MULTIMODAL STRUCTURED EXTRACTION FOR SELF-QUERYING MUSIC VIDEO RETRIEVAL AND PLAYLIST GENERATION

Kevin Dela Rosa
Aviary Labs
kdr@aviaryhq.com

ABSTRACT

In this study we introduce early results for a music video structured extraction framework designed to extract key metadata and descriptions such as genre, mood, video style, and summaries of general music, lyrical and visual narrative content. Leveraging video language models (VLM) and zero-shot prompting techniques, the system supports three key applications: entity discovery and browsing, multimodal self-querying retrieval, and playlist generation. The multimodal self-querying retrieval setup intelligently combines structured metadata filtering (e.g., video style, musical genre, emotion, visual elements) with lexical and semantic search, allowing users to query music videos using multiple facets. Additionally, the structured extraction powers entity discovery, enabling exploration of videos based on extracted metadata across the dataset. We provide qualitative examples of structured information extraction over an initial dataset of over 60K music videos to showcase the potential for search and video playlist generation.

1. MOTIVATION & METHOD

Structured information extraction using large language models has been a fairly common use case, but limited works have expanded this into multimodal domains. In the video domain, works like [1] and [2] have each illustrated frameworks for extracting structured information from large video corpora, and further [3] has illustrated their utility in specific content domains. In the music space, works like Song Describer dataset [4] are being used to drive innovations in the music captioning space.

In this work we explore using a video language model, Gemini 1.5 [5], in the music video domain. Specifically we use a VLM to extract a set of categorical and multi-value fields and targeted captions from music videos, by using a zero shot prompt specifying what fields and types of values to return on either the video or extracted audio content as input. Exact fields and example extracted values can

Field	**Category / Single Value**	**Mode**
Style	Narrative-Based	V
Substyle	Love Story	V
Tempo	Moderate	A
Vocalists	1	A, V
Decade	2010s	V
Danceability	Somewhat	A, V
Language	English	L
Field	**List / Multi-value**	**Mode**
Colors	Warm tones, Pink, Blue, Green	V
Genre	Pop, R&B	A
Emotion	Hopeful, Romantic, Joyful	A, L, V
Setting	Beach, Bedroom, Living room, Garden, Park	V
Things	Red flag, Waves, Pink dress, Flowers, Ring, Wine glasses, Trees, Car	V
Actions	Walking, Praying, Contemplating, Hugging, Talking, Smiling	V
Instruments	Acoustic Guitar, Drums, Bass, Vocals, Keys, Piano, Electric Guitar, Synth, Strings, Pop	A
Caption	**Summary / Description Value**	**Mode**
Music Caption	A powerful, emotional song with a driving beat and soaring vocals, expressing a deep love and commitment to a partner.	A, L
Visual Narrative	A woman is shown contemplating a proposal from her boyfriend. The video alternates between scenes of her reflecting on their relationship and scenes of her praying for guidance	V
Lyrics Summary	The lyrics express the woman's feelings of love and uncertainty as she considers a proposal. She seeks guidance from God before making a decision	L

Table 1. Example music video extraction and their associated modalities (A: Audio / Music Signal, L: Lyrics / Spoken Content, V: Visual / Video) from a single video in the dataset. Categorical and multi value fields can be useful in filtering subsets of the data, whereas the free formed targeted caption values can be helpful in summarization and retrieval use cases.

Late-Breaking / Demo Session Extended Abstract, ISMIR 2024 Conference

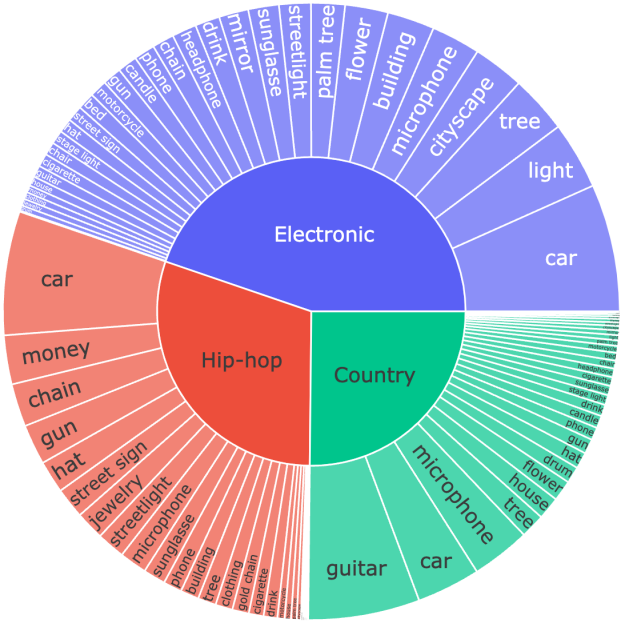

Figure 1. Example common visual props per music genre

be seen in Table 1. Music videos are particularly interesting as they are rich in acoustic / musical information, have vocally performed lyrics as opposed to speech, and often have narratives told primarily in the visual modality (e.g. actions and visual props) as opposed to spoken narration. For this study we curated a dataset of 61,865 music videos (176 days of total footage, 03:41 median duration), and performed structured information extraction.

