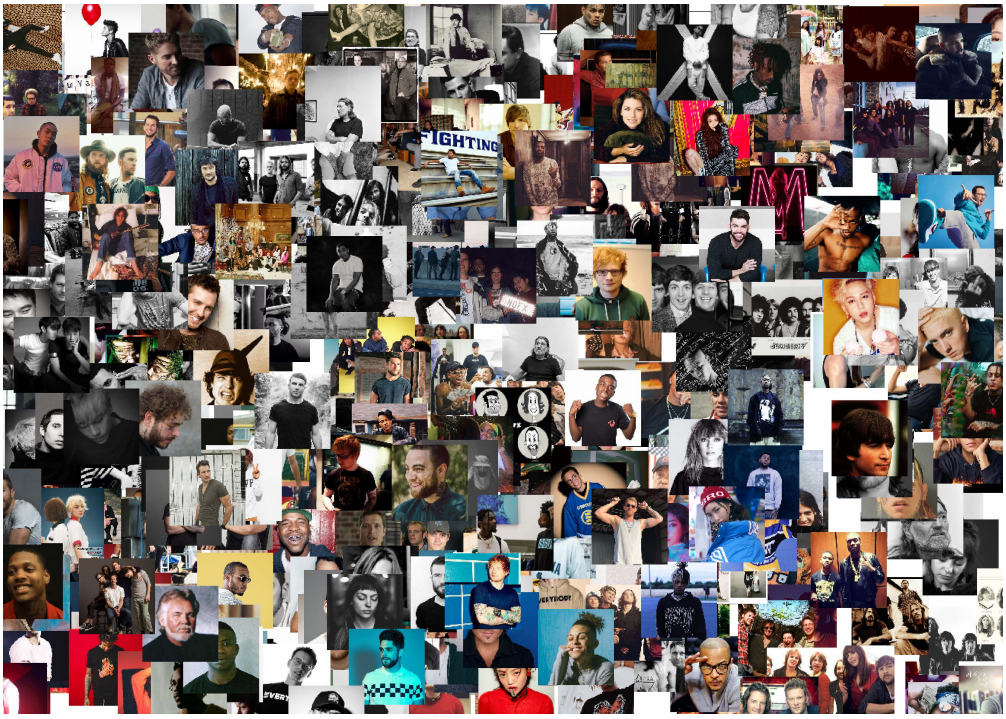


Fig. 4.19 Cluster of hip-hop images in t-SNE visualization (Fig. 4.17). Male artists, wearing relatively simple clothing such as T-shirts and sweaters, are present in many of the photos.

Figure 4.20 below displays the group of images outlined by the red box on the left side of Fig. 4.17, which consists of images mainly from the rock genre. As seen in the images, most of the artists are holding or playing instruments, with electric guitars being the most prominent. Since electric guitars are traditionally associated with rock music, it is easy to conclude that they were the object item that produced this cluster of images from the rock genre.

Fig. 4.20 Cluster of rock images in t-SNE visualization (Fig. 4.17). The majority of the photos contain acoustic or electric guitars.

Figure 4.21 below displays a cluster of images mostly from the K-pop genre, outlined by the orange box on the right side of Fig. 4.17. It is noticeable that the majority of the photos are of female artists who are wearing dresses or skirts with blouses. As such, it can be concluded that these clothing items produced this cluster of images from the K-pop genre.

Fig. 4.21 Cluster of K-pop images in t-SNE visualization (Fig. 4.17). The majority of the photos are of female artists, wearing dresses or skirts.

4.3 Discussion

The results of the genre classification using the clothing dataset, described in Section 4.2.1, produced F1 scores of just under 0.50, as shown in Table 4.3. While these results were not ideal, they did prove that the fashion items portrayed in the images were impressionable features in the task of genre classification, as they were significantly above chance (0.25). Using the dataset containing the ILSVRC objects, classification precision increased by at least 0.05 across all four genres, bringing the F1 scores above 0.5, establishing that the inclusion of items beyond clothing aids in this genre classification experiment, and that the correct genre can be predicted in the majority of cases.

Dissecting the items in the clothing dataset in Appendix A, as well as observing the top clothing items detected in Figs. 4.3–6, it can be seen that many of the clothing definitions can be very general and broad. Garments such as "street clothes", "man's clothing", "outerwear, overclothes", and "ensemble" are all clothing descriptions that encompass a variety of clothing items, and as such do not provide much detail in the fashion portrayed by the artists. It is very probable that this flaw contributed in lowering the performance of the classifiers.

On the other hand, the scatter plot generated using t-SNE in Fig. 4.11 and the evaluation of its clusters provided much insight into the clothing items that were impactful in defining a genre. Areas of the plot that were dominated by one specific genre were found to contain fashion-based themes, such as cowboy hats, bandanas, suits, and long hairstyles. While this held true for such areas, the majority of the scatter plot in Fig. 4.11 was mixed among the four genres, re-enforcing the difficulty in classification when using this dataset.

The performance increase using the objects dataset was the most pronounced in the K-pop genre, with an F1 score of 0.62, an increase of 0.13 compared to the clothing dataset (Fig. 4.6). Observing the t-SNE plot in Fig. 4.17 and studying the cluster containing K-

pop images, it was found that dresses and skirts clearly identified the genre from the rest. This was a deviation from the long hairstyles that formed the most obvious cluster of K-pop images within the clothing dataset. The inclusion of objects outside of clothing items introduced musical instruments into the dataset, which are commonly found in promotional photos. This was most obvious in the rock genre, where a cluster of images was clearly defined by acoustic and electric guitars. A musical instrument, such as an electric guitar, could be considered a fashionable accessory in some promotional photos, and such an addition aided in the classification of the genre. As different instruments are used much more prominently in specific genres of music, this result was not unexpected.

The noticeable performance improvement using the objects dataset over the clothing dataset provides some awareness about the impact of fashion alone for genre classification purposes. Since the pre-trained clothing model used in the experiment was fairly simple and not specifically created for the image dataset, it was possible that this shortcoming hindered the success of the classification. However, one could question the concrete impact that clothing and fashion has on one's ability to perform the same task. It could be conceivable that removing the human artist and all other background objects from the photos, leaving only just the clothing behind, does indeed make the prediction process quite difficult. It was key to the experiment that the gender and ethnicity of the artist were not taken into consideration, since these traits are not evenly distributed among genres, and as such would have significant influences on the predictions. Due to this factor, it was important that no physical human traits existed in the detection models. Also, by separating the clothing and object datasets, it was possible to observe that a more general and broad model, trained on a variety of object types, could outperform a specialized model trained on only clothing items and garments. Such difference might demonstrate that there are many factors within a promotional photo that provide hints on the musical genre of an artist, and fashion is just

one of them.

5 Conclusion

In this study, we attempted to link the fashion portrayed by popular artists with their musically associated genre. Promotional photos of current chart-topping artists were used as a dataset for the experiment, and both clothing items and non-clothing objects extracted from the images. The detected items were then used as a dataset for the second stage of the study, in which classification was performed to predict musical genre.

The results of the study displayed that the use of simple image recognition techniques, combined with basic classification algorithms, can recognize the genre of a promotional photo with a success rate greater than 50 percent. While unique fashion items such as cowboy hats aided in the prediction of the country genre, the simplicity of clothing in the rock genre made detection less precise, supporting the notion that imagery in this genre is not especially distinctive (Englis, Solomon, and Olofsson 1993). Based solely on clothing items, prediction results of the K-pop and R&B/hip-hop genres were higher than the others, revealing that it was easier to identify these two genres compared to country and rock. It was also found that the addition of object detection beyond clothing items, such as the recognition of musical instruments, significantly boosted the performance of the genre prediction task, with images from the K-pop genre gaining the most accuracy.

