Music to Video Matching Based on Beats and Tempo

Aleksandr Mikheev
Avito
Moscow, Russia
miheevalge@gmail.com

Ilya Makarov
AIRI, ISP RAS, Moscow, Russia
Research Center of the Artificial Intelligence Institute
Innopolis University, Innopolis, Russia
il.makarov@innopolis.ru

Abstract

The growing popularity of video-sharing platforms has intensified the need to harmonize video content with appropriate music. For creators, this pairing often demands considerable manual effort and repeated experimentation. This paper presents an automated solution that streamlines music selection by analyzing the raw content of both video and audio. Our approach evaluates a video's visual dynamics by measuring movement distribution over time via optical flow. Concurrently, for candidate music tracks, the system extracts the onset envelope to analyze rhythmic tempo evolution. We introduce novel similarity metrics, informed by these signals, to assess the suitability of a music-video pair. This method requires only raw sources, eschewing the need for supplemental metadata and distinguishing it from complex deep multi-modal fusion strategies. Experimental results, utilizing both continuous cross-correlation and discrete beat-distance metrics, demonstrate our model's efficacy across diverse video and music genres. Evaluations show a heightened accuracy in identifying the original video-music alignment compared to alternative pairings.

1 Introduction

The rise of video-sharing platforms like YouTube, TikTok, and Instagram has created an immense demand for engaging short-form content. A key element of a successful video is the synergistic pairing of visuals with a fitting soundtrack, a task that often requires professional editing skills and can be daunting for the average creator. Computational video editing offers a promising solution to automate this process and assist non-professionals.

A central challenge in this automation is developing a method to measure the compatibility between a video and audio clip. Many existing approaches rely on supervised learning with ground-truth datasets [1, 2, 3], which are difficult to scale and can introduce subjective biases from labelers. Other methods, such as emotion-based classification [4] or deriving associations from film music [5], are often computationally intensive, musically restrictive, and depend on highly subjective attributes. While unsupervised techniques like visual beat extraction [6] avoid these issues, they are typically suited only for very short clips.

In this study, we propose an unsupervised approach that uses digital signal processing (DSP) to define a similarity metric between video and audio based on their rhythmic structures. Our core hypothesis is that a video is more appealing to viewers when the on-screen motion is synchronized with the musical beats. This method is computationally efficient, avoids the pitfalls of dataset bias, and is designed for the popular short-video format.

We extract rhythmic information, specifically tempo, from both modalities and compute a distance metric between them. Soundtrack recommendations are then ranked by this distance, favoring

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: NeurIPS 2025 Workshop on AI for Music: Where Creativity Meets Computation.

audio clips with a tempo that aligns with the video's visual rhythm. We demonstrate that our model outperforms existing music recommendation models across various video domains, effectively identifying compelling pairings while filtering out inappropriate combinations that clash in genre or mood. Our objective is to validate that rhythmic alignment is a strong, measurable indicator of perceived music-video quality.

2 Related Work

Audio recommender systems for video are closely linked to Computational Video Editing (CVE), as both require robust feature extraction and alignment of audio-visual signals. Research in CVE has focused on enhancing video quality, generating meaningful frame sequences [7], and applying user-specified editing idioms [8].

A core challenge is synchronizing audio and video. Early work aligned visual and audio beats using optical flow and spectrogram analysis [6], suitable for short, rhythmically simple videos. The AlignNet model later improved resilience to non-uniform misalignments using attention and pyramidal processing, excelling in dance and speech alignment [9].

Feature extraction methods vary widely. Some approaches use low-level features (e.g., MFCCs, CNN outputs [4]), while others employ high-level semantic concepts like emotion [5, 10]. Techniques also differ in their processing: some studies use fixed-length segments and neural networks to learn audio-visual mappings [4], while others leverage music theory [5] or multi-task learning for beat and tempo estimation [11].

