



“Is Text-Based Music Search Enough to Satisfy Your Needs?” A New Way to Discover Music with Images

Jeongeun Park

parkje@hanyang.ac.kr

Division of Media, Culture and Design Technology,
Hanyang University
Ansan, Korea, Republic of

Changhoon Oh

changhoonoh@yonsei.ac.kr

Graduate School of Information, Yonsei University
Seoul, Korea, Republic of

Hyorim Shin

hyorim715@gmail.com

Graduate School of Information, Yonsei University
Seoul, Korea, Republic of

Ha Young Kim*

hayoung.kim@yonsei.ac.kr

Graduate School of Information, Yonsei University
Seoul, Korea, Republic of

ABSTRACT

Music is intrinsically connected to human experience, yet the plethora of choices often renders the search for the ideal piece perplexing, especially when the search terms are ambiguous. This study questions the viability of employing visual data, specifically images, in innovative queries for music search, and it aims to better align search results with users’ moods and situational context. We designed and evaluated three prototype systems for music search—TTTune (text-based), VisTune (image-based), and VTTune (hybrid)—to comparatively assess user experience and system usability. In a comprehensive user study involving 236 participants, each participant interacted with one of the systems and subsequently completed post-experimental surveys. A subset of participants also participated in in-depth interviews to further elucidate the potential and the advantages of image-based music retrieval (IMR) systems. Our findings reveal a marked preference for the user experience and usability offered by the IMR approach, as compared with the traditional text-based method. This underscores the potential of the image in an effective search query. Based on these findings, we discuss interface design guidelines tailored for IMR systems and factors affecting system performance, contributing to the evolving landscape of music search methods.

CCS CONCEPTS

• **Information systems** → **Information retrieval**; • **Human-centered computing** → **Empirical studies in interaction design**.

*Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CHI '24, May 11–16, 2024, Honolulu, HI, USA

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-0330-0/24/05

<https://doi.org/10.1145/3613904.3642126>

KEYWORDS

Image-to-music retrieval, music search, multimodal, user experience, system usability

ACM Reference Format:

Jeongeun Park, Hyorim Shin, Changhoon Oh, and Ha Young Kim. 2024. “Is Text-Based Music Search Enough to Satisfy Your Needs?” A New Way to Discover Music with Images. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*, May 11–16, 2024, Honolulu, HI, USA. ACM, New York, NY, USA, 21 pages. <https://doi.org/10.1145/3613904.3642126>

1 INTRODUCTION

Music is deeply entrenched in the human experience, and it transcends geographic, social, and cultural delineations. Its profound impact is so universal that some researchers have suggested it may predate language itself and assert that although some societies lack written or visual art forms, none exist without music [6, 33, 75]. As a cornerstone of human history, music has become an indelible part of our collective and individual identities.

The scope and diversity of music have expanded exponentially. The burgeoning digital era further exemplifies this phenomenon: Spotify, a leading music streaming platform, boasts a staggering catalog of 100 million songs, with an addition of approximately 60,000 new tracks each day [1, 14]. Alongside this proliferation, music search systems have evolved to incorporate sophisticated algorithms and recommendation systems. Beyond merely facilitating search using on concrete attributes such as artist name or song title, these systems now employ predictive algorithms to offer selections that are aligned with individual preferences, thereby amplifying the user’s listening experience [56, 63].

The form of query used in music search has also changed as technology has advanced. In the past, music could only be searched by entering precise text, but in the early 2000s, search methods emerged that utilized specific elements that constitute music, such as melody, beat, and lyrics [29, 48, 73]. However, despite the remarkable growth in methods used for music search, the ways in which music is searched are still directly or indirectly related to music [38]. Starting from the question “How does one search for music when one has no specific music in mind or when the search information is ambiguous and difficult to express in text?” We intend to discuss the need for new methods for exploring music.

There have been numerous studies on new music search methods [17, 21]. Among them, one study used visual data, such as images, as one of the new music search possibilities. Park et al. [55] suggested that images could be interpreted subjectively, and because of this, they could implicitly express the emotions of music. Indeed, from the perspective of expression, we often use simple yet effective visual content such as stickers, emoticons, and emojis, rather than multiple words to convey states or moods [66, 70]. Sometimes, a single intuitive image can convey multiple meanings more implicitly and effectively than can words, and it can produce more efficient results [18, 80]. In this study, considering that emotions or situations could be effectively conveyed through visual materials like images, we anticipated that users could utilize visual data as search queries even when they lacked information about music. Our research was conducted following the procedure outlined below.

First, by conducting a user study, we examined the behaviors of music search and gathered insights into users' perspectives, including pain points and needs related to existing music streaming services. Based on the derived results, we designed three interactive music search systems that can be used when the queries were not clear. These included (1) a natural language (text)-based music search system, which is a conventional method for expressing ambiguous thoughts or emotions, (2) an image-based music search system that uses images as queries, and (3) a hybrid music search system that uses images as queries and assists with text. Following this, interactive interfaces were designed to provide users with a list of searched music based on their input. Subsequently, we recruited participants to determine whether differences existed in the experience users had with these systems and to assess and compare the usability of each system. We recruited 236 participants, and they experienced one of the three systems and responded to a post-hoc survey. Lastly, we conducted interviews to explore and suggest various user scenarios in which our proposed image-based music retrieval (IMR) system could be applied. The findings from our user study, surveys, and interviews are summarized as follows:

- People have difficulty finding music that fits specific situations or moods when they do not have clear information about the music they are seeking, and they want to discover music that they have not encountered before.
- Text-based music search that can explicitly express one's emotion and state shows a higher level of alignment between queries and search results than do image-based music search methods.
- However, user experiences are more significantly positive when using images compared to using text in terms of the music search task.
- Image-based music searches provide higher satisfaction regarding system usability compared with text-based music search methods.
- In particular, image-based music search methods that intuitively express one's emotions or situations may be effective for individuals over the age of 40 years, as opposed to the younger generation aged 19–39 years.
- A hybrid music search system is suitable for enabling users to adapt to a new form of search when they are introduced to an image-based music search system.

- Image-based music search can be specialized for users who face challenges with the use of text, and it can be applied to scenarios such as searching for music that suits driving scenery or finding background music that is suitable for videos/images intended for social media platforms.

Based on these research findings, we present guidelines for the future development of interfaces and for the functionality of music search systems. The contributions of this study to the domain of music search and the human computer interaction (HCI) community are as follows:

- We investigated music search methods in specific scenarios in which input queries are unclear, as this kind of scenario has been little explored, and we discussed insights into the potential of music search systems that are mediated by images.
- We obtained empirical results related to new facets of user experiences with music search systems, through both quantitative and qualitative approaches.
- We designed and developed an interface for an image-based music search system, and we pioneered the user experience of a new kind of music search interface.

2 RELATED WORK

The aim of this study is to explore the need for new music search methods by comparing user experiences and evaluating the usability of different music search approaches. To achieve this, we review studies related to 1) users' music and multimedia experiences, and 2) music search methods.

2.1 Music and Multimedia Experiences

2.1.1 Music Listening Experience. Music, due to its influence on human emotions and physical responses, has been the subject of extensive effort that seeks to design and understand the experience of listeners [11, 32, 71]. Studies related to music listening can be categorized by objective: 1) designing the music listening experience, and 2) understanding the user experience. To examine studies related to the first objective, Bernhardt and Robinson [8] introduced a music mixing interface where users physically manipulated their bodies to blend music pieces, and they provided users with an emotionally immersive experience. In addition, one study introduced a new interactive agent that supported self-awareness, on the assumption that heart-warming melodies and lyrics could increase human self-awareness [11].

For the second objective of understanding musical experience, a pivotal study that delved into the fundamental reasons behind users' music search behaviors was research by Hosey et al. [34]. They categorized how and why people use search on music streaming platforms. They used four categories—listen, organize, share, and fact check—and they observed participants' behavior while the participants searched. The participants rated their search results in terms of success and effort and found that this was influenced by approaching music search with a focused, open, and exploratory mindset [34]. While previous studies have structured the experiences of music listeners, there have been also studies that proposed qualitative heuristic evaluation methods to assess the overall experience of music services [42]. Choi et al. [13] investigated the

usability of and interface satisfaction with YouTube. After analyzing YouTube users' experiences through a qualitative analysis, they proposed wireframes for a video streaming service to enhance the music listening experience. In another study, the user experience of recommendation and automatic curation features on a music streaming service was examined. The findings of this research revealed that users perceive an increase in algorithm personalization as dehumanizing [24]. These studies sought to understand music-listening users and fundamentally enhance their experiences, and they conducted research based on the search methods offered by existing music search platforms. However, the need for new music search methods that diverge from conventional approaches, and the potential impacts of such methods on users remain understudied. Therefore, this study compares proposed search methods for music with a conventional music search method to examine the utility of the new approach.

2.1.2 Music and Audio-Visual Experience. Some studies seek to connect the auditory element of music with visual experiences. Knees et al. [39] proposed a system called nepTune that visually organizes a music collection. They clustered musically similar songs through audio analysis and represented these clusters as three-dimensional islands, creating a virtual landscape. Users can explore this virtual landscape and listen to the music closest to their current location. Users are provided with descriptions and images related to the music, which are obtained through web search technology. While they did not offer an IMR system, they visually represented similar music as islands, allowing users to explore a vast music list effectively. Moreover, by providing visual information related to the music, they enhanced the visual representation of the music and offered users a new experience.

In another study, a music exploration service called Songrium was developed to allow users to visually explore the relationships between original songs and their derivative works in music video clips [31]. They mapped the original songs onto a two-dimensional (2D) space based on the audio similarity. Then, they categorized derivative works of the original songs into six categories, such as singing a song, dancing to a song, and performing a song on musical instruments. Users can explore music derivatives visually through Songrium based on the song characteristics.

Recently, Melchiorre et al. [51] introduced a study on an audio-visual mobile application interface that considers the user's emotional state. They introduced a system called emotion-aware music tower blocks (EmoMTB), which visualizes music as building blocks made of cubes through the t-student distributed stochastic neighborhood embedding algorithm. Similar songs are grouped into the same building block, and neighboring building blocks represent music of the same genre. Users can explore the landscape made of building blocks to select and listen to music from their preferred genre and receive personalized recommendations through customized building blocks. They provided a new user experience by visualizing music through EmoMTB.

Previous studies have grouped and visualized music based on the similarity of audio signals, allowing music exploration related to visual experiences. When there is no clear target music to search for, a new attempt to search for music using an exploratory method was made possible. However, the current study differs in that it does

not start from musical similarity. Instead, it explores the potential of using images as music search queries by extracting features embedded in the images and connecting them to related music. This approach marks a distinct difference from previous studies in the approach to music discovery.

2.2 Music Search Methods

Music information retrieval can be broadly divided into two modalities of search: text and audio samples [12, 55]. Both methods support predefined tag inputs such as artist names, song titles, and album titles (tag level), as well as diverse query inputs like sentence-level descriptions that are not predefined [20]. When music is searched for using tag-level inputs, the greatest challenge is to retrieve music that does not have specified tags from the database. Studies to address the drawbacks of tag-level music search have been conducted by applying metric learning [77] or by using multi-tags with multiple levels [76]. Doh et al. [20] comprehensively compared text-based music search methods at both the tag level and sentence level and proposed a novel stochastic sampling for text inputs.

