Analysing the Similarity of Album Art with Self-Organsing Maps

Rudolf Mayer ✉

Institute of Software Technology and Interactive Systems
Vienna University of Technology, Austria

Abstract. Digital audio has become an ubiquitously available medium, and for many consumers, it is the major distribution and storage form of music, accounting for a growing share of record sales. However, handling the ever growing size of both private and commercial collections becomes increasingly difficult. Users are often overwhelmed by the seemingly countless number of music tracks available. Computer algorithms that can understand and interpret characteristics of music, and organise and recommend them for and to their users can thus be of great assistance.

Therefore, a magnitude of research projects has been devoted in the last decade to automatically to make the sound characteristics of music machine interpretable, to e.g. allow for automatic categorisation of music, or to recommend track which are similar to the ones a user likes.

However, music is an inherently multi-modal type of data, and increasingly also other modalities of music have attracted interest from the community. The analysis of song lyrics and other textual data, such as websites or biographies associated with artists, together with social network data, has probably attracted most research in this area.

Album covers are another dimensionality characteristic to the music – they are often carefully designed by artists to convey a message consistent with the music and image of a band. Studies have shown that customers use album cover art as a visual cue when browsing music in regular record stores. We thus present a study on similarities in album covers, and their relations to certain styles and genres of bands. To this end, we employ Self-Organising Maps together with various visualisation techniques to automatically organise a music collection, and compare the results obtained when using both features from the music and the album covers.

1 Introduction and Related Work

Motivated by the vast spread of music in digital formats, Music Information Retrieval (MIR) has become a very important field to aid private and commercial users to organise their music collections. Important tasks are for instance automatic categorisation to organise music into predefined genres or moods, recommendation of music similar to a certain song (similarity retrieval), or the development of novel and intuitive interfaces to large music collections. A comprehensive overview of the research field is given in [12].

A strong focus on music's primary mode, the sound of a song, can be seen from the research in the last decade. A number of methods to extract descriptive features from the audio signal and to capture information such as rhythm, speed, amplitude or instrumentation have been proposed, ranging from low-level features describing the power spectrum to higher level ones. However, also other modalities associated with music have increasingly been employed for common MIR tasks.

Several research teams have been working on analysing textual information, often in the form of song lyrics and a vector representation of the term information contained in other text documents; an early example is a study on artist similarity via song lyrics [8]. Other cultural data is included in the retrieval process e.g. in the form of textual artist or album reviews [1].

The study in [2] suggests that 'an essential part of human psychology is the ability to identify music, text, images or other information based on associations provided by contextual information of different media'. It further suggests that a well-chosen cover of a book can reveal it's contents, or that lyrics of a familiar song can remind one of the song's melody. Album covers are generally carefully designed for specific target groups, as searching for music in a record shop is facilitated by browsing through album covers. There, album covers have to reveal very quickly the musical content of the album, and are thus used as strong visual clues [3]. Due to well-developed image recognition abilities of humans, this task can be performed very efficiently, much faster than listening to excerpts of the songs. This motivates and increased utilisation of this modality.

A multi-modal approach to query music, text, and images with a special focus on album covers is presented in [2]. In [5], a three-dimensional musical landscape via a Self-Organising Map is created and applied to small private music collections. Additional information like web data and album covers are used for labelling; album covers should facilitate the recognition of music known to the user. The covers are however not use in the SOM training itself.

The Self-Organising Map has also been applied to image data in the PicSOM project [6], for Information Retrieval in image databases, incorporating methods of relevance feedback.

In this paper, we want to empirically validate the hypothesis that album covers can provide cues to the type of music. We therefore organise a music collection with Self-Organising Maps using both music features and image features, and analysing the way the album covers are organised over the map. We investigate whether musical similarity and a similarity in the album cover art are correlated, and whether albums can really give a clue on the music they represent.

The remainder of this paper is structured as follows. Section 2 gives a brief outline over the SOM framework and visualisations employed, while Section 3 will introduce the feature sets employed to describe our music collection. Our experiments are then detailed in Section 5, before we conclude in Section 6.

2 SOM Framework

We employ the Java SOMToolbox framework[1], developed at the Vienna University of Technology, which provides methods for training SOMs. It further comprises an application for interactive, exploratory analysis of the map, allowing for zooming, panning and selection of single nodes and regions among the map. The application also allows to display digital images on top of the map grid, thus it can easily be used to visualise the album covers.