By using dimension reduction techniques and then plotting the results, the images were then clustered, with many areas of the plots consisting of mostly a single genre. In many of

these clusters, it was easy to identify the type of clothing that produced the genre grouping, such as cowboy hats in country, baseball caps in R&B/hip-hop, suits in rock, and dresses in the K-pop genre. The addition of objects outside of clothing also allowed us to observe that electric guitars were commonly used as an accessory in promotional photos of rock artists. As such, analyzing these results provided insight into the salient features that define the visual aesthetics of certain genres, and paths taken to differentiate them.

Recent U.S. pop music genre classification experiments on audio datasets have produced accuracy results exceeding 76 percent, with a greater number of genres (10) than the experiment described in this thesis.[1] However, the goal of this study was to observe the classification process purely on visual imagery in promotional photos, with emphasis on the clothing portrayed in the images. By splitting the process into two stages, object detection and genre classification, it was also possible to observe the clothing items that visually dominate each music category, providing awareness of the recent trends in fashion portrayal. A deeper dive into this topic would reveal invaluable information on market trends and promotional tactics in the music industry.

5.1 Future Work

The use of a pre-trained clothing model was adequate in producing classification results of significance. However, the performance could be improved by extensively training a model with a large, fashion-specific dataset of images. DeepFashion (Liu et al. 2016), a large-scale clothes database with annotated fashion images, could be useful in training a more accurate recognition model with a greater number of labels. This could then lead to the successful classification of more musical genres and subgenres than the ones used in this experiment. It

1. https://www.music-ir.org/nema_out/mirex2017/results/act/mixed_report

would be interesting to see if similar results could be produced with a variety of subgenres.

It would also be fascinating to run the same experiment procedure on promotional photos from different time periods. As this study was conducted on photos of current artists on top of their respective charts, the availability of older charts from previous decades provides the ability to evaluate the fashion stereotypes from multiple periods. This would provide additional insight into any recent disintegration of genre boundaries and its effect on fashion portrayal in photos, and if genres were historically more visually distinguishable.

Bibliography

Arbelaez, Pablo, Michael Maire, Charless Fowlkes, and Jitendra Malik. 2011. "Contour Detection and Hierarchical Image Segmentation." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33 (5): 898–916. doi:10.1109/TPAMI.2010.161.

Brochu, Eric, Nando De Freitas, and Kejie Bao. 2003. "The Sound of an Album Cover: Probabilistic Multimedia and IR." In *Proceedings of the 9th International Workshop on Artificial Intelligence and Statistics.*

Calefato, Patrizia. 2001. "Light My Fire: Fashion and Music." *Semiotica* 2001 (136): 491–503. doi:10.1515/semi.2001.094.

Deselaers, Thomas, Daniel Keysers, and Hermann Ney. 2008. "Features for Image Retrieval: An Experimental Comparison." *Information Retrieval* 11 (2): 77–107. doi:10.1007/s10791-007-9039-3.

Dodge, Samuel, and Lina Karam. 2016. "Understanding How Image Quality Affects Deep Neural Networks." In *Proceedings of the Conference on the Quality of Multimedia Experience,* 1–6. Red Hook, NY: Curran Associates.

Englis, Basil G., Michael R. Solomon, and Anna Olofsson. 1993. "Consumption Imagery in Music Television: A Bi-Cultural Perspective." *Journal of Advertising* 22 (4): 21–33. doi:10.1080/00913367.1993.10673416.

Fei-Fei, Li, and Pietro Perona. 2005. "A Bayesian Hierarchical Model for Learning Natural Scene Categories." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 524–531. Red Hook, NY: Curran Associates.

Fell, Michael, and Caroline Sporleder. 2014. "Lyrics-Based Analysis and Classification of Music." In *Proceedings of the International Conference on Computational Linguistics: Technical Papers,* 620–631. Dublin: Dublin City University / Association for Computational Linguistics.

Hadi Kiapour, M., Xufeng Han, Svetlana Lazebnik, Alexander C. Berg, and Tamara L. Berg. 2015. "Where to Buy It: Matching Street Clothing Photos in Online Shops." In *Proceedings of the IEEE International Conference on Computer Vision,* 3343–3351. Washington, DC: IEEE Computer Society.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. "Deep Residual Learning for Image Recognition." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 770–778. Red Hook, NY: Curran Associates.

Hochreiter, Sepp. 1998. "The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions." *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 6 (2): 107–116. doi:10.1142/S0218488598000094.

Joachims, Thorsten. 1998. "Text Categorization with Support Vector Machines: Learning with Many Relevant Features." In *Proceedings of the European Conference on Machine Learning,* 137–142. Berlin: Springer.

Khan, Shehroz S., and Daniel Vogel. 2012. "Evaluating Visual Aesthetics in Photographic Portraiture." In *Proceedings of the Eighth Annual Symposium on Computational Aesthetics in Graphics, Visualization, and Imaging,* 55–62. Geneva: Eurographics Association.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. 2012. "ImageNet Classification with Deep Convolutional Neural Networks." In *Advances in Neural Information Processing Systems 25,* 1097–1105. Red Hook, NY: Curran Associates.

Libeks, Janis, and Douglas Turnbull. 2010. "Exploring Artist Image Using Content-Based Analysis of Promotional Photos." In *Proceedings of the International Computer Music Conference,* 183–186. Ann Arbor, MI: Michigan Publishing.

———. 2011. "You Can Judge an Artist by an Album Cover: Using Images for Music Annotation." *IEEE MultiMedia* 18 (4): 30–37. doi:10.1109/MMUL.2011.1.

Liem, Cynthia, Meinard Müller, Douglas Eck, George Tzanetakis, and Alan Hanjalic. 2011. "The Need for Music Information Retrieval with User-Centered and Multimodal Strategies." In *Proceedings of the International ACM Workshop on Music Information Retrieval with User-Centered and Multimodal Strategies,* 1–6. New York: ACM.

Liu, Ziwei, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. 2016. "DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 1096–1104. Red Hook, NY: Curran Associates.

Maaten, Laurens van der, and Geoffrey Hinton. 2008. "Visualizing Data Using t-SNE." *Journal of Machine Learning Research* 9 (Nov): 2579–2605.

Makadia, Ameesh, Vladimir Pavlovic, and Sanjiv Kumar. 2008. "A New Baseline for Image Annotation." In *Proceedings of the 10th European Conference on Computer Vision,* 316–329. Berlin: Springer.

Mayer, Rudolf. 2011. "Analysing the Similarity of Album Art with Self-Organising Maps." In *Proceedings of the 8th International Workshop on Self-Organizing Maps,* 357–366. Berlin: Springer.

Mayer, Rudolf, and Andreas Rauber. 2010. "Multimodal Aspects of Music Retrieval: Audio, Song Lyrics – and Beyond?" In *Advances in Music Information Retrieval,* edited by Zbigniew W. Raś and Alicja A. Wieczorkowska, 333–363. Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-642-11674-2_15.

McCallum, Andrew, and Kamal Nigam. 1998. "A Comparison of Event Models for Naive Bayes Text Classification." In *AAAI-98 Workshop on Learning for Text Categorization,* 41–48. Palo Alto, CA: AAAI Press.

McKay, Cory, John Ashley Burgoyne, Jason Hockman, Jordan B. L. Smith, Gabriel Vigliensoni, and Ichiro Fujinaga. 2010. "Evaluating the Genre Classification Performance of Lyrical Features Relative to Audio, Symbolic and Cultural Features." In *Proceedings of the International Society for Music Information Retrieval Conference,* 213–218. Utrecht, Netherlands.