Music search methods that use audio samples such as melody, beat, or humming are based on the similarity between signals in query samples and in target samples [73]. Marolt [50] proposed a method for indexing 2D shift-invariant melody representations to search for music. Salamon et al. [61] designed a method to extract frequencies that corresponded to the pitches of primary melodies from music files, which can be utilized in search systems. There is also a study that applied genetic algorithms inspired by natural selection and survival of the fittest to audio-based music search methods to enhance search performance [59]. They segmented the acoustic signals precisely and converted them into musical notes, and then they used genetic algorithms to find music that had high relevance. In addition to these studies, music search methods using audio samples have achieved high search accuracy by utilizing query sources that were directly related to the target music; these search methods have been explored extensively [36, 48, 62]. Similar to methods that use text or voice, audio-based music search methods also require prior information about the music to be searched, or they require that parts of the target music be input. However, in this study, we use images as a method for searching music in situations where there is no prior information about the music and the search query is ambiguous.

There are only few studies that focus on how to search for music using the image modality as is done in this study. Some studies have used the emotion of music to search for images [78] or to construct databases of emotional synesthetic between images and music to search for image-music pairs based on emotional similarity [79]. However, these studies differ from the IMR described in the present study, where the input image is used to search for music content. Shang et al. [68] first developed a system that searches for music based on the implicit meaning embedded in images. They extracted captions from images and linked them to the metaphorical meanings embedded in classic poetry to find the most semantically similar musical pieces. However, their search method faces challenges in capturing users' daily emotions or moods due to relying on poetic metaphors. Park et al. [55] constructed large-scale datasets for IMR and used attributes such as mood and theme in images to

search for music. They presented an IMR framework and found that user satisfaction was improved when there was interaction according to the user intervention in IMR. In this study, we enhanced and employed the image-music mapping algorithm they proposed, applying it to system prototypes designed with user interaction elements that they found beneficial for improving satisfaction. Although they proposed an IMR algorithm, the empirical value an actual IMR system provides compared to existing text-based music search methods has not yet been studied. Therefore, our study compared the traditional text-based music search system with the IMR system to investigate the efficacy of the IMR system and verify the necessity of music search methods using image modality.

3 USER STUDY I

We organized the user studies into three steps according to each purpose to examine the following objectives: 1) Understanding users' music search behaviors and identifying their needs and pain points in music search; 2) Investigating the user experience and differences in usability between the systems to assess the need for the IMR system; 3) Exploring potential future usage scenarios for the IMR system. As a first user study, we conducted a focus group interview. For this, we recruited 12 users of music streaming services through online advertisements. There were no restrictions on the streaming platforms they used or on their average daily music listening time during recruitment. However, we sought individuals who had experienced difficulties when searching for music and who could actively express their opinions on the matter. The participants (9 males and 3 females) were all users of music streaming services, and they had an average age of 34. They used four types of music streaming services, and four participants utilized two or more services.

3.1 Study Procedure

Prior to the interviews, participants had an ice-breaking session that lasted approximately 10 minutes. During this time, they provided a brief self-introduction and described the music streaming service they used, when they listened to music, and in what situations. Afterward, a 40 minute session was held to answer questions and share opinions about usage patterns in music search. The facilitator asked open-ended questions based on participants' responses and inquired about when and in what situations they performed music searches, their criteria for selecting search results, any discomfort they experienced during music searches, and aspects they would like to see improved. All the interviews were recorded and transcribed with the consent of the participants. The study protocol was reviewed and approved by the IRB of our institution.

3.2 Findings

3.2.1 Differences in search methods according to the clarity of music search intent. Participants primarily searched for music either for the purpose of listening to music they knew or discovering something new. The former involved having a specific goal in mind for the search, whereas the latter was intended for exploring music from a broad range of options without a clear goal. When the search goal was clear, participants utilized specific pre-existing information such as song titles, artist names, and album titles in

their searches. The clearer the purpose of the search, the more the depth of the search was minimized by directly entering the target, which was mainly the title of the song. Conversely, when the search goal was unclear, they gradually narrowed down the scope based on surrounding information, like artist names and album titles, to pinpoint the precise target information. This is consistent with the findings from Hosey et al. [34], where the researchers observed that users who searched for desired music with a clear and focused mindset, and users who searched with a more open mindset exhibited differences in the entities they clicked on during their search, such as song titles or album names. However, when participants wanted to discover new music and did not have a specific target in mind, they employed two main approaches in their searches. The first approach involved narrowing down the search scope using keywords, such as genres or themes. The participants then explored curated playlists that were created by other users or playlists that were recommended by the system based on the participants' existing music preferences (from keywords to playlists; KP). The second approach was to directly input text (from text to playlists; TP). For instance, this method involved expressing the mood or the situations that were associated with the music, such as "calm jazz for working" or "healing music for rainy mornings," and searching for playlists that matched those descriptions. As such, there were differences between the search approaches in cases where the purpose of the music search was clear and in cases where the purpose was ambiguous. Furthermore, when the purpose was unclear, the methods for searching music could also be categorized into two distinct approaches.

3.2.2 Challenges in search when exact information about the music is lacking. Participants appeared to struggle to find the music they desired, unless they used precise information for their searches. The challenges in music search were evident in conversations with participants who used the KP method (P4) or the TP method (P9) to search for music, as well as those who used both methods (P1, P5). P5 discussed the difficulties that they encountered when using the keywords provided on the streaming platform, stating, "There are often times when the search results don't match the feeling I want. (Entering search terms) has to be very detailed. The more precisely I can express what I want, the higher the satisfaction with the search results." Agreeing with P5's opinion, P1 expressed frustration with the difficulty of selecting search terms, saying, "When I searched with as much detail as 'groovy music to listen to while driving', I did get lists that were somewhat similar to the music I wanted to hear. But it's a bit embarrassing when I can't even think of such descriptions. I just want to listen to music, and I don't even know what I want to listen to..., but it's difficult to search without any information." Responding to this, P4 mentioned, "But the platform I use doesn't support that kind of search (natural language input), right? I have to choose from pre-curated playlists." Similar to the observation made by P4, most commercialized music search systems rely on simple searches using metadata, and they offer limited song retrieval when queries are ambiguous [7, 53]. In addition to these participants, others also pointed out the inconvenience of constrained search methods that involve either text input or selection from curated playlists. Regarding the inconvenience of using the TP method for search, P9 mentioned, "I just type a lot of stuff hoping to find something. That's

what's inconvenient. There's a lot to type. (...) Clicking is convenient, but writing is a little annoying. I make a lot of typos when typing on my phone." According to previous studies [41, 74], more queries provide more stable and reliable search results. Moreover, including additional explanations regarding the user context improves the music search results. To ensure their satisfaction with the search results, participants using the TP method often input lengthy search queries. They reported that this lengthy input process is cumbersome, and they sometimes expressed difficulties, not knowing how to input their queries properly. Through conversations with the participants, we identified the following challenges and discomforts when there is no clear search target: **1) difficulty finding music that adequately reflects the listener's state or emotion, 2) inconvenience of writing long sentences for precise searches, 3) limitations due to constrained search methods.**

3.2.3 Difficulties in expanding musical preferences. Participants had a desire to discover new songs when they searched for music. However, many streaming services limited search results based on the music the users frequently listen to. The search methods that reflected user preferences were highly efficient in recommending music that was likely to satisfy users' musical tastes. Nevertheless, this approach was mentioned as a drawback due to the fact that participants ended up listening to songs of genres and moods that were similar to the ones they usually listened to, which narrowed down the breadth of their preferences. P8 said, *"I feel great when I discover good artists or songs that I didn't know before. But I keep getting search results that are similar to the artists I already like, or songs with similar vibes."* P11 also agreed, saying, *"I feel the same way. Like pre-censored...? There are times when newly found songs don't really feel new."* This phenomenon is similar to a filter bubble, where users are presented with tailored information that leads them to encounter only filtered content, which results in content bias [54]. Users desire serendipity, where significant discoveries arise from complete randomness, but search systems that fulfill this desire are lacking.

4 SYSTEM DESIGN

According to the findings of User Study I reported in Section 3, users encountered difficulties when they searched for songs without clear search terms, and there was a demand for new methods that could reflect their emotions or situations in music search. To satisfy these needs, users should be able to express their emotion or mood even without precise search terms, and there should be an intuitive means of capturing the situations or moments that trigger the desire to search. We intend to explore the use of images that can implicitly carry multiple meaning and serve as cues for evoking music. According to studies by Kellaris et al. [37] and Fraser [23], music can evoke images not only from past experiences but also from cues embedded within the music. In other words, just as music can evoke images, images can serve as catalysts for recalling music. Considering these perspectives, we conducted user-targeted experiments to design and implement a system that could be used to compare IMR methods with a traditional text-based approach. This section provides a detailed description of the process involved in our system design.

4.1 Design Goals

Considering the needs and pain points that the participants identified in User Study I on music search, we established the following three design goals for IMR systems. Table 1 presents the detailed descriptions of user needs and pain points related to setting these design goals and the insights derived from them.

Accordingly, we designed interactive system interfaces for three types of music search systems that could be used in exploratory music discovery contexts: text-based, image-based, and hybrid (image + text) systems. We named the systems TTTune, VisTune, and VTTune, respectively. TTTune used an existing method and served as a basis for comparing user responses to the new music search method systems that used images. VisTune employed only images as search terms for music, without the use of any text intermediaries. VTTune similarly incorporated images as intermediaries, but unlike VisTune, it used images as queries and supported them with text to represent the searched music. Detailed explanations of these approaches are provided in Section 4.3.

4.2 Dataset

This study, which was conducted online, sought to use publicly available datasets that did not present copyright problems to prevent the secondary use of the sound sources provided to the users. Therefore, we employed the open dataset MTG-Jamendo. The MTG-Jamendo dataset can be used for research purposes without copyright problems under Creative Commons licenses. The dataset comprises over 55,000 audio tracks, and each track is associated with multi-label information in 183 categories and is categorized into genres, instruments, and mood/themes. The genre category includes 87 tags, such as rock, pop, soundtrack, and jazz. The instrument category includes 40 tags, such as guitar, piano, and drums. The mood/theme category includes 56 tags, such as love, happy, energetic, and relaxing.

4.3 Interface

The three designed music search systems for this study were developed to satisfy design goals while operating interactively. In addition, in order to allow users to easily access and use the systems online, prototypes of the systems were developed based on the open-source app framework, Streamlit [4]. The main interface features of each system are shown in Figures 1, 2, and 3, for a more detailed overview of the systems, see Figures 8 to 13 in the Appendix.

Common: All three systems offer guidelines for using the system. Furthermore, we aimed to create a system that can be used when users are unsure of what music they want to search for and do not know where to start looking for it. Therefore, we designed the system to provide scenarios, as listed in Table 2, for various situations or moods in advance so that users can visualize the music they want. These scenarios were based on prior research and statistics about why and when people listen to music [47, 60]. Users select one of the provided scenarios and then, using the information from their mental imagery about music in that scenario, input text or images, depending on the system.

