16th International Conference on Machine Vision Applications (MVA)
National Olympics Memorial Youth Center, Tokyo, Japan, May 27-31, 2019.

Visual Rhythm Prediction with Feature-Aligning Network

Yutong Xie, Haiyang Wang, Yan Hao, Zihao Xu
Shanghai Jiao Tong University
Shanghai, China
{xxxxxyt,wanghaiyang,honeyhaoyan,shsjxzh}@sjtu.edu.cn

Abstract

In this paper, we propose a data-driven visual rhythm prediction method, which overcomes the previous works' deficiency that predictions are made primarily by human-crafted hard rules. In our approach, we first extract features including original frames and their residuals, optical flow, scene change, and body pose. These visual features will be next taken into an end-to-end neural network as inputs. Here we observe that there are some slight misaligning between features over the timeline and assume that this is due to the distinctions between how different features are computed. To solve this problem, the extracted features are aligned by an elaborately designed layer, which can also be applied to other models suffering from mismatched features, and boost performance. Then these aligned features are fed into sequence labeling layers implemented with BiLSTM [9] and CRF [10] to predict the onsets. Due to the lack of existing public training and evaluation set, we experiment on a dataset constructed by ourselves based on professionally edited Music Videos (MVs), and the F1 score of our approach reaches 79.6.

1 Introduction

Visual rhythm prediction has caught people's attention for years, since it can enable many valuable applications like automated video editing. This problem can be described as given a segment of videos, we want to decide whether each time point is an onset or not, as earlier works discussed [3][12]. But most of the previously proposed methods have a main disadvantage or disability: they primarily rely on human-crafted hard rules to compute the visual onsets [3][5][12][11], and can only perform well on a small set of specific videos, like dancing video without camera moving.

We admit that, the visual rhythm is hard to rigidly define with a simple formula, due to the variety and complexity of rhythm-related visual cues, including: 1. visual content changes; 2. movement of lens or camera; 3. environmental lighting conversion; 4. scene changes; 5. motion of performers, etc.

However, the visual rhythm can be indirectly reflected by the corresponding musical rhythm to some extent, especially in the professionally edited MVs, which enables us to learn how to predict visual onsets from MVs of high quality. So we propose a data-driven method for visual rhythm prediction with an end-to-end *Feature-Aligned Network* (FAN), and train it with sufficient data.

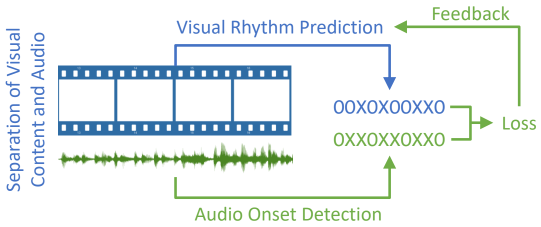

Figure 1. For a given video, we can predict the visual rhythm by deciding a time point to be an onset or not. While training the predictor with professionally edited MVs, the results of audio onset detection will be used as labels to rectify the predictor. The blue parts will be executed at both testing and training stage, while the green parts will only be executed at the training stage.

In our method, the visual rhythm is related with various visual cues mentioned above, we first extract the features closely linked with these cues, including 1. original frames; 2. frame residuals; 3. optical flow; 4. scene change; 5. body pose. The detailed reason why and by how we extract them will be further explained in Section 3.

Then these extracted visual features are fed into FAN to predict the onsets. Here, we observe that there are some slight misaligning between features over the timeline, and assume that this is due to the distinctions between how different features are computed. For example, the body pose is decided by only a single frame, while the frame residuals depend on two consecutive frames and scene change is influenced by a few continuous frames. So after the extraction and transformation, we align features with an elaborately designed layer, which can also be applied to other models suffering from mismatched features, and bring in performance improvement. Next, the aligned features are fed into final layers implemented with BiLSTM [9] and CRF [10] to predict the onsets, as we further formalize the visual rhythm prediction as the general sequence labeling problem. The architecture and details will be further explained in Section 4.

