Deep Video Audio Match Network

Junqi Liu
Shanghai Jiao Tong University
ljq435@sjtu.edu.cn

Lesheng Jin
Shanghai Jiao Tong University
king18817550378
@sjtu.edu.cn

Peiming Yang
Shanghai Jiao Tong University
yangpm1999@sjtu.edu.cn

ABSTRACT

Nowadays, multimedia is an important part of our life. The main part of multimedia is video and music, so how to match them is an important problem. But making computer understand human beings in video and music is very attractive. It is a task involving not only human being's motion beats but also their emotion conveyed by body. Music is a kind of data which has rich connotations and is in high dimension for it is very complicated. To handle this problem, we propose our Deep Video Audio Match Network(DVAMN) to match human pose videos and music. In DVAMN, we use Multiple-frame Fully Connected layers and Gated recurrent units to learn more information from other frames. And for less noise, we use human pose and music feature to train the model. Fortunately, our model not only has the ability of comprehending human being's motion and emotion, but also can understand music very well. For convenience we use dancing videos to build our model.

CCS Concepts

• **Computing methodologies** → **Artificial intelligence; Computer vision;**

Keywords

Deep Video Audio Match Network, Multi-frame Fully Connect, GRU, Human Pose videos, Music

1. INTRODUCTION

Matching music with a human pose video without sound is an important and challenging task. It can be a very good evidence for the existence of artistic attainments of artificial intelligence. If a model can deal with such a task, in another word, a model can analyze human being's emotion from body poses and motion beats, and comprehend music feeling to do a matching, it will be a great progress. Jumping out from this task, we can use the same model to measure human being's emotion from visual information and tell which video with music is more harmonious.

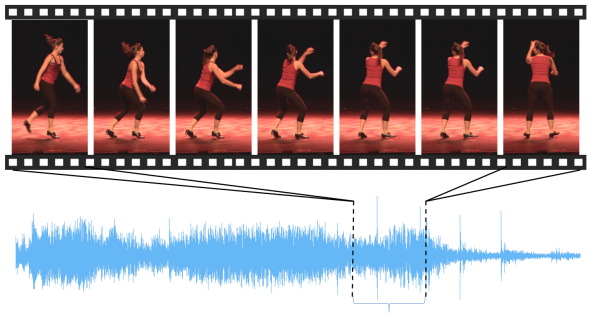

Figure 1: For dancing video, human action is highly related to the music beats. Select several frames from a human pose video, human can easily know the rhythm and emotion of its music.

With the help of deep learning, we can get much information from videos and music nowadays. But for human videos, the most important part is the human pose. Human often have some action corresponding to what he sound. So we use Alphapose [8] to extract the human pose and motion information to learn more information about the music.

Some works [18] have been done about music generation according to the video, but it may be hard and needs a lot data. For Visual rhythm prediction [18], they generate a simply drumbeat which can be represented by a binary string. It is not what we want to achieve, because music has much more colorful information than simply drumbeat. Undoubtedly, such colorful and complicated information bring much more trouble and it is really challengeable. So we decide to build a match system, that if you give a video, we can select a proper music from a list of music. It just like a recommendation system.

To handle this problem, we propose our DVAMN model, which base on Gated recurrent unit [6] and Multiple-frame Fully Connected layer(Sec. 3.3.2). The DVAMN use the pose of human being in the video to predict a sequence of music feature. Then it can be compare with musics in the database and find the best one. To extract the useful feature and remove the noise, we use VGGish [11, 12] to extract the

music feature for every 0.2 second with 1s window for our data as ground truth.

Since there is not much previous work and research on our topic, we need to build a new model and dataset. So we decide to focus on the dancing videos. Based on Youtube-8M Dataset [1], we collect our video data which was labeled 'solo dance'. we finally get more than 2500 minutes dancing video and music.

Our model and source codes are made publicly available to support reproducible research and discussion, we will keep updating if something new is discovered.

2. RELATED WORK

2.1 Visual Rhythm Prediction

Davis et al. propose a visual analogue for musical rhythm. [7] They state that the alignment of visual rhythm results in the visual information of dance. They want their model to be able to comprehends the concept of visual beats and patterns of motion. By warping visual beats and musical beats together, they can create or manipulate the visual information of dance in video.