Despite progress, comparing audio and video features directly remains underexplored [12]. Our work introduces unsupervised rhythmic similarity metrics to recommend soundtracks for any musical video.

3 Model Description

Given audio samples $\vec{A}$ and video frames $\vec{V}$, our goal is to construct scalar similarity metrics. We process them to generate 1-D onset signals: the **onset envelope** $Onset_a(n)$ for audio and the **visible impact** $Onset_v(n)$ for video. From these, we define a continuous similarity metric $CM(\vec{A}, \vec{V})$ as their cross-correlation. For a discrete similarity metric $DM(\vec{A}, \vec{V})$, we first extract beat sequences B_a and B_v from the onset signals using a dynamic programming approach [13]:

$$B_a = DP(Onset_a(n)), \quad B_v = DP(Onset_v(n)) \tag{1}$$

3.1 Extracting Information from Audio

Using the Python library *librosa,*[1] we derive a spectrogram $SP(n, k)$ via FFT. We then calculate spectral flux $SF(n, k)$ and aggregate positive values to get the onset strength signal:

$$SF(n, k) = |SP(n, k)| - |SP(n-1, k)| \tag{2}$$

$$Onset_a(n) = \sum_{k=0}^{K} \frac{SF(n, k) + |SF(n, k)|}{2} \tag{3}$$

3.2 Extracting Information from Video

We apply the Farneback optical flow algorithm to get velocity amplitude $\vec{Amp}$ and angles $\vec{Ang}$. We construct a directional distribution matrix $D(n, \theta)$ and compute a visible impact signal:

$$D(n, \theta) = \sum_{x,y} |Amp_n(x, y)| \mathbb{1}_\theta(Ang_n(x, y)) \tag{4}$$

$$DF(n, \theta) = |D(n, \theta)| - |D(n-1, \theta)|, \quad Onset_v(n) = \sum_{k=0}^{K} \frac{DF(n, k) + |DF(n, k)|}{2} \tag{5}$$

[1] `https://librosa.github.io/librosa/`

where $\mathbb{1}_\theta(\phi)$ is the indicator function for angle bin θ.

3.3 Continuous Video-Soundtrack Similarity

We compute the cross-correlation of the powered and normalized onset signals:

$$CM(\vec{A}, \vec{V}) = \max(G(Onset_a^5(n), Onset_v^5(n))) \tag{6}$$

Raising to the 5th power was found experimentally to maximize the difference between original and non-original pairs (Table 1).

Table 1: Difference in correlation coefficients for original and non-original pairs of onset signals.

Power of signals	Difference in %
1	-3
2	16
3	65
4	106
5	161
6	147

3.4 Discrete Video-Soundtrack Similarity

After extracting beats B_a and B_v, we calculate the sum of absolute differences to the nearest beat. The inverse of this sum gives the discrete similarity:

$$DM(\vec{A}, \vec{V}) = \frac{1}{\text{Bdistance}(B_a, B_v) + \epsilon} \tag{7}$$

To account for alignment, we test multiple starting point delays.

4 Experiments and Results

We tested our metrics on a custom dataset of 20 dance videos and 30 music tracks. For each video, we calculated similarity scores with all audios. Accuracy was measured by how often the original soundtrack received the highest score.

4.1 Discrete Video-Soundtrack Distance

This metric achieved 73% accuracy. Figure 1 shows an example of extracted beats, where alignment is strong for simple movements but degrades with complexity. The metric favors soundtracks with tempos similar to the original.