McKay, Cory, and Ichiro Fujinaga. 2006. "Musical Genre Classification: Is It Worth Pursuing and How Can It Be Improved?" In *Proceedings of the International Conference on Music Information Retrieval,* 101–106. Victoria, BC: University of Victoria.

Miller, Janice. 2011. *Fashion and Music.* Oxford: Berg Publishers.

Na, Youngjoo, and Tove Agnhage. 2013. "Relationship between the Preference Styles of Music and Fashion and the Similarity of Their Sensibility." *International Journal of Clothing Science and Technology* 25 (2): 109–118. doi:10.1108/09556221311298600.

Nanni, Loris, Yandre M. G. Costa, Alessandra Lumini, Moo Young Kim, and Seung Ryul Baek. 2016. "Combining Visual and Acoustic Features for Music Genre Classification." *Expert Systems with Applications* 45:108–117. doi:j.eswa.2015.09.018.

Negus, Keith. 1992. *Producing Pop: Culture and Conflict in the Popular Music Industry.* London: E. Arnold.

Neumayer, Robert, and Andreas Rauber. 2007. "Integration of Text and Audio Features for Genre Classification in Music Information Retrieval." In *Proceedings of the European Conference on Information Retrieval,* 724–727. Berlin: Springer.

Oramas, Sergio, Francesco Barbieri, Oriol Nieto, and Xavier Serra. 2018. "Multimodal Deep Learning for Music Genre Classification." *Transactions of the International Society for Music Information Retrieval* 1 (1): 4–21. doi:10.5334/tismir.10.

Oramas, Sergio, Oriol Nieto, Francesco Barbieri, and Xavier Serra. 2017. "Multi-Label Music Genre Classification from Audio, Text, and Images Using Deep Features." In *Proceedings of the International Society for Music Information Retrieval Conference,* 23–27. Suzhou, China: National University of Singapore.

Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. 2002. "Thumbs Up?: Sentiment Classification Using Machine Learning Techniques." In *Proceedings of the Conference on Empirical Methods in Natural Language Processing,* 79–86. Stroudsburg, PA: Association for Computational Linguistics.

Russakovsky, Olga, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, and Michael Bernstein. 2015. "ImageNet Large Scale Visual Recognition Challenge." *International Journal of Computer Vision* 115 (3): 211–252. doi:10.1007/s11263-015-0816-y.

Schindler, Alexander, and Andreas Rauber. 2015. "An Audio-Visual Approach to Music Genre Classification through Affective Color Features." In *Proceedings of the 37th European Conference on IR Research,* 61–67. Cham: Springer.

———. 2016. "Harnessing Music-Related Visual Stereotypes for Music Information Retrieval." *ACM Transactions on Intelligent Systems and Technology* 8 (2): 20:1–20:20. doi:10.1145/2926719.

Silver, Daniel, Monica Lee, and C. Clayton Childress. 2016. "Genre Complexes in Popular Music." *PloS one* 11 (5): e0155471. doi:10.1371/journal.pone.0155471.

Sproles, George B. 1981. "Analyzing Fashion Life Cycles-Principles and Perspectives." *Journal of Marketing* 45 (4): 116–124. doi:10.1177/002224298104500415.

Strähle, Jochen. 2017. *Fashion & Music.* Springer Series in Fashion Business. Singapore: Springer. http://link.springer.com/10.1007/978-981-10-5637-6.

Strinati, Dominic. 2004. *An Introduction to Theories of Popular Culture.* London: Routledge.

Sturm, Bob L. 2012. "A Survey of Evaluation in Music Genre Recognition." In *International Workshop on Adaptive Multimedia Retrieval,* 29–66. Cham: Springer.

Szegedy, Christian, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. "Going Deeper with Convolutions." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 1–9. Red Hook, NY: Curran Associates.

Tzanetakis, George, and Perry Cook. 2002. "Musical Genre Classification of Audio Signals." *IEEE Transactions on Speech and Audio Processing* 10 (5): 293–302. doi:10.1109/TSA.2002.800560.

Van Venrooij, Alex. 2009. "The Aesthetic Discourse Space of Popular Music: 1985–86 and 2004–05." *Poetics* 37 (4): 315–332. doi:10.1016/j.poetic.2009.06.005.

Whitman, Brian, and Paris Smaragdis. 2002. "Combining Musical and Cultural Features for Intelligent Style Detection." In *Proceedings of the International Conference on Music Information Retrieval,* 47–52. Paris, France: IRCAM.

Yamaguchi, Kota, M. Hadi Kiapour, and Tamara L. Berg. 2013. "Paper Doll Parsing: Retrieving Similar Styles to Parse Clothing Items." In *Proceedings of the IEEE International Conference on Computer Vision,* 3519–3526. Washington, DC: IEEE Computer Society.

Yamaguchi, Kota, M. Hadi Kiapour, Luis E. Ortiz, and Tamara L. Berg. 2012. "Parsing Clothing in Fashion Photographs." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 3570–3577. Red Hook, NY: Curran Associates.

Yang, Jun, Yu-Gang Jiang, Alexander G. Hauptmann, and Chong-Wah Ngo. 2007. "Evaluating Bag-of-Visual-Words Representations in Scene Classification." In *Proceedings of the International Workshop on Multimedia Information Retrieval,* 197–206. New York, NY: ACM.

Yang, Yi, and Deva Ramanan. 2011. "Articulated Pose Estimation with Flexible Mixtures-of-Parts." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 1385–1392. Red Hook, NY: Curran Associates.

Zheng, Alice. 2015. *Evaluating Machine Learning Models: A Beginner's Guide to Key Concepts and Pitfalls.* Sebastopol, CA: O'Reilly Media. https://www.oreilly.com/library/view/evaluating-machine-learning/9781492048756/.

Appendices

A Fashion Classes

The following is a list of the 304 synsets and their indices that were used in the clothing detection model (GoogLeNet) described in Section 3.2.1. The synsets were retrieved from the pre-trained image model available with the DeepDetect API.[1] The synsets were listed in the text file accompanying the model. The model was obtained in January 2019.

```
0   G-string, thong
1   sarong
2   wig
3   cocktail dress, sheath
4   sunhat, sun hat
5   trench coat
6   strapless
7   khimar
8   crown, diadem
9   military uniform
10  beret
11  man's clothing
12  hairpiece, false hair, postiche
13  pants suit, pantsuit
14  turban
15  scarf
16  hot pants
17  batting helmet
18  work-shirt
19  helmet
20  dinner dress, dinner gown, formal, evening gown
21  two-piece, two-piece suit, lounge suit
22  street clothes
23  ski cap, stocking cap, toboggan cap
24  doublet
25  abaya
26  ballet skirt, tutu
27  porkpie, porkpie hat
28  cloak
29  dress suit, full dress, tailcoat, tail coat, tails, white tie, white tie and tails
30  dashiki, daishiki
31  bolero
32  garter belt, suspender belt
33  jean, blue jean, denim
34  cavalier hat, slouch hat
35  bonnet, poke bonnet
36  hood
```