To facilitate the visual discovery of structures in the data, such as clusters, a wealth of approximatively 15 visualisations are provided, among them the U-Matrix [14] and Smoothed Data Histograms[13]. The former indicates distances between SOM nodes by colour-coding, and thus hints on cluster boundaries, while the latter visualises density in the data, also indicating clusters as nodes with high density. We also utilise the Thematic Classmap visualisation [9]. It which shows the distribution of meta-data labels or categories attached to the data vectors mapped on the SOM, by colouring the map in continuous regions, similar as e.g. a political map does for countries. To this end, it performs a Voronoi tessellation of the map space, and assigns colours to each Voronoi region to indicate how much a class contributes to the data items in that region.

To provide a partition of the map into separate clusters, the framework provides several clustering algorithms that can be applied on the vectors of the SOM nodes, such as Ward's linkage [4] algorithm.

3 Feature Sets

To obtain a vector representation of the music collection, we employ on the one hand methods that extract descriptive features from the audio snippets, as well as methods to extract features capturing information from the album covers.

3.1 Audio Features

The following descriptors are extracted from a spectral representation of an audio signal, partitioned into segments of 6 sec. Features are extracted segment-wise, and then aggregated for a piece of music computing the median (for RP and RH, see below) or mean (for SSD, see below) from multiple segments.

We describe the feature extraction algorithms very briefly, please refer to the references for further details.

Rhythm Patterns The feature extraction process for a Rhythm Pattern (RP) is composed of two stages. First, the specific loudness sensation on 24 critical frequency bands is computed through a Short Time Fast Fourier Transform (FFT). The resulting frequency bands are grouped according to the Bark scale, and successive transformation into the Decibel, Phon and Sone scales takes place. This

[1] `http://www.ifs.tuwien.ac.at/dm/somtoolbox/`

results in a psycho-acoustically modified Sonogram representation that reflects human sound perception. In the second step, a Discrete Fourier Transform is applied to this Sonogram, resulting in a spectrum of loudness amplitude modulation per modulation frequency for each critical band. After additional weighting and smoothing steps, a Rhythm Pattern exhibits magnitude of modulation for 60 modulation frequencies on the 24 critical bands [7].

Rhythm Histogram A Rhythm Histogram (RH) aggregates the modulation amplitude values of the critical bands computed in a Rhythm Pattern, and is thus a descriptor for general rhythmic characteristics in a piece of audio [7].

Statistical Spectrum Descriptor The first part of the algorithm for computation of a **Statistical Spectrum Descriptor** (SSD), the computation of specific loudness sensation, is equal to the Rhythm Pattern algorithm. Subsequently at set of statistical values (mean, median, variance, skewness, kurtosis, min and max) are calculated for each individual critical band. SSDs therby describe fluctuations on the critical bands; they capture both timbral and rhythmic information. In a number of evaluation studies, SSD have often shown to be superior for musical genre classification tasks [7].

3.2 Image Features

Colour Histogram This feature set computes the distribution of pixel values in the RGB colour space. For each colour channel, a histogram of values (from 0 to 255) is computed from all pixels in the image. To reduce the dimensionality, we employed binning of the values. 128 bins for each channels were determined as a good value through experimental evaluation in classification tasks. Thus the total dimensionality of such a feature vector is 384 dimensions.

Color Names Colour names [16] are a level of abstraction on top of a colour histogram – the colour space is divided in the 11 basic colours black, blue, brown, gray, green, orange, pink, purple, red, white and yellow. Each pixel is associated with one of these colours, and then, as before, a histogram of values for the whole image is computed. This feature vector thus has eleven dimensions.

SIFT – Bag of Visual Words Scale Invariant Feature Transform is a local feature descriptor which is invariant to certain transformations, such as scaling, rotation or brightness. The algorithm extracts interesting points in an image, which can then be used to identify similar objects. The points usually lie on high-contrast regions of the image, such as object edges. We utilise the algorithm presented in [15], which utilises a Harris corner detector and subsequently the Laplacian for scale selection. We created a 1024 dimensional codebook (Bag of Visual Words), capturing the relative distribution of the SIFT features.