TTTune: After a scenario is selected, the text-based TTTune allows users to input their search query in unstructured sentences

Table 1: System design goals designed based on insights derived from user needs and pain points.

No.	Users' needs and pain points	Derived insights	Design goals
1	(P2, 3, 8, 9) Want a simple and intuitive way to search for music.	The search should be straightforward, and the system should not require extensive explanations for use.	Systems should be intuitively usable and convenient, minimizing the depth of interactions.
2	(P1, 4, 7, 10, 11, 12) When the music users want to listen to is unclear, users do not know where to start looking for it.	Provide options for visualizing music that suits the desired situation or mood using examples for various contexts.	Systems should include examples of scenarios enabling users to imagine the type of music they need for specific situations or moods. This approach aims to create a method that allows users to make quick selections with minimal deliberation.
3	(P1, 5, 6, 8, 12) Want the search results to reflect users' current emotions or states more accurately.	The music envisioned by the user and the examples presented by the system may not always match; thus, allowing users to reflect their emotions or states directly is necessary for better alignment.	Systems should allow users to add tags or images, enhancing their selections to offer results that closely match their envisioned music, maximizing the alignment between the user's mental image of the desired music and the system-provided results.

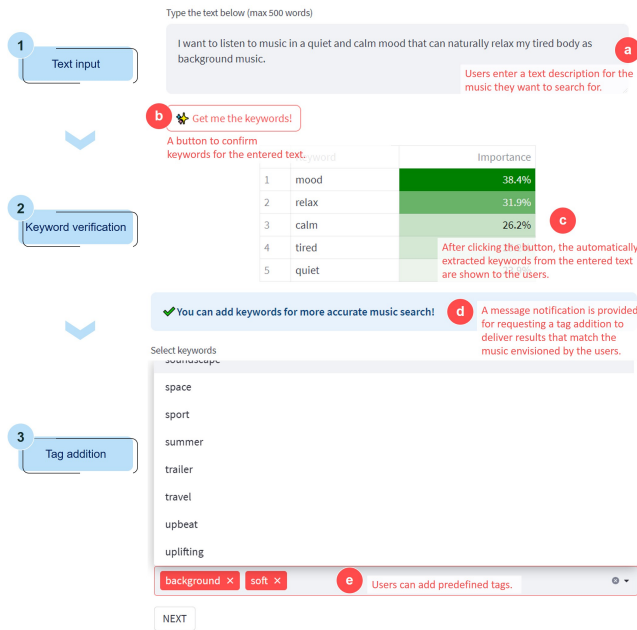


Figure 1: TTTune interface. (a) Users enter a text description of the music. (b) A button confirming the keywords for the text. (c) The automatically extracted keywords from the text are presented after clicking the button. (d) A notification is provided requesting a tag addition to deliver results that match the music envisioned by the user. (e) Users can add predefined tags.

Table 2: Scenarios used in the experiments.

Scenarios
Looking for music to listen to while feeling tired but unable to sleep.
Looking for music to listen to while doing my exercise.
Looking for music to listen to while preparing for a party.
Looking for music to discover new music.
Looking for music to listen to while playing with a child (children).
Looking for music to listen to while studying or working.
Looking for music that matches my current mood on the way to work.
Looking for music for listening to while driving.
Looking for music to listen to in the early morning.

using natural language. For instance, if a user selects a scenario like “*Feeling tired but unable to sleep*” and envisions calming music that can alleviate bodily tension, they can input a search query such as “*I want to listen to music with a tranquil and soothing mood that serves as background music to naturally relax my body.*” Once the query is entered, keywords that are automatically extracted from the user’s input sentence are presented. Users can review these extracted keywords and, if they want to, opt to include additional tags¹ from the predefined 56 mood/theme options. This procedure provides results that match the music users visualize. The 56 mood/theme tags are identical to the mood/theme tag labels used for classifying music in the MTG-Jamendo music dataset.

VisTune: Although TTTune enables the expression of mental imagery through natural language input, VisTune takes a different

¹In this paper, the term “tags” refers to the 56 types of mood/theme labels that constitute the music data, and “keywords” denotes the words automatically extracted from the input text by a keyword extraction model. In order to conceptually distinguish what the two words mean, they are described as tags and keywords, but in the actual system interface, they are unified and used as keywords to prevent confusion among users.

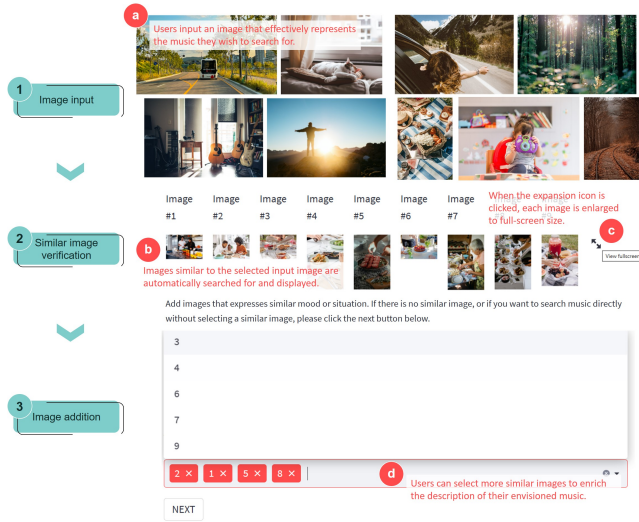


Figure 2: VisTune interface. (a) Users input an image representing the music they want. (b) Images similar to the selected input image are automatically searched for and displayed. (c) When the expansion icon is clicked, each image is enlarged to full-screen size. (d) Users can select more similar images to enrich the description of their envisioned music.

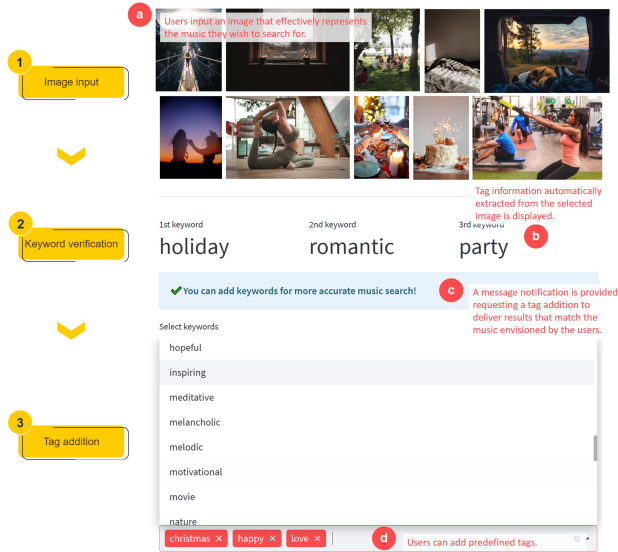


Figure 3: VTTune interface. (a) Users input an image representing the music they want. (b) Tag information automatically extracted from the selected image is displayed. (c) A notification is provided requesting a tag addition to deliver results matching the music envisioned. (d) Users can add predefined tags.

approach by allowing users to input queries through the selection of an image that aptly represents their mental imagery. The

researchers searched online for images that best depict the anticipated situations for each scenario. Subsequently, three to four representative images were selected for each scenario. During this process, care was taken to ensure that the images could encompass contrasting moods. For example, for a scenario related to exercise, images included both quiet exercises for body and mind training and powerful gym workouts. For a scenario related to early morning, images ranged from those conveying a calm and soothing atmosphere to those conveying a lively and cheerful one. This approach was taken to maximize the diversity of the range of situations associated with each scenario. Suppose a user selects a scenario of “Preparing for a party” and envisions a cheerful gathering of friends for a birthday celebration, VisTune presents a few example images to choose the one that best represents this mental imagery. After an image is selected, VisTune automatically searches and displays up to 10 images that capture a vibe and situation that are similar to those of the selected image. If desired, users can select additional similar images to obtain results more closely aligned with the music they envision.

VTTune: VTTune follows the same process as VisTune, and it starts with selecting a scenario and choosing a query image. However, while VisTune shows images similar to the selected query image and allows users to select additional images to better convey their mental imagery about the music, VTTune displays the tag information associated with the chosen query image. Subsequently, similar to TTTune, users can select additional mood/theme tags that they want to include from a predefined set of 56 tags, thereby enhancing the expression of their envisioned music.

4.4 Algorithms

The algorithms for the three music search systems designed for this study operate as illustrated in Figure 4, and their descriptions are as follows:

TTTune: When a user provides a natural language description of the desired music for search, a keyword extraction algorithm called KeyBERT is employed to extract keywords from the input sentences [28]. These extracted keywords must match the tag information in the music dataset used in this study. Therefore, a comparison between the extracted keywords and the tag information in the music dataset is carried out. If the extracted keywords are already present in the music tags, they are used as they are. However, if they are not present in the music tags, the keywords are replaced with words that are similar to the music tags based on the synonyms provided by the NLTK WordNet library [3]. For instance, “peaceful” substitutions to “calm,” and “unhappy” substitutions to “sad.” Ultimately, the entered tags and the substituted tags are combined to create the final tag selection.

VisTune and VTTune: Both VisTune and VTTune employ the same algorithm. We utilize the image dataset constructed by Park et al. [55], which is a large image database with multi-label tag information, and we apply their proposed IMR algorithm. This algorithm ensures that the images and the music in the MTG-Jamendo dataset share identical tag information, which allows for them to be mapped to one another. When an input image is provided, the algorithm retrieves similar images from the constructed database and combines the tag information associated with those similar images

to generate the final tags for the input image. The system operates by searching for music that has the same tags as do the generated final tags for the image. Park et al. only used the mood/theme tag information from the music dataset. However, for this study, the algorithm was modified to incorporate both the mood/theme and the genre tag information available in MTG-Jamendo in order to create a more detailed description of the music. VisTune and VTTune differ in terms of their interface and interaction methods. VisTune uses only images to search music and diversify tag information. In contrast, VTTune employs text, not images, to diversify the tag information.

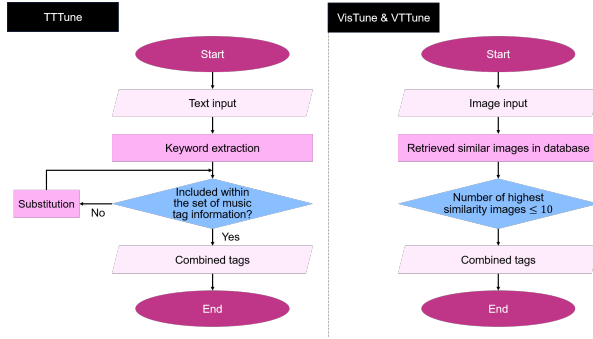


Figure 4: Operational algorithms for TTTune, VisTune, and VTTune. TTTune extracts keywords when text is entered and compares the keywords with the tag information from the music dataset. If the extracted keywords are not included in the music tag information, a substitution process takes place. This ensures alignment with the tag information in the music dataset, after which all refined words are combined. VisTune and VTTune search within the database for images with features similar to the input image when an image is entered. Among the retrieved images, the top 10 with the highest similarity are selected, and their tag information is combined.