1. https://www.deepdetect.com/applications/model

```
37  chemise, sack, shift
38  collar
39  slacks
40  pullover, slipover
41  holster
42  money belt
43  ball gown
44  hoopskirt, crinoline
45  grey, gray
46  spacesuit
47  camisole, underbodice
48  sweat pants, sweatpants
49  gaiter
50  caftan, kaftan
51  dress hat, high hat, opera hat, silk hat, stovepipe, top hat, topper, beaver
52  overall
53  undergarment, unmentionable
54  corset, girdle, stays
55  toupee, toupe
56  hose
57  overall, boilersuit, boilers suit
58  foul-weather gear
59  fur
60  burqa, burka
61  pajama, pyjama, pj's, jammies
62  hosiery, hose
63  maillot, tank suit
64  poncho
65  necktie, tie
66  cowboy hat, ten-gallon hat
67  hand-me-down
68  neckpiece
69  dirndl
70  formalwear, eveningwear, evening dress, evening clothes
71  legging, leging, leg covering
72  buckskins
73  beanie, beany
74  attire, garb, dress
75  short pants, shorts, trunks
76  beachwear
77  thong
78  pantie, panty, scanty, step-in
79  coat
80  frock coat
81  tam, tam-o'-shanter, tammy
82  toga
83  knee pad
84  millinery, woman's hat
85  bodice
86  sack coat
87  singlet, vest, undershirt
88  dress uniform
89  trouser
90  footwear
91  fur coat
92  capote, hooded coat
93  greatcoat, overcoat, topcoat
94  romper, romper suit
95  argyle, argyll
96  apparel, wearing apparel, dress, clothes
97  safety belt, life belt, safety harness
```

```
98  sable coat
99  bathrobe
100 fur hat
101 crinoline
102 knitwear
103 head covering, veil
104 kilt
105 mess jacket, monkey jacket, shell jacket
106 gown
107 trouser, pant
108 tudung
109 outerwear, overclothes
110 miniskirt, mini
111 golf glove
112 pinstripe
113 headpiece
114 coonskin cap, coonskin
115 pressure suit
116 Windsor tie
117 fatigues
118 slip-on
119 diaper, nappy, napkin
120 stretch pants
121 academic gown, academic robe, judge's robe
122 tights, leotards
123 bellbottom trousers, bell-bottoms, bellbottom pants
124 baseball cap, jockey cap, golf cap
125 jacket
126 Christmas stocking
127 pillbox, toque, turban
128 knee-high, knee-hi
129 costume
130 panty girdle
131 hard hat, tin hat, safety hat
132 chador, chadar, chaddar, chuddar
133 sundress
134 blue
135 cowl
136 bridal gown, wedding gown, wedding dress
137 chemise, shimmy, shift, slip, teddy
138 elbow pad
139 tea gown
140 fez, tarboosh
141 hat, chapeau, lid
142 black
143 drawers, underdrawers, shorts, boxers, boxershorts
144 spat, gaiter
145 lab coat, laboratory coat
146 pajama, pyjama
147 jersey, T-shirt, tee shirt
148 khakis
149 bathing cap, swimming cap
150 brassiere, bra, bandeau
151 cloth cap, flat cap
152 toga virilis
153 cap
154 muffler
155 stole
156 batting glove
157 garment
158 hijab
```

159 chasuble
160 sari, saree
161 bikini, two-piece
162 pea jacket, peacoat
163 fancy dress, masquerade, masquerade costume
164 outfit, getup, rig, turnout
165 raincoat, waterproof
166 nightgown, gown, nightie, night-robe, nightdress
167 tabi, tabis
168 gown, robe
169 dinner jacket, tux, tuxedo, black tie
170 underpants
171 mortarboard
172 coverall
173 shower cap
174 lederhosen
175 glove
176 Bermuda shorts, Jamaica shorts
177 cravat
178 disguise
179 Afro-wig
180 kid glove, suede glove
181 double-breasted suit
182 cloche
183 mask
184 clothing, article of clothing, vesture, wear, wearable, habiliment
185 diving suit, diving dress
186 sock
187 frock
188 roll-on
189 dress, frock
190 suit, suit of clothes
191 vest, waistcoat
192 separate
193 mink, mink coat
194 bloomers, pants, drawers, knickers
195 gauntlet, gantlet
196 apron
197 lingerie, intimate apparel
198 maxi
199 neckerchief
200 niqab
201 dressing gown, robe-de-chambre, lounging robe
202 shawl
203 mantilla
204 single-breasted suit
205 swallow-tailed coat, swallowtail, morning coat
206 athletic supporter, supporter, suspensor, jockstrap, jock
207 domino, half mask, eye mask
208 kurta
209 three-piece suit
210 ao dai
211 neckwear
212 sportswear, athletic wear, activewear
213 sweater, jumper
214 picture hat
215 nylons, nylon stocking, rayons, rayon stocking, silk stocking
216 bow tie, bow-tie, bowtie
217 shirt
218 bowler hat, bowler, derby hat, derby, plug hat
219 tunic

220 raglan
221 grass skirt
222 skirt
223 dirndl
224 Levi's, levis
225 surcoat
226 bikini pants
227 wraparound
228 face veil
229 athletic sock, sweat sock, varsity sock
230 academic costume
231 sheepskin coat, afghan
232 kepi, peaked cap, service cap, yachting cap
233 balldress
234 headscarf
235 stocking
236 swimming trunks, bathing trunks
237 jodhpurs, jodhpur breeches, riding breeches
238 jump suit, jumpsuit
239 briefs, Jockey shorts
240 salwar, shalwar
241 nightwear, sleepwear, nightclothes
242 negligee, neglige, peignoir, wrapper, housecoat
243 yarmulke, yarmulka, yarmelke
244 mitten
245 feather boa, boa
246 caftan, kaftan
247 uplift
248 wet suit
249 sweatshirt
250 polo shirt, sport shirt
251 jumper, pinafore, pinny
252 blouse
253 maillot
254 kameez
255 undies
256 snowsuit
257 single-breasted jacket
258 tiara
259 underwear, underclothes, underclothing
260 pith hat, pith helmet, sun helmet, topee, topi
261 bearskin, busby, shako
262 sweat suit, sweatsuit, sweats, workout suit
263 pedal pusher, toreador pants
264 fedora, felt hat, homburg, Stetson, trilby
265 cardigan
266 pantyhose
267 long trousers, long pants
268 headdress, headgear
269 swimsuit, swimwear, bathing suit, swimming costume, bathing costume
270 bomber jacket
271 belt
272 costume
273 ready-to-wear
274 football helmet
275 sombrero
276 balaclava, balaclava helmet
277 battle dress
278 petticoat, half-slip, underskirt
279 watch cap
280 snap-brim hat

281 business suit
282 black tie
283 woman's clothing
284 furnishing, trappings
285 vestment
286 boater, leghorn, Panama, Panama hat, sailor, skimmer, straw hat
287 kimono
288 turtleneck, turtle, polo-neck
289 skullcap
290 straitjacket, straightjacket
291 overgarment, outer garment
292 leotard, unitard, body suit, cat suit
293 halter
294 brace, suspender, gallus
295 cords, corduroys
296 parka, windbreaker, windcheater, anorak
297 seat belt, seatbelt
298 robe
299 ensemble
300 crash helmet
301 tricorn, tricorne
302 array, raiment, regalia
303 camisole

