Music and Image Classification by Using Artificial Neural Networks

Alexandra Machado
Universitat Politècnica de Catalunya
Jordi Girona #3, 08034 Barcelona
alexandra.machado@gmail.com

Mariela Machado
Universitat Politècnica de Catalunya
Jordi Girona #3, 08034 Barcelona
mariela.machado@gmail.com

Rafael E. Banchs
Barcelona Media Innovation Centre
Ocata #1, 08003 Barcelona
rafael.banchs@barcelonamedia.org

ABSTRACT
In this paper, we report some experimental results regarding music and image classification by using artificial neural networks. The classification problem is approached by using low level features that are directly extracted from audio signals and image data. The extracted features are used for classifying music signal into five different genre categories, and images into five different theme categories. State of the art classification performance was achieved. Classification results as well as future research lines are presented and discussed.

Keywords
Feature extraction, supervised classification, artificial neural networks.

1. INTRODUCTION
In the last few years, a great amount of multimedia data in digital format has been generated thanks to the great advances in information technologies and their accessibility to the common people. This is why there is an urgent need for finding new and efficient ways of managing, storing and accurately locating all this multimedia information so that it can be of utility when needed. For this information to be located when wanted, it first needs to be effectively indexed or described in order to facilitate the retrieval or query process. At the moment, there are effective methods that perform textual information indexing and retrieval in a very efficient manner, but there are not effective methods for direct audio/visual information indexing and retrieval.

In this work, we address the problem of classifying multimedia data by using low level features directly extracted from signals. Two specific problems are independently studied: music and image classification. Artificial neural networks are used as classifycation engines, and the problem of determining the optimal configuration for the given classification tasks is also studied.

The paper is structured as follows. First, a bibliographical review on feature extraction techniques for both music and images is presented. Then, a general description of artificial neural networks as classification engines is presented, along with a more detailed description for the multilayer perceptron architecture. Third, the evaluation metrics to be used to estimate classification quality are described, followed by a complete description of the used music and image datasets. Then, the experimental procedure and the obtained results are presented. And, finally, some conclusions and future research work lines are depicted.

2. FEATURE EXTRACTION
In any pattern recognition method, it is indispensable to characterize the data to be able to classify it. Within many applications, instead of raw data, attributes that provide sufficient information about the samples are used to represent the data and to resolve possible ambiguities in the classification. These data preprocessing is called attribute or feature extraction, and consist of taking the most significant attributes from data samples in order to obtain an alternative representation into a vector space of attributes. The features have to be *comprehensive* in the sense that they represent the music very well, *compact* in the sense that they require much smaller storage space than the raw acoustic data, and *efficient* in the sense that computation can be carried out efficiently [1]. In this section, the most commonly used features for representing audio and visual signals are described.

2.1 Music Related Features
Many of the features commonly used in music classification have been previously used with success on speech recognition, and other few features have been exclusively developed for music. Music related features have been traditionally divided into three basic sets of features: timbre, rhythm and pitch. In the following subsections, only those features considered the most relevant within each of these three sets are going to be explained. The presented review of features is based on the works of Tzanetakis and Cook [2], Li and Ogihara [3], Aucouturier and Pachet [3], and Herrera-Boyer et al. [4].

2.1.1 Timbre Features
These are features that have been previously used in speech recognition methods and music-speech discrimination. Because of this reason, these are the features that have been more studied and used. The musical timbre also called colour of a sound is a basic element that describes all the aspects of a musical sound that can not be described through pitch, loudness or length. The most important features belonging to this family are the following:

- Mel-Frequency Cepstral Coefficients (MFCC). It is a really popular acoustic feature that has proven its effectiveness in the speech recognition field. These coefficients provide a compact representation of the spectral envelope through less than 15 coefficients. It is actually a perceptually motivated feature, and even though only 13 coefficients are considered for speech representation, it has been found that the first five coefficients, excluding the coefficients corresponding to

the DC component, are the ones that provide the best genre classification performance [2].