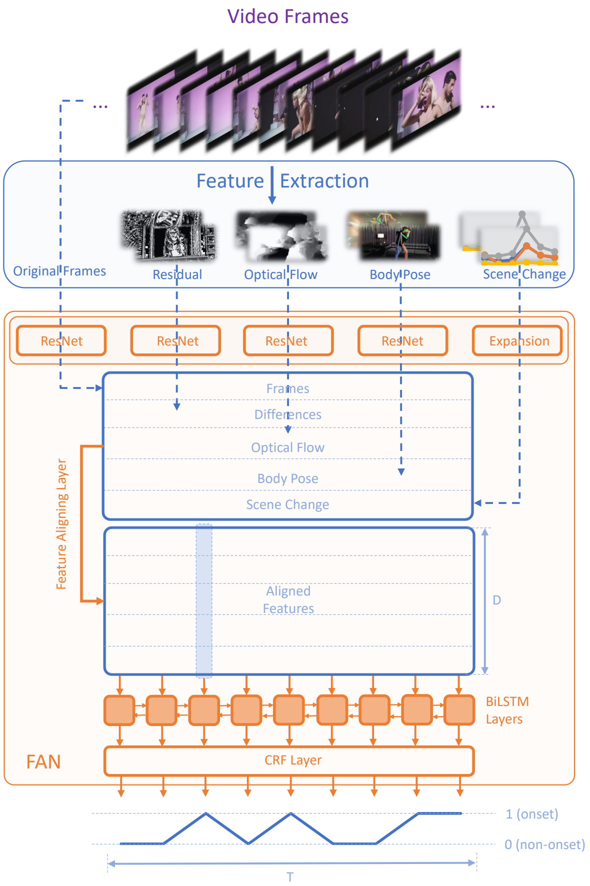

Figure 2. Our approach with the end-to-end Feature-Aligning Network (FAN). Extracted visual features are fed into the sequence labeling layers after transformation and aligning, which predicts the rhythm onsets.

Due to the lack of public training and evaluation set, we construct a *MV Visual Rhythm* (MVVR) dataset by ourselves based on MVs published on the Internet. In our dataset, as we assume that the visual and musical rhythm will properly match each other in professionally edited MVs, the results of musical onset detection is taken as ground truth. In the experiment, the F1 score reaches 79.6, which proves our method to be effective.

2 Related Work

2.1 Visual rhythm prediction

In this part, we will briefly review the previous work on visual rhythm prediction. Davis et al. [3] suggest that the sudden visible deceleration of the moving object can indicate the visual rhythm, and measure it by calculating the optical flow of videos. However, this rule-based method cannot distinguish the motion of central subject from the background or camera motion, and even minor camera motion disturbance can be a great interference for the detection, the motion patterns are too complicated to be well described with concise formulas.

Argüello et al. [5], Chu et al. [12] in another perspective, come up with the similar idea that visual rhythm can be treated as periodic patterns in actual motion, and then employ different motion detection methods to eventually extract visual rhythm onsets.

Chen et al. [11] describe visual rhythm in a more rough way as the occurrence frequency of rhythmic events like human movement and environment lighting change. Absolute frame difference and 2D angle-magnitude histogram of optical flows are used in this article to measure such frequency. However, it can not precisely predict the visual onset events timing and only estimate how intense the visual rhythm is.

The previously mentioned methods are all rule-based, and their key problem lies in that it can only perform well on a small set of specific videos, like dancing video without camera moving. Therefore, we refer to deep representation learning to handle more complicated patterns and enrich the expressiveness of our model.

2.2 Musical onset detection

Musical onset detection is an area that has been explored for years and can bring us much inspiration, since it is also modeled as a sequence labeling problem. Bello et al. [6] present a rule-based model and have achieved one of the best results by unsupervised methods. But this has soon been outperformed by methods employing deep neural networks [7][8][16], which convinces us to address the visual rhythm prediction problem with deep learning methods.

