SEETHEVOICE: LEARNING FROM MUSIC TO VISUAL STORYTELLING OF SHOTS

Wen-Li Wei [1], *Jen-Chun Lin*[1,3], *Tyng-Luh Liu*[1],
Yi-Hsuan Yang[2], *Hsin-Min Wang*[1], *Hsiao-Rong Tyan*[4], *and Hong-Yuan Mark Liao*[1]

[1]Institute of Information Science and [2]Research Center for IT Innovation, Academia Sinica, Taiwan
[3]Department of Electrical Engineering, Yuan Ze University, Taiwan
[4]Dept. of Information and Computer Engineering, Chung Yuan Christian Univ., Taiwan
{lilijinjin, jenchunlin}@gmail.com, liutyng@iis.sinica.edu.tw,
yang@citi.sinica.edu.tw, whm@iis.sinica.edu.tw, tyan@ice.cycu.edu.tw, liao@iis.sinica.edu.tw

ABSTRACT

Types of shots in the language of film are considered the key elements used by a director for visual storytelling. In filming a musical performance, manipulating shots could stimulate desired effects such as manifesting the emotion or deepening the atmosphere. However, while the visual storytelling technique is often employed in creating professional recordings of a live concert, audience recordings of the same event often lack such sophisticated manipulations. Thus it would be useful to have a versatile system that can perform video mashup to create a refined video from such amateur clips. To this end, we propose to translate the music into a near-professional shot (type) sequence by learning the relation between music and visual storytelling of shots. The resulting shot sequence can then be used to better portray the visual storytelling of a song and guide the concert video mashup process. Our method introduces a novel probabilistic-based fusion approach, named as multi-resolution fused recurrent neural networks (MF-RNNs) with film-language, which integrates multi-resolution fused RNNs and a film-language model for boosting the translation performance. The results from objective and subjective experiments demonstrate that MF-RNNs with film-language can generate an appealing shot sequence with better viewing experience.

Index Terms— Language of film, types of shots, live concert, recurrent neural networks

1. INTRODUCTION

Video mashup [1, 2, 3] is a process to generate a complete and high-quality video from a set of less-satisfactory clips, say, taken by audiences at different locations of the same event. To ensure the editing quality is close to *professional*, it is necessary to consider the *visual storytelling* of shots and explore the relation between music and visual storytelling. In the language of film, a shot (type) is a fundamental element of visual storytelling [4, 5]. It includes six types of shots, as described in Table 1. For example, the medium close-up emphasizes the subject from shoulders to the top of head, and the medium long shot instead highlights the subject from knees to the top of head completely. In film-making, an experienced director is skillful at switching different shot types to convey the emotion, ideas, and art through a visual storytelling process [4, 5]. Analogously, understanding how to properly employ shots is crucial in carrying out a concert video mashup process.

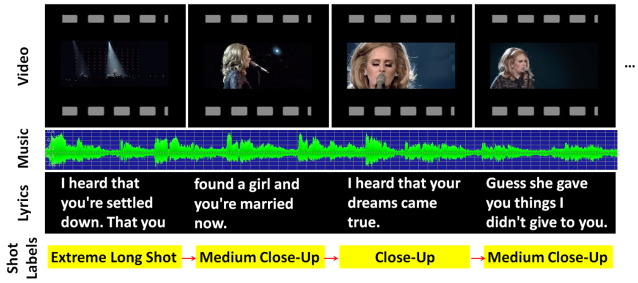

Fig. 1. An example of visual storytelling of an official concert video for the song Someone Like You by Adele live at Royal Albert Hall 2011. The video demo can be found at https://sites.google.com/site/music2shots/.

Exploring the relation between music and visual storytelling is also pivotal in making the resulting film/video strike a deep chord. In psychology, it is argued that a musical experience may evoke emotions when a listener conjures up images of things and events that have never occurred [6]. Indeed music and video are often accompanied to complement each other to enhance emotional resonance in official music videos, movies, television programs, and concert videos. Video can deepen our emotional response to music and expand the storytelling potential of songs, while music can heighten our emotional reaction to video and intensify our understanding of visuals. Figure 1 illustrates how visual storytelling is achieved in making an official concert video. In particular, it shows that the director sequentially uses the extreme long shot, medium close-up, close-up, and medium close-up in the beginning of the song to express the emotion and expand the storytelling of song. Motivated by the compelling relevance, we aim to

Table 1. The definition of six types of shots [4, 5].

