Information Retrieval and Recommendation using Emotion from Speech Signals

Alexander Iliev
University of Wisconsin
Stevens Point, WI 54481,USA
IMI - BAS
8, Akad. G. Bonchev, Str.
1113, Sofia, Bulgaria
ailiev@uwsp.edu

Peter L. Stanchev
Kettering University
Flint, MI 48504, USA
IMI -BAS
8, Akad. G. Bonchev, Str.
1113, Sofia, Bulgaria
pstanche@kettering.edu

Abstract

In this paper we describe a system of retrieving information from artwork based on textual cues, descriptive to relative art pieces, made available through the metadata itself. Large datasets of artwork can easily be mined by using alternative queries and search methodologies. In the most common search methodology a text-based query using a keyboard is performed. We are proposing a method for searching, finding and recommending digital media content based on pre-set metadata text queries organized in two categories, then mapped to speech sentiment cues extracted from the emotion layer of speech alone. We also account for the difference in sentiment expression for male and female speakers and further suggest that this differentiation may improve system performance.

Keywords: Data Search, Recommendation, Metadata, Speech, Sentiment, Emotion Recognition

1. Introduction

Today's smart systems aim to resolve the tremendous task of big data mining in various ways. Most of these are through the use of conventional text search [1]. The most common information retrieval is also performed through the use of text-based queries. In our work we used an art collection dataset from the Rijksmuseum in Amsterdam for the purpose of finding specific art based on descriptive cues, extracted from the metadata. To achieve our goal, we split our approach in two stages: *stage one* was to convert a small subset of the artwork in two pools: *joyful or happy* and *calming, melancholic or saddening*. In *stage two* we detected the emotion of a speaker, which was later mapped to the cues predetermined in stage one. The text information was firstly converted to a specific period and then to art items from that period. The conversion is based on rules using local image features such as harmony and contracts [2]. Once this was done, we entered the second stage. It is well known that emotions are conveyed through different visual cues such as body language or facial expression [3], as well as physiological changes such as sweating or change of skin coloration [4-6]. Our focus was on speech signals alone. More specifically, we based our approach on detecting the emotional status of a person and then applying the results in finding relevant artwork based on mood. This is to say that most of our work went into the second stage. Hence applying the results of emotions extracted from speech was performed on the art collection for practical purposes only.

In our work we used an art collection with 900 paintings from the following periods:

- Baroque period, era in the history of the Western arts roughly coinciding with the 17th century. Some of the qualities most frequently associated with the Baroque are grandeur, sensuous richness, drama, vitality, movement, tension, emotional exuberance, and a tendency to blur distinctions between the various arts. The big presence of dark colors and dark-light contrast is typical for Baroque. This is connected to using techniques of oil painting that gives very deep and dark effects in the artwork from one side and light-dark contrast on the other.
- Cubism, highly influential visual arts style in the 20th century that was created principally by the painters Pablo Picasso and Georges Braque in Paris between 1907 and 1914. Cubist painters were not bound to copying form, texture, color, and space; instead, they presented a new reality in paintings that depicted radically fragmented objects, whose several sides were seen simultaneously.

0-7695-6354-6/18/$31.00 ©2018 IEEE
DOI 10.1109/MIPR.2018.00054

- Expressionism, artistic style in which the artist seeks to depict not objective reality but rather the subjective emotions and responses that objects impose on the viewer. This is accomplished through distortion, exaggeration, primitivism, and fantasy and through the vivid, jarring, violent, or dynamic application of formal elements.
- Impressionism, is a study of sunlight, which alters the local tones of natural objects, and study of light in the atmospheric world of landscape, provided the Impressionist painters with new essential patterns shows the distribution of lightness in paintings from different movements.

2. Proposed approach

Information retrieval methodology varies depending on the type of media it is applied to. In this work we propose an approach based on emotion recognition from speech signals with application to artwork recommendation using text-descriptive metadata. The latter represents a dataset of the artwork denoted by *A*. For practical purposes we created a smaller subset B. We then performed our search on the textual content of the metadata portion in subset B. In order for this approach to result in effective outcome, we had to create mapping of some of the keywords from set B into the chosen general mood categories [*joyful, happy*] and [*calming*, *melancholic*, *saddening*]. If we represent the set of keywords from the art descriptive metadata for both datasets as vectors and set vector *A* to represent all metadata elements, then set *B* to represent a smaller subset in *A*, then we can express them as follows: $A \in \{0,1,\ldots m\}$, $B \in \{0,1,\ldots n\}$, where $B \subseteq A$ since m > n.

The next step was to extract sentiment from speech as described in [7-10]. As suggested by Cowie and Cornelius [11-12] there is a set of six basic emotions. These are the so-called "big-six" emotions: *anger, happiness, sadness, fear, surprise,* and *disgust*. For practical reasons and to simplify the task, we have chosen to only work with two out of the 'big-six' emotions. Considering the dataset and the application it was found pertinent that *happy* and *sad* were going to be the most closely applicable to the problem at hand so they were the chosen emotion categories. Another reason for our selection was the fact that they are generally more distinct and reflected the expectation for this application. To clarify, we assumed that most people are not fearful or angry when they observe artwork, but rather joyful or melancholic mood is sought in this scenario.