B ImageNet ILSVRC Classes

The following is a list of the 1,000 ILSVRC synsets and their indices that were used in the object detection model (ResNet-50) described in Section 3.2.1. The synsets were retrieved from the pre-trained image model available with the DeepDetect API.[1] The synsets were listed in the text file accompanying the model. The model was obtained in January 2019.

```
0   tench, Tinca tinca
1   goldfish, Carassius auratus
2   great white shark, white shark, man-eater, man-eating shark, Carcharodon carcharias
3   tiger shark, Galeocerdo cuvieri
4   hammerhead, hammerhead shark
5   electric ray, crampfish, numbfish, torpedo
6   stingray
7   cock
8   hen
9   ostrich, Struthio camelus
10  brambling, Fringilla montifringilla
11  goldfinch, Carduelis carduelis
12  house finch, linnet, Carpodacus mexicanus
13  junco, snowbird
14  indigo bunting, indigo finch, indigo bird, Passerina cyanea
15  robin, American robin, Turdus migratorius
16  bulbul
17  jay
18  magpie
19  chickadee
20  water ouzel, dipper
21  kite
22  bald eagle, American eagle, Haliaeetus leucocephalus
23  vulture
24  great grey owl, great gray owl, Strix nebulosa
25  European fire salamander, Salamandra salamandra
26  common newt, Triturus vulgaris
27  eft
28  spotted salamander, Ambystoma maculatum
29  axolotl, mud puppy, Ambystoma mexicanum
30  bullfrog, Rana catesbeiana
31  tree frog, tree-frog
32  tailed frog, bell toad, ribbed toad, tailed toad, Ascaphus trui
33  loggerhead, loggerhead turtle, Caretta caretta
34  leatherback turtle, leatherback, leathery turtle, Dermochelys coriacea
35  mud turtle
36  terrapin
```