4 Collection

Music information retrieval research in general suffers from a lack of standardised benchmark collections, being mainly attributable to copyright issues. Nonetheless, some collections have been used frequently in the literature. These were howeber not usable for the study in this paper, as none of these collection comes with a complementary set of album covers, and additionally most collections either miss information about song title and artist, or are royalty free music from relatively unknown artists – for both cases, automated fetching album covers from the web is not feasible.

Therefore, we composed our own test collection containing both audio snippets and album covers, by crawling data from the webshop *amazon.com*, which provides rich information for their music shop. Considering the best-selling list from several different genres, for each album (or maxi-single) found, we downloaded the cover, and the 30 second audio snippet of the first song. We thereby skipped entries for which either the cover was of too poor quality (below 400×400 pixels), or the 30 second song snippets was missing. Amazon organises the contents of it its music shop into 25 top-level genres, with many sub-categories; songs may, and frequently are, assigned to multiple genres. We aimed at selecting rather diverse and non-overlapping genres, to achieve distinctive styles in the cover art, and thus chose genres such as 'Goth and Industrial Rock', 'Rap and Hip-Hop', 'Reggae', 'Country', 'Electronic', 'Classical music' and 'Blues'. Overall, the collection comprises more than 900 songs.

5 Experimental Analysis

We trained maps of the size 22×18 nodes, i.e. a total of 352 nodes, with each of the audio features. From a manual inspection, the map trained with SSD features seems to provide the best arrangement of music according to the authors perception, superior to RP and RH features.

This map is depicted in Figure 1, with the result of a clustering of the nodes superimposed on the map lattice.

It can be observed that the classical music (indicated by light-grey colour) is separated rather well from the other genres, being mostly located in the upper-right corner. This area also matches the boundary detected via the clustering of the map nodes using the Ward's linkage method. Gothic and Alternative rock music, indicated in green, is mostly located in the lower-left corner, though a few pieces are distributed on other areas as well. These pieces are mostly slow songs, using a lot of instrumentation found also in e.g. classical music, such as violins, and therefore most of these mapping patterns appear logical from a musical point of view. Reggae music (red) can be mostly found in the upper-left corner and upper-centre, often together with Hip-Hop (blue), with which it shares a lot of rhythmic and tempo characteristics. Jazz/Blues (dark-grey), which borrows many styles from other genres, is organised in a number of smaller, but in itself rather consistent, clusters. The distribution of these clusters all over the map

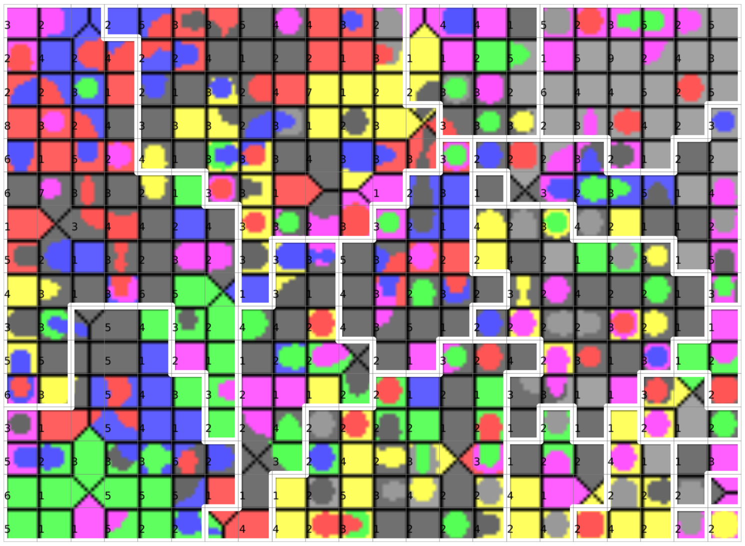


Fig. 1. Distribution of genres over the map with SSD audio features. Clusters obtained via Ward's linkage clustering of the nodes is indicated by white lines

is motivated by the nature of this genre, which is a confluence of several music traditions, and has incorporated many aspects of popular music. Electronic music (pink) shows no clear pattern, distributed in small groups all over the map.

In Figure 2, a Smoothed Data Histogram (SDH) visualisation of this map is depicted, with the 'Islands of Music' [13] metaphor, where islands represent areas with high density. It can be seen that the arrangement of different genres correlates to some degree with the SDH, such as in the area of high-density in the upper right, which represents the cluster of classical music.