5 USER STUDY II

To explore the potential of using IMR systems in situations where the searched music is not clear, we compared the user experience and system usability of a text-based music search system (TTTune) with IMR systems (VisTune and VTTune). We examined the task-related aspects connected to the music search objective and the attractiveness of the interface to assess the suitability of the image modality for music searching.

5.1 Participants

In this experiment, adults without auditory or visual impairments were selected as participants because they were required to listen to the searched music, and those VisTune and VTTune users had to select images. In addition, the experimental criteria for participants were that they had to be people who were familiar enough with digital devices to participate in the online experiment and complete the survey. Given that music is consumed across generations, we tried to recruit not only the younger generation (aged 19–39 years),

Table 3: Demographic information of the all participants.

Attribute	Variable range	Sample size (N = 236)
Gender	Male	155 (65.7%)
	Female	81 (34.3%)
Age	19-29	53 (22.4%)
	30-39	88 (37.3%)
	40-49	45 (19.1%)
	50-59	29 (12.3%)
	Over 60	21 (8.9%)
	American Indigenous	1 (0.4%)
Ethnicity/Race	Asian	53 (22.5%)
	Black	18 (7.6%)
	Latinx/Hispanic	8 (3.4%)
	Middle Eastern/North African	1 (0.4%)
	Multi Race/Ethnicity	3 (1.3%)
	White/Caucasian	146 (61.9%)
	Not to disclose	6 (2.5%)

but also individuals from older age groups to ensure a balanced generational range. The participants for the study were recruited through the online crowdsourcing platform Amazon Mechanical Turk and offline advertisements. We recruited 247 participants. However, excluding those who dropped out or completed the experiment in less than five minutes, leading to unreliable responses, only 236 participants (155 males and 81 females) were included in this study. The number of respondents for each system, TTTune, VisTune, and VTTune, was 80, 77, and 79, respectively. The participants' ages ranged from 19 years to over 60 years, where 59.7% of participants were in the age range 19 to 39 years, and 40.3% were 40 years or older. Table 3 summarizes the participants' demographic information.

5.2 Study Procedure

This study was conducted online. Participants were randomly assigned to one of the three prepared systems and were provided with a URL to access the online app. Upon accessing the provided URL, participants read an introductory text before proceeding to the main experiment. The introductory text explained the research procedure and methods, and participants were asked to click the Agree button if they agreed to participate. If they did not agree, they were excluded from the study. The experiment consisted of the following steps:

1) Scenario selection: Participants choose their preferred scenario from a set of predefined scenarios describing situations or moods for which they may want to search for music. These scenarios describe a situation or mood associated with listening to music, such as “Looking for music that matches my current mood on the way to work” or “Looking for music to listen to in the early morning.” Nine scenarios were used in this study, as shown in Table 2. **2) Query Input:** After selecting a preferred scenario, TTTune users input a text description elaborating on their chosen scenario. VisTune and VTTune users, on the other hand, select an image sample that best represents the chosen scenario. **3) Confirmation:** After the text input or image selection is completed, TTTune users are shown the keywords extracted from their input text. VisTune users are presented with similar images that match the chosen mood or situation, based on the selected image. VTTune users can confirm the tag

information associated with the selected image. **4) Addition:** After the keywords are confirmed, TTTune and VTTune users can add additional tag information. VisTune users, after reviewing similar images, can also add supplementary images that further convey a similar vibe or situation as their chosen image. **5) Listening to retrieved music:** Based on the input and additional information provided by the user, a set of five music tracks is presented as the search result. Participants listen to the presented music tracks and assess how well the retrieved music matches their mental imagery. **6) Iteration:** Steps 1 to 5 are repeated a total of three times with different scenarios that are chosen by participants. **7) Survey response:** After using the system three times, the participants respond to a survey in the final step. The survey measures are detailed in a later section. We also added an open response section to the survey so that participants could freely express their thoughts on the systems.

5.3 Survey Measures

To explore the potential of our new search system using the image modality, we designed an experiment with a between-subject design to compare 1) the degree of match between user input queries and retrieved music across the designed systems (music suitability), 2) the overall user experience with the systems, and 3) the perceived usability of the systems. For this purpose, we employed widely used metrics in the HCI field, namely the user experience questionnaire (UEQ-S) and post-study usability questionnaire (PSSUQ), to compare the three systems in this study [40, 44, 45]. We examined the statistical significance through the analysis of variance (ANOVA) and post-hoc tests.

After listening to five searched tracks, participants rated how well the music matched what they envisioned to measure music suitability. The study used a 7-point Likert scale, where 1 indicated "strongly disagree" and 7 indicated "strongly agree." The experiment was repeated three times; thus, the music suitability assessment occurred three times. UEQ-S measures a total of eight items, and it examines both task-related pragmatic aspects and non-task-related hedonic aspects. Participants express their preference for one of the two presented words using a score from 1 to 7. For example, if 1 indicates "not interesting" and 7 indicates "interesting," scores closer to 7 indicate that the experience of using the system is more interesting, whereas scores closer to 1 indicate that the experience is less interesting. PSSUQ comprises 16 items that are measured on a 7-point Likert scale, and they measure overall usability, system usefulness, information quality, and interface quality. In PSSUQ, 1 typically represents strong positive sentiments, whereas 7 represents strong negative sentiments. However, in our survey, the labels were changed so that higher scores indicate strong positivity [72]. Reverse scoring was used to maintain the original scale.

The UEQ-S and PSSUQ provide results from large benchmark datasets measured across various applications and websites [64, 65]. For example, UEQ-S can be compared to mean scores of 486 studies involving data from 21,175 people, enabling internal validation of the designed system. Therefore, we compared the three systems and provided benchmark scores as a reference to determine the level of each system's evaluation scores.

5.4 Results

5.4.1 Music matching suitability. The evaluation of the suitability of music matching by users for the TTTune, VisTune, and VTTune systems resulted in scores of 5.73, 4.44, and 4.89, respectively. During a total of three trials, participants mentally envisioned their desired music based on the scenarios they personally selected, and they expressed them using either text or images. Afterward, the matching suitability between the searched and envisioned music was evaluated, and these results were averaged for each system. The results showed that participants who used TTTune provided higher evaluations of music matching suitability compared with the other two systems. This could be attributed to the fact that using text as a medium allows users to express their desired music more elaborately. On the other hand, systems that use images as intermediaries for music search tend to convey intentions indirectly rather than directly. Therefore, the meaning that is intended may vary based on the selected image, and it may differ from the individual's envisioned musical imagery. This is closely related to the nature of images, as they can be subjectively interpreted, which leads to variations in how people perceive the same image [27, 57]. Due to these factors, VTTune, which supplements users' desired imagery feelings with text-based tags, is likely to have higher music matching suitability than is VisTune.

5.4.2 User experience evaluation. Table 4 presents the UEQ-S scores for TTTune, VisTune, and VTTune. The UEQ-S scores range from -3 to +3, where -3 represents negative responses, 0 indicates neutral responses, and +3 represents highly positive responses. From the results, we can see that the overall user experience score is highest for VTTune, followed by VisTune and then TTTune. The ANOVA found no significant difference in the overall evaluation of user experience ($F(2, 233) = 2.443, p = .089$) or non-task-oriented aspects, such as the system's appeal ($F(2, 233) = .488, p = .615$). However, a significant difference was found in the pragmatic quality items that examined task-oriented aspects ($F(2, 233) = 8.031, p < .001, \eta_p^2 = .064$). These results were further analyzed using Scheffe's post-hoc analysis, finding a significant difference between the traditional text-based system TTTune and the image-based system VTTune for the pragmatic items ($p < .001$). This outcome indicates a difference in user experience in performing tasks between TTTune and VTTune. While no significant difference was found in hedonic quality unrelated to the task among the three systems, all three systems require improvement in the interface aspect. According to the feedback from participants, one male participant in his 40s (P17) who experienced VisTune stated, "It was good to be able to search for music quickly in an intuitive way, but the method of displaying images seems to need improvement." Some other participants also mentioned discomfort with the way images were displayed. It seems that further research is needed on how VisTune and VTTune can more efficiently present images.

Nevertheless, the systems showed a high user experience score overall, which is a highly encouraging result. When compared with the benchmark dataset, the overall scores of all systems exceeded the Benchmarks' average UEQ-S score, as shown in Figure 5. In particular, VTTune's overall and pragmatic quality achieved a "Good" score compared with the benchmark dataset. "Good" here means

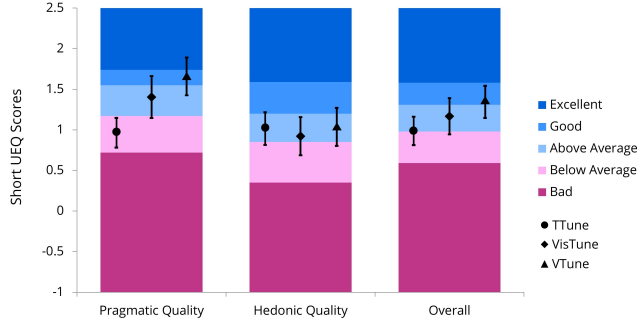


Figure 5: Comparison of UEQ-S and benchmark scores for each system. From the left, the graph presents the pragmatic quality, hedonic quality, and overall categories, with the UEQ-S scores for each system displayed with the confidence interval.

that 10% of the evaluations of the pragmatic quality of the benchmark dataset are better than VTTune's UEQ-S, and 75% are worse. These findings indicate that the method of music search with images can provide users with a positive experience, and that the best user experience score is achieved when users combine familiar text-based approach with an image-based approach.

Table 4: UEQ-S scores for each system. The UEQ-S scores are presented for the categories of pragmatic quality, hedonic quality, and overall. The IMR systems, VisTune and VTTune, received overall higher user experience scores compared to TTTune, with a particularly significant difference in the pragmatic quality.

	TTTune		VisTune		VTTune		<i>p</i>
	Score	SD	Score	SD	Score	SD	
Pragmatic Quality	0.98	0.86	1.40	1.16	1.65	1.06	.000***
Hedonic Quality	1.03	0.91	0.92	1.05	1.05	1.03	.615
Overall	1.00	0.81	1.17	1.00	1.35	0.92	.089

5.4.3 Usability evaluation of the system. The PSSUQ was measured to investigate the usability aspect of the system in more detail, and Table 5 lists the scores for each system. The closer the PSSUQ score is to 1, the more satisfactory it is. From the results, we can see that the systems that employ images as intermediaries, which are VisTune and VTTune, demonstrated higher satisfaction than did TTTune, which uses text alone as an intermediary. The ANOVA results revealed a significant difference in the overall assessment of system usability among the three systems ($F(2, 233) = 5.661$, $p < .01$, $\eta_p^2 = .046$). Moreover, PSSUQ can be divided into subsets, which include system usefulness, information quality, and interface quality. When analyzed by subset, significant differences were observed among the three systems in system usefulness ($F(2, 233) = 12.642$, $p < .001$, $\eta_p^2 = .098$), information quality ($F(2, 233) = 3.572$, $p < .05$, $\eta_p^2 = .030$), and the scores were in the order of highest satisfaction: VTTune, VisTune, and TTTune. The results of the post-hoc analysis found a significant difference between TTTune and the IMR

systems VisTune ($p < .05$) and VTTune ($p < .05$). The difference in system usefulness was noticeable, with significant differences between TTTune and VisTune ($p < .001$), and TTTune and VTTune ($p < .001$), indicating that the IMR systems provide higher satisfaction regarding system usability than TTTune.