1. https://www.deepdetect.com/applications/model

```
37  box turtle, box tortoise
38  banded gecko
39  common iguana, iguana, Iguana iguana
40  American chameleon, anole, Anolis carolinensis
41  whiptail, whiptail lizard
42  agama
43  frilled lizard, Chlamydosaurus kingi
44  alligator lizard
45  Gila monster, Heloderma suspectum
46  green lizard, Lacerta viridis
47  African chameleon, Chamaeleo chamaeleon
48  Komodo dragon, Komodo lizard, dragon lizard, giant lizard, Varanus komodoensis
49  African crocodile, Nile crocodile, Crocodylus niloticus
50  American alligator, Alligator mississipiensis
51  triceratops
52  thunder snake, worm snake, Carphophis amoenus
53  ringneck snake, ring-necked snake, ring snake
54  hognose snake, puff adder, sand viper
55  green snake, grass snake
56  king snake, kingsnake
57  garter snake, grass snake
58  water snake
59  vine snake
60  night snake, Hypsiglena torquata
61  boa constrictor, Constrictor constrictor
62  rock python, rock snake, Python sebae
63  Indian cobra, Naja naja
64  green mamba
65  sea snake
66  horned viper, cerastes, sand viper, horned asp, Cerastes cornutus
67  diamondback, diamondback rattlesnake, Crotalus adamanteus
68  sidewinder, horned rattlesnake, Crotalus cerastes
69  trilobite
70  harvestman, daddy longlegs, Phalangium opilio
71  scorpion
72  black and gold garden spider, Argiope aurantia
73  barn spider, Araneus cavaticus
74  garden spider, Aranea diadematа
75  black widow, Latrodectus mactans
76  tarantula
77  wolf spider, hunting spider
78  tick
79  centipede
80  black grouse
81  ptarmigan
82  ruffed grouse, partridge, Bonasa umbellus
83  prairie chicken, prairie grouse, prairie fowl
84  peacock
85  quail
86  partridge
87  African grey, African gray, Psittacus erithacus
88  macaw
89  sulphur-crested cockatoo, Kakatoe galerita, Cacatua galerita
90  lorikeet
91  coucal
92  bee eater
93  hornbill
94  hummingbird
95  jacamar
96  toucan
97  drake
```

```
98  red-breasted merganser, Mergus serrator
99  goose
100 black swan, Cygnus atratus
101 tusker
102 echidna, spiny anteater, anteater
103 platypus, duckbill, duckbilled platypus, duck-billed platypus, Ornithorhynchus anatinus
104 wallaby, brush kangaroo
105 koala, koala bear, kangaroo bear, native bear, Phascolarctos cinereus
106 wombat
107 jellyfish
108 sea anemone, anemone
109 brain coral
110 flatworm, platyhelminth
111 nematode, nematode worm, roundworm
112 conch
113 snail
114 slug
115 sea slug, nudibranch
116 chiton, coat-of-mail shell, sea cradle, polyplacophore
117 chambered nautilus, pearly nautilus, nautilus
118 Dungeness crab, Cancer magister
119 rock crab, Cancer irroratus
120 fiddler crab
121 king crab, Alaska crab, Alaskan king crab, Alaska king crab, Paralithodes camtschatica
122 American lobster, Northern lobster, Maine lobster, Homarus americanus
123 spiny lobster, langouste, rock lobster, crawfish, crayfish, sea crawfish
124 crayfish, crawfish, crawdad, crawdaddy
125 hermit crab
126 isopod
127 white stork, Ciconia ciconia
128 black stork, Ciconia nigra
129 spoonbill
130 flamingo
131 little blue heron, Egretta caerulea
132 American egret, great white heron, Egretta albus
133 bittern
134 crane
135 limpkin, Aramus pictus
136 European gallinule, Porphyrio porphyrio
137 American coot, marsh hen, mud hen, water hen, Fulica americana
138 bustard
139 ruddy turnstone, Arenaria interpres
140 red-backed sandpiper, dunlin, Erolia alpina
141 redshank, Tringa totanus
142 dowitcher
143 oystercatcher, oyster catcher
144 pelican
145 king penguin, Aptenodytes patagonica
146 albatross, mollymawk
147 grey whale, gray whale, devilfish, Eschrichtius gibbosus, Eschrichtius robustus
148 killer whale, killer, orca, grampus, sea wolf, Orcinus orca
149 dugong, Dugong dugon
150 sea lion
151 Chihuahua
152 Japanese spaniel
153 Maltese dog, Maltese terrier, Maltese
154 Pekinese, Pekingese, Peke
155 Shih-Tzu
156 Blenheim spaniel
157 papillon
158 toy terrier
```

```
159 Rhodesian ridgeback
160 Afghan hound, Afghan
161 basset, basset hound
162 beagle
163 bloodhound, sleuthhound
164 bluetick
165 black-and-tan coonhound
166 Walker hound, Walker foxhound
167 English foxhound
168 redbone
169 borzoi, Russian wolfhound
170 Irish wolfhound
171 Italian greyhound
172 whippet
173 Ibizan hound, Ibizan Podenco
174 Norwegian elkhound, elkhound
175 otterhound, otter hound
176 Saluki, gazelle hound
177 Scottish deerhound, deerhound
178 Weimaraner
179 Staffordshire bullterrier, Staffordshire bull terrier
180 American Staffordshire terrier, Staffordshire terrier, American pit bull terrier, pit
    bull terrier
181 Bedlington terrier
182 Border terrier
183 Kerry blue terrier
184 Irish terrier
185 Norfolk terrier
186 Norwich terrier
187 Yorkshire terrier
188 wire-haired fox terrier
189 Lakeland terrier
190 Sealyham terrier, Sealyham
191 Airedale, Airedale terrier
192 cairn, cairn terrier
193 Australian terrier
194 Dandie Dinmont, Dandie Dinmont terrier
195 Boston bull, Boston terrier
196 miniature schnauzer
197 giant schnauzer
198 standard schnauzer
199 Scotch terrier, Scottish terrier, Scottie
200 Tibetan terrier, chrysanthemum dog
201 silky terrier, Sydney silky
202 soft-coated wheaten terrier
203 West Highland white terrier
204 Lhasa, Lhasa apso
205 flat-coated retriever
206 curly-coated retriever
207 golden retriever
208 Labrador retriever
209 Chesapeake Bay retriever
210 German short-haired pointer
211 vizsla, Hungarian pointer
212 English setter
213 Irish setter, red setter
214 Gordon setter
215 Brittany spaniel
216 clumber, clumber spaniel
217 English springer, English springer spaniel
218 Welsh springer spaniel
```

```
219 cocker spaniel, English cocker spaniel, cocker
220 Sussex spaniel
221 Irish water spaniel
222 kuvasz
223 schipperke
224 groenendael
225 malinois
226 briard
227 kelpie
228 komondor
229 Old English sheepdog, bobtail
230 Shetland sheepdog, Shetland sheep dog, Shetland
231 collie
232 Border collie
233 Bouvier des Flandres, Bouviers des Flandres
234 Rottweiler
235 German shepherd, German shepherd dog, German police dog, alsatian
236 Doberman, Doberman pinscher
237 miniature pinscher
238 Greater Swiss Mountain dog
239 Bernese mountain dog
240 Appenzeller
241 EntleBucher
242 boxer
243 bull mastiff
244 Tibetan mastiff
245 French bulldog
246 Great Dane
247 Saint Bernard, St Bernard
248 Eskimo dog, husky
249 malamute, malemute, Alaskan malamute
250 Siberian husky
251 dalmatian, coach dog, carriage dog
252 affenpinscher, monkey pinscher, monkey dog
253 basenji
254 pug, pug-dog
255 Leonberg
256 Newfoundland, Newfoundland dog
257 Great Pyrenees
258 Samoyed, Samoyede
259 Pomeranian
260 chow, chow chow
261 keeshond
262 Brabancon griffon
263 Pembroke, Pembroke Welsh corgi
264 Cardigan, Cardigan Welsh corgi
265 toy poodle
266 miniature poodle
267 standard poodle
268 Mexican hairless
269 timber wolf, grey wolf, gray wolf, Canis lupus
270 white wolf, Arctic wolf, Canis lupus tundrarum
271 red wolf, maned wolf, Canis rufus, Canis niger
272 coyote, prairie wolf, brush wolf, Canis latrans
273 dingo, warrigal, warragal, Canis dingo
274 dhole, Cuon alpinus
275 African hunting dog, hyena dog, Cape hunting dog, Lycaon pictus
276 hyena, hyaena
277 red fox, Vulpes vulpes
278 kit fox, Vulpes macrotis
279 Arctic fox, white fox, Alopex lagopus
```

```
280 grey fox, gray fox, Urocyon cinereoargenteus
281 tabby, tabby cat
282 tiger cat
283 Persian cat
284 Siamese cat, Siamese
285 Egyptian cat
286 cougar, puma, catamount, mountain lion, painter, panther, Felis concolor
287 lynx, catamount
288 leopard, Panthera pardus
289 snow leopard, ounce, Panthera uncia
290 jaguar, panther, Panthera onca, Felis onca
291 lion, king of beasts, Panthera leo
292 tiger, Panthera tigris
293 cheetah, chetah, Acinonyx jubatus
294 brown bear, bruin, Ursus arctos
295 American black bear, black bear, Ursus americanus, Euarctos americanus
296 ice bear, polar bear, Ursus Maritimus, Thalarctos maritimus
297 sloth bear, Melursus ursinus, Ursus ursinus
298 mongoose
299 meerkat, mierkat
300 tiger beetle
301 ladybug, ladybeetle, lady beetle, ladybird, ladybird beetle
302 ground beetle, carabid beetle
303 long-horned beetle, longicorn, longicorn beetle
304 leaf beetle, chrysomelid
305 dung beetle
306 rhinoceros beetle
307 weevil