- Spectral Centroid. This feature calculates the frequency in which the center of mass of the spectral power distribution lies on. It represents the center of gravity of the magnitude spectrum for each frame (STFT), which is a measure of the spectral shape, and more specifically of the brightness. With higher frequencies, the texture will be brighter.
- Spectral Rolloff. This is another measure of spectral shape. For a given frame, it indicates the frequency value in the spectrum so that 85% of the power of the magnitude distribution is concentrated below it. This value decreases with percussive sounds.
- Spectral Flux. This feature measures the spectral difference between two consecutive frames, so it naturally measures the amount of local spectral change in the spectral shape. It has been mainly used in monophonic instrument recognition.
- Zero crossing rate. This feature indicates how many time-domain zero crossings are there in a given frame. It is a measure of the noisiness of the signal.
- Spectral Flatness. Theoretically, the spectral flatness is defined as the ratio between the geometric mean and the arithmetic mean. This feature indicates how flat the spectrum of a sound is and how is the power distributed in the spectrum. A high spectral flatness indicates that the spectrum has a similar amount of power in all spectral bands, and the spectrum would appear flat and smooth like for example the white noise. While the low spectral flatness indicates that the spectral power is concentrated in a small number of bands and the spectrum is less flat [4].
- Spectral Magnitude Kurtosis. This feature gives an idea of the degree of peakedness of a distribution. It is a statistical tool which can indicate the presence of a series of transients and their locations in the frequency domain.
- Spectral Magnitude Skewness. This feature provides an idea of the asymmetry of the variance of the values. The function will return larger positive values when there are more extreme values above the media, larger negative values when there are more values below the media, and zero when the distribution around the media is equilibrated.
- Low frequency energy relation. It is computed as the ratio between the spectrum energy below 100Hz and the total energy of the spectrum.
- High Frequency Content. It measures the presence of high-frequency contents in a signal frame, through a linear weighting function to the power magnitude spectrum.
- Spectral Mean. It is simply the spectral power mean value of the FFT values of the audio signal.
- Energy. Is another simple feature that denotes the energy of the signal, which is the sum of the squared amplitude within a frame.
- Maximal Magnitude Frequency. This feature is simply the frequency corresponding to the spectral bin with maximal energy within the frame

2.1.2 Rhythm and Pitch Features

Most of the studies in music classification and speech recognition use timbre descriptors, but recently there has been an agreement on the necessity of taking also rhythm into account. The features developed to detect rhythm are based on beat detection, which measures the presence of high-frequency contents in a signal frame through a linear weighting function to the power magnitude spectrum.

One way to calculate the set of rhythm related features is based on using a beat analysis algorithm to build a beat histogram and get an idea of the strength and complexity of the beat in the music sample [2]. Another approach is based on using second order statistics like angular second moment, correlation, and entropy, among others [3].

On the other hand, pitch related features have been even less studied than rhythm features, but it has been considered important to mention them here. Tzanetakis and Cook [2] have used pitch related features computed by means of a multipitch detection algorithm. This algorithm works with the same procedure as beat detection, but it varies in some points. More specifically, it divides the signal into two frequency bands and performs the envelope extraction like in the beat detection process, then applies an enhanced autocorrelation function (SACF) to detect the prominent peaks that will correspond to the main pitches for that short segment of sound.

Rhythm and pitch related features will not be used in the present study.

2.2 Image Related Features

Image feature extraction and classification have long been investigated in the past decade. In this section, some of the most used and effective feature extraction methods will be explained. Three different families of primary or low-level features can be identified: color, texture and shape. Although there are many other features, these three families of features are the most basic and commonly used ones. The presented review of features is based on the works of Manjunath and Deng [5], Myrka Hall-Beyer [6], Chen et al. [7], Grigorescu et al. [8], and Zhang [9].

2.2.1 Color Features

These constitute by far the most used features for image retrieval and classification, being actually one of the most used matching techniques the one based on *color histogram.* This method represents the distribution of colors in an image. The RGB[1] color space of the image is quantized into N number of bins, then the image is analyzed and each pixel is classified according to its color in one of these bins and at the same time, the number of pixels in each bin is counted providing the proportion of color in

[1] R = red, G=green, B=blue

the image [5]. This method can be implemented in two different ways: with global histograms or with local histograms. In the first case, the image is analyzed as a whole, and in the second case, the image is analyzed in segments resulting in a more detailed and complete representation. However, the second case has the disadvantage of requiring much more time of calculation which in many cases can be unpractical.

