Supplementary Material

October 20, 2022

1 Proposed MELON dataset statistics

This section describes the full dataset statistics in terms of music and image-caption sets.

1.1 Music audio samples study

- **Mood based audio sample distribution:** Figure 1 shows the distribution of the Jamendo music dataset [] audio samples for the selected moods. The plot is sorted in descending order of number of audio samples present in a particular mood. The moods are categorized based on emotions as well as themes where emotions include relaxing, ambiental, emotional, sad, happy, funnny, romantic, melancholic, calm, positive etc. and examples of themes are christmas, adventure, action, children, commercial etc. There are few moods associated specific to sound e.g slow, fast are based on tempo/speed of the music based on beats per minute. Similarly, soft, cool ,upbeat are based on the pitch of the music.

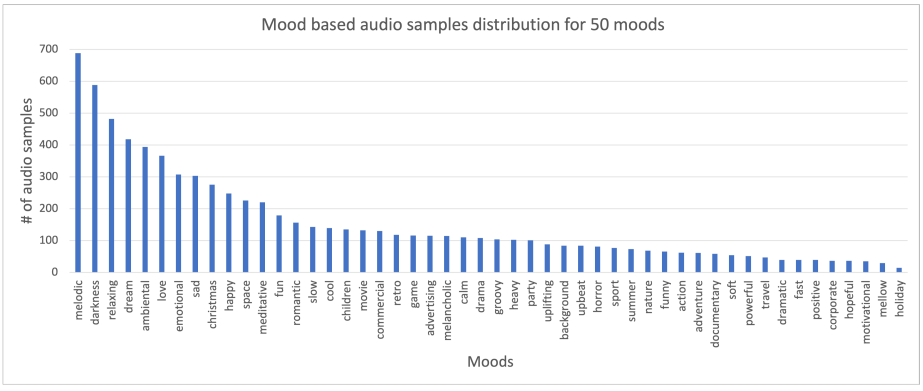

Figure 1: Histogram plot of selected 50 moods based on number of audio samples present in the mood.

- **Audio correlation based on mood** To analyse the correlations among the moods, chroma features are used. Chroma features are an interesting and potent representation for music audio. They project the whole spectrum into 12 bins to represent the 12 unique semitones (or chroma) of the musical octave. Knowing the distribution of chroma, even without the absolute frequency (i.e., the original octave), can provide useful musical information about the audio and may even reveal perceived musical similarity that is not visible in the original spectra because in music, notes that are exactly one octave apart are perceived as being particularly similar. Therefore, from an audio of 20sec duration chroma features are extracted and constant-Q transform [] is applied to get chroma cq features. These time-frequency features are averaged across time to generate a 12d feature vector for the entire audio snippet. Then to compare the feature correlation between any pair of moods from the mood subset = $\{happy, sad, upbeat, calm\}$, we compute average of euclidean distances of between samples of same mood (inter-class distance) and similarly between samples of different moods (intra-class distance). Figure 2 shows the confusion matrix generated by converting the euclidean distance into similarity by inverting and mapping into $[0, 1]$.

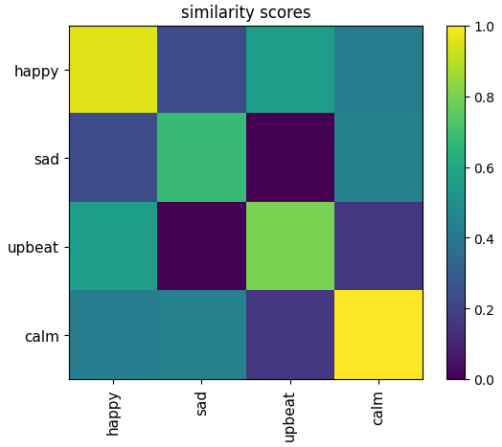

Figure 2: Confusion matrix of moods happy, sad, upbeat and calm based on euclidean distance between features.

1.2 Image-text data distribution

This section will discuss about the diversity of the dataset in terms of different design elements, objects and themes. Figure 3 shows samples of images from different moods. It can be observed that the images are very much relevant to each mood. Since the associated caption also describes the image content, we plot the word cloud for each mood. Figure 4 demonstrates the diversity and variability of our dataset using word clouds of 18 different moods. Each subplot shows words present in the mood extracted from the captions with varying size and colour. The size is representative of the frequency of the word in the captions corresponding to that mood. One can note that the word cloud of any particular mood contains significant variations in tags while being restricted to the relevant theme. For instance, for the `sport` word cloud, there are words like "football", "soccer", "fitness", "basketball" that one would typically associate with `sport`.