308 fly
309 bee
310 ant, emmet, pismire
311 grasshopper, hopper
312 cricket
313 walking stick, walkingstick, stick insect
314 cockroach, roach
315 mantis, mantid
316 cicada, cicala
317 leafhopper
318 lacewing, lacewing fly
319 dragonfly, darning needle, devil's darning needle, sewing needle, snake feeder, snake
    doctor, mosquito hawk, skeeter hawk
320 damselfly
321 admiral
322 ringlet, ringlet butterfly
323 monarch, monarch butterfly, milkweed butterfly, Danaus plexippus
324 cabbage butterfly
325 sulphur butterfly, sulfur butterfly
326 lycaenid, lycaenid butterfly
327 starfish, sea star
328 sea urchin
329 sea cucumber, holothurian
330 wood rabbit, cottontail, cottontail rabbit
331 hare
332 Angora, Angora rabbit
333 hamster
334 porcupine, hedgehog
335 fox squirrel, eastern fox squirrel, Sciurus niger
336 marmot
337 beaver
338 guinea pig, Cavia cobaya
339 sorrel
```

```
340 zebra
341 hog, pig, grunter, squealer, Sus scrofa
342 wild boar, boar, Sus scrofa
343 warthog
344 hippopotamus, hippo, river horse, Hippopotamus amphibius
345 ox
346 water buffalo, water ox, Asiatic buffalo, Bubalus bubalis
347 bison
348 ram, tup
349 bighorn, bighorn sheep, cimarron, Rocky Mountain bighorn, Rocky Mountain sheep, Ovis
    canadensis
350 ibex, Capra ibex
351 hartebeest
352 impala, Aepyceros melampus
353 gazelle
354 Arabian camel, dromedary, Camelus dromedarius
355 llama
356 weasel
357 mink
358 polecat, fitch, foulmart, foumart, Mustela putorius
359 black-footed ferret, ferret, Mustela nigripes
360 otter
361 skunk, polecat, wood pussy
362 badger
363 armadillo
364 three-toed sloth, ai, Bradypus tridactylus
365 orangutan, orang, orangutang, Pongo pygmaeus
366 gorilla, Gorilla gorilla
367 chimpanzee, chimp, Pan troglodytes
368 gibbon, Hylobates lar
369 siamang, Hylobates syndactylus, Symphalangus syndactylus
370 guenon, guenon monkey
371 patas, hussar monkey, Erythrocebus patas
372 baboon
373 macaque
374 langur
375 colobus, colobus monkey
376 proboscis monkey, Nasalis larvatus
377 marmoset
378 capuchin, ringtail, Cebus capucinus
379 howler monkey, howler
380 titi, titi monkey
381 spider monkey, Ateles geoffroyi
382 squirrel monkey, Saimiri sciureus
383 Madagascar cat, ring-tailed lemur, Lemur catta
384 indri, indris, Indri indri, Indri brevicaudatus
385 Indian elephant, Elephas maximus
386 African elephant, Loxodonta africana
387 lesser panda, red panda, panda, bear cat, cat bear, Ailurus fulgens
388 giant panda, panda, panda bear, coon bear, Ailuropoda melanoleuca
389 barracouta, snoek
390 eel
391 coho, cohoe, coho salmon, blue jack, silver salmon, Oncorhynchus kisutch
392 rock beauty, Holocanthus tricolor
393 anemone fish
394 sturgeon
395 gar, garfish, garpike, billfish, Lepisosteus osseus
396 lionfish
397 puffer, pufferfish, blowfish, globefish
398 abacus
399 abaya
```

```
400 academic gown, academic robe, judge's robe
401 accordion, piano accordion, squeeze box
402 acoustic guitar
403 aircraft carrier, carrier, flattop, attack aircraft carrier
404 airliner
405 airship, dirigible
406 altar
407 ambulance
408 amphibian, amphibious vehicle
409 analog clock
410 apiary, bee house
411 apron
412 ashcan, trash can, garbage can, wastebin, ash bin, ash-bin, ashbin, dustbin, trash
    barrel, trash bin
413 assault rifle, assault gun
414 backpack, back pack, knapsack, packsack, rucksack, haversack
415 bakery, bakeshop, bakehouse
416 balance beam, beam
417 balloon
418 ballpoint, ballpoint pen, ballpen, Biro
419 Band Aid
420 banjo
421 bannister, banister, balustrade, balusters, handrail
422 barbell
423 barber chair
424 barbershop
425 barn
426 barometer
427 barrel, cask
428 barrow, garden cart, lawn cart, wheelbarrow
429 baseball
430 basketball
431 bassinet
432 bassoon
433 bathing cap, swimming cap
434 bath towel
435 bathtub, bathing tub, bath, tub
436 beach wagon, station wagon, wagon, estate car, beach waggon, station waggon, waggon
437 beacon, lighthouse, beacon light, pharos
438 beaker
439 bearskin, busby, shako
440 beer bottle
441 beer glass
442 bell cote, bell cot
443 bib
444 bicycle-built-for-two, tandem bicycle, tandem
445 bikini, two-piece
446 binder, ring-binder
447 binoculars, field glasses, opera glasses
448 birdhouse
449 boathouse
450 bobsled, bobsleigh, bob
451 bolo tie, bolo, bola tie, bola
452 bonnet, poke bonnet
453 bookcase
454 bookshop, bookstore, bookstall
455 bottlecap
456 bow
457 bow tie, bow-tie, bowtie
458 brass, memorial tablet, plaque
459 brassiere, bra, bandeau
```

```
460 breakwater, groin, groyne, mole, bulwark, seawall, jetty
461 breastplate, aegis, egis
462 broom
463 bucket, pail
464 buckle
465 bulletproof vest
466 bullet train, bullet
467 butcher shop, meat market
468 cab, hack, taxi, taxicab
469 caldron, cauldron
470 candle, taper, wax light
471 cannon
472 canoe
473 can opener, tin opener
474 cardigan
475 car mirror
476 carousel, carrousel, merry-go-round, roundabout, whirligig
477 carpenter's kit, tool kit
478 carton
479 car wheel
480 cash machine, cash dispenser, automated teller machine, automatic teller machine,
    automated teller, automatic teller, ATM
481 cassette
482 cassette player
483 castle
484 catamaran
485 CD player
486 cello, violoncello
487 cellular telephone, cellular phone, cellphone, cell, mobile phone
488 chain
489 chainlink fence
490 chain mail, ring mail, mail, chain armor, chain armour, ring armor, ring armour
491 chain saw, chainsaw
492 chest
493 chiffonier, commode
494 chime, bell, gong
495 china cabinet, china closet
496 Christmas stocking
497 church, church building
498 cinema, movie theater, movie theatre, movie house, picture palace
499 cleaver, meat cleaver, chopper
500 cliff dwelling
501 cloak
502 clog, geta, patten, sabot
503 cocktail shaker
504 coffee mug
505 coffeepot
506 coil, spiral, volute, whorl, helix
507 combination lock
508 computer keyboard, keypad
509 confectionery, confectionary, candy store
510 container ship, containership, container vessel
511 convertible
512 corkscrew, bottle screw
513 cornet, horn, trumpet, trump
514 cowboy boot
515 cowboy hat, ten-gallon hat
516 cradle
517 crane
518 crash helmet
519 crate
```

```
520 crib, cot
521 Crock Pot
522 croquet ball
523 crutch
524 cuirass
525 dam, dike, dyke
526 desk
527 desktop computer
528 dial telephone, dial phone
529 diaper, nappy, napkin
530 digital clock
531 digital watch
532 dining table, board
533 dishrag, dishcloth
534 dishwasher, dish washer, dishwashing machine
535 disk brake, disc brake
536 dock, dockage, docking facility
537 dogsled, dog sled, dog sleigh
538 dome
539 doormat, welcome mat
540 drilling platform, offshore rig
541 drum, membranophone, tympan
542 drumstick
543 dumbbell
544 Dutch oven
545 electric fan, blower
546 electric guitar
547 electric locomotive
548 entertainment center
549 envelope
550 espresso maker
551 face powder
552 feather boa, boa
553 file, file cabinet, filing cabinet
554 fireboat
555 fire engine, fire truck
556 fire screen, fireguard
557 flagpole, flagstaff
558 flute, transverse flute
559 folding chair
560 football helmet
561 forklift
562 fountain
563 fountain pen
564 four-poster
565 freight car
566 French horn, horn
567 frying pan, frypan, skillet
568 fur coat
569 garbage truck, dustcart
570 gasmask, respirator, gas helmet
571 gas pump, gasoline pump, petrol pump, island dispenser
572 goblet
573 go-kart
574 golf ball
575 golfcart, golf cart
576 gondola
577 gong, tam-tam
578 gown
579 grand piano, grand
580 greenhouse, nursery, glasshouse
```

```
581 grille, radiator grille
582 grocery store, grocery, food market, market
583 guillotine
584 hair slide
585 hair spray
586 half track
587 hammer
588 hamper
589 hand blower, blow dryer, blow drier, hair dryer, hair drier
590 hand-held computer, hand-held microcomputer
591 handkerchief, hankie, hanky, hankey
592 hard disc, hard disk, fixed disk
593 harmonica, mouth organ, harp, mouth harp
594 harp
595 harvester, reaper
596 hatchet
597 holster
598 home theater, home theatre
599 honeycomb
600 hook, claw
601 hoopskirt, crinoline
602 horizontal bar, high bar
603 horse cart, horse-cart
604 hourglass
605 iPod
606 iron, smoothing iron
607 jack-o'-lantern
608 jean, blue jean, denim
609 jeep, landrover
610 jersey, T-shirt, tee shirt
611 jigsaw puzzle
612 jinrikisha, ricksha, rickshaw
613 joystick