3 Visual Feature Extraction

In our method, we first extract rhythm-related features from the visual content, and the results will be subsequently fed into the end-to-end network — FAN, to predict visual onsets.

Considering the visual content changes, movement of lens or camera, environmental lighting conversion, scene changes and motion of performers as visual rhythm cues in videos, we extract features as following:

Original Frames From original frames, we can know what is shown in the video, which is helpful to the rhythm prediction when the video is periodically changing the performing content. Here we simply take frames from RGB channels.

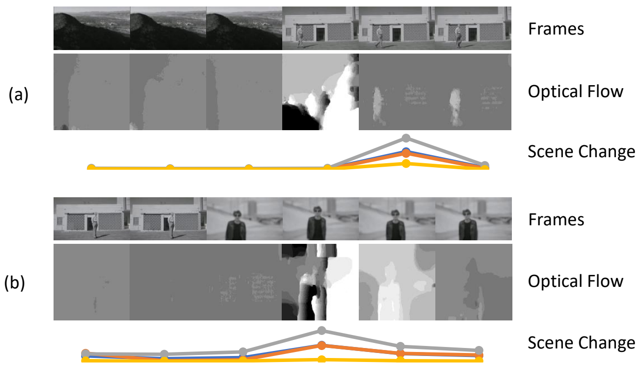

Figure 3. Two segments drawn from one video. In (a), the optical flow exactly matches the change in the frames, but the scene change detection result seems to be delayed. In (b), “peaks” of the optical flow and scene change are in the same position, but both fall behind the original frames.

Frame Residuals The residual of frames implies the movement of the lens or objects, which can also reflects the visual rhythm [11]. In this part, we directly compute the residual between two adjacent frames with subtraction.

Optical Flow Detection Optical flow is the pattern of apparent motion of image objects between two consecutive frames caused by the movement of objects or the camera. It reflects the intensity of the action in videos, which offers us some visual rhythm cues. In this part, we adapt the Lucas-Kanade method [14], which solves the basic optical flow equations for pixels in a neighbourhood by the least squares criterion, to gain a feature map whose size is the same as original frames.

Scene Change Detection Scene change detection divides a video into physical shots, which are video sequences that consists of continuous number of video frames for particular action. In this part, we make use of a video sequence detection method [15] based on the three dimensional histogram of color images, which gets the pattern of scene change by comparing the difference of histograms and their size between consecutive frames.

Body Pose Detection Most musical videos contain human movement (In the searching results of “MV” on YouTube, 99% of the videos contain human figures and their body motion), and human’s regular move pattern is also a good indicator for visual rhythm [5][12]. For this part, we refer to the state-of-art work of pose estimation [1] to extract body movement in a key point level.

4 FAN: Feature-Aligning Network

In the end-to-end Feature-Aligning Network, we take the previously extracted features as inputs, and first transform them into a common feature vector space (Section 4.1). After this, to alleviate the observed mismatching problem, we align them with an elaborately designed layer (Section 4.2). These aligned features are next fed into sequence labeling layers (Section 4.3) to predict the onsets. The whole prediction process is illustrated as Figure 2.

4.1 Feature Transformation

For the original frames and feature maps extracted by frame residuals, optical flow detection and body pose detection, we further transform them into feature vectors with ResNet-34 [17].

As for the low-dimensional feature extracted by scene change detection, we expand it into a high-dimensional feature space with fully connected layers.

4.2 Feature Aligning Layer

We observe that there are some slight misaligning between features over the timeline, as shown in Figure 3. Since the input video frames are selected by a common criteria, we assume that this problem is due to the distinctions between how different features are computed. Body pose is decided by only a single frame, corresponding to the original frame. As for the frame residuals and optical flow, they depend on two consecutive frames, while scene changes are influenced by a few continuous frames. So on a single time point, frames that are related with each feature vector are different, possibly leading to the misaligning problem.