For a more detailed inspection, Figure 3 depicts 24 nodes in the upper-right corner of the map, the area containing mostly classical music; this section of the map contains a total of 64 songs. To indicate the genre, the class visualisation [9] from Figure 1 is also used in this illustration, using the same colours as background for the different genres as in Figure 1. On a first glance, there seems to be a certain coherence between the album covers. The most striking shared characteristics between the classical music album covers seems to be the frequent use of photos of people, in some cases the artists themselves, in other cases the musician interpreting the piece of music. These album covers generally follow a rather simple pattern for the background, consisting of few colours, and none or few objects. Many of the albums also simply feature a completely white background. The album covers on the top-edge of the figure mostly belong to the electronic genre; most of them share very similar instrumentation as the classical pieces, mostly the use of a piano or flutes. However, the album art seems to differ quite strongly, with a stronger use of dark colours, and more complex themes.

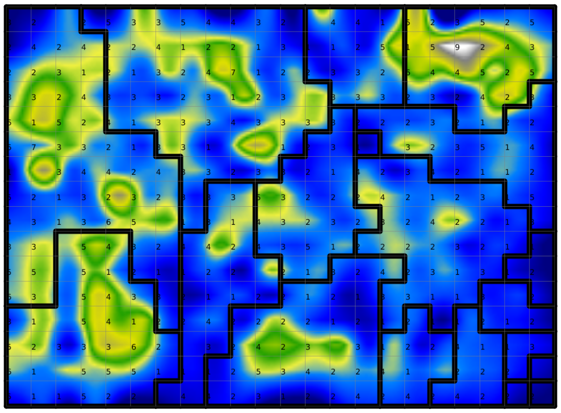


Fig. 2. Smoothed Data Histograms of the map with SSD audio features

We can make similar observations for areas with Jazz and Country music, such as the area on the lower-right of the map, shown in Figure 4(a). Again, most of the covers feature portraits of the artists; however, there is a slightly different pattern in the background, using more darker colours, and thus allowing a subtle differentiation between the previous examples. Similar observations can be made for many Reggae songs.

A cluster with songs from the Gothic and Alternative Rock genre is shown in Figure 4(b). Out of a total of 13 covers, only six show people, and in most cases, these portraits are heavily altered and appear more artificial. Noteworthy is also the use of many dark and flashy colours, which create a dark appearance.

While other areas of the map do not show that clear patterns, it can be concluded that at least to a certain degree, musical similarity as determined by the SSD audio features and the vector projection of the SOM also coincides with some similarity in album cover art.

When organising the map with the image features, we build again on the assumption that album covers carry some clues about the music characteristics, and thus similar music should be located in neighbouring regions of the map. However, when using the simple features such as colour histograms or color names, the latter being depicted in Figure 5(a), this assumption is not fulfilled. While the organisation of album covers along the colour properties gives a nice overview, this arrangement does not match with the genres they belong to, as can be seen in Figure 5(a). There is basically no region in the map that shows a continuous area of similar music. We can thus conclude that for an interface to music, simple features such as the ones derived from colours are not sufficient.

Figure 5(b) depicts a section of a map trained with the SIFT BoV features. This section holds covers that, with a very few exceptions, depict people; further, most of the songs are from the Hip-Hop genre. It could thus be concluded that SIFT BoV features can be useful to detect shapes of faces, which we identified earlier as an important aspect for several genres. We can also observe in some