Types of Shots	Description	Example
Close-Up (CU)	A Close-Up is used to show emotion on the subject's face. That is, the ***face occupies most of the screen (image)***.	
Medium Close-Up (MCU)	A Medium Close-Up contains a subject's ***head and shoulders*** completely.	
Medium Shot (MS)	A Medium Shot contains a subject from the ***waist to the top of the head***.	
Medium Long Shot (MLS)	A Medium Long Shot would contain a subject from his/her ***knees to the top of the head***.	
Long Shot (LS)	A Long Shot would contain a subject's entire body from the ***top of the head to the bottom of the feet***.	
Extreme Long Shot (XLS)	An Extreme Long Shot covers ***a large area or landscape***. It would be hard to see any reactions/emotion from people in the shot since they are too far away.	

develop techniques to translate music into a near-professional shot sequence for better portraying the visual storytelling of a song to guide the concert video mashup process.

The task to interpret the music with an appropriate and near-professional shot sequence is challenging. Its difficulty can be further complicated by considering the various musical elements (*e.g.*, genre, emotion, *etc.*) and the style of the director. To the best of our knowledge, research on translating music into visual storytelling of shots has not been actively explored. The most relevant works might be those for video captioning [7, 8, 9], which translates an *image* sequence (visual type information) into a *word* sequence (sentence). The sequence-to-sequence modeling is often implemented with recurrent neural network (RNN) frameworks, such as those with long short-term memory (LSTM) units [10]. However, as we observe in our experiments, existing RNN-based models may not effectively learn the long-term dependencies over large and varied temporal intervals in the music video.

To handle the aforementioned difficulties, we introduce a novel probabilistic fusion model, termed as multi-resolution fused RNNs (MF-RNNs) with film-language. (See Figure 2.) The main idea is to simultaneously consider various temporal resolution units, including low-resolution, middle-resolution, and high-resolution, to capture the temporal structure with different ranges. For instance, we can use the low-resolution unit (4-sec per frame) to capture the long-range temporal structure, and the high-resolution unit (1-sec per frame) for the detailed local temporal structure. Such an idea is consistent with the concept of storyboard [11]. That is, in film-making/video production, the director will first sketch out a rough shot (*i.e.*, low-resolution unit) according to the script and then expand the storytelling by adding more high-contrast shots (*i.e.*, high-resolution unit). Driven by these findings, the proposed MF-RNNs is constructed to not only model the various temporal resolutions independently but also relate the statistical dependencies between them. In addition, statistical shot transitions based on a film-language model are further integrated with the MF-RNNs to boost the translation performance. In Figure 2, we sketch an overview of the proposed music to visual storytelling of shots framework. First, an input music signal is framed with various temporal resolution clips and transformed into log-scaled mel-spectrogram. To better represent the music content and to build the relation between the music and the shot, we encode the music as the composition of music tags [12], and then construct a feature representation from the rich feature hierarchies of convolutional neural network (CNN) to model the shot types. Finally, the MF-RNNs with film-language is used to translate the music into shot sequence (defined in the language-of-film). We note that besides the six types of shots listed in Table 1, we further consider two additional variants, each of which focuses on either the audience shot (ADS) or musical instrument shot (MIS), to enrich the visual storytelling in a concert video. In summary, our main contributions include:

- Our work is the first to translate music into a shot sequence for understanding the visual storytelling in concert videos and facilitating the video mashup process.
- We propose a novel probabilistic fusion model, MF-RNNs with film-language, to improve the translation performance.

2. MUSIC REPRESENTATIONS

Each music content is represented by applying JY-net [12] to obtain the hierarchical features of music tags for log-scaled mel-spectrogram with low, middle, and high temporal resolution unit (4-sec, 2-sec, and 1-sec per frame), respectively. The JY-net is trained on the MagnaTagATune music dataset [13], which includes 25,863 29-second clips over 188 tags. The class of genre includes the tags of jazz, new age, etc., while the class of musical instrument includes the tags of guitar, flute, etc. Using forward propagation in JY-net, we extract features from the output layer and the two fully-connected layers as the music tag representations for each input frame with 128-bin log-scaled mel-spectrogram, where the feature dimensions are 188-D, 512-D and 512-D, respectively. These features are concatenated to construct a combined vector of (188+512+512)-D for each frame as the music representation.