Besides, misaligned features can seriously harm the performance of the prediction network. This is because all features along a time point will be mixed and then mapped to a new feature space during the subsequent feature transformation, which turns the misaligned features to worthless even adverse noise. In addition, the offset is position-sensitive, which means that it cannot be eliminated by simply working on the process of feature extraction.

Therefore, we propose to alleviate this problem by a feature aligning layer, which can automatically learn how to align features over the timeline. More specifically, it employs the attention mechanism over sequence [13], with which great progress has been made in Natural Language Processing (NLP) area, especially on the machine translation task. For a group of features G, the aligning layer rearrange it with the scaled dot-product attention as Figure 4.

Denote $t = 1, \ldots, T$ as the time indicator with T the length of this segment of video, $d \in G$ as the feature indicator of this group, $\boldsymbol{X} \in \mathbb{R}^{T \times D}$ as the concatenated extracted features with D the total dimension, p as the maximum offset.

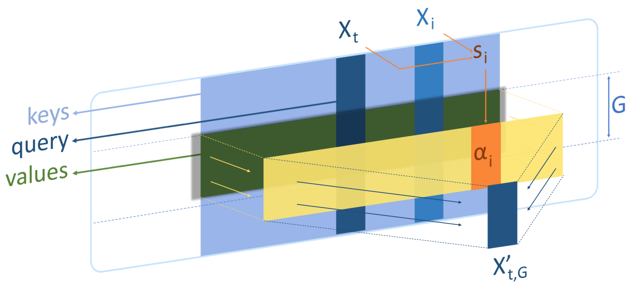

Figure 4. The results of rearranging a group of features by feature aligning layer is the weighted sum of values (a range of time points with this feature group), with attention weights computed by scoring the relevance of the query (this time point with all feature) and keys (a range of time points with all features).

Then by applying two group-specific transformation matrices $\boldsymbol{W}^q, \boldsymbol{W}^K \in \mathbb{R}^{D\times D}$, we can define the scaled dot-product attention scores $\boldsymbol{s} \in \mathbb{R}^{2p+1}$ and the normalized attention weight $\boldsymbol{\alpha} \in [0,1]^{2p+1}$, from which the rearranged results $\boldsymbol{X'} \in \mathbb{R}^{T\times|G|}$ can be computed as following

$$s_i = \frac{(\boldsymbol{W}^q\boldsymbol{X}_t)(\boldsymbol{W}^K\boldsymbol{X}_i)^{\mathrm{T}}}{\sqrt{D}} \in \mathbb{R} \tag{1}$$

$$\alpha_i = \frac{\exp(s_i)}{\sum_{j=t-p}^{t+p}\exp(s_j)} \in [0,1] \tag{2}$$

$$X'_{t,d} = \sum_{i=t-p}^{t+p} \alpha_i X_{i,d} \in \mathbb{R} \tag{3}$$

Here $\boldsymbol{X}_t$ plays the role of query, $\boldsymbol{X}_i$ plays the role of key, and $X_{t,d}$ plays the role of value. The result of rearranging a group of features by aligning layer is the weighted sum of values, with attention weights computed by scoring the relevance of the query and keys. Furthermore, by examining the attention weights, we can know how features are rearranged to align with each other.

The feature aligning layer is also an organic part of the end-to-end network, since the aligned results are next fed into subsequent layers, in which the training loss is back propagated. Thus the parameter matrices $\boldsymbol{W}^q, \boldsymbol{W}^K$, determining how values are weighted, can be optimized by taking gradient on the loss.

4.3 Sequence Labeling Layers

Given a sequence of frames, deciding whether one of it is an onset or not is indeed a sequence labeling problem. To utilize the information of adjacent frames, we employ the Bidirectional Long Short Term Memory (BiLSTM) [9] modules instead of separately predicting on each time point.