According to participant feedback, some found TTTune cumbersome because it sometimes requires entering lengthy text when they did not know the exact information about the music, similar to findings in previous studies [41, 74]. In contrast, participants considered VisTune and VTTune more convenient, as participants needed only to choose images that matched the mood or situation of their envisioned music. Differences were also evident in the overall experiment participation time. Participants using VTTune took about 20 minutes to complete the entire process, including the survey, while those using VisTune took 22 minutes, and those using TTTune took 33 minutes. This indicates that VTTune enabled the fastest music search, and systems that employed images as intermediaries offered a more intuitive and convenient search experience than did systems relying solely on text. Compared with the average scores of the benchmark dataset, all three systems demonstrated high overall usability in all categories, as is shown in Figure 6. Among the systems, the music search method in particular used images, which implies that it could offer users a convenient music search system.

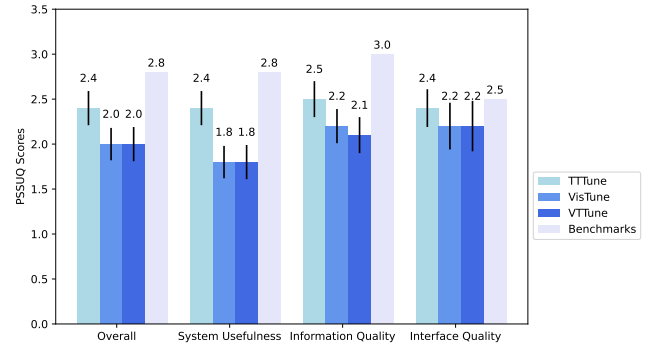


Figure 6: Comparison of PSSUQ and benchmark scores for each system. From the left, the graph displays the overall, system usefulness, information quality, and interface quality categories, with the PSSUQ scores for each system displayed with the confidence interval. The usability evaluations for TTTune, VisTune, and VTTune were more positive in all categories than the benchmark scores.

5.4.4 Comparison of results by age group. Although the systems were not designed for commercial application, but instead to explore the potential of a new music search method, all three systems received overall better evaluations on all aspects compared with the benchmark dataset's average scores, with respect to both user experience score and perceived usability. To delve deeper into the results by age group, we divided participants into two groups: those aged 19–39 years and those aged 40 years and above, and we compared their UEQ-S and PSSUQ scores. While there was no statistically significant difference, Table 6 reveals that all three systems generally

Table 5: PSSUQ scores for each system. The PSSUQ scores are presented for the categories of overall, system usefulness, information quality, and interface quality. The overall scores were more positively evaluated for the IMR systems, VisTune and VTTune, compared with the text-based TTTune, with significant differences in the overall, system usefulness, and information quality categories.

PSSUQ Classification	Item number	TTTune		VisTune		VTTune		<i>p</i>
		Score	SD	Score	SD	Score	SD	
Overall	1–16	2.41	0.85	2.02	0.79	2.02	0.88	.004**
System Usefulness	1–6	2.40	0.87	1.82	0.80	1.81	0.87	.000***
Information Quality	7–12	2.48	0.93	2.16	0.85	2.14	0.91	.030*
Interface Quality	13–15	2.37	0.97	2.15	1.16	2.22	1.29	.468

*Lower PSSUQ score indicates higher usability.

provided a more positive experience in terms of user experience score for the 40 years and above age group than for the 19–39 years age group. As shown in Table 7, the usability score for the systems differed depending on the system. In the case of TTTune, the 19–39 years age group rated higher than the 40 years and above age group. Although VTTune did not exhibit notable differences between age groups for PSSUQ scores, VisTune showed relatively higher satisfaction from the 40 years and above age group. There was also an opinion related to this in the participant feedback. One female participant in her 60s (P57) who used VisTune provided the feedback, “*I would love to continue using this system because it’s easy to express my mood with pictures.*” Another male participant aged 60 and above (P62) said, “*Searching for music is not easy, but it was convenient, especially for older people like me, because you only need to select pictures that match your current mood.*” Through these results, it can be inferred that image-based music search, such as VisTune, could be particularly useful for users aged 40 years and above.

6 USER STUDY III

To explore potential scenarios where the IMR systems we suggest could be utilized, we conducted semi-structured interviews. Subsequently, we integrated the scenarios derived from these interviews in three cases. This section provides descriptions of these scenarios.

6.1 Methodology

In order to summarize the advantages and potential usage scenarios of the IMR systems, we recruited 12 participants (5 males, 7 females) from the pool of 236 participants who had experienced VisTune or VTTune in User Study II, with 6 participants in each group, on a first-come, first-served basis. Their average age was 38.75 years, with a standard deviation of 11.64 years. Because the participants had previously experienced either VisTune or VTTune, they were guided in advance to think about situations where they might use such systems for a smooth interview. Before the interviews began, participants provided their informed consent, and the overall procedure lasted approximately 40 minutes. All interview content was recorded and transcribed with the participants’ consent.

6.2 Results

Based on the participants’ interview results, we constructed a total of three IMR usage scenarios, as depicted in Figure 7. The three

scenarios are as follows: 1) Music search for users who have difficulty writing text, 2) Searching for music that matches the scenery while the subject is driving, and 3) Music search for background music that suits visual data to be uploaded on social media. Detailed descriptions of these scenarios are provided in the following section.

6.2.1 Case 1: Music search for users who have difficulty writing text. The IMR systems designed in this study, VisTune and VTTune, can function as music search tools for users who have difficulty writing text. A female participant in her 20s (PI09) who used VisTune stated, “*I had a finger fracture not long ago and had a splint. It was really inconvenient to type with my left hand, and it slowed down my typing speed and caused many typos. Even now, I’m being careful. (...) With this system, it was convenient because I could search for music with just a few clicks.*” A female participant aged 60 or older (PI05) who used VTTune stated, “*People like me (who are older) often struggle when trying to search for music, especially when I can’t remember the song titles. I don’t even know where to search, and I’m not confident about typing things repeatedly. (With this system), I just chose an image that matches my mood, so it wasn’t difficult.*” She also mentioned that she thought these systems could also serve as entertainment devices. Considering the participants’ responses, it can be concluded that for users who have impaired eyesight or physical discomfort, which may cause difficulty in text input, VisTune and VTTune could provide support for exploratory music search.

6.2.2 Case 2: Searching for music that matches the scenery while the subject is driving. A male participant in his 40s (PI01) who used VTTune mentioned “*After trying this, I thought it could be applied to cars. Nowadays, cars have cameras, so why not capture the external scenery while I’m driving and play music that matches it?*” Just as he described, an IMR system could be utilized in conjunction with cameras that capture the surroundings. Because this system searches for music based on tag information related to the mood or the situation embedded in images, it searches for music that aligns with the surrounding environment. This advantage of an IMR system can be effectively utilized in driving environments. Furthermore, in future scenarios such as autonomous driving, where AI becomes the driver, entertainment elements that can be engaged in instead of driving will become even more significant for people. In such cases, an IMR system could serve as a service that searches for music that complements the driving environment.

Table 6: UEQ-S scores by age group for each system. Compared with the 19–39 years age group, the group aged 40 years and older generally rated all aspects of the UEQ-S more positively.

	TTTune (N = 80)	19–39 y (N = 48)	40 y and above (N = 32)	VisTune (N = 77)	19–30 y (N = 46)	40 y and above (N = 31)	VTTune (N = 79)	19–30 y (N = 47)	40 y and above (N = 32)
Pragmatic Quality	0.98	0.84 (↓)	1.17 (↑)	1.40	1.34 (↓)	1.50 (↑)	1.65	1.69 (↑)	1.59 (↓)
Hedonic Quality	1.03	0.93 (↓)	1.16 (↑)	0.92	0.89 (↓)	0.98 (↑)	1.05	0.95 (↓)	1.20 (↑)
Overall	1.00	0.89 (↓)	1.17 (↑)	1.17	1.12 (↓)	1.23 (↑)	1.35	1.32 (↓)	1.40 (↑)

Table 7: PSSUQ scores by age group for each system. VisTune showed relatively higher satisfaction among the group aged 40 years and older compared to the other two systems.

PSSUQ Classification	TTTune (N = 80)	19–39 y (N = 48)	40 y and above (N = 32)	VisTune (N = 77)	19–39 y (N = 46)	40 y and above (N = 31)	VTTune (N = 79)	19–39 y (N = 47)	40 y and above (N = 32)
Overall	2.41	2.32 (↓)	2.55 (↑)	2.02	2.04 (↑)	2.01 (↓)	2.02	2.00 (↓)	2.06 (↑)
System Usefulness	2.40	2.28 (↓)	2.57 (↑)	1.82	1.87 (↑)	1.74 (↓)	1.81	1.73 (↓)	1.93 (↑)
Information Quality	2.48	2.40 (↓)	2.59 (↑)	2.16	2.16 (–)	2.15 (↓)	2.14	2.14 (–)	2.14 (–)
Interface Quality	2.37	2.30 (↓)	2.48 (↑)	2.15	2.12 (↓)	2.20 (↑)	2.22	2.25 (↑)	2.17 (↓)

*Lower PSSUQ score indicates higher usability.

6.2.3 Case 3: Music search for background music that suits visual data to be uploaded on social media. The younger generation nowadays has become highly accustomed to using social media like Instagram, TikTok, and YouTube to share images and videos. Furthermore, there is a trend of using social media platforms for search, rather than using traditional search engines [2, 35]. In line with this generational trend, there is some potential for utilizing IMR systems that search for background music that suits data to be uploaded on social media. In relation to this, a female participant in her 30s (PI04) who used VisTune mentioned, “I work on editing YouTube videos, and about 80% of my work involves finding music to insert into videos. There’s a synergetic effect when the right music blends with the video. I have my own know-how (for finding music), but if there is something like this system that matches images and music, I think I can search for music using the representative scene in my video.” As she suggested, in the future, a system could evolve that searches for background music that is suitable for videos and images, and that uses representative scenes or thumbnails of visual data as a query.

7 DISCUSSION

In this section, we present guidelines for designing the user interface of the IMR systems and discuss how the quality and specificity of input queries affect performance. We also examine the limitations of this study and future work.

7.1 Interface Design for Image-based Music Search

Considering the feedback provided by participants and the opinions given in the interviews, it can be concluded that participants desired a variety of image selection options with a simple format in the IMR systems. Moreover, the participants preferred viewing larger images one by one rather than viewing multiple images at once. When we considered this insight, we concluded that applying a swiping mechanism for quick image navigation could prove advantageous from an interface design perspective [19, 25].

In addition, we presented image selection options to facilitate the intuitive expression of the desired music atmosphere through chosen images, which eliminated the time spent searching for images to use as input. Although input images were provided to users in a supportive way, previous research results suggest that users might be less proactive when pre-suggested inputs are available [43]. Therefore, in the future, it may be necessary to consider providing users with the option to directly upload their own desired images in addition to offering a supportive approach. Another approach to consider could involve utilizing images from the surrounding environment that are automatically captured through a user device.