Figure 2. Most common emotions / tones per video style

2. APPLICATIONS & EARLY RESULTS

The structured extraction system described in Section 1 supports key applications in music video analysis, enabling both content exploration and retrieval.

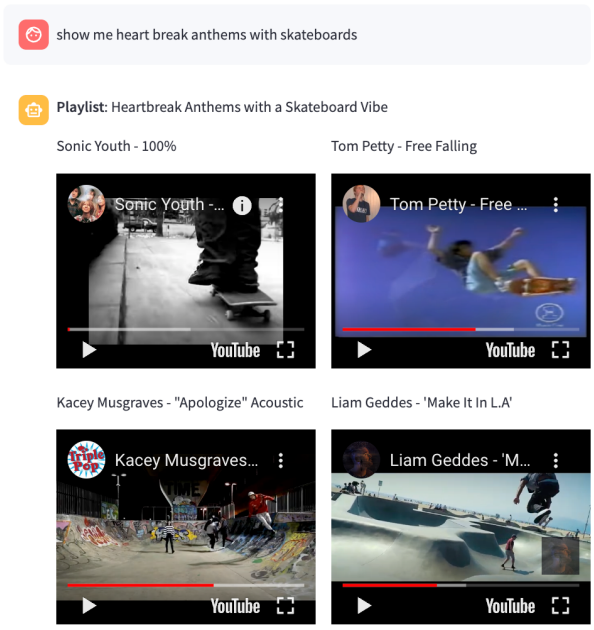

Figure 3. Generated music video playlist

2.1 Entity Discovery and Summarization

Figure 1 highlights common visual props across three genres (Country, Electronic, Hip-Hop), showcasing the system's ability to summarize key elements. These summaries help users explore large music video collections by focusing on genre-specific patterns. Similarly, Figure 2 shows a breakdown of emotions per video style, supporting content discovery based on emotional themes.

2.2 Multimodal Retrieval and Playlist Generation

The system enables advanced multimodal retrieval by combining structured metadata with semantic search. It uses a self-querying retrieval [7] mechanism that processes both natural language queries and metadata fields (e.g., video style, genre, and visual elements). For example, a query like "heart break anthems with skateboards" is split into two parts: "heart break anthems," which is compared for semantic similarity against lyric summaries, and "skateboards," which is matched to visual metadata.

This approach allows for flexible, cross-modality searches that combine text, audio, and visual information. Figure 3 illustrates a playlist generated from such a query, demonstrating how the system retrieves videos that meet both lyrical and visual criteria. The integration of structured metadata filtering with semantic search provides a powerful method for retrieval and playlist generation.

2.3 Future Work

Future work will focus on implementing more extensive filtering and quality measures prior to publishing the dataset, and developing quantitative benchmarks for playlist generation and other potential downstream tasks.

3. REFERENCES

[1] K. D. Rosa, "Automated multimodal entity discovery for large-scale video collection understanding," 2024, under review / in submission.

[2] M. Farré, A. Marafioti, L. Tunstall, L. Von Werra, and T. Wolf, "Finevideo," https://huggingface.co/datasets/HuggingFaceFV/finevideo, 2024.

[3] K. D. Rosa, "Structured entity extraction from travel videos using vision-language models," in *Workshop on Recommenders in Tourism (RecTour 2024), co-located with the 18th ACM Conference on Recommender Systems*, Bari, Italy, 2024.

[4] I. Manco, B. Weck, S. Doh, M. Won, Y. Zhang, D. Bogdanov, Y. Wu, K. Chen, P. Tovstogan, E. Benetos, E. Quinton, G. Fazekas, and J. Nam, "The song describer dataset: a corpus of audio captions for music-and-language evaluation," 2023. [Online]. Available: https://arxiv.org/abs/2311.10057

[5] DeepMind, "Gemini flash: Lightweight model with long context understanding," https://deepmind.google/technologies/gemini/flash, 2024, accessed: 2024-09-10.

[6] OpenAI, "New embedding models and api updates," https://openai.com/index/new-embedding-models-and-api-updates/, 2024, accessed: 2024-09-27.

[7] LangChain, "Self query: How to use," https://python.langchain.com/docs/how_to/self_query/, 2024, accessed: 2024-09-27.