Aside from considering the information adjacent frames, it is also beneficial to jointly predict over the whole sequence. For example, in a smoothing video, it is unlikely to have two consecutive onsets in a very short time. Thus we apply a final Conditional Random Filed (CRF) [10] layer to make the prediction aware of consecutive predictions.

5 Experiment

5.1 Dataset

To our best knowledge, there exists no public visual rhythm prediction dataset. So we construct our own dataset based on YouTube-music-video-5M[1], a collection offers various styles of MVs, and published professionally edited MVs on YinYueTai[2], a popular music website. By manually filtering out poorly edited MVs, where the visual and musical rhythm don't match each other, we obtained a video set of size 800.

Fixing the segment length at $T = 20$, for each piece of video segment, we first separate the visual and audio contents, then

1. Extract video frames from RGB channels as input $\boldsymbol{X} \in \mathbb{R}^{T\times H\times W\times 3}$ where $H, W = 224$ is the normalized height and width. The extraction is at 4fps, to balance the labeling sensitivity and human's tolerance of visual rhythm deviation;
2. Take the musical onset detection result as the ground truth $\boldsymbol{y} \in \{0,1\}^T$.

Moreover, we wash out some excessively intense or smooth segments whose audio onset ratio (number of frames containing audio onsets divided by the segment length T in the sense of 4fps) beyond the range $[0.2, 0.8]$.

Finally, we form an MV Visual Rhythm (MVVR) dataset whose statistic information is listed as Table 1, where the counting is taken on musical onset detection results, i.e. the ground truth.

Table 1. MV Visual Rhythm Dataset

	Counting
Total Frames	850,620
Onset Labels	318,932
Non-onset Labels	531,688

[1] YouTube-music-video-5M: https://github.com/keunwoochoi/YouTube-music-video-5M

[2] YinYueTai: http://www.yinyuetai.com/

5.2 Audio Onset Detection

As for the audio onset detection, we follow the work of Sebastian Böck et al. [2], in which the spectral flux onset strength envelope is computed, and onset events are located by picking peaks in the envelope.

5.3 Implementation Details

In our implementation achieving the best performance, we employ ResNets of 34 layers and 2 layers fufully connected layers to first transform extracted features, and the dimension of transformed feature space D is 500. Then in the aligning layer, the maximum offset p is set as 2, and features are grouped by the extraction. In the sequence labeling layers, we use 2 layers of BiLSTMs and the CRF to predict the visual onsets, where the hidden state is of dimension 256. During training, the Adam optimizer is utilized, and the learning rate is set as 3×10^{-5}.

5.4 Results and Analysis

We estimate our prediction results with the metric of F1 score, in which the precision is defined as the true onset predictions count divided by all onset predictions count, and recall is defined as the true onset predictions count divided by all onsets count in the ground truth.

We experiment on the performance achieved by different features. Here all components in FAN are involved. The main results are shown in Table 2.

Table 2. Performance improvement with different visual features.

	Precision	Recall	F1 Score
Frames+Residuals	61.7	**99.4**	76.1
Optical Flow	61.7	99.2	76.1
Scene Change	56.3	88.1	68.7
Body Pose	61.2	87.4	72.0
All Features	**78.2**	81.0	**79.6**

As the above table shows, the combination of frames and residuals obtain the highest recall as 99.4, probably because of the frequent object or camera motion which leads to the predictor's tending to label frames as visual onsets. Frame residuals and optical flow gain similar results, probably because they both measure the differences between two consecutive frames. However, scene change and body pose perform not as ideal as the former features. This is possibly because, scene changes and the body pose may be sparse in a small number of video segments, due to the variety of our videos, since not every frame contains changing of view or figures to be detected. Above all, the combination of all features reaches the best precision of 78.2 and F1 score of 79.2.