Chen et al. state that true multimedia experience involves intrinsic interactions. [5] [3] They describe visual rhythm problem as the frequency of rhythmic events which is a more rough task. However, it can not predict the events timing very precisely and can only do the prediction basing on how intense the visual rhythm is.

The methods mentioned above are all rule based. Therefore, we aim at deep representation learning to deal with more complicated task like music recommendation and enrich the our model's expressiveness.

2.2 Audio Analysis

In machine listening, there are many kinds of features and methods. One of most important features is Mel-frequency cepstral coefficients(MFCCs) which commonly used in music informational retrieval [13] or speech recognition [9]. MFCCs are coefficients that collectively make up an mel-frequency cepstrum(MFC) which is representation of the short-term power spectrum of a sound. MFCCs can be extracted by a quick algorithm and perform well on most tasks.

The methods based on deep learning gain increasing popularity. Parascandolo et al. [16] use recurrent neural networks on sound processing. In their model, a multilabel BLSTM RNN is trained to build the relation between acoustic features and class's indicators. The result of their work outperforms previous methods.

Deep convolutional neural networks(CNNs) achieve great success on image classification. Takahashi et al. [11] apply CNNs on audio classification tasks and the model also shows a good performance.

2.3 Audio-Visual Correspondence

[14, 15, 4] are a series of research on networks making use of the concurrence or correlation between visual and audio. They train audio networks to synthesize sounds like hitting or scratching sounds from silent videos [14] or generate ambient sounds from a video frame [15]. They also build a deep audio network SoundNet [4] with a student-teacher training procedure to learn sound representations from unlabeled videos.

Zhao et al. [19] create a end-to-end learnable system that captures motions to separate musical instrument sounds. The model they build captures the natural synchronization of audio-visual signals.

In recent years, a new "Audio-Visual Correspondence" task arise and Arandjelović et al. proposes a novel and powerful method by training visual and audio networks on raw unconstrained videos with a unsupervision method [2]. A large video set is used for training in order to leverage the inherent coherence between vision and sound. And the representations generated by the model gain a great performance on two sound classification benchmarks.

3. METHOD

In this section, we will introduce our approach to learn the relationship between visual and auditory information in a human pose video. Because video and audio both have much noise, we use some method to extract the frame level video feature and audio feature. we use dancing video to train our model. For dancing video , we use human pose detection model AlphaPose [8] for frames through the whole video (Sec. 3.1). And for the music, we use VGGish [11, 12] model to get the frame level audio features(Sec. 3.2). Then we use the **Deep Video Audio Match Network(DVAMN)** to build the relationshap of these features(Sec. 3.3).

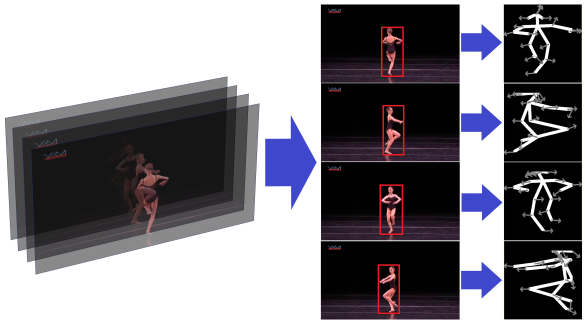

Figure 2: Pose data Pipeline: From a human pose video we capture frames with 10 fps, then extract human being's pose and do normalization on it to get the white skeleton showed in the figure. Besides, we need to get residual of human being's key points as motion information showed by gray arrows in the figure

3.1 Human Pose Detection

Still image captured from video has RGB information, which includes background, brightness, etc. Our task may be not so relevant with such things. To reduce the variance brought by them, we simply pay our attention to human being's action. With the help of AlphaPose[8], we can get skeleton of human being and abandon all the other parts in image to extract the effective information. Let $\mathbf{P}_i$ be the coordinates of 17 key points of human body in the i-th frame in a video.

