This CVPR workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the version available on IEEE Xplore.

Visual Rhythm and Beat

Abe Davis
Stanford University
abedavis.com

Maneesh Agrawala
Stanford University

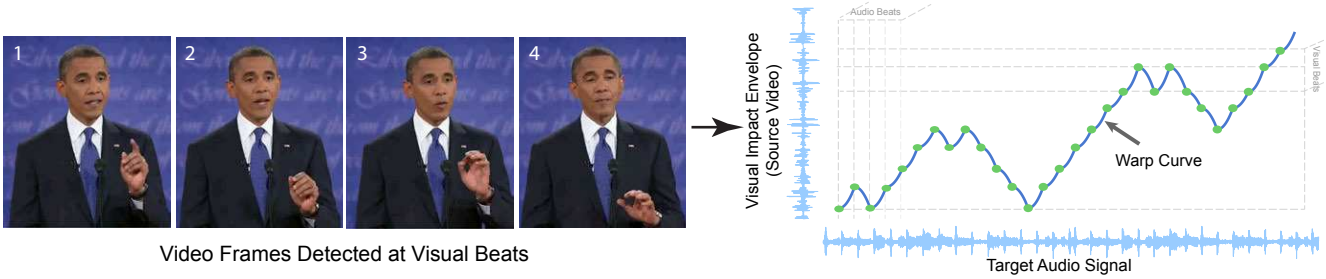

Figure 1: ***Accidental Dance and Unfolding*** - by analyzing visual rhythm in a collection of video, we can search for segments of unintentionally-rhythmic motion and use them to synthesize dance performances. (Left) We see video frames taken at the moment of four consecutive visual beats detected in video from a 2012 presidential debate. These visual beats lie at the high and low points of a repetitive up-and-down hand gesture. (Right) The warp curve on the right shows the process of unfolding. After finding a short dance-like segment of video, we generate a random walk through its corresponding sequence of visual beats to synthesize a longer dance video.

Abstract

We present a visual analogue for musical rhythm derived from an analysis of motion in video, and show that alignment of visual rhythm with its musical counterpart results in the appearance of dance. Central to our work is the concept of visual beats — patterns of motion that can be shifted in time to control visual rhythm. By warping visual beats into alignment with musical beats, we can create or manipulate the appearance of dance in video. Using this approach we demonstrate a variety of retargeting applications that control musical synchronization of audio and video: we can change what song performers are dancing to, warp irregular motion into alignment with music so that it appears to be dancing, or search collections of video for moments of accidentally dance-like motion that can be used to synthesize musical performances. (This paper is a workshop preview of Davis et al. 2018 [8].)

1. Introduction

Music and dance are closely related through the concept of rhythm, which describes how events—e.g., the sound of an instrument or the movement of a body—are distributed in time. Rhythm is in some sense a very intuitive concept: infants can recognize and follow basic rhythms at as early as six months of age [27, 6], and even some animals—certain parrots and elephants, for example—are known to move in time with simple music [24, 25]. However, the task of quantifying rhythm is not trivial, and has been the topic of extensive research in the context of both music [12, 17, 14, 13, 11, 9, 1] and dance [3, 10]. Our work builds on that research to explore a visual analogue for rhythm—which we call ***visual rhythm***—in video. Just as musical rhythm captures the temporal arrangement of sounds, visual rhythm captures the temporal arrangement of visible motion. We focus on analyzing that motion to identify structure related to dance.

Our central hypothesis is that music and dance are characterized by complementary rhythmic structure in audible and visible signals. Our exploration of that structure builds on the concept of ***visual beats***—visual events that, when temporally aligned with musical beats, create the appearance of dance. The relationship between visual and musical beats provides a starting point from which we derive visual analogues for other rhythmic concepts, including onset strength and tempo. Visual beats also give us a recipe for manipulating rhythmic structure in video: we first identify visual beats, then time-warp those beats into alignment with a specified target. Provided we are able to identify the necessary beats, we show that it is possible to warp video into dance-like alignment with any song of our choice.

1.1. Applications

The quantification of visual rhythm enables many applications. We focus primarily on those related to video retargeting, which combines analysis and synthesis of dance. In addition to motivating our work, these applications serve to test our basic assumptions about visual rhythm and dance.