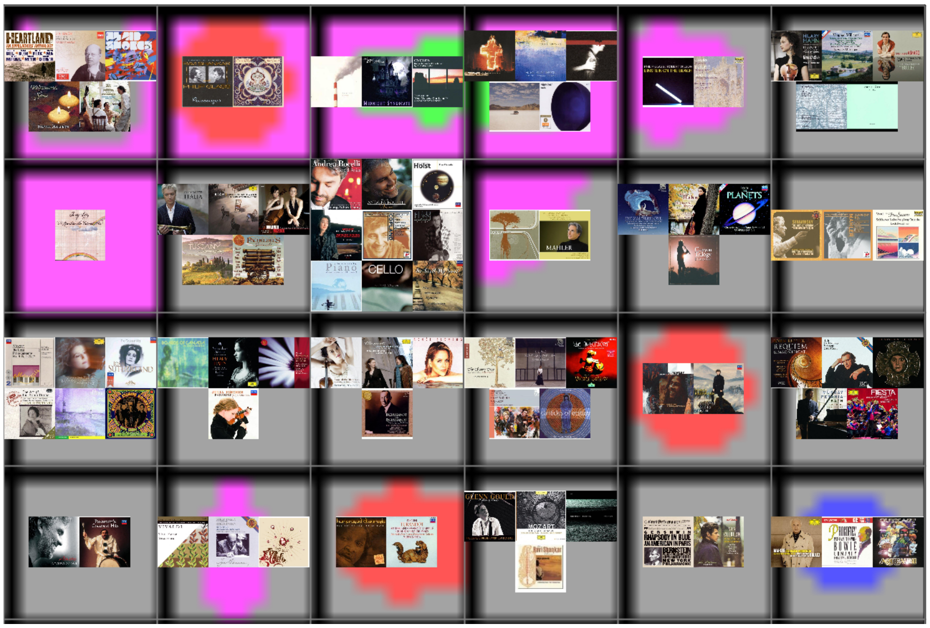


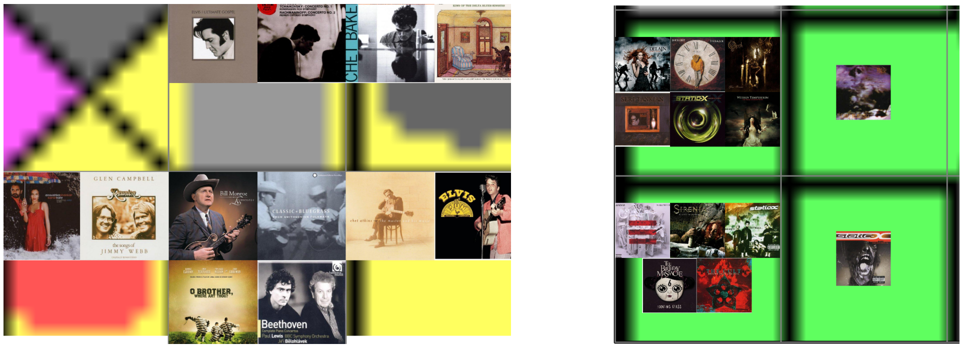
Fig. 3. Album covers in the cluster of classical music (SSD audio features, top-right corner)

other areas that these features are very well working on depicting outliers, mostly albums with very complex cover art. However, similar observations as for the map with color names hold true – the features don't seem to be able to capture the complex similarities in the covers very well.

Finally, we applied the method described in [11], which allows for a analytical comparison of SOMs. It enables to identify differences in mappings obtained by different SOM trainings, by indicating which data items are mapped closely together in both maps. It can also be used to compare two maps trained on different features, for example on the music and song lyrics, as in [10]. Applying this method to the maps trained with the album cover features and the ones extracted from the music, we notice only a very small percentage of matches in the two different mappings – most of the songs that were mapped together in the music SOM are mapped to divergent areas in the album cover SOM.

6 Conclusions

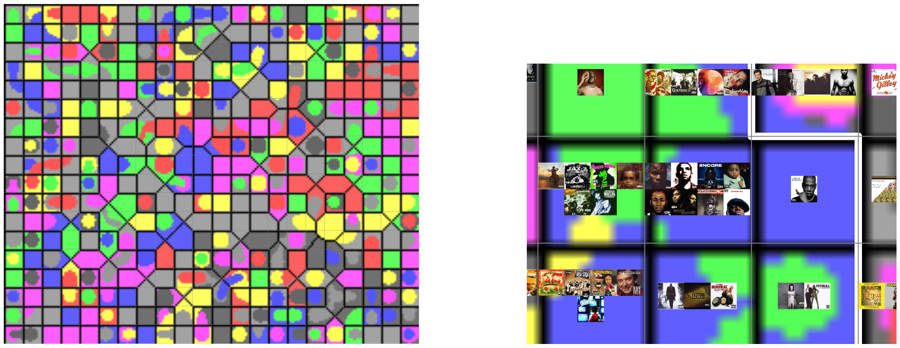
We performed an analysis of the similarity of album art and the music they represent. To this end, we extracted audio features from the music, and image features from the album covers, and trained a set of SOMs with it. The SOM trained with the audio features revealed that in a number of cases, the musical similarity of the music is also reflected in the album covers, e.g. by the use of



(a) Country, Jazz and Classical

(b) Gothic and Alternative Rock

Fig. 4. Album covers in the map with SSD audio features



(a) Color names

(b) SIFT features

Fig. 5. Music maps trained on the image features from the album covers

portraits or rather abstract objects, and also partly by the colours. The maps trained with the image features could, however, only reconfirm some of these similarities, when using the SIFT features to describe the visual content.

