Automatic composition of descriptive music: A case study of the relationship between image and sound

Lucía Martín-Gómez, Javier Pérez-Marcos, María Navarro Cáceres

BISITE Research Group, University of Salamanca, Calle Espejo, S/N, 37007 Salamanca, Spain

Abstract. Human beings establish relationships with the environment mainly through sight and hearing. This work focuses on the concept of descriptive music, which makes use of sound resources to narrate a story. The Fantasia film, produced by Walt Disney was used in the case study. One of its musical pieces is analyzed in order to obtain the relationship between image and music. This connection is subsequently used to create a descriptive musical composition from a new video. Naive Bayes, Support Vector Machine and Random Forest are the three classifiers studied for the model induction process. After an analysis of their performance, it was concluded that Random Forest provided the best solution; the produced musical composition had a considerably high descriptive quality.

Keywords: Descriptive music, automatic composition, image, video, classification

1 Introduction

Human beings establish all kind of relationships with their environment. These relationships are made thanks to the human senses, such as sight and hearing, which extract information from the surroundings. Cognitive processes allow us to assimilate the information [32]. Throughout history, images and music have been two of the most common means of interaction between humans and their surroundings.

Some authors have already studied the way the human being uses the senses of vision and audition to establish relationships. Thus, numerous approaches concerning the perception and cognition of music can be found in the literature. There are researches in the field of Psychology that study human reactions when a person is listening to music [1, 24]. Furthermore, some techniques such as sentiment analysis [29] and Brain Computer Interfaces (BCIs) [34] are used with the same purpose. On the other hand, there are some techniques that study the human perception of images. Specifically, a work called Eyetracking [7] measures eye positions and analyzes its movements with the purpose of understanding the way a person processes an image. In [33], some attentional regions are discovered in each image after the extraction of some descriptors.

Some works have already intended to capture human behavior in this regard, by using different approaches to compose music, with the aid of techniques such as synesthesia [11,20]. After a careful examination of the literature, we developed a new approach that aims to generate a sequence of sounds from a preliminary video. The final goal of the proposed system is to obtain a musical result that will describe the image provided by a user. To achieve this, we propose to divide the video into frames. Afterwards, the visual and auditory characteristics of these frames are extracted. Finally, a pattern must be established from this information by applying some data mining techniques.

The animated film Fantasia [3] produced by Walt Disney has been chosen for the case study. This film is made up of a concatenation of eight pieces of classical music, described with the illustrations of professional animators. Due to the deep analysis made by expert people, Fantasia will be used in this work in order to create a model that relates some characteristics of the images to sound. This pattern will then be applied to a new image in order to compose music. A deep analysis of 483 movie frames was carried out in order to extract a set of image descriptors; it will be used to perform the translation in the reverse direction. Furthermore, a musical analysis has been performed in order to extract information about the relationship between each of the frames studied and its sound. Then, the selected classifiers were applied to the data set, and due to the quality of the results Random Forest (RF) [2] was chosen for this proposal. Thus, a model which represented the relationship between the characteristics of the image and the sound was obtained. This model will then be applied in the last stage of this work, where a new video is divided into frames, and each one of them is translated into a sound. The concatenation of all the resulting sounds will compose the final melody.

Section 2 reviews some image-to-sound approaches found in the literature. Section 3 outlines the workflow of the system and Section 4 analyzes different techniques used for image processing and descriptive music composition that are used in this work. In Section 5, all the details of this case study are explained: firstly, a deep analysis of the animated film Fantasia is made and the obtained data are presented. Then, classification techniques are applied for the purpose of defining the relationship between the characteristics of image and sound. Different classification techniques are applied in this work, and their results are discussed in Section 6. This section also presents the obtained musical results, where the previously created model is used to compose a sequence of sounds that describes new images provided by users. Finally, the last section presents the conclusions drawn upon the completion of this work. Moreover, future lines of research are discussed.

2 Image to music conversion

The interest to fuse visual and musical art is not new. Some researchers have tried to find a psychological link between images and sounds through synesthesia or cognitive audition. Drawing on these phenomena, different approaches have been

developed, some related to the synesthesia, the spectograms or the descriptive music.