614 kimono
615 knee pad
616 knot
617 lab coat, laboratory coat
618 ladle
619 lampshade, lamp shade
620 laptop, laptop computer
621 lawn mower, mower
622 lens cap, lens cover
623 letter opener, paper knife, paperknife
624 library
625 lifeboat
626 lighter, light, igniter, ignitor
627 limousine, limo
628 liner, ocean liner
629 lipstick, lip rouge
630 Loafer
631 lotion
632 loudspeaker, speaker, speaker unit, loudspeaker system, speaker system
633 loupe, jeweler's loupe
634 lumbermill, sawmill
635 magnetic compass
636 mailbag, postbag
637 mailbox, letter box
638 maillot
639 maillot, tank suit
640 manhole cover
641 maraca
```

```
642 marimba, xylophone
643 mask
644 matchstick
645 maypole
646 maze, labyrinth
647 measuring cup
648 medicine chest, medicine cabinet
649 megalith, megalithic structure
650 microphone, mike
651 microwave, microwave oven
652 military uniform
653 milk can
654 minibus
655 miniskirt, mini
656 minivan
657 missile
658 mitten
659 mixing bowl
660 mobile home, manufactured home
661 Model T
662 modem
663 monastery
664 monitor
665 moped
666 mortar
667 mortarboard
668 mosque
669 mosquito net
670 motor scooter, scooter
671 mountain bike, all-terrain bike, off-roader
672 mountain tent
673 mouse, computer mouse
674 mousetrap
675 moving van
676 muzzle
677 nail
678 neck brace
679 necklace
680 nipple
681 notebook, notebook computer
682 obelisk
683 oboe, hautboy, hautbois
684 ocarina, sweet potato
685 odometer, hodometer, mileometer, milometer
686 oil filter
687 organ, pipe organ
688 oscilloscope, scope, cathode-ray oscilloscope, CRO
689 overskirt
690 oxcart
691 oxygen mask
692 packet
693 paddle, boat paddle
694 paddlewheel, paddle wheel
695 padlock
696 paintbrush
697 pajama, pyjama, pj's, jammies
698 palace
699 panpipe, pandean pipe, syrinx
700 paper towel
701 parachute, chute
702 parallel bars, bars
```

```
703 park bench
704 parking meter
705 passenger car, coach, carriage
706 patio, terrace
707 pay-phone, pay-station
708 pedestal, plinth, footstall
709 pencil box, pencil case
710 pencil sharpener
711 perfume, essence
712 Petri dish
713 photocopier
714 pick, plectrum, plectron
715 pickelhaube
716 picket fence, paling
717 pickup, pickup truck
718 pier
719 piggy bank, penny bank
720 pill bottle
721 pillow
722 ping-pong ball
723 pinwheel
724 pirate, pirate ship
725 pitcher, ewer
726 plane, carpenter's plane, woodworking plane
727 planetarium
728 plastic bag
729 plate rack
730 plow, plough
731 plunger, plumber's helper
732 Polaroid camera, Polaroid Land camera
733 pole
734 police van, police wagon, paddy wagon, patrol wagon, wagon, black Maria
735 poncho
736 pool table, billiard table, snooker table
737 pop bottle, soda bottle
738 pot, flowerpot
739 potter's wheel
740 power drill
741 prayer rug, prayer mat
742 printer
743 prison, prison house
744 projectile, missile
745 projector
746 puck, hockey puck
747 punching bag, punch bag, punching ball, punchball
748 purse
749 quill, quill pen
750 quilt, comforter, comfort, puff
751 racer, race car, racing car
752 racket, racquet
753 radiator
754 radio, wireless
755 radio telescope, radio reflector
756 rain barrel
757 recreational vehicle, RV, R.V.
758 reel
759 reflex camera
760 refrigerator, icebox
761 remote control, remote
762 restaurant, eating house, eating place, eatery
763 revolver, six-gun, six-shooter
```

```
764 rifle
765 rocking chair, rocker
766 rotisserie
767 rubber eraser, rubber, pencil eraser
768 rugby ball
769 rule, ruler
770 running shoe
771 safe
772 safety pin
773 saltshaker, salt shaker
774 sandal
775 sarong
776 sax, saxophone
777 scabbard
778 scale, weighing machine
779 school bus
780 schooner
781 scoreboard
782 screen, CRT screen
783 screw
784 screwdriver
785 seat belt, seatbelt
786 sewing machine
787 shield, buckler
788 shoe shop, shoe-shop, shoe store
789 shoji
790 shopping basket
791 shopping cart
792 shovel
793 shower cap
794 shower curtain
795 ski
796 ski mask
797 sleeping bag
798 slide rule, slipstick
799 sliding door
800 slot, one-armed bandit
801 snorkel
802 snowmobile
803 snowplow, snowplough
804 soap dispenser
805 soccer ball
806 sock
807 solar dish, solar collector, solar furnace
808 sombrero
809 soup bowl
810 space bar
811 space heater
812 space shuttle
813 spatula
814 speedboat
815 spider web, spider's web
816 spindle
817 sports car, sport car
818 spotlight, spot
819 stage
820 steam locomotive
821 steel arch bridge
822 steel drum
823 stethoscope
824 stole
```

```
825 stone wall
826 stopwatch, stop watch
827 stove
828 strainer
829 streetcar, tram, tramcar, trolley, trolley car
830 stretcher
831 studio couch, day bed
832 stupa, tope
833 submarine, pigboat, sub, U-boat
834 suit, suit of clothes
835 sundial
836 sunglass
837 sunglasses, dark glasses, shades
838 sunscreen, sunblock, sun blocker
839 suspension bridge
840 swab, swob, mop
841 sweatshirt
842 swimming trunks, bathing trunks
843 swing
844 switch, electric switch, electrical switch
845 syringe
846 table lamp
847 tank, army tank, armored combat vehicle, armoured combat vehicle
848 tape player
849 teapot
850 teddy, teddy bear
851 television, television system
852 tennis ball
853 thatch, thatched roof
854 theater curtain, theatre curtain
855 thimble
856 thresher, thrasher, threshing machine
857 throne
858 tile roof
859 toaster
860 tobacco shop, tobacconist shop, tobacconist
861 toilet seat
862 torch
863 totem pole
864 tow truck, tow car, wrecker
865 toyshop
866 tractor
867 trailer truck, tractor trailer, trucking rig, rig, articulated lorry, semi
868 tray
869 trench coat
870 tricycle, trike, velocipede
871 trimaran
872 tripod
873 triumphal arch
874 trolleybus, trolley coach, trackless trolley
875 trombone
876 tub, vat
877 turnstile
878 typewriter keyboard
879 umbrella
880 unicycle, monocycle
881 upright, upright piano
882 vacuum, vacuum cleaner
883 vase
884 vault
885 velvet
```

```
886 vending machine
887 vestment
888 viaduct
889 violin, fiddle
890 volleyball
891 waffle iron
892 wall clock
893 wallet, billfold, notecase, pocketbook
894 wardrobe, closet, press
895 warplane, military plane
896 washbasin, handbasin, washbowl, lavabo, wash-hand basin
897 washer, automatic washer, washing machine
898 water bottle
899 water jug
900 water tower
901 whiskey jug
902 whistle
903 wig
904 window screen
905 window shade
906 Windsor tie
907 wine bottle
908 wing
909 wok
910 wooden spoon
911 wool, woolen, woollen
912 worm fence, snake fence, snake-rail fence, Virginia fence
913 wreck
914 yawl
915 yurt
916 web site, website, internet site, site
917 comic book
918 crossword puzzle, crossword
919 street sign
920 traffic light, traffic signal, stoplight
921 book jacket, dust cover, dust jacket, dust wrapper
922 menu
923 plate
924 guacamole
925 consomme
926 hot pot, hotpot
927 trifle
928 ice cream, icecream
929 ice lolly, lolly, lollipop, popsicle
930 French loaf
931 bagel, beigel
932 pretzel
933 cheeseburger
934 hotdog, hot dog, red hot
935 mashed potato
936 head cabbage
937 broccoli
938 cauliflower
939 zucchini, courgette
940 spaghetti squash
941 acorn squash
942 butternut squash
943 cucumber, cuke
944 artichoke, globe artichoke
945 bell pepper
946 cardoon
```

```
947 mushroom
948 Granny Smith
949 strawberry
950 orange
951 lemon
952 fig
953 pineapple, ananas
954 banana
955 jackfruit, jak, jack
956 custard apple
957 pomegranate
958 hay
959 carbonara
960 chocolate sauce, chocolate syrup
961 dough
962 meat loaf, meatloaf
963 pizza, pizza pie
964 potpie
965 burrito
966 red wine
967 espresso
968 cup
969 eggnog
970 alp
971 bubble
972 cliff, drop, drop-off
973 coral reef
974 geyser
975 lakeside, lakeshore
976 promontory, headland, head, foreland
977 sandbar, sand bar
978 seashore, coast, seacoast, sea-coast
979 valley, vale
980 volcano
981 ballplayer, baseball player
982 groom, bridegroom
983 scuba diver
984 rapeseed
985 daisy
986 yellow lady's slipper, yellow lady-slipper, Cypripedium calceolus, Cypripedium
    parviflorum
987 corn
988 acorn
989 hip, rose hip, rosehip
990 buckeye, horse chestnut, conker
991 coral fungus
992 agaric
993 gyromitra
994 stinkhorn, carrion fungus
995 earthstar
996 hen-of-the-woods, hen of the woods, Polyporus frondosus, Grifola frondosa
997 bolete
998 ear, spike, capitulum
999 toilet tissue, toilet paper, bathroom tissue
```