Development of web service for generating musical composition from an image

1st Nikita Nikitin
Software Engineering Department
Volgograd State Technical University
Volgograd, Russia
set.enter@mail.ru

2nd Vladimir Rozaliev
Software Engineering Department
Volgograd State Technical University
Volgograd, Russia
vladimir.rozaliev@gmail.com

3nd Yulia Orlova
Software Engineering Department
Volgograd State Technical University
Volgograd, Russia
yulia.orlova@gmail.com

***Abstract*—This paper describes the process of designing and developing a web service for the automated generation of sounds from an image. Also describes the main method for synthesizing music from an image, which is based on the joint use of neural networks and light-music theory. It also contains testing of developed service.**

***Keywords*—recurrent neural network, light-music theory, automated music generation, schemes for correlating colors and notes, Keras, Flask.**

I. Introduction

Nowadays, more and more papers are being published aimed at automating the process of creating musical compositions, however, this process is creative, depends on many factors, starting from the experience and mood of the composer, ending with the area of residence and other external factors, so music cannot be created in automatically mode, therefore, the role of the user-composer is very high and we can only talk about the automation of this process. The emotionality that is conveyed by music and paintings is difficult to recognize [1]. Although the process of creating music is based on clearly defined musical rules, it cannot be completely formalized. To reduce the role of the user in the process of choosing the characteristics of a musical composition, as well as to take into account the emotional component (for example, the emotional state of the user-composer), in this work it is supposed to obtain the characteristics of the composition from an image.

In the framework of this work, automation of the process of creating music by means of automated generation of sounds from an image is assumed. In other words, the generation of sounds from an image is the process of converting an image into one or more sequences of notes, with a certain fundamental tone and duration [2].

II. Basic method

The first step in developing a service is to determine the main method of the program - the method of generating musical material from an image. This method consists of two component:

- an algorithm for correlating color and musical characteristic;
- algorithm for generating melodic parts using neural networks.

A. Algorithm for correlating color and musical characteristic

The main parameters of the resulting musical composition is tonality and tempo. These parameters describe the emotional component of the composition, and should be determined by analyzing the color of the image. For this, first of all, it is necessary to determine the ratio of color and musical characteristics [3]:

- the color hue correlates with the height of the note;
- the color group with the musical mood;
- the brightness with the octave of note;
- the saturation with the duration.

Then, it is necessary to determine the correlation scheme between the color and the pitch. At the moment, there are a large number of such schemes, however, the schemes of I. Newton, Louis-Bertrand Castel, A. Wallace Rimmington, A. Eppley and L. J. Belmont were implemented in this work [4].

The algorithm for determining tonality is based on image analysis and consists of 4 steps:

- The first step is to convert the input image from the RGB color space to HSV. This step allows to convert the image to a more convenient form, since the HSV space already contains the necessary characteristics - the color name (determined by the hue parameter), saturation and brightness parameters [5];
- The second step - analyzing the whole image to determine the predominant color;
- The third step is to determine the hue and color group of the primary color;
- The fourth step - according to the selected scheme of correlation between colors and notes, as well as the results obtained in the previous steps, it is necessary to determine the tonality of the composition. To determine the tempo, it is necessary to obtain the brightness and saturation of the primary color, and calculate the tempo according to these parameters;

B. algorithm for generating melodic parts using neural networks

In this paper, the following algorithm is proposed for obtaining a composition from an image (algorithm for generating the melodic part using neural networks):

- according to the method of correlating color and musical characteristics, the tonality of the work and the sequence of the first 20 percents of the notes read from the image are obtained [6];
- then, according to the obtained sequence of notes, the continuation of the work using the trained model and neural network is predicted;
- according to the final sequence of notes and tonality, and also according to the method of correlation of color and musical characteristics, the harmonic part of the work is built.

III. Developing of the web service

To develop the web site for sound generation based on image color spectrum, the following architecture was proposed (“Fig. 1”):

- the main subsystem to work with user input (also includes the image analysis module);
- subsystem to work with neural networks;
- the subsystem for sound synthesising.

A. The main subsystem

This subsystem provides functionality to work with user input, analysis the given image and then pass the results to other modules. It consists of the 5 modules:

- playback control module - provides functionality to save, play and stop music;
- composition parameter definition module - this module provides functionality to set the characteristics of the result music composition, such as musical instrument (piano or guitar) and correlation scheme (correlation between colors and pitches);
- image uploader - provides ability to choose and load the image from the user’s PC;
- image analysis module - this module analysis the image and determines the characteristics of the result musical composition from an image, such as predominant color, color sequence in suitable format and other image characteristics;
- module for generating xml text - this module generates xml text based on image characteristics provided by previous module.