Dance Retargeting: By time-warping the visual beats of existing dance footage into alignment with new music, we can change the song that a performer is dancing to.

This is a special case of retargeting where we can assume that visual beats are already aligned with musical beats in the source video, allowing us to find them with simple audio beat tracking. We leverage this to test our central hypothesis about visual beats and dance separately from any computer vision algorithms.

Dancification: Our visual beat hypothesis allows for the existence of visual beats in non-dance video, but implies such visual beats should not be distributed according to any discernible tempo. If we can find such beats through purely visual means, we can use them to transform non-dance video into dance video. We call this *dancification*. We can also use this strategy to improve bad or off-tempo dancing, providing a kind of "auto-tune" for dance.

Accidental Dance: We can adapt our strategy for identrifying visual beats into a search criteria, which we can use to find segments of dance-like or near dance-like motion in large collections of video. If only short segments of such video can be found, we generate random walks through the visual beats of those segments to synthesize an arbitrary length of output dance video.

Visual Instrument: Visual beats provide temporal control points that can also be used for more general manipulation of video. For example, by warping visual beats into alignment with the notes of a musical instrument (e.g., recorded MIDI or a transcribed performance) we can use that instrument as a musical interface for editing video.

1.2. Beat Saliency

We begin by factoring the perception of beat — both for music and dance — into different types of saliency, drawing on observations from literature on the arts [4, 2, 7, 22, 29] as well as heuristics used by related work on audio beat tracking [12, 21, 17, 14, 11, 13] and the computational analysis of dance [16, 5, 18, 23, 10, 3]. The saliency metrics described here guide our design of heuristics for visual beat tracking and a dance-specific strategy for time-warping video, which we describe in our full paper [8].

Musical beats are often defined as moments where a listener would clap or tap their feet in accompaniment with music. This definition relies on an implied measure of saliency, with different sounds affecting the perception of beats in different ways. Most work on rhythmic analysis approximates this saliency implicitly through the use of a heuristic objective for finding beats in audio. Typically that objective is expressed as a combination of two functions: one temporally local function that measures musical ***onset strength*** (indicating the start of musical notes), and another function that measures adherence to a particular ***tempo***, as indicated by periodic patterns in the distribution of onset strength over time.

Our definition of visual beats implies a related type of saliency, rooted in the perception of dance. We assume this saliency can also be factored into local and rhythmic components, from which we will derive visual complements for onset strength and tempo. Note that the local component of visual beat saliency is different from classic image saliency [19, 26, 15] in that it is a function of visible motion, and should reflect some measure of our ability to localize events in time.

We refer to the rhythmic components of visual and musical beat saliency as ***rhythmic saliency*** and the local components as ***local saliency***.

1.3. Synchro-Saliency

The perception of dance is greatly influenced by musical accompaniment. This is why a dance can appear synchronized with one piece of music, and out of place with another. We discuss this synchronization in terms of what we call ***synchro-saliency***, which measures the perceived strength of relationships between visible and audible events.

We describe any two functions $h_a(t_a)$ over audible events and $h_v(t_v)$ over visible events as ***synchro-salient complements*** if their product approximates synchro-saliency h_s:

$$h_a(t_a)h_v(t_v) \approx h_s(t_a, t_v) \tag{1}$$

In other words, synchro-salient complements are corresponding functions over audio and video that indicate high synchro-saliency when large values are aligned in time.

In Davis 2018 [8] we design heuristics for the local and rhythmic saliency of video to be synchro-salient complements of corresponding heuristics used in audio beat detection. This lets us express dancification as the alignment of rhythmic saliency with a target.