In this study, participants aged 60 years and above who took part in the interviews experienced difficulties in searching for music with the use of music streaming services [67]. In general, older adults take longer to complete tasks and use fewer apps than do younger adults due to cognitive decline [26]. Taking these factors into consideration, IMR systems like VisTune could be appealing to seniors, so it is necessary to consider developing a system that is specialized for them. Additionally, efforts should be made to accommodate users who face physical difficulties with inputting text. When systems are specialized for seniors who struggle with music search and for individuals who face physical difficulties with inputting text, the following considerations should be considered: enlarge image sizes to ensure users can perceive images without difficulty and increase the margin between images to reduce click errors [5, 16, 22]. Moreover, the simplification and intuitive design of the interface should be considered, as previous studies have shown that conveying too much information on a single page is not effective for older adults [52, 58]. However, as preferences for interface design may differ among cultural groups [30], further in-depth research on the design aspect will be necessary to develop systems that are specialized for older adults in the future.

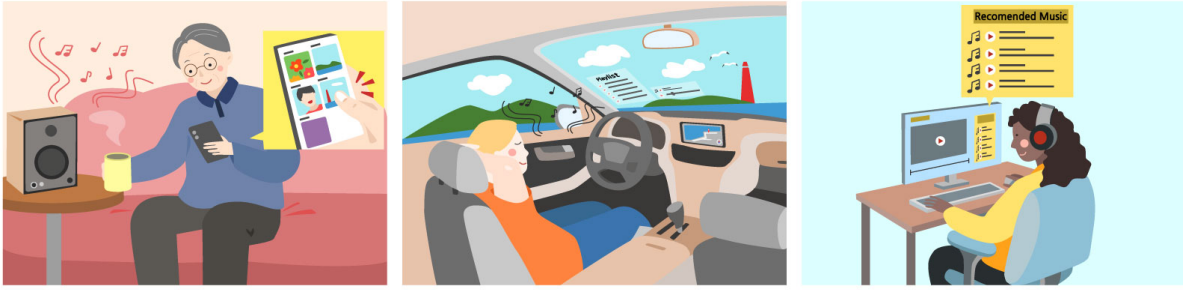


Figure 7: Three potential usage scenarios for IMR systems. The first scenario demonstrates the IMR system functioning as a specialized system for users having difficulty with text input. The second scenario portrays the IMR system as a music recommendation feature that matches music with the outside scenery while a car is driven; this includes autonomous driving. The third scenario presents an IMR system automatically searching for music that matches videos or images uploaded to social media.

7.2 Factors Influencing the Performance and User Satisfaction with Image-based Music Search

The satisfaction with music search results can vary depending on the input query quality. Just as search results can be more closely aligned with users' desired music when the meaning of the text used as a query is accurately understood [10], search performance can be enhanced when the meaning extracted from an image is universally accepted. Unlike text, images can be interpreted in various and subjective ways by different people [27, 57]. Therefore, it may be helpful to use an image as a query in which the meaning can be extracted within an expected range by the public to increase the accuracy of the music search. Thus, providing users with previously prepared image examples can help them quickly envision the music they want and positively affect the management of the search result quality. However, to embrace users' subjective interpretations and diversity, a hybrid method, such as VTTune, which uses the image as a query but supplements it with text, could be effective. This method includes a complementary process in which the system ensures that the meaning it extracts from the entered query matches what the user wants. Considering that user experience and system usability satisfaction for VTTune scored the highest in this study, a method similar to VTTune is a rational approach for employing IMR systems. Thus, the performance and satisfaction of image-based music search can vary depending on how users specifically visualize music and use queries that clearly express it. This principle is applicable to IMR and other music search systems [9].

7.3 Limitations and Future Work

Our study has several limitations. First, there is a need to enhance and personalize the image tagging method. It is necessary not only to improve the extraction of implied meanings within images but also to develop methods that are capable of extracting personalized tag information. Second, the interface functionality and the optimization of the three systems used in this study need improvement.

When the app prototypes were designed for the experiments, we employed open-source libraries. Although these libraries provide various user interface components, there were limitations in the customizability of certain elements. Therefore, when we implemented the interfaces of the three systems that we created for this study, there were challenges in customizing features such as the arrangement of images or music lists, and in fine-tuning font sizes. Third, in this study, pre-designed scenarios were provided to assist users in visualizing their target music. Although these scenarios were based on previous studies and statistics related to why and when people listen to music, they do not encompass all possible cases [47, 60]. Thus, it is difficult to analyze in which specific scenario the IMR systems will be effective, which is a limitation. It is necessary to expand the experiment in the future based on more diverse scenarios.

For future work, we plan to focus more on specific age groups or scenarios to examine the effectiveness of the IMR systems and intend to conduct comparisons involving a wider range of retrieval methods and various measures. Furthermore, designing music listening experiences with the innovative approach of directly generating the music users want to hear using the emerging generative models would also be enlightening [15, 46, 49, 69].

8 CONCLUSION

This study investigated the user experience and system usability of a novel music search method using images as intermediaries, with a particular emphasis on situations where there is no specific music to search for. We designed three music search prototypes—TTTune (text-based), VisTune (image-based), and VTTune (hybrid)—and we conducted user studies that employed both quantitative and qualitative approaches. The research findings indicate that 1) users demonstrated positive user experiences and high usability ratings for IMR systems, 2) the user experience with the music search system mediated solely by images was high in the over-40 participant group, and 3) hybrid search methods that use images as queries and

that are assisted by text were preferred over text-only or image-only methods. Based on these results, we saw the potential of a new music search method that uses images. We collaborated with participants to ideate scenarios for the application of IMR methods, and we provided insights into future directions for interface development. We hope this study contributes fresh ideas to music search systems and serves as an opportunity to explore the broader potential of visual modalities, like images.

ACKNOWLEDGMENTS

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2023R1A2C200337911 and No. RS-2023-00220762).

REFERENCES

- [1] 2018. Spotify — about. <https://newsroom.spotify.com/company-info/>. Accessed: 2023-8-15.
- [2] 2022. Brainstorm tech 2022: Organizing the world's information. <https://fortune.com/videos/watch/Brainstorm-Tech-2022-Organizing-The-Worlds-Information/934585a6-7fb6-41a5-8ef3-e497f8ca2986>. Accessed: 2023-8-21.
- [3] 2023. NLTK: Sample usage for wordnet. <https://www.nltk.org/howto/wordnet.html>. Accessed: 2023-8-22.
- [4] 2023. Streamlit. <https://streamlit.io/>. Accessed: 2023-8-22.
- [5] Muna S Al-Razgan, Hend S Al-Khalifa, Mona D Al-Shahrani, and Hessah H AlAjmi. 2012. Touch-based mobile phone interface guidelines and design recommendations for elderly people: A survey of the literature. In *Neural Information Processing: 19th International Conference, ICONIP 2012, Doha, Qatar, November 12-15, 2012, Proceedings, Part IV* 19. Springer, 568–574.
- [6] Philip Ball. 2010. *The music instinct: how music works and why we can't do without it*. Random House.
- [7] Adam Berenzweig, Beth Logan, Daniel PW Ellis, and Brian Whitman. 2004. A large-scale evaluation of acoustic and subjective music-similarity measures. *Computer Music Journal* (2004), 63–76.
- [8] Daniel Bernhardt and Peter Robinson. 2008. Interactive control of music using emotional body expressions. In *CHI'08 extended abstracts on Human factors in computing systems*. 3117–3122.
- [9] Michele Buccoli, Massimiliano Zanon, Augusto Sarti, and Stefano Tubaro. 2013. A music search engine based on semantic text-based query. In *2013 IEEE 15th International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 254–259.
- [10] Donald Byrd and Tim Crawford. 2002. Problems of music information retrieval in the real world. *Information processing & management* 38, 2 (2002), 249–272.
- [11] Wanling Cai, Yucheng Jin, Xianglin Zhao, and Li Chen. 2023. "Listen to Music, Listen to Yourself": Design of a Conversational Agent to Support Self-Awareness While Listening to Music. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–19.
- [12] Michael A Casey, Remco Veltkamp, Masataka Goto, Marc Leman, Christophe Rhodes, and Malcolm Slaney. 2008. Content-based music information retrieval: Current directions and future challenges. *Proc. IEEE* 96, 4 (2008), 668–696.
- [13] Ahyeon Choi, Eunsik Shin, Haesun Joong, Joongseek Lee, and Kyogu Lee. 2023. Towards a New Interface for Music Listening: A User Experience Study on YouTube. *arXiv preprint arXiv:2307.14718* (2023).
- [14] Brian Clark. 2022. How Many Songs are There in the World? (2023). <https://www.musicianwave.com/how-many-songs-are-there-in-the-world/>. Accessed: 2023-8-15.
- [15] Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. 2023. Simple and Controllable Music Generation. *arXiv preprint arXiv:2306.05284* (2023).
- [16] Ana Correia De Barros, Roxanne Leitão, and Jorge Ribeiro. 2014. Design and evaluation of a mobile user interface for older adults: navigation, interaction and visual design recommendations. *Procedia Computer Science* 27 (2014), 369–378.
- [17] James J Deng and Clement HC Leung. 2013. Music retrieval in joint emotion space using audio features and emotional tags. In *Advances in Multimedia Modeling: 19th International Conference, MMM 2013, Huangshan, China, January 7-9, 2013, Proceedings, Part I* 19. Springer, 524–534.
- [18] Wei Di, Neel Sundareshan, Robinson Pirmuthu, and Anurag Bhardwaj. 2014. Is a picture really worth a thousand words? - on the role of images in e-commerce. In *Proceedings of the 7th ACM international conference on Web search and data mining*. 633–642.
- [19] Morgan Dixon, Gierad Laput, and James Fogarty. 2014. Pixel-based methods for widget state and style in a runtime implementation of sliding widgets. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2231–2240.
- [20] SeungHeon Doh, Minz Won, Keunwoo Choi, and Juhan Nam. 2023. Toward Universal Text-to-Music Retrieval. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.
- [21] Yazhong Feng, Yueting Zhuang, and Yunhe Pan. 2003. Music information retrieval by detecting mood via computational media aesthetics. In *Proceedings IEEE/WIC international conference on web intelligence (WI 2003)*. IEEE, 235–241.
- [22] Arthur D Fisk, Sara J Czaja, Wendy A Rogers, Neil Charness, and Joseph Sharit. 2020. *Designing for older adults: Principles and creative human factors approaches*. CRC press.
- [23] Cynthia Fraser. 2014. Music-evoked images: Music that inspires them and their influences on brand and message recall in the short and the longer term. *Psychology & Marketing* 31, 10 (2014), 813–827.
- [24] Sophie Freeman, Martin Gibbs, and Bjorn Nansen. 2023. Personalised But Impersonal: Listeners' Experiences of Algorithmic Curation on Music Streaming Services. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–14.
- [25] Luke K Fryer and Kaori Nakao. 2020. The Future of Survey Self-Report: An Experiment Contrasting Likert, VAS, Slide, and Swipe Touch Interfaces. *Frontline Learning Research* 8, 3 (2020), 10–25.
- [26] Mitchell L Gordon, Leon Gatys, Carlos Guestrin, Jeffrey P Bigham, Andrew Trister, and Kayur Patel. 2019. App usage predicts cognitive ability in older adults. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.
- [27] Howard Greisdorf and Brian O'Connor. 2002. Modelling what users see when they look at images: a cognitive viewpoint. *Journal of documentation* 58, 1 (2002), 6–29.
- [28] Maarten Grootendorst. 2020. KeyBERT: Minimal keyword extraction with BERTKeyBERT: Minimal keyword extraction with BERT.
- [29] Zhiyuan Guo, Qiang Wang, Gang Liu, Jun Guo, and Yueming Lu. 2012. A music retrieval system using melody and lyric. In *2012 IEEE International Conference on Multimedia and Expo Workshops*. IEEE, 343–348.
- [30] Shathel Haddad, Joanna McGrenere, and Claudia Jacova. 2014. Interface design for older adults with varying cultural attitudes toward uncertainty. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1913–1922.
- [31] Masahiro Hamasaki, Masataka Goto, and Tomoyasu Nakano. 2014. Songrium: a music browsing assistance service with interactive visualization and exploration of protect a web of music. In *Proceedings of the 23rd International Conference on World Wide Web*. 523–528.
- [32] Hasmina Hassan, Zunairah Haji Murat, Valerie Ross, and Norlida Buniyamin. 2012. A preliminary study on the effects of music on human brainwaves. In *2012 International Conference on Control, Automation and Information Sciences (ICCAIS)*. IEEE, 176–180.
- [33] Simon Holland, Katie Wilkie, Paul Mulholland, and Allan Seago. 2013. Music interaction: understanding music and human-computer interaction. In *Music and human-computer interaction*. Springer, 1–28.
- [34] Christine Hosey, Lara Vujović, Brian St. Thomas, Jean Garcia-Gathright, and Jennifer Thom. 2019. Just give me what I want: How people use and evaluate music search. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–12.
- [35] Kalley Huang. 2022. For Gen Z, TikTok is the new search engine. *The New York Times* (2022).
- [36] Mohammad Khairul Islam, Hyung-Jin Lee, Anjan Kumar Paul, and Joong-Hwan Baek. 2007. Content-based music retrieval using beat information. In *Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007)*, Vol. 3. IEEE, 317–321.
- [37] James J Kellaris, Anthony D Cox, and Dena Cox. 1993. The effect of background music on ad processing: A contingency explanation. *Journal of Marketing* 57, 4 (1993), 114–125.
- [38] Peter Knees and Markus Schedl. 2015. Music retrieval and recommendation: A tutorial overview. In *Proceedings of the 38th International ACM SIGIR conference on research and development in information retrieval*. 1133–1136.
- [39] Peter Knees, Markus Schedl, Tim Pohle, and Gerhard Widmer. 2007. Exploring music collections in virtual landscapes. *IEEE multimedia* 14, 3 (2007), 46–54.
- [40] Bettina Laugwitz, Theo Held, and Martin Schrepp. 2008. Construction and evaluation of a user experience questionnaire. In *HCI and Usability for Education and Work: 4th Symposium of the Workgroup Human-Computer Interaction and Usability Engineering of the Austrian Computer Society, USAB 2008, Graz, Austria, November 20-21, 2008. Proceedings* 4. Springer, 63–76.
- [41] Jin Ha Lee. 2010. Analysis of user needs and information features in natural language queries seeking music information. *Journal of the American Society for Information Science and Technology* 61, 5 (2010), 1025–1045.
- [42] Jin Ha Lee and Rachel Price. 2016. User experience with commercial music services: An empirical exploration. *Journal of the Association for Information Science and Technology* 67, 4 (2016), 800–811.
- [43] Ariel Levy, Monica Agrawal, Arvind Satyanarayan, and David Sontag. 2021. Assessing the impact of automated suggestions on decision making: Domain