B. The neural network subsystem

This subsystem provides functionality to work with neural network. The aim of this subsystem is to predict the melodic part of the result musical composition and determine the harmony part. It consists of the following parts:

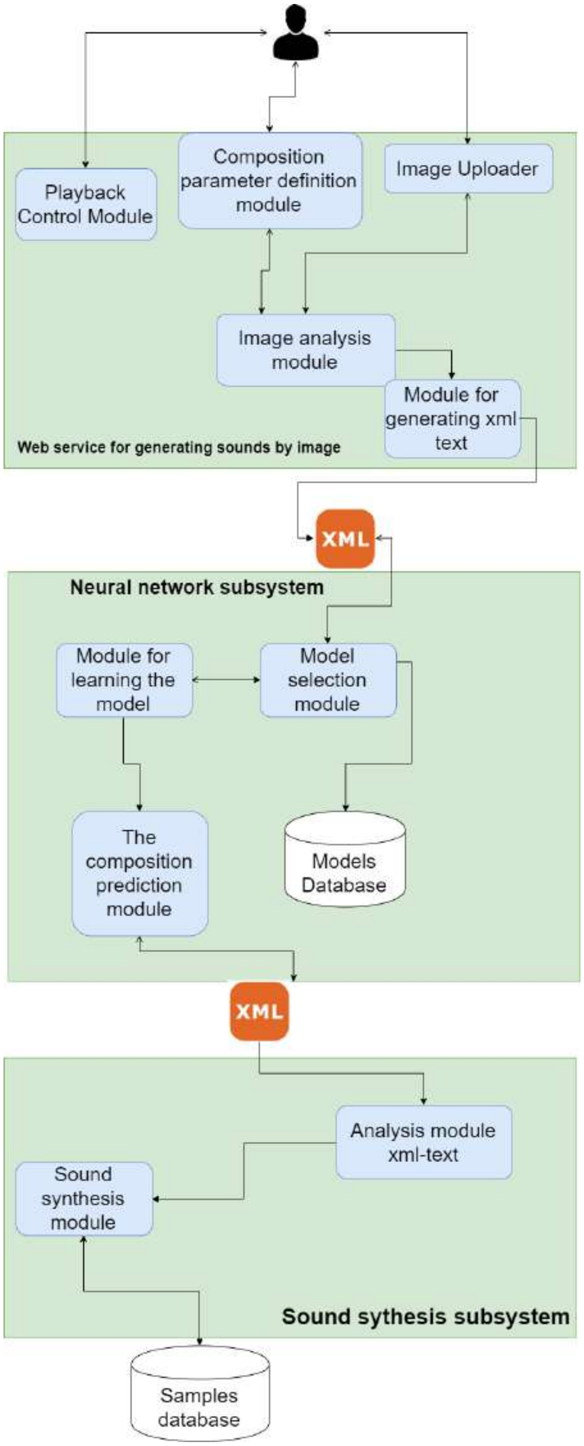


Figure 1. Architecture of the web site.

- learn module - this module is the helper part, this means that the user does not directly interact with this module. It provides the ability to learn the model based on given corpus of musical compositions in midi format, After learning, the module saves model in HDF5 file and saves in the database

(Models database) [7];

- model selection module - this modules determines suitable model from the database (with coordination with the previous module), based on given image characteristics and loads it into memory to subsequent prediction;
- the composition prediction module - this module predicts the melodic part and determines the harmonic part based on certain musical rules [8]. It uses the model has been loaded by the previous module, and also checks the image characteristics and sequence of colors. Also, it forms the xml text to send the composition in internal notation for the subsequent sound synthesis.

C. The sound synthesis subsystem

This subsystem generates the sound in .mp3 format based on given xml text with the result musical composition. It consists of two main modules and one database:

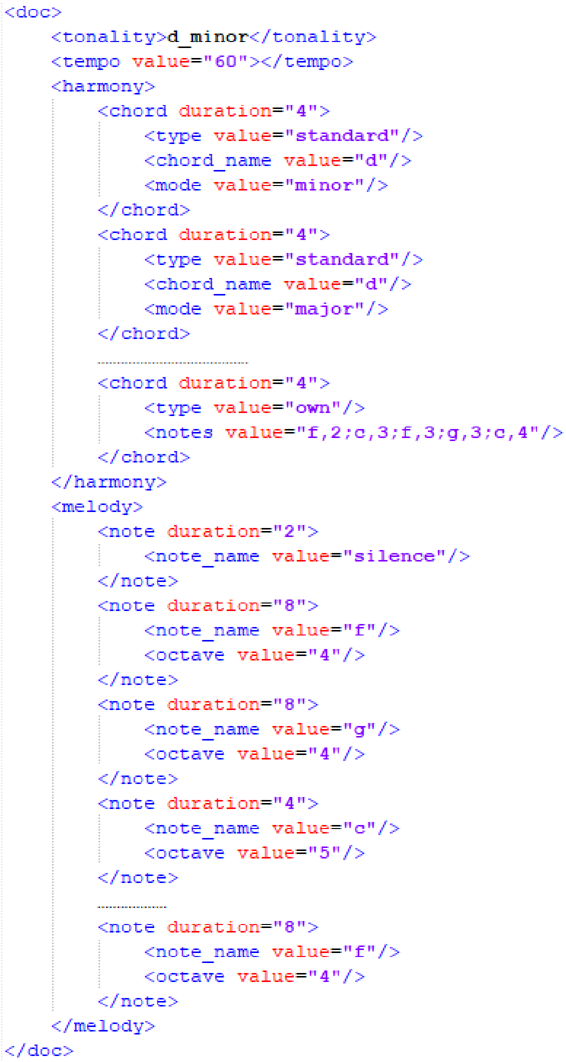
- xml-text analysing module - this module analysis the given xml text from the previous subsystem and translates this text to suitable format for subsequent synthesis;
- sound synthesis module - it combines different pieces of music (samples) [9] from the database with overall mp3 sound file based on the result from the previous module. As the result, this module saves the mp3 file to the server's internal folders and passes the path to the playback control module.

D. Description of the structure of xml text

For the interaction of subsystems, a proprietary xml file structure was developed. It contains the following tags:

- tonality - tag indicating the tonality of the composition;
- tempo - tag indicating the tempo of the composition;
- harmony - tag containing a description of harmony. Inside this tag is the *chord* tag;
- chord - the tag inside which *type*, *chord_name* and *mode* tags are located. The property of the tag is *duration*, indicating the duration of the chord;
- type - the tag, an important part of which is the *value* property, which indicates the type of chord - *standard* or *own*;
- chord_name - the tag denoting the name of the key note (C, D, E, F, G, A, H, etc.);
- mode - tag describing the musical mode of the chord (major or minor);
- if the chord type is *own*, then instead of the *type*, *chord_name* and *mode* tags, the *notes* tag is written inside the *chord* tag, which contains the value properties for *notes* and *durations* of the chord to create. Notes are separated by commas, the pairs are separated by a semicolon;
- after closing the *chord* tag, the *melody* tag follows, which indicates the melodic part of the composition. This tag, as well as the *chord* tag, requires explicit closing with the <*/chord*> and <*/melody*> tags respectively;
- inside the *melody* tag the *note*, *note_name* and *octave* tags are located. The first contains the *duration* property, which indicates the duration of the note, the second contains the *value* property, which indicates the name of the pitch (C, D, E, F, G, A, H etc.). Finally, the third *octave* tag has a *value* property that represents the octave of the note;

Example of xml text with harmonic and melodic parts is represented at "Fig. 2"

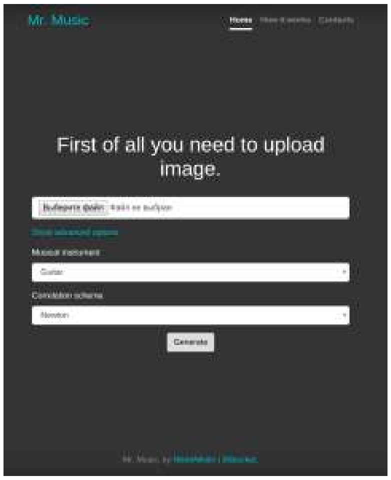


```
<doc>
    <tonality>d_minor</tonality>
    <tempo value="60"></tempo>
    <harmony>
        <chord duration="4">
            <type value="standard"/>
            <chord_name value="d"/>
            <mode value="minor"/>
        </chord>
        <chord duration="4">
            <type value="standard"/>
            <chord_name value="d"/>
            <mode value="major"/>
        </chord>
        ..........................
        <chord duration="4">
            <type value="own"/>
            <notes value="f,2;c,3;f,3;g,3;c,4"/>
        </chord>
    </harmony>
    <melody>
        <note duration="2">
            <note_name value="silence"/>
        </note>
        <note duration="8">
            <note_name value="f"/>
            <octave value="4"/>
        </note>
        <note duration="8">
            <note_name value="g"/>
            <octave value="4"/>
        </note>
        <note duration="4">
            <note_name value="c"/>
            <octave value="5"/>
        </note>
        ............
        <note duration="8">
            <note_name value="f"/>
            <octave value="4"/>
        </note>
    </melody>
</doc>
```

Figure 2. Example of xml text with harmonic and melodic parts.