References

[1] S. Böck and Gerhard Widmer. Maximum filter vibrato suppression for onset detection. 2013. 1

[2] T. L. Bolton. Rhythm. *The American Journal of Psychology*, 6(2):145–238, 1894. 2

[3] T. R. Brick and S. M. Boker. Correlational methods for analysis of dance movements. *Dance Research*, 29(supplement):283–304, 2011. 1, 2

[4] M. Chion, C. Gorbman, and W. Murch. *Audio-vision: Sound on Screen*. Film and Culture. Columbia University Press, 1994. 2

[5] H. chul Lee and I. kwon Lee. Automatic synchronization of background music and motion. In *in Computer Animation, in*

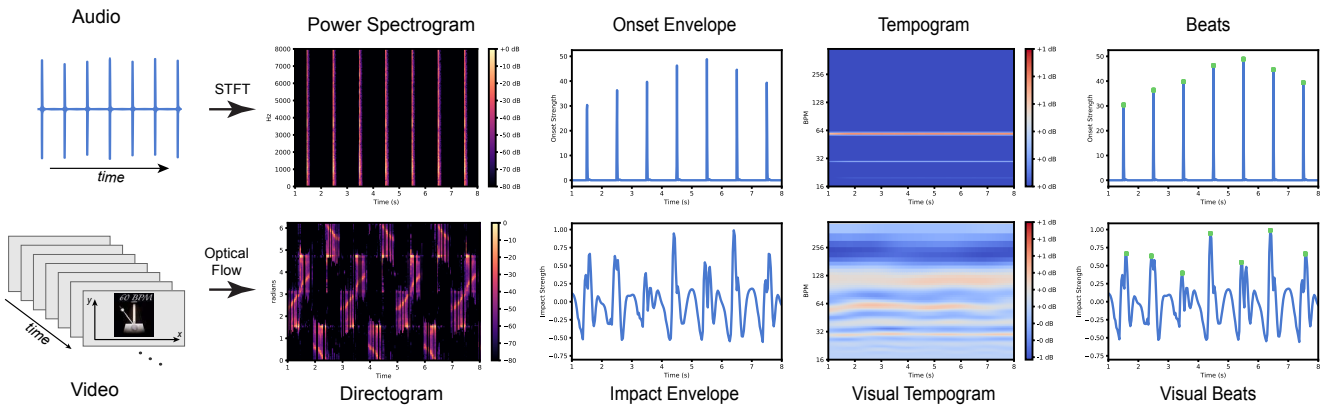

Figure 2: ***Rhythmic Features in Audio and Video*** – The top row shows features used to quantify metric structure in audio. The bottom row shows the synchro-salient complements that we use to quantify metric structure in video. The visualizations here correspond to the audio and video from footage of a simple metronome [20]. As we would expect from footage of a metronome, the detected visual metric structure is aligned with its synchro-salient complement in audio.

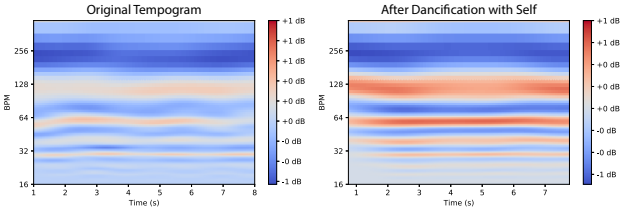

Figure 3: ***Dance-Specific Interpolation for Time Warping*** - Here we see the effect of our interpolation strategy on the metronome video [20] from Figure 2 when visual beats in the original video are warped to themselves. The timing of visual beats does not change in this case, but timing around those beats is changed, producing acceleration and deceleration to emphasize rhythmic structure already in the video. Note that both linear and cubic interpolation would have no effect on the input in this example.

Computer Graphics Forum, Volume 24, Issue 3 (2005, pages 353–362, 2005. 2

[6] L. K. Cirelli, C. Spinelli, S. Nozaradan, and L. J. Trainor. Measuring neural entrainment to beat and meter in infants: Effects of music background. *Frontiers in Neuroscience*, 10:229, 2016. 1

[7] H. Cowell and D. Nicholls. *New Musical Resources*. Cambridge University Press, 1996. 2

[8] A. Davis and M. Agrawala. Visual rhythm and beat. *ACM Trans. Graph.*, 37(4), Aug. 2018. 1, 2, 4

[9] S. Dixon. Onset detection revisited. In *In Proceedings of the 9th international conference on digital audio effects*, pages 133–137, 2006. 1

[10] V. Dyaberi, H. Sundaram, T. Rikakis, and J. James. The computational extraction of spatio-temporal formal structures in the interactive dance work '22'. In *2006 Fortieth Asilomar Conference on Signals, Systems and Computers*, pages 59–63, Oct 2006. 1, 2