2.2.2 Texture Features

This is another important family of features for image retrieval. The texture features capture the changes in brightness or grey level values in an image, which provides useful information for the recognition or classification tasks. For example, if images of the sky and the sea are to be classified, the differences will not be as obvious from color histograms as from texture features. Indeed, in this particular case, successful classification will be possible only by including texture features. Texture features quantify: grey level differences (contrast), defined size of the area where change occurs, and existence or lack of directionality [6].

There are several methods for extracting texture features from images. However, all recent studies and their results show that there are mainly three methods for the analysis of this particular and important type of features: co-occurrence (or gray level co-occurrence) matrix (GLCM), Gabor filters, and wavelet transforms.

- Co-occurrence matrix (GLCM). This is the most commonly used texture descriptor. The features are extracted from a matrix that measures the occurrence values at a given offset, considering the spatial relationship of the pixels. "The GLCM is a tabulation of how often different combinations of pixel brightness values (grey levels) occur in a pixel pair in an image" [6]. From this matrix it is possible to calculate many important features of the image that described the texture in a statistical sense. The main features belonging to this family are: homogeneity, contrast, correlation, variance, entropy, energy, etc

- Gabor filters. This type of texture feature is a filter represented by a group of wavelets as explained by Chen et al. [7]. The image is processed by the filter with different orientations and scales resulting in a set of filtered images. The product of these convolutions generates Gabor wavelet transforms that then are summed up to generate an array of magnitudes representing the transforms at every scale and orientation wanted. Then the standard deviations and the mean value of the magnitude of the array are calculated given a certain orientation and scale which constitute the feature texture vector representing the image.

- Wavelet transforms. This constitutes an alternative approach, which is based on complex transformations of the image into a multi-scale representation including both spatial and frequency information. From these transform coefficients several wavelet-based texture features can be extracted, being the wavelet moment the more important one. Although this seems to be a very promising approach for texture feature computation, it has the disadvantage of its computational complexity.

2.2.3 Shape Features

Along with texture and color, these are the most used features for content-based image retrieval and classification. Shape extraction has been traditionally approached through segmentation or *edge detection.* There are two main groups of shape descriptors: contour-based and region-based descriptors [9]. The first class of descriptors only uses shape boundary information, while the second class of descriptors uses both boundary information and interior information.

The main disadvantage of any shape feature is that for it to be useful, it has to be invariant to rotation, scale and translation. This has been actually the main drawback in the development of techniques for shape feature extraction. This is because in order to recognize a flower in an image, for example, the method has to be able to detect flowers in general but not only flowers in certain positions. Regarding this problem, many methods for shape feature extraction have been proposed and studied, from which the most common ones are: edge detection, chain code, polygonal approximations, curvature, Fourier descriptors and moment descriptors. In this work, only two of these techniques will be further explained.

- Fourier descriptors (FD). These are the most commonly used contour-based shape descriptors and, in general, the most commonly used technique for shape detection. They correspond to the complex Fourier coefficients of the image which provide information about the orientation and symmetry of the regions, as well as the position of regions in the image. The main advantage of this kind of features is that they are invariant to common transformations like rotation, translation, and changes in scale [9].

- Edge detection. This procedure marks or detects the abrupt changes in the image brightness or luminous intensity. Edges in images are defined by regions with jumps of intensity from one pixel to the next and. The objective of edge detection is to reduce the amount of data while preserving only the structural properties of the image. There are many methods for edge detection, which are mostly based on using filters that detect these changes in the image, ignoring the useless information. The most commonly used filters for edge detection purposes are the *Canny filter* and the *Prewitt filter*. The result of edge detection is a simpler image containing the relevant information from which the shape extraction task becomes easier.

In this work, a very simple approximation for shape features will be used. We will restrict shape feature to simple color histograms of the filtered images resulting from applying Canny and Prewitt filters.

3. ARTIFICIAL NEURAL NETWORKS

This section will present a theoretical review of artificial neural networks, which are the artificial intelligence systems selected to carry out the classification tasks. First, a general description of artificial neural networks as classification engines is presented. Then, a more detailed description on the feed-forward multilayer perceptron architecture is given. For a more comprehensive discussion on this topic, refer to Haykin [10].