We also need to polish the skeleton information into some more effective form. Pose is the most important part of human being action. However, it is hard to learn if we use the

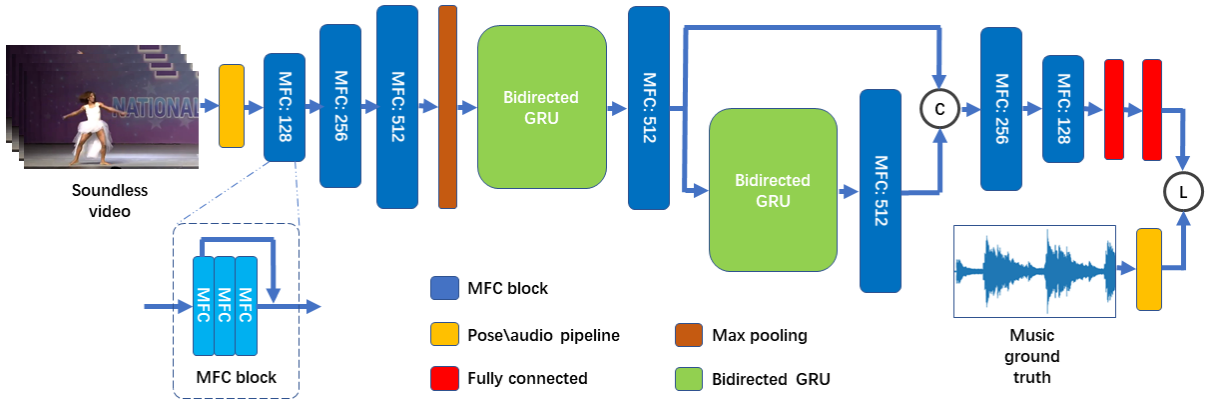

Figure 3: DVAMN overview: In DAVMN, we use MFC blocks to extend features or downsample features according to near by frames. Each MFC block have 3 MFC layers with short-cut. The MFC blocks are in different channel which will be explained latter(Sec. 4.2). And two Biderected GRU units are used to learn long term information of other frames and do a concatenation at last. A max-pooling layer is used to compress frames information to match audio features. Finally, we get a loss function according to music audio ground truth's feature.

absolute coordinate directly. Thus we have to do normalization to move the position of human being to center and zoom into a standard scale. We can get $\hat{\mathbf{P}} \in \mathbb{R}^{17*2*T}$ as key points after normalization.

Meanwhile it may bring some trouble that some information lost after normalization so that we cannot acquire the motion of human being. Thus, beside normalizing pose, we also need to do subtraction between adjacent frames of original poses to get the residual to represent movement information of human body. We consider the residual as human being's motion:

$$\mathbf{M}_i = \mathbf{P}_{i+1} - \mathbf{P}_i$$

The whole structure is performed in Fig. 2

3.2 Music Feature Embedding

The background music of human pose videos contains lots of information like rhythm and speed. Hence, we need to derive representations from video audios. As deep learning has achieved excellent performance, we choose VGGish to extract the audio embeddings. VGGish generates 128-dimensional embeddings of each audio frame based on a modified VGGNet [17] and VGGish is trained on a large audio dataset AudioSet [16].

VGGish resamples the input audio to 16kHz, then applies Short-Time Fourier Transform(STFT) to the audio with a selected window size and window hop size. The magnitudes of STFT is used for generating the log mel spectrogram. 1 second audio's features are framed into vectors of 0.96 seconds, where are 96 frames of 10 ms each and the rest 4 frames are abandoned as the head frame and tail frame.

VGGish's model part is a variant of VGGNet model. VGGish consists of four groups of convolution and maxpool layers. After convolution and maxpool layers, there are three Fully-Connected(FC) layers. And all hidden layers' nonlinearity function are Rectified Linear Unit(ReLU). Compared to the original VGGNet, VGGish uses fewer convolution/maxpool layers and smaller FC layers. The input of

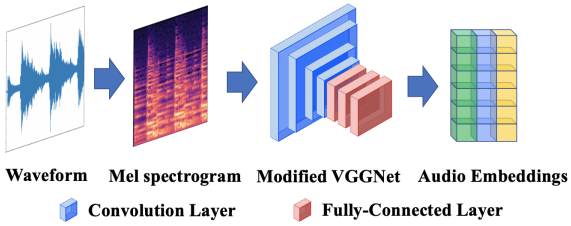

Figure 4: VGGish Pipeline: Get mel spectrogram from audio. Then input features to VGGish network and the output is audio embeddings.

VGGish's network is those feature vectors and the output is the embedding of the audio segmentation. The whole procedure is performed in Fig. 4

3.3 Matching Network

Now we have $\hat{\mathbf{P}}$, $\mathbf{M}$ and $\mathbf{A}$, we need to find a method the get the relationship between them. So we propose **Deep Video Audio Match Network(DVAMN)** to solve this problem.