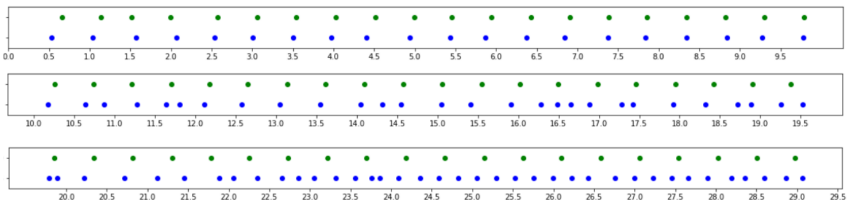
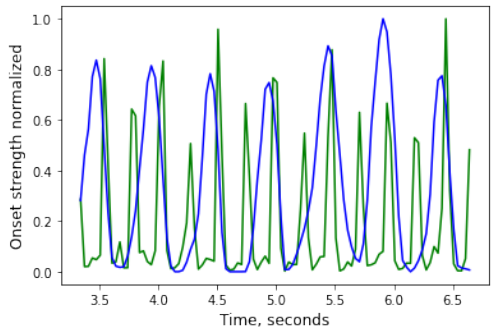


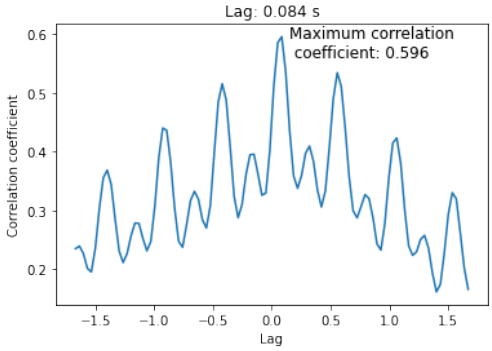
Figure 1: Audio (green dots) and video (blue dots) beat visualization over time. Top band shows strong correspondence for simple movements; complexity increases in second and third bands.

4.2 Continuous Video-Soundtrack Similarity

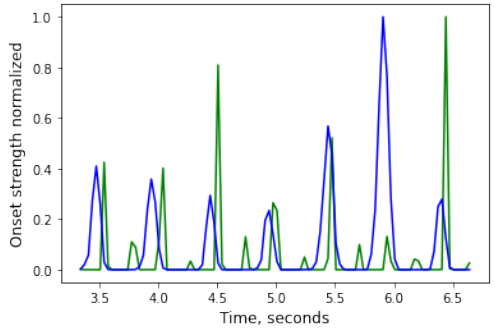
This metric achieved 87% accuracy, outperforming the discrete method. Figure 2 shows the alignment of onset signals, their cross-correlation, and how raising the signals to a power makes peaks more distinguishable. The highest correlation was for the original video-soundtrack pair, with a bias toward similar tempos.



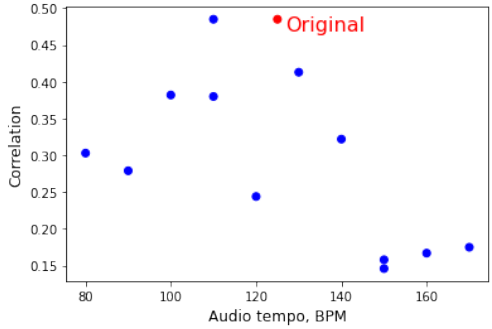
(a) Onsets of video (blue) and audio (green) over time, showing rhythmic alignment with a slight lag.



(b) Cross-correlation of visible impacts (video) and onset envelope (audio) signals.



(c) Cross-correlation of signals raised to the 5th power, reducing noise and emphasizing peaks.



(d) Correlation between a fixed video's visible impacts and various audio onset envelopes.

Figure 2: Continuous video-soundtrack similarity metrics: (a) onset alignment, (b) cross-correlation analysis, (c) power-enhanced correlation, and (d) correlation comparison across different soundtracks.

5 Discussion and Comparison

The continuous similarity metric outperformed the discrete one, offering more flexibility through parameters like the signal power. On videos algorithmically aligned using [6]'s pipeline, our accuracy reached 93%, exceeding that for human-generated content. This suggests our metric could serve as a powerful objective function for automated video editing systems. Our method significantly outperformed the approach in [4], which reported 58% accuracy. The higher performance of our unsupervised, rhythm-based approach highlights the significance of temporal alignment as a primary factor for perceived quality, potentially surpassing more subjective measures like emotion. Furthermore, its computational efficiency makes it suitable for real-time application on popular platforms.