3.1 General Description

Artificial neural networks were born with the pioneering work of neurophysiologist, in which they tried to find a mathematical model to represent neurons and recreate, with artificial methods, the capacity of the brain to interpret information. Recently, artificial neural networks have become innovative classification methods which imitate the learning process in the human brain.

In this sense, an artificial neural network is an information processing structure that is composed of a large number of highly interconnected processing elements working in parallel to solve specific problems. They may also be defined as a massively parallel and distributed processor that resembles the brain in two aspects: knowledge is acquired by the network through a learning process, interneuron connection strengths (known as synaptic weights) are used to store the knowledge.

The basic element in any artificial neural network is the *artifial neuron or perceptron*, which is a processing unit with many inputs and one output. The output is computed from the inputs by using a linear combiner formed by a set of synapses or connecting links, an adder that sum up the weighted, and an activation function that performs a non-linear transformation of the weighted sum. A simplified model of the perceptron is illustrated in figure 1. This model is known as the McCulloch and Pitts model.

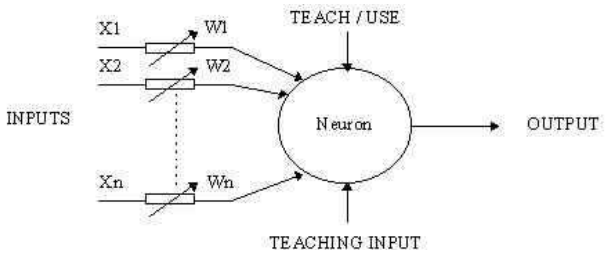

Figure 1. Scheme of an artificial neuron or perceptron.

The learning process consists in adjusting the synaptic weights of the network elements, or perceptrons, in an orderly fashion such that a desired error function is minimized. According to Haykin [10], there are mainly three different possible architectures for an artificial neural network: single-layer or multi-layer feed-forward networks, recurrent networks and lattice structures. The most widely used architecture is the multi-layer feed-forward network, so it is the one used in this work and will be explained next.

3.2 Multilayer Perceptrons

In a multi-layer feed-forward network the processing units are arranged into layers interconnected in a consecutive fashion. The network is formed by an input layer, one or more hidden or middle layers, and an output layer. While the number of units in the hidden layers is variable, the number of units in the input and output layer are fixed by the problem conditions. In fact, in the input layer it is defined by the number of variables or features used to represent the data samples, and in the output layer it is defined by the number of output or classes required. On the other hand, the amount of units in the hidden layers constitutes a parameter to be adjusted, and the accuracy of the classification results strongly depends on this number.

After having defined the network architecture, the other important element is the learning algorithm. The most popular one is the so called back-propagation algorithm [10]. The back propagation algorithm updates the connection weights by minimizing the network's estimation error over the training set. This training method is inspired from the idea that the system learns from its mistakes, in a process in which it minimizes the classification error over a previously defined training dataset.

The steps that the back-propagation algorithm follows are:

- First, it computes the network outputs for a previously defined training set of inputs.
- Second, it computes an error value or cost function by comparing the obtained outputs with the desired results.
- Third, the error is propagated backward form the output layer to the input layer and the weights of the units at each layer are adjusted so that the local error at each layer is minimized.
- This procedure is repeated in an iterative fashion until the global output error is reduced to acceptable values.

The main advantage of artificial neural networks is that they have the properties of adaptive-learning, self-organization and fault-tolerance capabilities. And its applications are innumerable, with known success in applications such as machine diagnostics, financial forecasting, fraud detection, quality control, credit rating, intelligent searching and many more. Artificial neural network applications are generally categorized into three different types of problems: classification, function approximation, and time series prediction.

In this work, we will use artificial neural networks as classification engines for performing the tasks of music and image classification. A multi-layer feed-forward architecture with only one hidden layer will be used, and the optimal number of units in the hidden layer will be determined by trial and error for each of the two classification tasks under consideration. The back-propagation algorithm will be used for training the classifiers.