6 Conclusion

In this paper, by formalizing the visual rhythm prediction as a sequence labeling problem, we proposed a data-driven method with an end-to-end network named Feature-Aligning Network, which utilizes the visual features including original frames and their residuals, optical flow, scene change and body pose, to predict visual onsets with sequence labeling layers consisting of BiLSTM and CRF. Here, we observed the slight misaligning between features over the timeline, and assumed that this is due to the distinctions between how different features are computed. Then we addressed this problem with an elaborately designed aligning layer, which can also be applied to other models suffering from mismatched features, and bring in performance improvement. Lastly, we constructed a MV Visual Rhythm dataset based on professionally edited MVs to fill the vacancy of the training set and public evaluation set. In the experiment on this dataset, the F1 score of our approach reached 79.6, which proved our method to be effective.

References

[1] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, Cewu Lu. RMPE: Regional Multi-person Pose Estimation In *Computer Vision and Pattern Recognition*, 2016

[2] Sebastian Böck and Gerhard Widmer. Maximum Filter Vibrato Suppression For Onset Detection In *Proc. of the 16th Int. Conference on Digital Audio Effects* (DAFx-13), Maynooth, Ireland, September 2-6, 2013

[3] Davis, Abe, and Maneesh Agrawala. Visual Rhythm and Beat. In *International Conference on Computer Graphics and Interactive Techniques*, vol. 37, no. 4, 2018, p. 122.

[4] Chen, Trista P., et al. visual rhythm prediction and Its Applications in Interactive Multimedia. In *IEEE MultiMedia*, vol. 18, no. 1, 2011, pp. 8895.

[5] Argüello, Camilo, and Marcela Iregui. Exploring Rhythmic Patterns in Dance Movements by Video Analysis. In *International Conference on Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management*, 2016, pp. 123131.

[6] Juan Pablo Bello, Chris Duxbury, Mike Davies, and Mark Sandler. On the use of phase and energy for musical onset detection in the complex domain. *IEEE Signal Processing Letters*, 11(6):553–556, 2004.

[7] Rong Gong and Xavier Serra. Towards an efficient deep learning model for musical onset detection. *arXiv preprint arXiv:1806.06773*, 2018.

[8] Jan Schlüter and Sebastian Böck. Improved musical onset detection with convolutional neural networks. In *ICASSP*, pages 6979–6983, 2014.

[9] Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5-6):602–610, 2005.

[10] Xuezhe Ma, Eduard H. Hovy. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. In *Meeting of the Association for Computational Linguistics*, 2016.

[11] Trista P Chen, Ching-Wei Chen, Phillip Popp, and Bob Coover. Visual rhythm detection and its applications in interactive multimedia. In *IEEE MultiMedia*, (1):88–95, 2011.

[12] Wei-Ta Chu and Shang-Yin Tsai. Rhythm of motion extraction and rhythm-based cross-media alignment for dance videos. In *IEEE Transactions on Multimedia*, 14(1):129–141, 2012.

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin. Attention is All you Need. In *Neural Information Processing Systems*, pp 5998-6008, Jan 1st 2017.

[14] Jean-Yves Bouguet Pyramidal Implementation of the Affine Lucas Kanade Feature Traker: Description of the algorithm In *Intel Corporation*, 2001.

[15] Igor S. Gruzman, Anna S. Kostenkova Algorithm of scene change detection in a video sequence based on the threedimensional histogram of color images In *International Conference on Actual Problems of Electronics Instrument Engineering (APEIE)*, 2012.

[16] Florian Eyben, Sebastian Böck, Björn Schuller, and Alex Graves. Universal onset detection with bidirectional long-short term memory neural networks. In *Proc. 11th Intern. Soc. for Music Information Retrieval Conference, ISMIR, Utrecht, The Netherlands*, pages 589–594, 2010.

[17] K He, X Zhang, S Ren, and J Sun. Deep residual learning for image recognition. corr, vol. abs/1512.03385, 2015.