Synesthesia has been widely used for this purpose over the course of history [6]. It is a neurological phenomenon that occurs when one sense is being stimulated and an automatic experience arises in another one. This perceptual process has led to many creations that relate different fields of art such as poetry, dance, painting and music [15, 23]. One of the most recurrent associations in the synesthesia phenomenon is the one which unites color and sound. Computationally, there are many works that propose an automatic translation; in [22] a visual color notation for music is proposed and the Monalisa App [11] is a standalone software that converts images into sounds and vice versa with the aid of an intermediate step in which the information is treated as binary numbers.

Newton established a model called Musical Color Wheel [4] where each one of the 7 colors of the prism was related to one of the 7 musical notes. However, this theory is weakened by the fact that the chromatic scale has 12 notes. Lagresille used this model as a starting point and proposed a new broaded translation system by establishing a relationship between the properties of color and those of sound [25]. Specifically, he proposed that the saturation of color and the volume of sound were directly proportional, as well as the brightness of color and the sharpness of sound. His theory also involves obtaining notes on the basis of synesthesia. He divides the chromatic circle into 12 sectors and assigns a musical note to each one [20].

However, the use of synesthesia as a mean of conversion between images and music leads to several problems. On the one hand, this phenomenon entails a high level of subjectivity due to the huge differences in the way that people perceive information. On the other hand, only 1% of population experiences synesthesia, which makes it difficult to establish a proven relationship between color and sound.

In order to address the subjectivity problems, there is a visualization technique that derives from the signal processing theory: spectrograms, which are visual representations of the spectrum of some signal frequencies such as sound [19]. It is widely used in music classification problems because of the large amount of information that it provides. In [26] a musical onset detection problem is solved by applying a Convolutional Neural Network (CNN) to a set of spectrograms. [14] proposes to extract information from the spectrogram to increase the robustness of convolutional neural networks and to decrease noisy and degraded channel conditions. Nevertheless, although this technique makes use of visual representations and sound, it would be very difficult to use it as a tool for the translation of images to music.

Another approach that involves the concept of descriptive music, [28], which could be considered as a music genre that tries to create certain images, scenes or moods in the listener's mind. Throughout history, many relevant composers have contributed with their creations to this genre [28]. A classic example could be Vivaldi's *The Four Seasons*, where each one of the four violin concerts is related to one season (spring, summer, autumn and winter) and it attempts to

musically describe some typical details and climatic conditions such as thunders, the song of the birds, the rain and the wind. Another example could be Saint-Saëns' *The Carnival of the Animals*, where each movement characterizes an animal (the swan, the elephant, some tortoises, etc.). A set of this kind of musical compositions has been compiled in the film Fantasia, produced by Walt Disney [3]. Some professional animators have been chosen for the creative task; they created a model that related colors to musical emotions and analyzed how to design the characters and their movements on the basis of musical details [3]. Figure 1 shows a frame of the video fragment related to the musical piece "The Sorcerer's Apprentice" where Mickey Mouse learns to do magic. The creation of the model, as well as the process of composing music will be explained in more detail in Section 5.

Fig. 1. Frame related to the video fragment that describes the musical piece "The Sorcerer's Apprentice" from the Fantasia film

3 Method

The final goal of this work is the automatic composition of descriptive music. Figure 2 shows a visual scheme of this creative process in which the workflow is divided into two stages. On the one hand, in the training stage, the preliminary video is divided into a set of frames. A set of image descriptors is extracted from each one of them, as well as the main note that sounds at the moment when the frame is being played. Afterwards, a classifier is applied to the data in order to extract a model that relates both, image descriptors and sounds. On the other hand, the test stage where the musical composition is created takes place. A new video is chosen and divided into a set of frames. All the considered image descriptors are extracted again, and the model is applied to each one of them in order to obtain a sound. The concatenation of all the sounds makes up a sequence of notes which describes the characteristics of the frames.