We thus conclude that while there is potential in using album covers for music information related tasks, there is a need for more powerful image feature descriptors. Such descriptors could be face detectors, more advanced use of points of interest features, and a combination of these features into a single descriptor.

References

1. Stephan Baumann, Tim Pohle, and Shankar Vembu. Towards a socio-cultural compatibility of MIR systems. In *Proceedings of the 5th International Conference of Music Information Retrieval (ISMIR'04)*, pages 460–465, Barcelona, Spain, October 10-14 2004.

2. Eric Brochu, Nando de Freitas, and Kejie Bao. The sound of an album cover: Probabilistic multimedia and IR. In Christopher M. Bishop and Brendan J. Frey, editors, *Proceedings of the 9th International Workshop on Artificial Intelligence and Statistics*, Key West, FL, USA, January 3-6 2003.
3. Sally Jo Cunningham, Nina Reeves, and Matthew Britland. An ethnographic study of music information seeking: implications for the design of a music digital library. In *Proceedings of the 3rd ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 5–16, Washington, DC, USA, 2003. IEEE Computer Society.
4. Joe H. Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, March 1963.
5. Peter Knees, Markus Schedl, Tim Pohle, and Gerhard Widmer. An Innovative Three-Dimensional User Interface for Exploring Music Collections Enriched with Meta-Information from the Web. In *Proceedings of the ACM 14th International Conference on Multimedia (MM'06)*, pages 17–24, Santa Barbara, California, USA, October 23-26 2006.
6. Jorma Laaksonen, Markus Koskela, Sami Laakso, and Erkki Oja. PicSOM—content-based image retrieval with self-organizing maps. *Pattern Recogn. Lett.*, 21(13-14):1199–1207, 2000.
7. T. Lidy and A. Rauber. Evaluation of feature extractors and psycho-acoustic transformations for music genre classification. In *Proc. ISMIR*, pages 34–41, London, UK, September 11-15 2005.
8. Beth Logan, Andrew Kositsky, and Pedro Moreno. Semantic analysis of song lyrics. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME'04)*, pages 827–830, Taipei, Taiwan, June 27-30 2004.
9. Rudolf Mayer, Taha Abdel Aziz, and Andreas Rauber. Visualising class distribution on self-organising maps. In Joaquim Marques de Sá, Luís A. Alexandre, Włodzisław Duch, and Danilo Mandic, editors, *Proceedings of the International Conference on Artificial Neural Networks (ICANN'07)*, volume 4669 of *LNCS*, pages 359–368, Porto, Portugal, September 9 - 13 2007. Springer.
10. Rudolf Mayer, Jakob Frank, and Andreas Rauber. Analytic comparison of audio feature sets using self-organising maps. In *Proceedings of the Workshop on Exploring Musical Information Spaces, in Conjunction with ECDL 2009*, pages 62–67, Corfu, Greece, October 2009.
11. Rudolf Mayer, Robert Neumayer, Doris Baum, and Andreas Rauber. Analytic comparison of self-organising maps. In *Proceedings of the 7th Workshop on Self-Organizing Maps*, pages 182–190, St. Augustine, Fl, USA, June 8–10 2009.
12. Nicola Orio. Music retrieval: A tutorial and review. *Foundations and Trends in Information Retrieval*, 1(1):1–90, September 2006.
13. Elias Pampalk, Andreas Rauber, and Dieter Merkl. Using Smoothed Data Histograms for Cluster Visualization in Self-Organizing Maps. In *Proceedings of the International Conference on Artifical Neural Networks (ICANN'02)*, pages 871–876, Madrid, Spain, August 27-30 2002. Springer.
14. Alfred Ultsch and H. Peter Siemon. Kohonen's Self-Organizing Feature Maps for Exploratory Data Analysis. In *Proceedings of the International Neural Network Conference (INNC'90)*, pages 305–308, Dordrecht, The Netherlands, 1990. Kluwer Academic Press.
15. K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1582–1596, 2010.

16. J. Van De Weijer and C. Schmid. Applying color names to image description. In *IEEE International Conference on Image Processing (ICIP 2007)*, volume 3. IEEE, 2007.