- experts mediate model errors but take less initiative. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.
- [44] James R Lewis. 1992. Psychometric evaluation of the post-study system usability questionnaire: The PSSUQ. In *Proceedings of the human factors society annual meeting*, Vol. 36. Sage Publications Sage CA: Los Angeles, CA, 1259–1260.
- [45] James R Lewis. 1995. IBM computer usability satisfaction questionnaires: psychometric evaluation and instructions for use. *International Journal of Human-Computer Interaction* 7, 1 (1995), 57–78.
- [46] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. 2023. Audioldm: Text-to-audio generation with latent diffusion models. *arXiv preprint arXiv:2301.12503* (2023).
- [47] Adam J Lonsdale and Adrian C North. 2011. Why do we listen to music? A uses and gratifications analysis. *British journal of psychology* 102, 1 (2011), 108–134.
- [48] Lie Lu, Hong You, HongJiang Zhang, et al. 2001. A Newapproach To Query By Humming In Music Retrieval.. In *ICME*. 22–25.
- [49] Rishi Madhok, Shivali Goel, and Shweta Garg. 2018. SentiMozart: Music Generation based on Emotions. In *ICAART* (2). 501–506.
- [50] Matija Marolt. 2008. A mid-level representation for melody-based retrieval in audio collections. *IEEE Transactions on Multimedia* 10, 8 (2008), 1617–1625.
- [51] Alessandro B Melchiorre, David Penz, Christian Ganhör, Oleg Lesota, Vasco Frago, Florian Fritzl, Emilia Parada-Cabaleiro, Franz Schubert, and Markus Schedl. 2022. EmoMTB: Emotion-aware music tower blocks. In *Proceedings of the 2022 International Conference on Multimedia Retrieval*. 206–210.
- [52] Eun-Shim Nahm, Jennifer Preece, Barbara Resnick, and Mary Etta Mills. 2004. Usability of health Web sites for older adults: a preliminary study. *CIN: Computers, Informatics, Nursing* 22, 6 (2004), 326–334.
- [53] Alexandros Nanopoulos, Dimitrios Rafailidis, Maria M Ruxanda, and Yannis Manolopoulos. 2009. Music search engines: Specifications and challenges. *Information Processing & Management* 45, 3 (2009), 392–396.
- [54] Eli Pariser. 2011. *The filter bubble: What the Internet is hiding from you*. penguin UK.
- [55] Jeongeun Park, Minchae Kim, and Ha Young Kim. 2023. Image Is All for Music Retrieval: Interactive Music Retrieval System Using Images with Mood and Theme Attributes. *International Journal of Human-Computer Interaction* (2023), 1–15.
- [56] Michael J Pazzani and Daniel Billsus. 2007. Content-based recommendation systems. In *The adaptive web: methods and strategies of web personalization*. Springer, 325–341.
- [57] Lev Poretsky, Joel Lanir, and Ofer Arazy. 2019. Feel the image: The role of emotions in the image-seeking process. *Human-Computer Interaction* 34, 3 (2019), 240–277.
- [58] Andrew E Reed, Joseph A Mikels, and Kosali I Simon. 2008. Older adults prefer less choice than young adults. *Psychology and aging* 23, 3 (2008), 671.
- [59] Seungmin Rho, Byeong-jun Han, Eenjun Hwang, and Minkoo Kim. 2008. MUSEM-BLE: A novel music retrieval system with automatic voice query transcription and reformulation. *Journal of Systems and Software* 81, 7 (2008), 1065–1080.
- [60] Felix Richter. 2019. Any Time, Any Place: How People Listen to Music.
- [61] Justin Salamon, Emilia Gómez, Daniel PW Ellis, and Gaël Richard. 2014. Melody extraction from polyphonic music signals: Approaches, applications, and challenges. *IEEE Signal Processing Magazine* 31, 2 (2014), 118–134.
- [62] Justin Salamon, Joan Serra, and Emilia Gómez. 2013. Tonal representations for music retrieval: from version identification to query-by-humming. *International Journal of Multimedia Information Retrieval* 2, 1 (2013), 45–58.
- [63] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. 2001. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*. 285–295.
- [64] Jeff Sauro and James R Lewis. 2016. *Quantifying the user experience: Practical statistics for user research*. Morgan Kaufmann.
- [65] Martin Schrepp, Jörg Thomaschewski, and Andreas Hinderks. 2017. Construction of a benchmark for the user experience questionnaire (UEQ). *International Journal of Interactive Multimedia and Artificial Intelligence* 4 (2017).
- [66] Philip Sargeant. 2019. *The Emoji Revolution: How technology is shaping the future of communication*. Cambridge University Press.
- [67] Neil Selwyn. 2004. The information aged: A qualitative study of older adults’ use of information and communications technology. *Journal of Aging studies* 18, 4 (2004), 369–384.
- [68] Lanyu Shang, Daniel Zhang, Jialie Shen, Eamon Lopez Marmion, and Dong Wang. 2021. CCMR: A classic-enriched connotation-aware music retrieval system on social media with visual inputs. *Social Network Analysis and Mining* 11 (2021), 1–14.
- [69] Xiaodong Tan, Mathis Antony, and H Kong. 2020. Automated Music Generation for Visual Art through Emotion.. In *ICCC*. 247–250.
- [70] Ying Tang and Khe Foon Hew. 2019. Emoticon, emoji, and sticker use in computer-mediated communication: A review of theories and research findings. *International Journal of Communication* 13 (2019), 27.
- [71] Hans-Joachim Trappe. 2012. The effect of music on human physiology and pathophysiology. *Music and medicine* 4, 2 (2012), 100–105.
- [72] Thomas Tullis and William Albert. 2010. *Measuring the user experience: Collecting, analyzing, and presenting usability metrics*. Morgan Kaufmann.
- [73] Rainer Typke, Frans Wiering, Remco C Veltkamp, Joshua D Reiss, Geraint A Wiggins, et al. 2005. A survey of music information retrieval systems. In *Proc. 6th international conference on music information retrieval*. Queen Mary, University of London, 153–160.
- [74] Julián Urbano, Markus Schedl, and Xavier Serra. 2013. Evaluation in music information retrieval. *Journal of Intelligent Information Systems* 41, 3 (2013), 345–369.
- [75] Nils L Wallin, Bjorn Merker, and Steven Brown. 2001. *The origins of music*.
- [76] Ju-Chiang Wang, Meng-Sung Wu, Hsin-Min Wang, and Shyh-Kang Jeng. 2011. Query by multi-tags with multi-level preferences for content-based music retrieval. In *2011 IEEE International Conference on Multimedia and Expo*. IEEE, 1–6.
- [77] Minz Won, Sergio Oramas, Oriol Nieto, Fabien Gouyon, and Xavier Serra. 2021. Multimodal metric learning for tag-based music retrieval. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 591–595.
- [78] Baixi Xing, Kejun Zhang, Shouqian Sun, Lekai Zhang, Zenggui Gao, Jiaxi Wang, and Shi Chen. 2015. Emotion-driven Chinese folk music-image retrieval based on DE-SVM. *Neurocomputing* 148 (2015), 619–627.
- [79] Baixi Xing, Kejun Zhang, Lekai Zhang, Xinda Wu, Jian Dou, and Shouqian Sun. 2019. Image-music synesthesia-aware learning based on emotional similarity recognition. *IEEE Access* 7 (2019), 136378–136390.
- [80] Robert Zinko, Paul Stolk, Zhan Furner, and Brad Almond. 2020. A picture is worth a thousand words: How images influence information quality and information load in online reviews. *Electronic Markets* 30 (2020), 775–789.