The DVAMN is a seq-to-seq model that build base on Multiple-frame Fully Connected layer and Gated recurrent unit(GRU) [6]. We use the human pose and motion information as input, and predict the audio feature for each audio frames.

3.3.1 Deep Video Audio Match Network

Because the human body motion is highly related to the changes of music beats, we use pose and motion information of human to predict the corresponding music features, and find the best matched music of it. For convenience, we concatenate $\hat{\mathbf{P}}_i$ and $\mathbf{M}_i$ for each frame and get the moved-pose feature sequence $\mathbf{V_i}$ for video frames.

To predict the audio frames, we first extend the pose

features. We use Multiple-frame Fully Connected(MFC) layer(Sec. 3.3.2) to extend the pose and motion information of each pose according to the neighboring frames. And in order to get a larger receptive field, we use 3 MFC layers as an MFC block. Because our net work is a deep net work, we also use short cut like ResNet [10] in each block to avoid gradient vanishing and gradient exploding. we use several MFC blocks in our DVAMN to extend the pose feature to 512 dimension. To reduce the size of model and match the long window audio frame, we use a max pool to transform the pose feature from 10fps to 5fps after MFC blcoks.

With MFC layers, we can learn from nearby frames. But for most human pose videos, we also need long term information to know the style and rhythm of the video. Thus we use Gated recurrent unit(GRU) [6] to make our model learn the long term information. Because human pose is very complex, and the body motion not only related the previous pose, but affected by the later pose, we use bidirectional GRU to gain the both side features. To improve the performance, we use 2 GRU layers and concatenate the output of them.

After GRU, several MFC blocks are downsample to 256 dimension. And before output the result, we use a frame-wise fully connected layer and a MFC layer to resize the prediction to match the size of the audio feature. The whole structure can be see at Figure 3.

3.3.2 *Multiple-frame Fully Connected Layer*

Our inputs of each frame only have 68 dimension highly abstract after normalization, so we need to extend the feature map of the pose information. But traditional fully connected layer only can get information from one frame. It may not get enough pose motion information, which is the most important thing to generate the music beats. Convolution may solve this problem. But for coordinate features, it may not have strong relation with the near by elements. So we finally propose the MFC layer.

We use a special kernel to do the convolution in MFC layer. Using convolution kernels to have a sight among a few of continue frames can synthesis some motion information in the frame data stream. But the features in same frame do not have a strong spatial relationship, and convolution along the feature direction makes no sense. Thus, we just do convolution along the frame direction and make kernel to cover all the features in each frame. With this Multiple-frame Fully Connected layer, we can extract information more efficiently to help our model understand human being's motion.

3.3.3 *Loss*

We define A as the output of DVAMN network, and A_t as the ground truth of audio feature. Then we use *logcosh* function to construct our loss function:

$$L = log(cosh(|A - A_t|))$$

It is simple, but work. If few dimension of A is not match A_t well, it will not lead to a very large loss. But when A is close to A_t, the reduce of L will be slow and cautious.

4. EXPERIMENT

4.1 Dataset

We collect 968 videos with label 'Solo Dance' in Youtube-8M [1]. By By capturing 10 frames for every second from these videos, we get 1518716 frames with well polished pose and motion data. Then, we fix the length of video to 12s, and select 706657 eligible fragments with 120 continuous frames.

And for video of these fragments, we extract the corresponding audio feature for every 0.2s with VGGish. And because the VGGish calculate the feature for 1 second every time, the audio is overlapped.

Finally, we divide data with ratio of 4 : 1 to generate training set and validation set.

4.2 Implement Details

For 68 dimension pose and motion input, we used 3 MFC blocks with output size are 128, 256, 512. And after 2 GRU layers with output size 512, we used 2 MFC blocks downsample the output to 256 dimension. Then a frame-wise fully connected layer were followed to downsample it to 128 dimension. Finally used a MFC layer to reshape the size of output corresponding to the audio feature. By the way, before GRU layers, we used a maxpool to reduce 10fps video to 5fps to match the audio feature.

For all MFC layers, we used bias and Rectified Linear Unit(ReLU) activation function. We used Nadam optimizer with learn rate 0.0002, $\beta_1 = 0.9$, $\beta_2 = 0.999$. The batch size was setted to 64. we used NVIDIA 2080Ti to train our model.