4. EVALUATION METRICS

Three different evaluation metrics were used to measure the success of each classification experiment. These metrics, which are briefly described in this section, are the standard evaluation metrics of precision, recall and error rate. Figure 2 provides a graphical illustration of what each of these metrics is actually measuring.

- Precision. This metric evaluates the number of right classifications from one genre, out of the total number of samples that were classified as that genre.
- Recall. This metric evaluates the number of right classifications from one genre, out of the total number of samples that are actually from that genre.
- Error rate. This metric corresponds to the total number of erroneous classifications from each genre (sum of the number of samples that were classified as that genre but were from other genres, and the number of samples that were from that genre but were classified as from other

genres), out of the total number of samples that were classified.

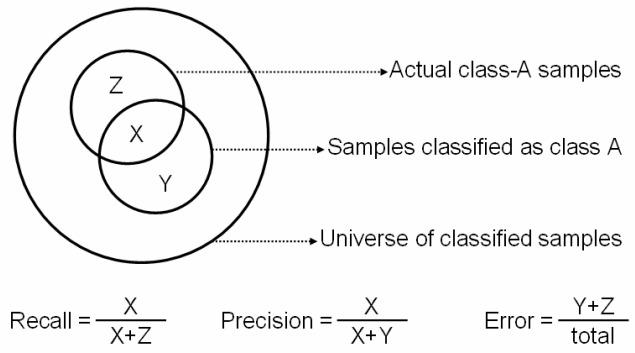

Figure 2. Metrics used to evaluate classification success.

The three previously mentioned evaluation metrics were computed by only taking into account the universe of classified samples, i.e. those samples not assigned to any class or assigned to more than one class by the classifier were not considered for precision, recall and error rate computation. In this sense, an additional rate for unclassified samples was also measured. This rate corresponds to the number of unclassified samples out of the total number of samples provided in the test input dataset.

5. COLLECTION OF THE DATASETS

This section describes the specific data collections, music and images, that were used for training and evaluating the artificial neural network classifiers. These datasets were manually collected from the web and, in addition to the audio and visual contents, all the available related text information was also downloaded. This provides a rich set of metadata descriptors for the audio and visual contents. It is important to mention, that in this work only the audio and visual data are used for classification purposes. The metadata information was collected for future research projects.

5.1 The Music Dataset

The songs and genres were carefully selected in order to obtain a representative set of samples for each genre. All songs were downloaded manually from http://www.allmusic.com/ and then converted from ASX into the MP3 format. The metadata was extracted by copying all available information on the webpage, and a text file with the same name of the MP3 archive was created, following an established format for the structured metadata and not structured metadata. After collecting 2060 music samples of 30 seconds long each, the database was created and made public through a text-based search engine, which is available at http://varovani.barcelonamedia.org/search_music/, to make accessible the dataset for future projects.

In summary, the dataset is composed of 2060 30-second music segments from nine different genres, from which only five genres have been selected for the classification experiments: classical, rock, jazz, latin and electronica. Additionally, for preparing the data for classification experiments, each 30-second sample was divided into six 5-second samples, considering that the samples that were going to be passed onto the classifier to train it and test it were thought to have only five seconds of duration. According to previous works, five seconds of musical signal is enough to extract the nature of the signal.

Finally, 1000 and 500 5-second samples were selected for constructing the train and test data sets, respectively. Such sets included the same amount of samples for each of the five selected music genres. And a total of 174 features were finally used for characterize them. These features included MFCC, spectral mean, energy and spectral centroid, which were computed over consecutive frames of approximately 11 ms; and flatness, rolloff and low frequency content, which were computed over the whole 5-second samples.

5.2 The Image Dataset

Differently from the music data collection, the image dataset was originally intended to be automatically extracted from videos. Initially, pages with scientific content, or oriented to research, as open resources were looked for. But results of this initial search did not satisfy many of the requirements mentioned before, so the search had to be re-directed, and many other options were considered. In this new search, other problems were encountered that were directly related with the availability of the videos, since in many cases some videos were not directly downloadable or had an economical cost. This became the main problem during this phase of the work.