On the other hand, a specific set of emotions had to be selected for the particular application, which also depended on the available emotion sets relevant to the art-descriptive metadata textual information.

Extraction of the emotional set may happen by using methods established in [4]. In addition, a gender-based system was realized for a finite set of emotions of male and female speakers. In particular, the voiced sections of speech signals were used and the amount of the opening and the closing of the epiglottis were determined through inverse filtering techniques [4]. The voices of five male and five female speakers were used for this exercise.

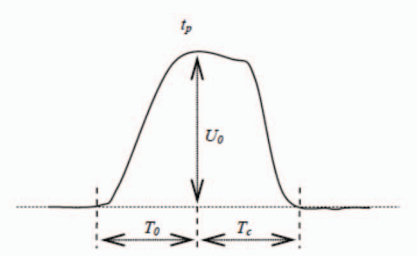

Figure 1: Glottal shape model as established by Fant 1979 [4]

The ratio of the closing over the opening phase of the glottis, as shown in Figure 1, was used and is referred here as the Glottal Symmetry (GS), whish is represented in this form:

$$GS = \left[\frac{T_c}{T_o}\right]$$

where *GS* represents the glottal symmetry, T_c is the closing phase, T_o is the opening phase, U_o is the peak volume velocity of the glottal pulse that occurs at the t_p instance.

The system was trained on 80% of the observations and was tested on 20%. Different methods for classification were considered such as: *Bayes Classifier*, *C4.5*, *SVM*, and *k-NN*. One of the most successful and prominent relevantly new classification method used was the Optimum-Path Forest (OPF) [13]. The OPF algorithm is measuring the connectivity between samples in the feature space. In this regard, it is an improved method over the k-NN classifier since it is not only taking into account the simplest distance from the samples to their corresponding feature vectors, but also it is considering the distance among the samples. In addition, OPF classification rule is performing an optimal search of the entire feature space, thus preventing possible misclassifications as a result of using local decision functions. The OPF method is fairly simple, it is fast, and it works well in multi-class scenarios, as is parameter independent. The method does not make any guesses about the classes, as is the case in SVM or ANN-MLP. The main idea behind OPF is to find strong connections between prototypes from each class, so that new samples are assigned to a class with the shortest prototype connection. In other words, it is the one that offers a minimum-cost path, while taking into account all possible

paths from a given prototype. In this way it works well when classes are overlapping, since these prototypes are defending the class in the overlapping areas of the attribute space.

3. Discussion of results

As can be seen in figures 2 and 3, not only we can determine the emotion for any given subject, but being able to establish a particular emotion by gender can lead to more tailored approach and better data retrieval and as a result improved recommendation. The difference between *happy* and *sad* is quite visible as they exhibit quite different levels of energy in their respective expression. We observe that there is a very different emotional expression between *happy* and *sad* for male speakers as compared to the same expressions in female speakers. The average line of the male representation of *sad* emotion can be visibly drawn around the 1.8 ratio line, where the same for *happy* can be drawn around the 1.2-1.3 ratio line. The difference for female subjects was not as prominent although the overall performance was successful. It is easy to observe that in general the ratios for *sad* are higher than the ones for *happy*, which is because there is a much more gradual opening phase for *sad* as opposed to *happy*. Finally, we can observe some spikes in the extraction of data for female speaker for *happy*.

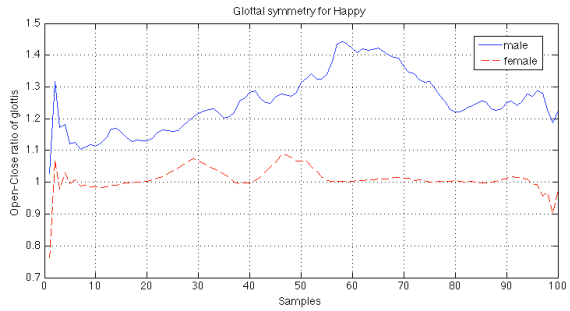

Figure 2: Glottal symmetry for male and female speakers – *happy* emotion

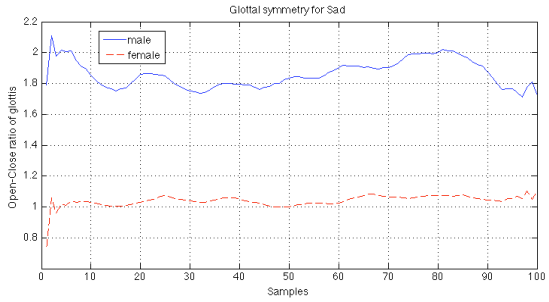

Figure 3: Glottal symmetry for male and female speakers – *sad* emotion

4. Conclusions

In this work we used sentiment from speech in the application of data retrieval based on artwork. It was established that emotion as extracted from speech could be of great value when it comes to finding and recommending art. This problem has an easy solution when the main media (in our case the artwork) has a pretty well described metadata, which can easily be queried. Moreover, the descriptive metadata of any multimedia related datasets could also benefit from this approach. This type of methodology can be successfully applied to various solutions related to multimedia, broadcasting and any kind of sophisticated data retrieval procedure. Accounting for gender specific emotion search may increase the accuracy and performance of data retrieval. Part of the problem was not only to create an accurate mapping between art pieces and emotions, but also to find a fine transition when combining these fields together.