E. Screenshots of the web service

Screenshots of a web service for generating a musical sequence from an image are shown in "Fig. 3" and "Fig. 4".



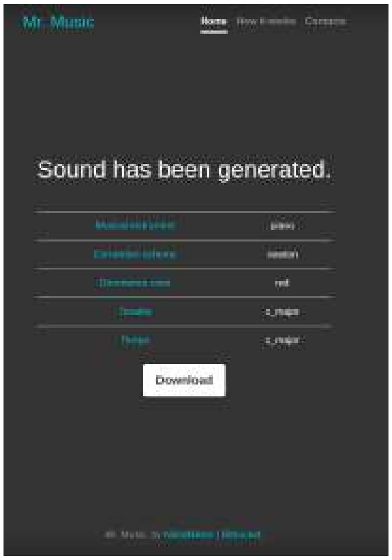
First of all you need to upload image.

Выберите файл Файл не выбран

Musical instrument: Guitar

Correlation scheme: Newton

Figure 3. Image sound generation page.



Sound has been generated.

Musical instrument	piano
Correlation scheme	newton
[illegible]	red
[illegible]	c_major
[illegible]	c_major

Figure 4. Generated sounds download page.

IV. Testing of neural network parameters

A. Number of LSTM layers

The size of LSTM layers has a big impact on the quality of the resulting musical composition. The larger the size, the faster the network converges, however, this leads to retraining. This may be due to the fact that a larger number of neurons allows the network to store more training set data in the scales. It was found that the most optimal configuration for the proposed model for generating musical compositions is the presence of 4 LSTM levels, each of which contains 512 cells [10].

B. Batch size for network training

The training data should be divided into small parts, according to which the neural network is trained. After each such batch, the state of the LSTM network is reset and the weights are updated. For the testing, various batch sizes were tested and it was found that smaller batch sizes make the model more efficient. Nevertheless, the presence of small batch sizes makes training very long. It was found that a batch size of 512 gives the best result while maintaining relatively fast training [10].

V. Conslusion

In this work, the main stages of developing a web service for generating musical compositions from an image was described: the main method of the program, then the application architecture, and finally, the testing process.

Acknowledgment

The work is partially supported by the Russian Foundation for Basic Research and the administration of the Volgograd region (projects 18-07-00220, 19-47-340003, 19-47-343001, 19-47-340009 project).

References

[1] V. Rozaliev Methods and Models for Identifying Human Emotions by Recognition Gestures and Motion. The 2013 2nd International Symposium on Computer, Communication, Control and Automation 3CA 2013, 2013, pp. 67–71.

[2] Wu.Xiaoying A study on image-based music generation. Master's thesis, Simon Fraser University, 2008.

[3] D. Cope Computer Models of Musical Creativity, Cambridge, MIT Press, Cambridge Mass., 2005.

[4] D. Chernyshev Tsveta i noty. Available at: http://mi3ch.livejournal.com/2506477.html (accessed 2019, Dec)

[5] R. Szeliski Computer Vision: Algorithms and Applications, New York, Springer Science & Business Media, 2011. 979 p.

[6] J. Fernández, F. Vico AI Methods in Algorithmic Composition: A Comprehensive Survey. Journal of Artificial Intelligence Research, 2013, no 48, pp. 513-582

[7] J. Schmidhuber Deep learning in neural networks: An overview, Neural Networks, 2015, vol 61, pp 85-117

[8] V. Vakhromeev, Elementarnaya teoriya muzyki [Elementary music theory], Moskow, Gosudarstvennoe muzykal'noe izdatel'stvo, 1962. 248 p.

[9] M. Russ Sound Synthesis and Sampling, London, Taylor & Francis Group, 2012. 568 p.

[10] N. Hewahi, S. AlSaigal and S. AlJanahi Generation of music pieces using machine learning: long short-term memory neural networks approach. Arab Journal of Basic and Applied Sciences, vol. 26, no. 1, 2019, pp. 397-413

Разработка веб сервиса для генерации музыкальной композиции по изображению

Никитин Н.А., Розалиев В.Л., Орлова Ю.А.

В данной статье описывается процесс проектирования и разработки веб-сервиса для автоматической генерации звуков (музыкальной композиции) по изображению. Также описывается основной метод синтеза музыки из изображения, который основан на совместном использовании нейронных сетей и светомузыкальной теории. В первом разделе кратко описано развитие данной области и описана постановка задачи. Второй раздел посвящён определению основного метода работы программы - преобразования худодественных характеристик в музыкальные. Третий раздел содержит детальное описание процесса разработки веб-сервиса. Последний раздел описывает процесс тестирования.

Received 30.12.2019