6 Conclusion

We introduced unsupervised similarity metrics for video-soundtrack recommendations based on rhythmic alignment. Our analysis confirmed a stronger rhythmic harmony in human-generated clips than in random pairs, validating our core hypothesis. The high performance achieved demonstrates that rhythmic coherence is a powerful and effective signal for measuring audio-visual compatibility. Future work involves developing a music indexing system using rhythm and tempo features for efficient search across large libraries. We also would like to examine benefits from feature extraction with depth-aware models like [14, 15, 16].

Acknowledgments and Disclosure of Funding

The work was supported by the Ministry of Economic Development of the RF (agreement No. 139-10-2025-034 dd. 19.06.2025, IGK 000000C313925P4D0002).

References

[1] Pedro Morgado, Yi Li, and Nuno Nvasconcelos. Learning representations from audio-visual spatial alignment. *Advances in Neural Information Processing Systems*, 33:4733–4744, 2020.

[2] Sadiq H Abdulhussain, Basheera M Mahmmod, M Iqbal Saripan, SAR Al-Haddad, Thar Baker, Wameedh N Flayyih, Wissam A Jassim, et al. A fast feature extraction algorithm for image and video processing. In *2019 international joint conference on neural networks (IJCNN)*, pages 1–8. IEEE, 2019.

[3] Vadim Kantorov and Ivan Laptev. Efficient feature extraction, encoding and classification for action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2593–2600, 2014.

[4] Jen-Chun Lin, Wen-Li Wei, James Yang, Hsin-Min Wang, and Hong-Yuan Mark Liao. Automatic music video generation based on simultaneous soundtrack recommendation and video editing. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 519–527, 2017.

[5] Fang-Fei Kuo, Meng-Fen Chiang, Man-Kwan Shan, and Suh-Yin Lee. Emotion-based music recommendation by association discovery from film music. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 507–510, 2005.

[6] Abe Davis and Maneesh Agrawala. Visual rhythm and beat. *ACM Trans. Graph.*, 37(4):122–1, 2018.

[7] Vineet Gandhi, Remi Ronfard, and Michael Gleicher. Multi-clip video editing from a single viewpoint. In *Proceedings of the 11th European Conference on Visual Media Production*, pages 1–10, 2014.

[8] Mackenzie Leake, Abe Davis, Anh Truong, and Maneesh Agrawala. Computational video editing for dialogue-driven scenes. *ACM Trans. Graph.*, 36(4):130–1, 2017.

[9] Jianren Wang, Zhaoyuan Fang, and Hang Zhao. Alignnet: A unifying approach to audio-visual alignment. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3309–3317, 2020.

[10] Rajiv Ratn Shah, Yi Yu, and Roger Zimmermann. Advisor: Personalized video soundtrack recommendation by late fusion with heuristic rankings. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 607–616, 2014.

[11] Sebastian Böck, Matthew EP Davies, and Peter Knees. Multi-task learning of tempo and beat: Learning one to improve the other. In *ISMIR*, pages 486–493, 2019.

[12] Timothy R Brick and Steven M Boker. Correlational methods for analysis of dance movements. *Dance Research*, 29(supplement):283–304, 2011.

[13] Daniel PW Ellis. Beat tracking by dynamic programming. *Journal of New Music Research*, 36(1):51–60, 2007.

[14] Ilya Makarov and Gleb Borisenko. Depth inpainting via vision transformer. In *2021 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pages 286–291. IEEE, 2021.

[15] Ilya Makarov, Maria Bakhanova, Sergey Nikolenko, and Olga Gerasimova. Self-supervised recurrent depth estimation with attention mechanisms. *PeerJ Computer Science*, 8:e865, 2022.

[16] Dmitrii Maslov and Ilya Makarov. Online supervised attention-based recurrent depth estimation from monocular video. *PeerJ Computer Science*, 6:e317, 2020.