[11] D. P. W. Ellis. Beat tracking by dynamic programming. *Journal of New Music Research*, 36(1):51–60, 2007. 1, 2

[12] M. Goto. An audio-based real-time beat tracking system for music with or without drum-sounds. 30, 09 2002. 1, 2

[13] P. Grosche, M. Muller, and F. Kurth. Cyclic tempogram – a mid-level tempo representation for musicsignals. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5522–5525, March 2010. 1, 2

[14] X. Hu, J. H. Lee, D. Bainbridge, K. Choi, P. Organisciak, and J. S. Downie. The mirex grand challenge: A framework of holistic user-experience evaluation in music information retrieval. *J. Assoc. Inf. Sci. Technol.*, 68(1):97–112, Jan. 2017. 1, 2

[15] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *IEEE International Conference on Computer Vision (ICCV)*, 2009. 2

[16] T.-h. Kim, S. I. Park, and S. Y. Shin. Rhythmic-motion synthesis based on motion-beat analysis. *ACM Trans. Graph.*, 22(3):392–401, July 2003. 2

[17] A. Lerch. *An Introduction to Audio Content Analysis: Applications in Signal Processing and Music Informatics*. Wiley-IEEE Press, 1st edition, 2012. 1, 2

[18] Z. Liao, Y. Yu, B. Gong, and L. Cheng. audeosynth: Music-driven video montage. *ACM Trans. Graph. (SIGGRAPH)*, 34(4), 2015. 2

[19] F. Liu, Y. Niu, and M. Gleicher. Using web photos for measuring video frame interestingness. In *Proceedings of the 21st International Jont Conference on Artifical Intelligence*, IJCAI'09, pages 2058–2063, San Francisco, CA, USA, 2009. Morgan Kaufmann Publishers Inc. 2

[20] LumBeat. 60 bpm metronome, Feb 2013. 3

[21] B. McFee, C. Raffel, D. Liang, D. P. W. Ellis, M. McVicar, E. Battenberg, and O. Nieto. librosa: Audio and music signal analysis in python. 2015. 2

[22] K. McPherson. *Making Video Dance: A Step-by-step Guide to Creating Dance for the Screen*. Routledge, 2006. 2