The final solution was to use tools that enable the capture of video streaming. A video data collection was finally constructed by suing several different sources. The final collection includes 732 videos belonging to five different themes. This collection is available at http://varovani.barcelonamedia.org/search_video/, and the dataset is accessible for future projects.

Nevertheless, the original objective of using these videos to extract images for the classification systems was not met, mainly because of the lack of time and the impossibility of finding an adequate method for segmenting the video into a number of images. For this reason, another image data set was downloaded for carrying out the image classification experiments. A total of 1050 images equally distributed into five categories: animations, animals, landscape, sports and history, were downloaded from the web; mainly form http://www.flickr.com/.

Finally, the 1050 image collection was divided into two sets of 700 and 350 images for constructing the train and test data sets, respectively. Such sets included the same amount of samples for each of the five image categories. And a total of 147 features were finally used for characterize them. These features included RGB color histograms, texture measurements extracted from both the co-occurrence matrix and the Gabor filters, and a very simple approximation for shape features that were based on the color histograms of the filtered images resulting from applying Canny and Prewitt filters.

6. EXPERIMENTS AND RESULTS

This section describes in detail all music and image classification experiments and their corresponding results. First, the main experimental procedure is described in detail. Then, experimental results for both music and image classification are presented and discussed.

6.1 Experimental Procedure

For both classification problems, the same artificial neural network architecture was used: a feed-forward back-propagation network. Each classifier was designed to have as many input units as the number of features used for representing the samples (174 in the case of the music classification task, and 147 in the case of the image classification task). Each classifier had five units in the output layer, each of one represented one sample class. In this sense, the classifier implemented the one-hot encoded output method, i.e. only the output corresponding to the identified class should be equal to one (and all others equal to zero).

On the other hand, only one hidden layer was considered, and the total number of units in the hidden layer was varied in order to determine the best classifier configuration. According to this, different experiments were performed to find the most appropriate number of units for the hidden layer. Each of these experiments consisted on a total of 10 independent simulations, from which evaluation metrics were averaged. Once the optimal network configuration was determined (total number of units in the hidden layer), a final experiment for evaluating the overall system performance was conducted.

In all the cases, evaluation metrics were computed over the set of classified outputs. This means that unclassified samples were removed from the results, and the incidence of unclassified samples is reported by means of the unclassified sample rate. Two types of unclassified samples were considered: first, those cases in which the classifier did not assigned any class to the sample (all network outputs are equal to zero); and, second, those cases in which the classifier assigned more than one class to a sample (more than one output are equal to one).

6.2 Classification of Music by Genre

For determining the best system configuration, a total of 10 simulations were conducted for five different configurations: 7, 10, 15, 20 and 24 units in the hidden layer. Figure 3 presents curves for recall vs. number of hidden units for each of the five considered music genres.

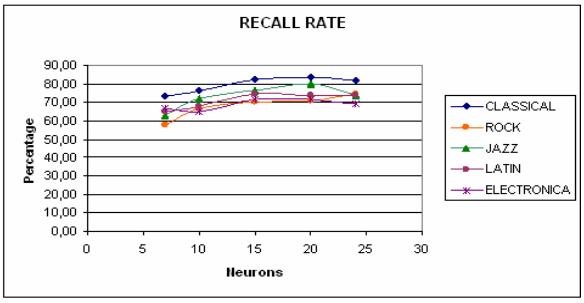

Figure 3. Recall vs. number of units in the hidden layer.

From figure 3, it can be observed that best recalls always occur between 15 and 24 hidden units for the rock, jazz, latin and electronica genres. A more detailed analysis of all evaluation metrics determined that there is a clear tendency for having either the best precision or the best recall when considering 20 units in the hidden layer. For this reason this configurations was selected as the best one for the music classification task.

Table 1 summarizes the precision, recall and error rates obtained for each genre by using the optimal system configuration. It is important to mention that for this system configuration the measured unclassified sample rate was 21%.

Table 1. Evaluation metrics for optimal system configuration.