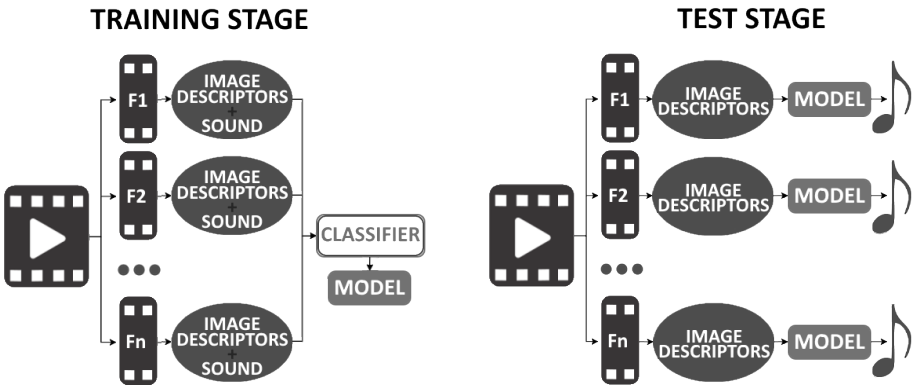

Fig. 2. Overview of the method applied in the system

4 Techniques

The characteristics that define each frame have been divided into two groups: shape features and color features. To obtain the first ones, Scale-Invariant Feature Transform method has been used together with Bag of Visual Words in order to achieve a dimensional reduction of the SIFT descriptor vector. These techniques are described in sections 4.1 and 4.2 respectively. For the second group of characteristics, color histogram has been used. It is described in section 4.3.

4.1 Scale-Invariant Feature Transform

Scale-Invariant Feature Transform (SIFT) is a method for extracting distinctive invariant features from images that can be used to perform reliable matching between different views of an object or scene [16]. SIFT descriptors are invariant to translations, rotations and scaling transformations and partially perspective transformations and illumination variations.

SIFT descriptors comprise a method for detecting points of interest from a grey-level image at which statistics of local gradient directions of image intensities were accumulated to give a summarizing description of the local image structures in a local neighborhood around each interest point. The steps for obtaining these image descriptors are as follows [16]:

1. **Scale-space extrema detection**: First step consist in searching over all scales and image locations. In this phase, keypoints are detected using a cascade filtering method in order to identify candidate locations that are invariant to scale and orientation.
2. **Keypoint localization**: The next step is to perform a detailed fit to the nearby data for location, scale, and ratio of principal curvatures. This information allows to reject points that have low contrast or are poorly localized along an edge.

3. **Orientation assignment**: In this stage one or more orientations are assigned to each keypoint location based on local image gradient directions. By assigning a consistent orientation to each keypoint based on local image properties. The keypoint descriptor can be represented relative to a consistent orientation, achieving invariance to image rotation.
4. **Keypoint descriptor**: The last step is to compute a descriptor for the local image region that is highly distinctive yet is as invariant as possible to remaining variations. The local image gradients are measured at the selected scale in the region around each keypoint. These are transformed into a representation that allows to be invariant to significant levels of local shape distortion and change in illumination.

4.2 Bag of Visual Words

As seen in the previous section, an image can be defined from a series of descriptors that contain important information about the features of the image (like SIFT descriptors). These keypoints can be grouped into clusters, so that each cluster is considered a visual word. Bag of Visual Words (BoVW) is a technique by which images are represented from visual words that symbolize the clusters where keypoints are grouped together, by obtaining a vector containing the (weighted) count of each visual word in that image [35]. This characteristic vector is used in the classification task. BoVW is a representation of images analogous to the Bag of Words (BoW) representation of text documents.

4.3 Color Histogram

A color histogram $H(M)$ is a vector $(h_l, h_2, \ldots, h_n)$, where each element h_J represents the number of pixels falling in bin j in image M [17]. The color space chosen for this work has been RGB, one of the most used in color histograms. Each of the channels has been divided into 256 bins, which correspond to the color intensity in the range $[0 - 255]$. For each of the channels a histogram has been obtained, i. e. R, G and B. In this way, a representation of the color of the image is obtained as three vectors, one for each channel of the RGB space.

5 Case study

We will detail our proposal by making use of the Fantasia film as a preliminary case study. As we explained before, the methodology follows a two stage flow. In Section 5.1 the task of extracting data from the initial video and its processing are explained. Section 5.2 describes the process in which the model that relates color, shape and the disposition of elements with sound is obtained.

5.1 Data description

When a composer creates music he needs a source of inspiration. In Computational Creativity, the generation of creative content is carried out by making

machines to imitate the way in which people create art [31]. In this work we design a system that composes descriptive music on the basis of the visual characteristics of a preliminary video. The first step was to analyze the fragment of "The Nutcracker Suite" played in the Fantasia film. The aim of this proposal is to carry out a reverse process to how the Fantasia film was created: we create a model that relates sounds to some features of the images such as color, shape and disposition of the elements.