Our model and source codes will be publicly available to support reproducible research and discussion.

4.3 Result

We evaluate our model by the loss value, Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). Because there are few works focus on dancing video and music matching, so we just compare the MFC layer and traditional fully connected layer.

With all fragments of validation set, we get the mean value of Loss, RMSE and MAE in Table 1.

Table 1: Results of NVAMN

	Loss	RMSE	MAE
DVAMN(FC)	45.78	59.85	46.31
DVAMN(MFC)	41.87	57.34	42.20

We can see MFC layer help a lot in matching problem for video and music. In the future, we may test our model on a larger dataset and use some new metrics to measure the error of matching.

5. CONCLUSION

In this paper, we propose a matching model between soundless video and music called DVAMN. We represent the video information with detailed human being's body pose and residual between nearby frame as human being's motion. We also extract some features for audio to represent its style. Our network mainly bases on Multiple-frame Fully Connected layers and GRU to extract information well from frame stream and make it be sequence to sequence. After training and validation on Youtube-8M data, we can do the matching very well, and we build a music library instead of original sound of videos to test.

6. REFERENCES

[1] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *CoRR*, abs/1609.08675, 2016.

[2] R. Arandjelović and A. Zisserman. Look, Listen and Learn. *arXiv e-prints*, page arXiv:1705.08168, May 2017.

[3] C. Argüello and M. Iregui. Exploring rhythmic patterns in dance movements by video analysis. In V. G. Duffy, editor, *Digital Human Modeling: Applications in Health, Safety, Ergonomics and Risk Management*, pages 123–131, Cham, 2016. Springer International Publishing.

[4] Y. Aytar, C. Vondrick, and A. Torralba. SoundNet: Learning Sound Representations from Unlabeled Video. *arXiv e-prints*, page arXiv:1610.09001, Oct 2016.

[5] T. P. Chen, C. Chen, P. Popp, and B. Coover. Visual rhythm detection and its applications in interactive multimedia. *IEEE MultiMedia*, 18(1):88–95, Jan 2011.

[6] K. Cho, B. van Merrienboer, Ç. Gülçehre, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *CoRR*, abs/1406.1078, 2014.

[7] A. Davis and M. Agrawala. Visual rhythm and beat. *ACM Trans. Graph.*, 37(4):122:1–122:11, July 2018.

[8] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu. RMPE: Regional multi-person pose estimation. In *ICCV*, 2017.

[9] T. Ganchev, N. Fakotakis, and G. Kokkinakis. Comparative evaluation of various mfcc implementations on the speaker verification task. In *Proceedings of the SPECOM*, volume 1, pages 191–194, 2005.

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[11] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. Channing Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson. CNN Architectures for Large-Scale Audio Classification. *arXiv e-prints*, page arXiv:1609.09430, Sep 2016.

[12] A. Kumar and B. Raj. Deep CNN Framework for Audio Event Recognition using Weakly Labeled Web Data. *arXiv e-prints*, page arXiv:1707.02530, Jul 2017.

[13] M. Müller. *Information retrieval for music and motion*, volume 2. Springer, 2007.

[14] A. Owens, P. Isola, J. McDermott, A. Torralba, E. H. Adelson, and W. T. Freeman. Visually Indicated Sounds. *arXiv e-prints*, page arXiv:1512.08512, Dec 2015.

[15] A. Owens, J. Wu, J. H. McDermott, W. T. Freeman, and A. Torralba. Ambient Sound Provides Supervision for Visual Learning. *arXiv e-prints*, page arXiv:1608.07017, Aug 2016.

[16] G. Parascandolo, H. Huttunen, and T. Virtanen. Recurrent neural networks for polyphonic sound event detection in real life recordings. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6440–6444. IEEE, 2016.

[17] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv e-prints*, page arXiv:1409.1556, Sep 2014.

[18] Y. Xie, h. Wang, Y. Hao, and Z. Xu. Visual rhythm prediction with feature aligning network. 2019.

[19] H. Zhao, C. Gan, W.-C. Ma, and A. Torralba. The sound of motions. *arXiv preprint arXiv:1904.05979*, 2019.

Authors' background

Your Name	Title*	Research Field	Personal website
Peiming Yang	Undergraduate student	Computer Vision	
Jiuqi Liu	Undergraduate student	Computer Vision	
Leshang Jin	Undergraduate student	Natural Language Processing	