Genre	Precision	Recall	Error
Classical	87%	83%	8%
Rock	69%	71%	11%
Jazz	75%	80%	10%
Latin	77%	74%	9%
Electronica	73%	71%	10%

From table 1, it can be observed, that the overall precision rates were over 69% of success which can be considered a good classification performance. The best rate achieved is 87% for the classical samples. The second best precision rate was achieved for latin music with 77%. The worst precision rates were for rock with 69% and electronica with 73%. These results were expected as these two genres have a broad nature, including rhythms that vary a lot according to the sub-genre, while jazz, latin and classical have more specific instruments and rhythms.

The recall rates had an overall performance similar to the precision rates but slightly better, being the lowest 70%. It has also its best performance for classical music, but the second best recall rate was obtained for jazz with 80%, instead of latin that this time is the third best rate. Both, rock and electronica, keep their places as the worst classified genres. Finally, and regarding error rate, it can be observed that consistently the lowest error rate was for classical music, while the worst performance was for rock, which is not surprising. Latin music shows the second lowest error rate, and electronica is the following best error rate.

6.3 Classification of Images by Theme

In this case, for determining the best system configuration, a total of 10 simulations were conducted for four different configurations: 5, 10, 15, and 20 units in the hidden layer. Figure 4 presents curves for recall vs. number of hidden units for each of the four considered music genres.

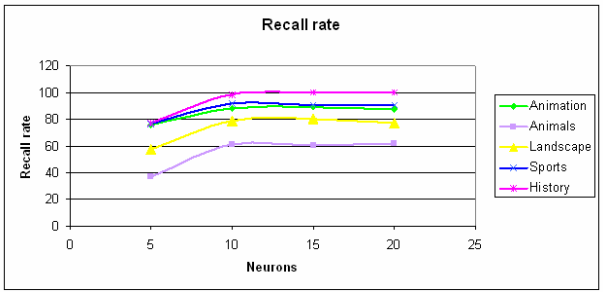

Figure 4. Recall vs. number of units in the hidden layer.

From figure 4, it can be observed that best recalls always occur between 10 and 20 for all the considered image categories. A more detailed analysis of all evaluation metrics determined that

there is a clear tendency for having either the best precision or the best recall when considering 15 units in the hidden layer. For this reason this configurations was selected as the best one for the image classification task.

Table 2 summarizes the precision, recall and error rates obtained for each category by using the optimal system configuration. It is important to mention that for this system configuration the measured unclassified sample rate was 10%.

Table 2. Evaluation metrics for optimal system configuration.

Genre	Precision	Recall	Error
Animations	97%	89%	2%
Animals	72%	60%	10%
Landscape	75%	85%	11%
Sports	89%	91%	3%
History	85%	97%	1%

From table 2, it can be observed that the history category was the one with the best recall and the lowest error rate. The next category with best recall was sports, then animation followed by landscape and animals. And the lowest error rates, after history, were achieved by animation, sports, animals and landscape. Regarding precision, the best rate was achieved for animation, result that was also expected given the particularities of color and texture exhibited by this kind of images. It can be seen also, from the table, that sports follow animation with the second best precision rate, followed then by history, landscape and finally animals.

As also observed from table 2, the worst overall performance was for the nature related categories, both landscape and animals, with this last one being the one that presented the worst results of all. This fact was also expected given the similarities in color, textures and shapes that can occur between these two categories, which are generally confused with each other. And history and animation were the categories more easily classified.

7. CONCLUSIONS AND FUTURE WORK

This work presented some experimental results on the tasks of automatic classification of musical signals and images. The main objective of the work was implementing an evaluating music and image classifiers based on low-level features directly extracted from the audio and visual data contents of the signals. The classification systems were implemented by means of artificial neural networks, which allows for implementing a supervised classification procedure by considering multiple data classes simultaneously into the same classification engine.

In the case of the music classification experiments, five different musical genres were considered. Comparing the obtained results to the most prominent related works, the error rates are low, the precision and recall rates high, and the general performance for all the genres has never a precision and recall rate below 50%. This implies that the developed classifier is providing state-of-the-art classification performances. The accuracy rate for the rock genre reported in other works is around 40%, which was significantly improved within this research work. However, it is important to note that results are generally not strictly comparable unless the same data collection is used and the same set of music genres is considered.

In the case of the classification experiments, five different image categories were considered too, and state-of-the-art classification performances were also achieved for all considered categories. It also could be observed that there are some specific categories for which better performances are always achieved independently of the system configuration used when classifying them, as well as there are some other categories for which the systems have problems to identify.