First, an analysis of the video frames is carried out. The only seconds of the film that are considered are those that present changes in color, shape or disposition of elements. The film is shot at 24 frames per second and for our study the first one out of every 8 is chosen. Thus, three frames are considered per each selected second. This results in a collection of 483 images for analysis. From each one of them we obtain 1168 image descriptors which gather up their main visual features. In addition, an auditory analysis is performed by a musician to extract the most important note from all those that sound simultaneously at the moment the frame is being played in the video. Data relative to image descriptors and the most important sound of each frame can be found at [18].

We establish a relationship between image and sound by extracting patterns from their basic features. The image descriptors that are considered in this work have been divided into two groups: the first one describes the characteristics of the shape and disposition of the elements in the image, and the second one collects information related to color.

The first group of data is comprised of the first 400 values and they correspond to the frequency vector of the visual words obtained from the application of BoVW to the SIFT descriptors. 400 visual words have been chosen to form the visual vocabulary. The second group of data is comprised of the last 798 values and they correspond to the color histogram for RGB color space, 256 values for each channel. Figure 3 shows the SIFT descriptors and color histogram of frame 1442.11. Each SIFT descriptor represents a border or shape of the elements that make up the image, such as the head and arms of the flowers. The color histogram shows that the three RGB channels have a high number of pixels in the 50 to 75 bins, which corresponds to the colors of the flowers where the three R, G and B channels are present.

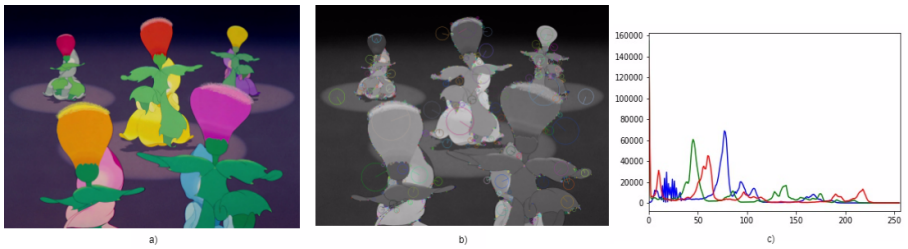

Fig. 3. Image features from frame 1442.11. Figure a) shows the original frame. In Figure b) SIFT descriptors can be found, and Figure c) exposes the color histogram

Finally, the class of the data set is the most important note that is heard in each frame. Sometimes, this sound will be the fundamental pitch of the chord that is being reproduced, but in other cases it will be the note that stands out from the sounds cloud. The 12 notes of the chromatic scale are considered for this task. The format used for its specification is MIDI [8], restricting the record to the octave that corresponds to the central C of the piano to simplify the task. Table 1 shows each one of the 12 notes and its corresponding MIDI encoding.

C	C#	D	D#	E	F	F#	G	G#	A	A#	B
60	61	62	63	64	65	66	67	68	69	70	71

Table 1. Numerical notation of the musical notes in the data set

As a last step, after obtaining the visual and sound information of each one of the selected frames and collecting it in the data set, a brief analysis is performed. The number of attributes, which are all descriptors of the image, are a total of 1168 for each instance. Additionally, in Table 2 the number of frames that is classified for each label is shown. This information is also presented as the percentage of frames per label out of the total number of the cases considered.

	C	C#	D	D#	E	F	F#	G	G#	A	A#	B
Frequency	6	24	93	12	39	21	30	66	0	78	42	72
Frequency(%)	1.24	5	19.25	2.48	8.07	4.35	6.21	13.66	0	16.15	8.7	14.9

Table 2. Frequency of each class in the data set

From the values shown in the Table 2, two conclusions can be drawn. On the one hand, there is no data pertaining to class "G#". This means that, in the video fragment that has been studied, there is no frame in which the most important note is "G#". Thus, no appraisals can be made for this note in future cases. On the other hand, the number of frames related to each class or note is not the same. Some classes like "D" and "A" have more occurrences than others like "C" and "G#". In data mining this problem is known as a class imbalance, and in this case it is due to the analysis of a single musical piece with a defined tonality that determines the important notes over the others.

5.2 Model induction

Supervised Machine Learning [13] infers a mathematical function that relates a set of attributes. This technique builds a model that classifies instances based on a set of previously labeled data. In this work, classification is used to extract a pattern that relates the image descriptors to the main sound in a video frame.