A APPENDIX

A.1 System Design

Figures 8 to 10 each represent the main interface design screens of the TTTune, VisTune, and VTTune systems, respectively, as presented to actual users. Figure 11 shows the screen for displaying searched music, a feature that all three systems commonly employ. Figures 12 and 13 illustrate the survey screens that appear in the final stage of the experiment.

Streamlit

group-a.streamlit.app/?path=%2Fmount%2Fsrc%2Fgroupa_keyword%2Fresults...

시크릿 모드 업데이트 완료

Let's find music!

✔ STEP 1: Please select your preferred scenario from the three provided options.

Please select a scenario that you preferred the most.

☒ Feeling tired but unable to sleep.

☐ During exercise (yoga or fitness, etc.).

☐ Preparing for a party.

Example: a quiet, peaceful song good for sleeping.

✔ STEP 2: Enter a short descriptive text that suits the chosen scenario to search for music. We will find music that matches the text you have typed.

✔ STEP 3: After entering the text, click the button below and wait for a while until the next process.

⚠ The system may take a little time to initialize for the first run. From the second time onwards, it will be instant, so don't worry!

Type the text below (max 500 words)

peaceful classical music suitable for background sound

🌟 Get me the keywords!

	Keyword	Importance
1	classical	55.7%
2	peaceful	43.5%
3	sound	28.6%
4	background	25.5%
5	suitable	12.6%

✔ You can add keywords for more accurate music search!

Select keywords

calm × | × ▾

NEXT

Figure 8: Interface design screen of TTTune. Based on predefined scenarios, users visualize the music they want and then express the music they want to search for in text.

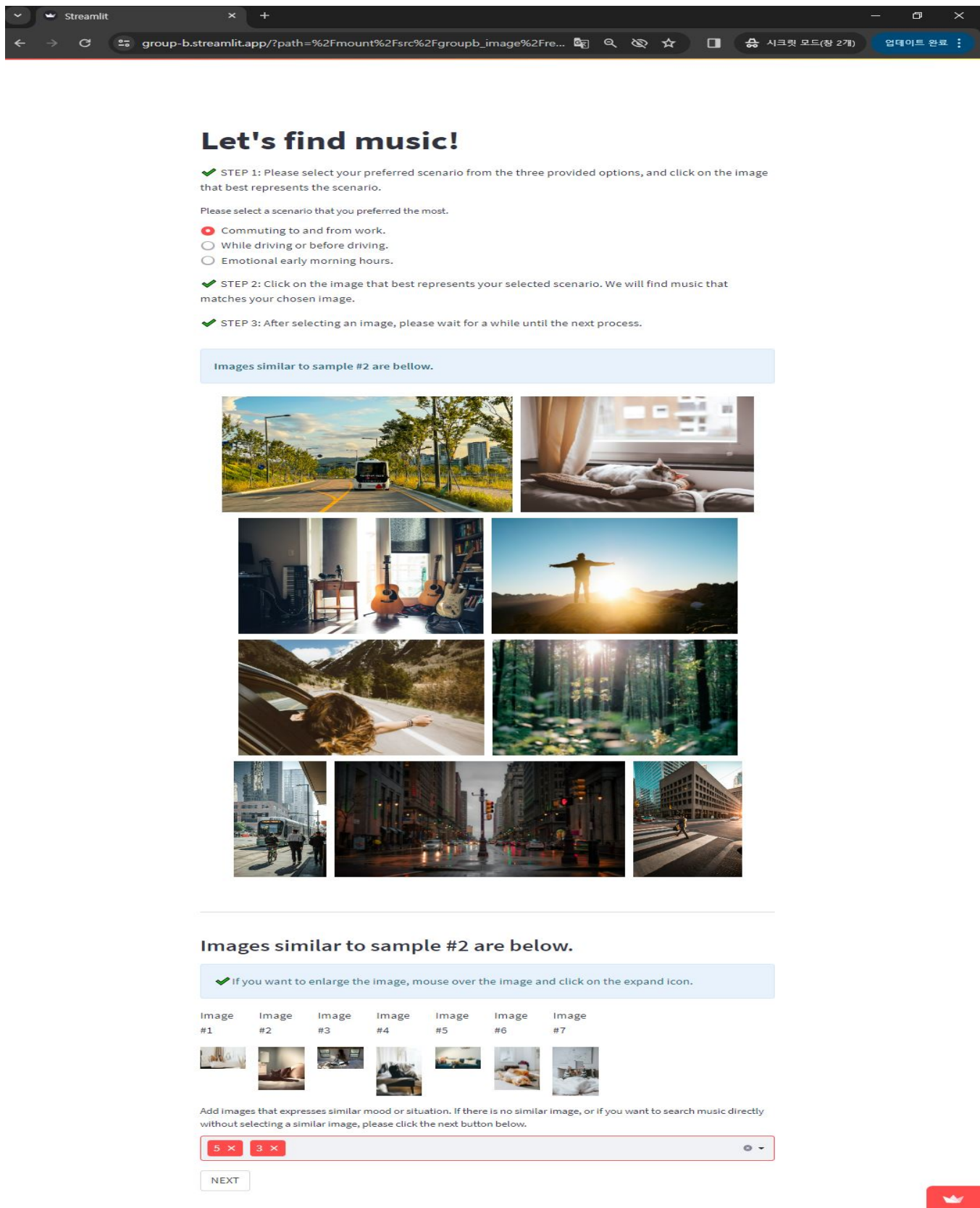


Figure 9: Interface design screen of VisTune. Based on predefined scenarios, users visualize the music they want and then express the music they want to search using only images.

Streamlit

group-c.streamlit.app/?path=%2Fmount%2Fsrc%2Fgroupc_hybrid%2Fre...

시크릿 모드(참 3개)

업데이트 완료

Let's find music!

✓ STEP 1: Please select your preferred scenario from the three provided options, and click on the image that best represents the scenario.












Please select a scenario that you preferred the most.

☒ Feeling tired but unable to sleep.
☐ During exercise (yoga or fitness, etc.).
☐ Preparing for a party.

✓ STEP 2: Click on the image that best represents your selected scenario. We will find music that matches your chosen image.

✓ STEP 3: After selecting an image, please wait for a while until the next process.

Keywords of Image #7 Are Below.

1st keyword

2nd keyword

3rd keyword

calm

relaxing

-

✓ You can add keywords for more accurate music search!

Select keywords

dream × epic × meditative ×

NEXT

Figure 10: Interface design screen of VTTune. Based on predefined scenarios, users visualize the music they want and then express the music they want to search using both images and text.

Streamlit



group-c.streamlit.app/?path=%2Fmount%2Fsrc%2Fgroupc_hybrid%2Fres...

시크릿 모드(창 3개)업데이트 완료

Music Finder



Now, we find music lists that match the image and keywords!

- The music searched in this study is a copyright-free sound sources provided for research purposes.
- Therefore, we inform you that it may be different from the latest music you are familiar with.



 Please enjoy the music and answer the questions below. 

- Listen to music for at least 30 seconds and answer the question (slide bar) below.



▶ 0:00 / 3:05



▶ 0:00 / 2:34



▶ 0:00 / 4:47

▶ 0:00 / 0:50

▶ 0:00 / 1:38

Overall, do you think the retrieved music matches the selected images and added keywords well?

Strongly disagree

Neither agree nor disagree

Strongly agree

- Note: Please evaluate how well the selected image and added keywords represent the music, rather than providing a 'like' or 'dislike' rating for the provided music.

NEXT

Figure 11: Music provision interface. Users can listen to up to five pieces of searched music. After listening to the provided music, they respond to how well the searched music matches the music they visualized.

Overall, I am satisfied with how easy it is to use this system.

☐ Strongly Agree
☐ Agree
☐ Somewhat Agree
☒ Neither Agree Nor Disagree
☐ Somewhat Disagree
☐ Disagree
☐ Strongly Disagree

It was simple to use this system.

☐ Strongly Agree
☐ Agree
☒ Somewhat Agree
☐ Neither Agree Nor Disagree
☐ Somewhat Disagree
☐ Disagree
☐ Strongly Disagree

I was able to complete the tasks and scenarios quickly using this system.

☐ Strongly Agree
☐ Agree
☒ Somewhat Agree
☐ Neither Agree Nor Disagree
☐ Somewhat Disagree
☐ Disagree
☐ Strongly Disagree

I felt comfortable using this system.

☐ Strongly Agree
☒ Agree
☐ Somewhat Agree
☐ Neither Agree Nor Disagree
☐ Somewhat Disagree
☐ Disagree
☐ Strongly Disagree

It was easy to learn to use this system.

☒ Strongly Agree
☐ Agree
☐ Somewhat Agree
☐ Neither Agree Nor Disagree
☐ Somewhat Disagree
☐ Disagree
☐ Strongly Disagree

I believe I could become productive quickly using this system.

☐ Strongly Agree
☐ Agree
☐ Somewhat Agree
☒ Neither Agree Nor Disagree
☐ Somewhat Disagree
☐ Disagree
☐ Strongly Disagree

The system gave error messages that clearly told me how to fix problems.

☐ Strongly Agree
☒ Agree
☐ Somewhat Agree
☐ Neither Agree Nor Disagree
☐ Somewhat Disagree
☐ Disagree
☐ Strongly Disagree

Figure 12: Survey Screen. Participants sequentially respond to demographic questions, PSSUQ items, and UEQ items. This screenshot illustrates a part of the PSSUQ items.

☐ Neither Agree Nor Disagree
☐ Somewhat Disagree
☐ Disagree
☐ Strongly Disagree

If this system becomes commercially available, in what situations do you think you will use it? (Example: When I want to listen to new music, but it is difficult to express my search terms in text.)

Please kindly provide a response to the question.

What aspects of the system you have used would you like to see improved?

Please kindly provide a response to the question.

Please read the question and choices carefully before providing your answer.

Was the system obstructive or supportive?

Obstructive: 0 Supportive: 7

Was the system complicated or easy?

Complicated: 0 Easy: 7

Was the system inefficient or efficient?

Inefficient: 0 Efficient: 7

Was the system confusing or clear?

Confusing: 0 Clear: 7

Was the system boring or exciting?

Boring: 0 Exciting: 7

Was the system not interesting or interesting?

Not interesting: 0 Interesting: 7

Was the system conventional or inventive?

Conventional: 0 Inventive: 7

Was the system usual or leading edge?

Usual: 0 Leading edge: 7

Here is your ID: 9243ea8a-75f9-49f4-bc6f-b5e792b735ca

Copy this value to paste into MTurk.

When you have copied this ID, please click the check box below to submit your survey

☒ Do you want to move to the next page?

END

Figure 13: Survey Screen. Participants sequentially respond to demographic questions, PSSUQ items, and UEQ items. This screenshot displays the UEQ items and a part of some open-ended response items.