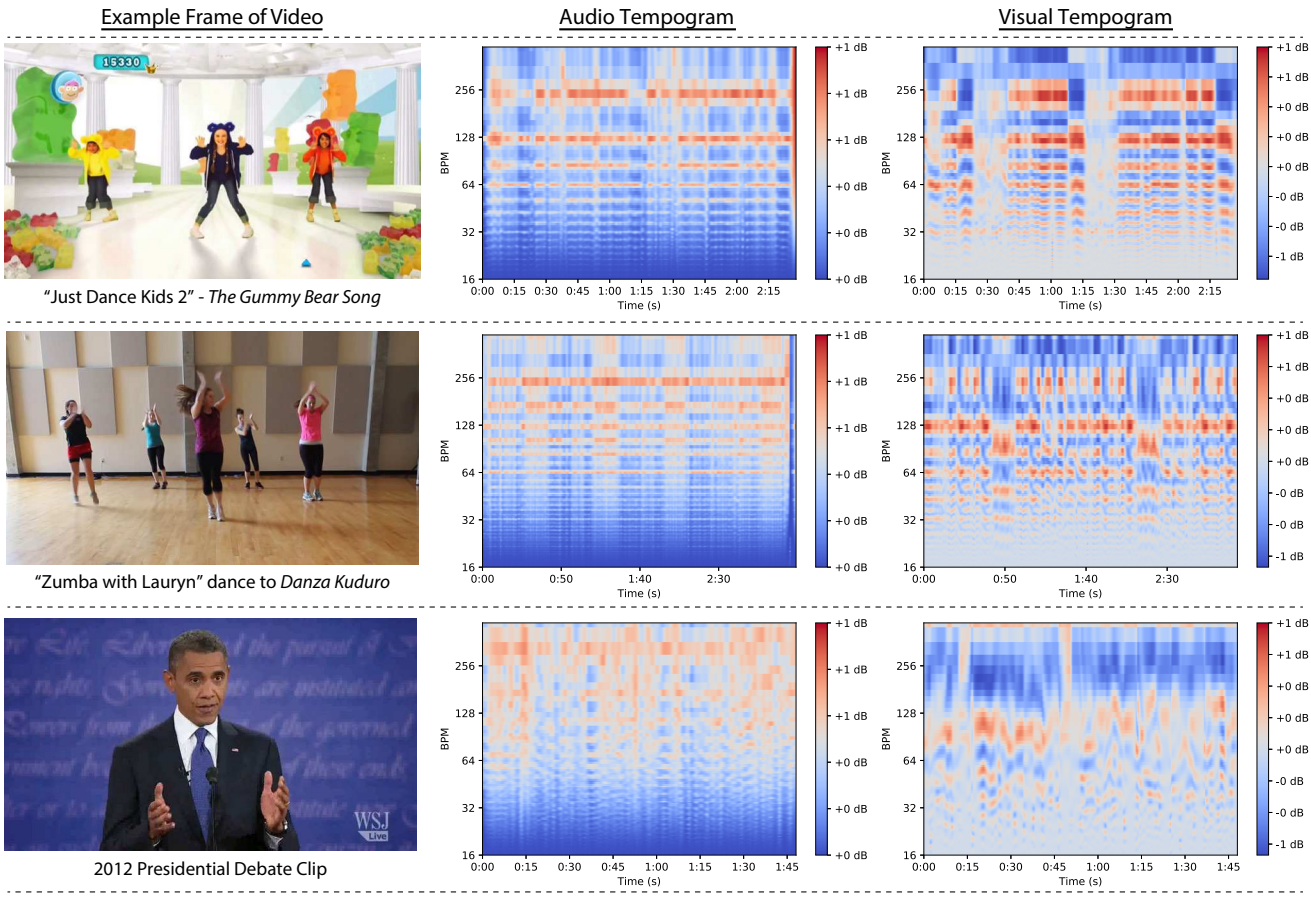

Figure 4: ***Comparison of Regular (Audio) and Visual Tempograms for For Dance and Non-Dance Video*** – Here we compare audio (middle column) and visual (right column) tempograms for three videos, with representative frames shown on the left. The top row visualizes dance video from a videogame made for children [28]. The middle row shows the visual tempogram for a zumba dance routine [31] set to the song *Danza Kuduro* by Don Omar (Note: here we calculated the audio tempogram on an aligned version of the original track, as the audio recorded with the video was low quality). The bottom row shows tempograms for a clip from the first 2012 Presidential Debate [30] (the same source video is shown in Figure 1 and featured in our supplemental video). In the dance examples (top two rows), we see high energy across matching harmonic tempos for both audio and video. In the non-dance video (bottom row), local tempo is more ambiguous and less consistent. Our full paper [8] describes a method for warping videos like the one on the bottom row in order to create rhtyhmic structure and align it with that of a target piece of music.

[23] T. P. Chen, C.-W. Chen, P. Popp, and B. Coover. Visual rhythm detection and its applications in interactive multimedia. 18:88–95, 01 2011. 2

[24] A. D. Patel and S. M. Demorest. 16 - comparative music cognition: Cross-species and cross-cultural studies. In D. Deutsch, editor, *The Psychology of Music (Third Edition)*, pages 647 – 681. Academic Press, third edition edition, 2013. 1

[25] A. D. Patel, J. R. Iversen, M. R. Bregman, and I. Schulz. Experimental evidence for synchronization to a musical beat in a nonhuman animal. *Current Biology*, 19(10):827–830, 2017/11/14. 1

[26] Y. Pritch, A. Rav-Acha, and S. Peleg. Nonchronological video synopsis and indexing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(11):1971–1984, Nov 2008. 2

[27] B. H. Repp and Y.-H. Su. Sensorimotor synchronization: A review of recent research (2006–2012). *Psychonomic Bulletin & Review*, 20(3):403–452, Jun 2013. 1

[28] Ubisoft. Just dance kids 2 i am a gummy bear, May 2013. 4

[29] C. Vernallis. *Experiencing Music Video: Aesthetics and Cultural Context*. Columbia University Press, 2004. 2

[30] WSJDigitalNetwork. Best moments of first obama/romney debate, Oct 2012. 4

[31] Zumba with Layryn. "danza kuduro" zumba routine, Jun 2014. 4