After a theoretical analysis of the classification task, in search of those techniques that would provide the best results, three algorithms were selected: Naive Bayes (NB), Support Vector Machine (SVM) and Random Forest (RF).

NB classifier is a probabilistic classifier based on the Bayes theorem and the hypothesis of independence between predictive variables. It assumes that the predictive attributes are conditionally independent given the class, and it posits that no hidden or latent attributes influence the prediction process [12].

SVM classifier is a supervised technique that finds a linear separating hyperplane with the maximal margin in a higher dimensional space [10]. In this work, the Minimum Sequential Optimization (SMO) training algorithm has been used [27]. This algorithm divides the global problem into a series of problems that are as small as possible and which are solved independently.

RF classifier is an ensemble algorithm which consists of tree predictors, such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest [2].

In order to determine the final choice, three classifiers were applied to the data set and the obtained results were discussed as detailed in Section 6. In conclusion, RF outperformed the other classifiers and for this reason it was used to create the model.

6 Discussion

In this section two types of results will be discussed. On the one hand, Section 6.1 presents a comparison of the performance of the algorithms applied in the classification task. On the other hand, the descriptive quality of the music obtained is analyzed in Section 6.2.

6.1 Performance of the classifiers

The performance indicators of the three classifiers (NB, SVM and RF) applied in the model induction process are shown in Table 3. Each row shows all the information on the performance of a classifier and the columns correspond to different quality measures [9]. Precision and recall are measures of exactness in the classification task: they look at how many of the returned instances are correct and how many positives the model returned respectively. F-score combines precision and recall information and determine a weighted single value and Kappa measures the agreement of the evaluations on the same samples. Root-Mean-Square Error (RMSE) is a metric that gives information about the concentration of the data around its best fit. Finally, Receiver Operating Characteristic (ROC) represents the exchanges between true positives and false positives.

	Precision	Recall	F-Score	Kappa	RMSE	ROC
NB	0, 474	0, 429	0, 421	0.3489	0.3224	0, 770
SVM	0, 795	0, 793	0, 791	0.7627	0.2654	0, 950
RF	0, 843	0, 832	0, 832	0.807	0.1855	0, 983

Table 3. NB, SVM and RF performance indicators

As shown in Table 3, the accuracy of the classifiers varies greatly. Based on the metrics of precision, recall and F-score, NB is the classifier with the least exactness. Conversely, RF has a good performance. The Kappa metric shows that NB (0.3489) and SVM (0.7627) classify affected by a random agreement factor unlike RF (0.807). The RF obtains the best value for the RMSE metric (0.1855); it means that the misclassified sounds are close to the good one. The optima value for ROC is 1; thus, RF almost reaches the maximum value for this metric too, outperforming the other classifiers again. Since it performed better than the other classifiers, RF has been chosen for this work.

6.2 Music composition

To evaluate the descriptive quality of the musical composition process, a new video is selected. The video fragment, that corresponds to the musical piece known as "The Firebird" of Ígor Stravinski in the film Fantasia 2000 [5], is chosen for this task due to its different color ranges and the movement of its characters. Three frames per second were extracted and feature extraction was applied. This data can be found at [18]. Finally, the model obtained in the classification task was applied to its image descriptors. As a result, a sequence of notes was created. To do a better analysis, the sound of the preliminary video was removed and the sequence of sounds composed by the system was played instead. The sequence of sounds can also be found at [18].

As in the previous case, the musical composition was not inspired by the video but by the images that compose it. For this purpose, the video was divided into a set of frames; specifically, in this case three frames per second were extracted. The next step is feature extraction for each one of the frames. In the same way as in the creation of the data set, the information about the shape, disposition and color was obtained.

In a final step, the model extracted by the classifier is applied to the new data. As a result, each one of the frames is translated into a sound that describes it according to the previously defined pattern. The concatenation of the obtained sounds gives rise to a sequence of notes. The duration of each of these sounds is 0.33 seconds. When two or more consecutive frames are translated into the same note, the musical result is a single sound with a duration equal to the sum of those of the isolated sounds. Thus, the musical composition is endowed with rhythm. Moreover, when there is no visual change in a set of consecutive frames, the sound that describes them is a continuous note.