At the end, our experiments show that sentiment recognition as extracted from speech signals can be successfully applied to any media related service that need some level of automation in the quest for smart system development and in addition, can be implemented in a gender separated fashion. Any such system may also be considered as a step towards solving a more global technology development solution in a larger digital asset ecosystem.

Our main contribution is in applying speaker sentiment in order to find and recommend digital content. This is an example of how useful methodology can be implemented in various enabling technologies. Finally, the current research aims to initiate the construction of a conceptual model of a multifunctional digital culture ecosystem based on the latest concepts and solutions in the area of data retrieval and recommendation.

5. Future work

In any further development on the topic we could potentially extract features from speech in two different areas. The first area can deal with recognizing a specific word or set of words from speech by using conventional speech recognition technology using Google Speech API as well as pos tags from the Natural Language Toolkit (nltk). In this case, we could map the words extracted from the speech recognizer to the predetermined mood categories with close approximation to the words extracted from the metadata of the artwork. The second area can be extracting sentiment from speech as described in this paper, hence combining the two and achieving better results.

In order to make this approach closer to the real world, another goal can be set to test the implication of this application in more realistic scenario and analyze user responses in order to verify the validity of our method.

6. Acknowledgments

This work is partly funded by the Bulgarian NSF under the research project № DN02/06/15.12.2016 "Concepts and Models for Innovation Ecosystems of Digital Cultural Assets", WP2 - Creating models and tools for improved use, research and delivery of digital cultural resources, WP3 - Designing a model of a multifunctional digital culture ecosystem.

6. References

[1] Stanchev, P., Multimedia Standards. History. State of Art, T.-h. Kim et al. (Eds.): FGIT 2011, Springer-Verlag Berlin Heidelberg 2011, LNCS 7105, pp. 39-42, 2011

[2] Ivanova K., Stanchev P., Color Harmonies and Contrasts Search in Art Image Collections, First International Conference on Advances in Multimedia, MMEDIA 2009, July 20-25, 2009, Colmar, France, pp. 180-187

[3] Spiros V. Ioannou, Amaryllis T. Raouzaiou, Vasilis A. Tzouvaras, Theofilos P. Mailis, Kostas C. Karpouzis, Stefanos D. Kollias, "Emotion recognition through facial expression analysis based on a neurofuzzy network", Neural Networks, ELSAVIER, 18(4):423-35, May 2005.

[4] Iliev, A., Monograph "Emotion Recognition from Speech", Lambert Academic Publishing, ISBN-10: 3847377604, ISBN-13: 978-3847377603, 2012.

[5] Iliev, A.I., Scordilis, M.S., "Spoken Emotion Recognition Using Glottal Symmetry", EURASIP Journal on Advances in Signal Processing, Volume 2011, Article ID 624575, 2011.

[6] Igor Bisio, Alessandro Delfino, Fabio Lavagetto, Mario Marchese, and Andrea Sciarrone, "Gender-Driven Emotion Recognition Through Speech Signals for Ambient Intelligence Applications", IEEE Transactions On Emerging Topics In Computing, Vol. 1, No. 2, pp. 244-257, December 2013.

[7] Iliev, A.I., "Emotion Recognition in Speech using Inter-Sentence Time-Domain Statistics", IJIRSET International Journal of Innovative Research in Science, Engineering and Technology, Vol. 5, Issue 3, pp. 3245-3254, March 2016

[8] Iliev, A.I., "Feature vectors for emotion recognition in speech", National Informatics Conference, Sofia, Bulgaria, 2016, pp. 225-238

[9] Iliev, A.I., Scordilis, M.S., "Emotion Recognition in Speech using Inter-Sentence Glottal Statistics", Proceedings of the 15th International Conference on systems, Signals and Image Processing, IEEE-IWSSIP 2008, Bratislava, Slovakia, June 25-28, 2008, pp. 465-468

[10] Iliev, A.I., Zhang, Y., Scordilis, M.S., "Spoken Emotion Classification Using ToBI Features and GMM", Proceedings of the 14th International Workshop on Signals and Image Processing 2007 and the 6th EURASIP Conference focused on Speech and Image Processing, Multimedia Communications and Services. IEEE-IWSSIP 2007, Maribor, Slovenia, June 27-30, 2007, pp. 495-498

[11] Cornelius R., 1996. The Science of Emotion. Research and Tradition in the Psychology of Emotion. Upper Saddle River, NJ: Prentice-Hall, pp. 260

[12] Cowie R. and Cornelius R., 2003. Describing the Emotional States that are Expressed in Speech. Speech Communication, Vol. 40, pp. 5-32

[13] Iliev, A.I., Scordilis, M.S., Papa J.P., Falcão A.X., "Spoken emotion recognition through optimum-path forest classification using glottal features", Journal on Computer Speech and Language, ELSEVIER, Vol. 24, Issue 3, 2010, pp. 445-460