The categories that had always excellent performance were history and animation, and the category that always exhibited low performance was animals. The categories with best performances are characterized for exhibiting some particularities in their color and texture features, which differentiate them from the rest of the categories. For example, all images in the history category where in black and white format, and those in the animation category were all cartoons or drawings. So, for that case of animation and history it was already expected for the systems to classify them correctly as it was confirmed.

On the other hand, for the animal and landscape categories, lower classification rates were expected because they both are nature-related categories that can be easily confused.

As future work, several lines of research could be identified. In the case of music classification, future research should focus on the following topics:

- to consider other musical genres from the database created, varying the number and combination of musical genres, to study their results and compare them with the ones obtained in this project,
- to vary the set of features used and study its influence in the classification task for each of the considered genres,
- to study the feasibility for a classifier of this kind for distinguishing the language in which a given song is sung,
- to detect emotion in the musical signals; in this case, the classifier would have to rely on a feature vector that can capture very well some characteristics able to efficiently represent different classes of emotions.

On the other hand, in the case of image classification, future research should focus on the following topics:

- to study the problems of shot detection and key-frame extraction, which allows for automatically extracting the set of most relevant images from the videos in order to classify the videos by using their images,
- to study more complex and informational shape-related features with the objective of improving image classification performance in general, and more specifically for the animal category (in this particular case, the only difference with the landscape category, or other nature-related categories, is the presence of animals, which could be actually recognized by means of an efficient shape extraction method).

Finally, regarding the general problem of audiovisual content classification and management, the following research lines are proposed:

- to study other state-of-the-art classification methods such as, for example, support vector machines, and compare their performance with artificial neural network classification systems,
- to combine two different classification systems (for example, artificial neural networks and support vector machines) into a more powerful classification engine,
- another interesting project would be to combine text-based and signal-based classification paradigms, by using the available metadata and the corresponding audiovisual contents.

8. ACKNOWLEDGMENTS

The authors would like to thank to Barcelona Media Innovation Centre for supporting this research, as well as the permission to publish these results.

Additionally, the authors would like to thank Rodrigo Meza and Andreas Kaltenbrunner for their valuable help at some of the phases of this project.

9. REFERENCES

[1] Li, T. and Ogihara, M. (2006, June). Toward Intelligent Music Information Retrieval. IEEE Transactions on Multimedia, Vol. 8, (Nº 3). Pp 564-574

[2] Tzanetakis, G. and Cook, P. (2002, July). Musical Genre Classification of Audio Signals. IEEE Transactions on Speech and Audio Processing, Vol. 10, (Nº 5). Pp 293-302

[3] Aucouturier, J-J. and Pachet, F. (2003). Representing Musical Genre: A State of the Art. Journal of New Music Research, Vol. 32, (Nº 1), pp 83-93

[4] Herrera-Boyer, P., Peeters, G, Dubnov, S. (2003). Automatic Classification of Musical Instrument Sounds. Journal of New Music research, 32:1, pp 3-21

[5] Manjunath B. & Deng Y. (1997). Content-Based Search of Video Using color, Texture, and Motion. International Conference on Image Processing (ICIP'97) Volume 2, p. 534.

[6] Myrka Hall-Beyer (2006). Tutorial: GLMC Texture. Available online at: http://www.fp.ucalgary.ca/mhallbey/tutorial.htm.

[7] Chen, L.; Lu, G & Zhang D. (2004). Effects of Different Gabor Filters Parameters on Image Retrieval by Texture. Multimedia Modelling Conference, 2004. Proceedings. 10th International Issue, 5-7. p. 273 – 278

[8] Grigorescu S.E, Petkov N & Kruizinga P. (2002) Comparison of Texture Features Based on Gabor filters. IEEE Trans. on Image Processing, Vol 10, p.1160-1167.

[9] Zhang D. (2002). Image retrieval based on shape. Doctoral thesis for the degree of Doctor of Philosophy. Monash University. Australia.

[10] Haykin, S., (1994) Neural Network: a comprehensive foundation. USA, Mcmillan College publishing company, pp3-57