From the conducted evaluation we can identify two aspects. On the one hand, the sequence of sounds is substantially different depending on the classifier that is used in the model induction process. On the other hand, in all cases, the relationship between the sound and the visual characteristics of the frames can be spotted easily. When two or more frames are visually similar, their corresponding sounds are the same; however, when two frames have some differences in color, shape or disposition the sound obtained is also different.

7 Conclusions and future work

This proposal successfully builds a system for automatic musical composition that describes a preliminary video. First, a deep analysis of the film Fantasia was conducted and the relationship between the color, the shape and the disposition of the elements of an image and its sound was established. Afterwards, a new video is divided into a set of frames, and the previously extracted model is applied to their image descriptors for the creation of a sequence of sounds.

One of the steps of the data set creation process is the labeling task: each frame is related to an isolated sound, which is the most important of all those that are sounding at the moment when it is being visually played. Despite the existence of some automated techniques for obtaining the main pitch of a sounds cloud or a chord which are based on the Fast Fourier Transform [21, 30], in our data set the label is analytically obtained by a music expert. This makes the labeling a slower, more expensive and more complex process. However, according to the final results it is a well-invested effort.

After the creation of the data set that is used to build the model, a brief study on the distribution of information is conducted. Due to the analysis of a single musical piece there is a class with no representation in the data set and a class imbalance problem: in music, each composition has a note which is considered the most important, and there is a set of sounds that are more related to it than the others. However, the accuracy of 83% proves that the classifier behaves appropriately despite the difference of examples for each class considered. It means that there is a clear relationship between the image descriptors and the sound in this case study. With regard to the classifiers, RF is the one that has the best result, outperforming NB and SVM. Additionally, despite their similar success rates, the three models produce substantially different musical results.

Furthermore, the fragment of "The Firebird" from the film Fantasia 2000 [5] was used in the test stage. It belongs to Animation genre as well as the film Fantasia. For this reason, there is a considerable similarity between the frames of both videos, what entails that the descriptors extracted from the images are also analogous in both cases. Thus, due to the application of the model to a data set which is similar to the training one, the results are quite good.

In this work, the movement of objects and characters is not specifically analyzed. However, a video could be considered as a concatenation of several frames or images. Thus, an isolated frame was translated into an isolated sound, and as the movement is a continuous change of position, it was implicitly studied in the

progression of several consecutive frames. Additionally, the result is not a melody with a great musical quality based on some harmony rules, but a sequence of sounds that describes the concatenation of frames.

In this proposal, only the video section corresponding to "The Nutcracker Suite" is analyzed. An extension of the data set could be performed by studying some more fragments of the film. This deeper analysis could lead to the solution of the unbalanced data problem; the tonality is different for each musical piece that sound in the film, and consequently the frequency of the notes also varies. Moreover, the data set would consist of several pieces of different styles, giving rise to a more robust model constructed with the data of all classes.

Acknowledgments

This work was supported by the Spanish Ministry, Ministerio de Economía y Competitividad and FEDER funds. Project. SURF: Intelligent System for integrated and sustainable management of urban fleets TIN2015-65515-C4-3-R.

References

1. Babtrakinova, O.I., Voloshko, A.A., Snistaryova, P.A.: Influence of modern music on young generation. Human and society (2), 4–6 (2017)
2. Breiman, L.: Random forests. Machine learning 45(1), 5–32 (2001)
3. Clague, M.: Playing in'toon: Walt disney's" fantasia"(1940) and the imagineering of classical music. American Music 22(1), 91–109 (2004)
4. Collopy, F.: Color, form, and motion: Dimensions of a musical art of light. Leonardo 33(5), 355–360 (2000)
5. Culhane, J.: Fantasia 2000: Visions of Hope. Disney Editions (1999)
6. Cytowic, R.E.: Synesthesia: A union of the senses. MIT press (2002)
7. Duchowski, A.T.: Eye tracking methodology. Theory and practice 328 (2007)
8. England, R.: Standard midi file production as the focus of a broad computer science course. Journal of Computing Sciences in Colleges 32(5), 4–10 (2017)
9. Guillet, F., Hamilton, H.J.: Quality measures in data mining, vol. 43. Springer (2007)
10. Hsu, C.W., Chang, C.C., Lin, C.J., et al.: A practical guide to support vector classification (2003)
11. Jo, K., Nagano, N.: Monalisa:" see the sound, hear the image". In: NIME. pp. 315–318 (2008)
12. John, G.H., Langley, P.: Estimating continuous distributions in bayesian classifiers. In: Proceedings of the Eleventh conference on Uncertainty in artificial intelligence. pp. 338–345. Morgan Kaufmann Publishers Inc. (1995)
13. Kotsiantis, S.B., Zaharakis, I., Pintelas, P.: Supervised machine learning: A review of classification techniques (2007)
14. Kovács, G., Tóth, L., Van Compernolle, D., Ganapathy, S.: Increasing the robustness of cnn acoustic models using autoregressive moving average spectrogram features and channel dropout. Pattern Recognition Letters (2017)
15. Lerdahl, F.: The sounds of poetry viewed as music. Annals of the New York Academy of Sciences 930(1), 337–354 (2001)

16. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International journal of computer vision 60(2), 91–110 (2004)
17. Lu, G., Phillips, J.: Using perceptually weighted histograms for colour-based image retrieval. In: Signal Processing Proceedings, 1998. ICSP'98. 1998 Fourth International Conference on. vol. 2, pp. 1150–1153. IEEE (1998)
18. Martín-Gómez, L., Pérez-Marcos, J.: Data repository of fantasia case study (Nov 2017), `https://github.com/lumg/FantasiaDisney_data`
19. Müller, M.: Book: Fundamentals of music processing. Signal 1, 0–5
20. Navarro-Cáceres, M., Bajo, J., Corchado, J.M.: Applying social computing to generate sound clouds. Engineering Applications of Artificial Intelligence 57, 171–183 (2017)
21. Peeters, G.: Music pitch representation by periodicity measures based on combined temporal and spectral representations. In: Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on. vol. 5, pp. V–V. IEEE (2006)
22. Poast, M.: Color music: Visual color notation for musical expression. Leonardo 33(3), 215–221 (2000)
23. Ranjan, A., Gabora, L., OConnor, B.: The cross-domain re-interpretation of artistic ideas. arXiv preprint arXiv:1308.4706 (2013)
24. Sakka, L.S., Juslin, P.N.: Emotional reactions to music in depressed individuals. Psychology of Music p. 0305735617730425 (2017)
25. Sanz, J.C.: Lenguaje del color:(sinestesia cromática en poesía y arte visual). El autor (1981)
26. Schluter, J., Bock, S.: Improved musical onset detection with convolutional neural networks. In: Acoustics, speech and signal processing (icassp), 2014 ieee international conference on. pp. 6979–6983. IEEE (2014)
27. Schölkopf, B., Burges, C.J., Smola, A.J.: Advances in kernel methods: support vector learning. MIT press (1999)
28. Seeger, C.: Prescriptive and descriptive music-writing. The Musical Quarterly 44(2), 184–195 (1958)
29. Tsoumakas, G., Katakis, I., Vlahavas, I.: Mining multi-label data. In: Data mining and knowledge discovery handbook, pp. 667–685. Springer (2009)
30. Tzanetakis, G., Ermolinskyi, A., Cook, P.: Pitch histograms in audio and symbolic music information retrieval. Journal of New Music Research 32(2), 143–152 (2003)
31. Varshney, L.R., Pinel, F., Varshney, K.R., Schörgendorfer, A., Chee, Y.M.: Cognition as a part of computational creativity. In: Cognitive Informatics & Cognitive Computing (ICCI* CC), 2013 12th IEEE International Conference on. pp. 36–43. IEEE (2013)
32. Vishton, P.M., Vishton, P.M.: Understanding the secrets of human perception. Teaching Company (2011)
33. Wang, Z., Chen, T., Li, G., Xu, R., Lin, L.: Multi-label image recognition by recurrently discovering attentional regions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 464–472 (2017)
34. Yanagimoto, M., Sugimoto, C.: Convolutional neural networks using supervised pre-training for eeg-based emotion recognition. In: 8th International Workshop on Biosignal Interpretation (BSI) (2016)
35. Yang, J., Jiang, Y.G., Hauptmann, A.G., Ngo, C.W.: Evaluating bag-of-visual-words representations in scene classification. In: Proceedings of the international workshop on Workshop on multimedia information retrieval. pp. 197–206. ACM (2007)