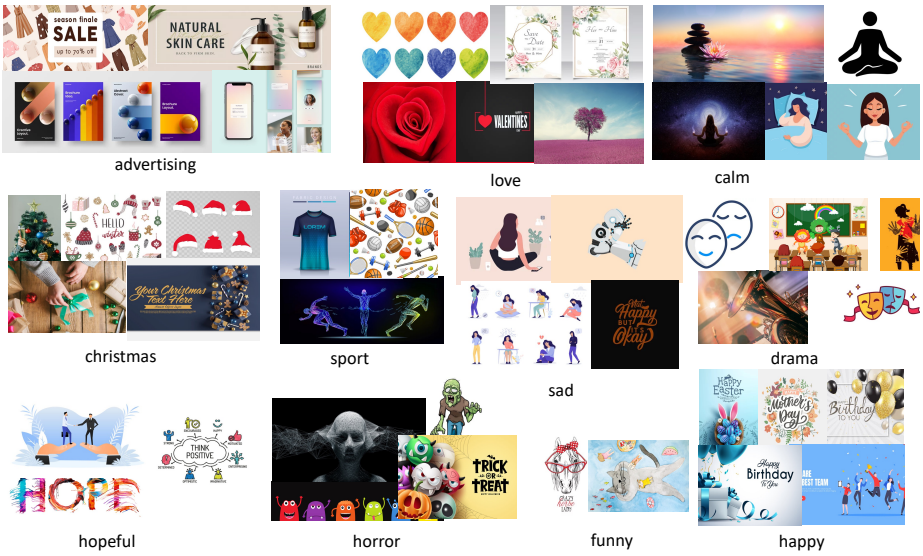

Figure 3: Sample images from different mood categories

Figure 4: The plot shows word cloud for different mood categories as (a) advertising (b) ambiental (c) christmas (d) commercial (e) drama (f) emotional (g) fun (h) game (i) happy (j) hopeful (k) melodic (l) relaxing (m) sad (n) space (o) sport (p) upbeat (q) summer (r) retro

2 Comparison of proposed "MELON" dataset with other datasets

Figure 5 shows the comparison of proposed dataset with other existing datasets for multi-modal audio retrieval. It can be seen that the orange block containing images from dataset Shuttersong [1] contain human photos and VGGSound dataset [2] contains human action images. Similarly, IMEMNet [3] and Audioset [4] also contain images of people whereas MUGEN dataset [5]comprises of video games images and videos. The right side (blue colour block) shows images from "MELON" dataset which comprises of design elements like shapes like flowers, triangles representing mountains and writing in various fonts and colours. Therefore, this is one-of-a-kind dataset which have not existed before.

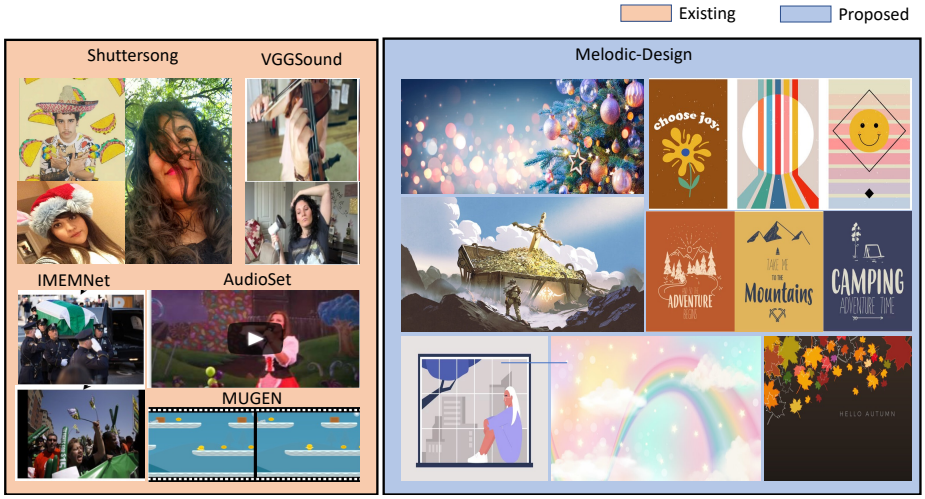

Figure 5: Orange block represents existing datasets visual contents and blue block represent "Melodic-Design" dataset (proposed) visual content. Existing datasets focussed on real-world photos and videos whereas Melodic-Design focussed on design and creative elements.

References

[1] Xuelong Li, Di Hu, and Xiaoqiang Lu, "Image2song: Song retrieval via bridging image content and lyric words," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5649–5658.

[2] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman, "Vggsound: A large-scale audio-visual dataset," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 721–725.

[3] Sicheng Zhao, Yaxian Li, Xingxu Yao, Weizhi Nie, Pengfei Xu, Jufeng Yang, and Kurt Keutzer, "Emotion-based end-to-end matching between image and music in valence-arousal space," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 2945–2954.

[4] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[5] Thomas Hayes, Songyang Zhang, Xi Yin, Guan Pang, Sasha Sheng, Harry Yang, Songwei Ge, Isabelle Hu, and Devi Parikh, "Mugen: A playground for video-audio-text multimodal understanding and generation," *arXiv preprint arXiv:2204.08058*, 2022.