To observe whether the music tag helps in translating shots, we compare each of the music tag over the eight shot

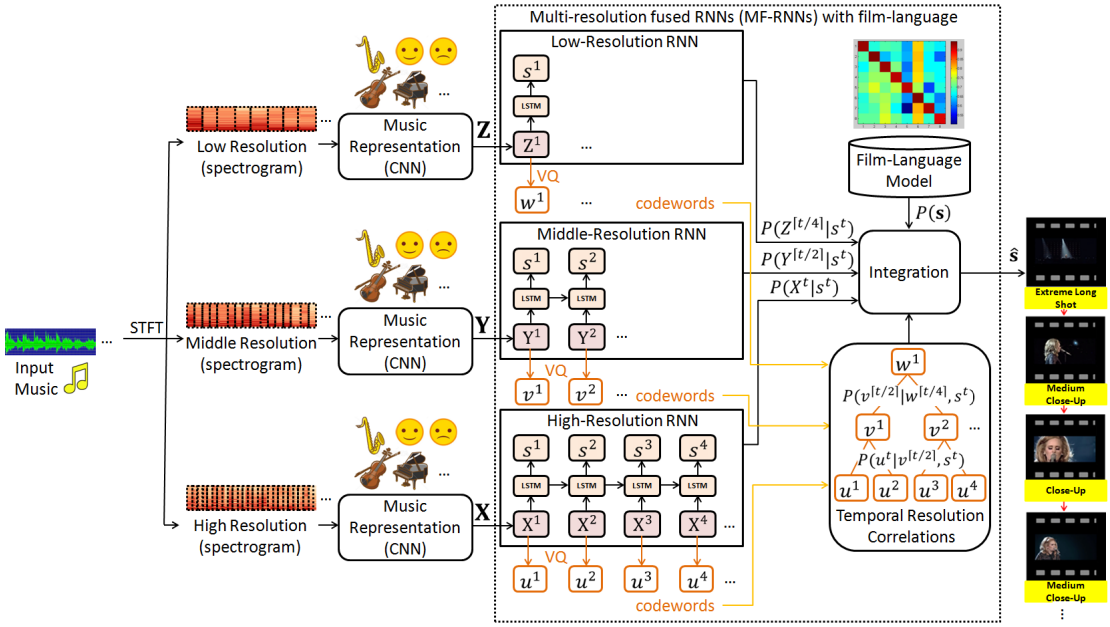

Fig. 2. Illustration of the proposed MF-RNNs with film-language framework for music to visual storytelling of shots translation.

types from (average) music tag distribution of the output layer. Notice that only the highest score for each music tag is kept in the corresponding shot type. Figure 3 shows these plots obtained from our training data. The substantial variations suggest the usefulness of adopting music tags as features for translating the shot types. For example, one can readily infer that the medium shot emphasizes (*i.e.*, with the highest score) the music tags of technical, fast, and electronic, and the audience shot instead highlights the music tags of vocal, noise, and weird. The insightful information indicates the representative music tags between different shot types.

3. MUSIC TO VISUAL STORYTELLING OF SHOTS

To model the shot-based visual storytelling of a piece of music, MF-RNNs is constructed to not only individually model the extracted music representations of various temporal resolutions but also relate the statistical dependencies between them. We further integrate MF-RNNs with a film-language model to translate the music into a near-professional shot sequence as shown in Figure 2. Given the music representation sequences of high, middle, and low temporal resolutions $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{Z}$, the task to decide the shot sequence $\hat{\mathbf{s}}$, corresponding to the high temporal resolution, can be casted as follows:

$$\hat{\mathbf{s}} = \arg\max_{\mathbf{s}\in S} P(\mathbf{s}\,|\,\mathbf{X},\mathbf{Y},\mathbf{Z}) \propto P(\mathbf{X},\mathbf{Y},\mathbf{Z}\,|\,\mathbf{s})P(\mathbf{s}) \quad (1)$$

where S denotes the set of all possible high temporal resolution shot sequences. $P(\mathbf{X},\mathbf{Y},\mathbf{Z}\,|\,\mathbf{s})$ is the joint probability of the representation sequences $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{Z}$, and $P(\mathbf{s})$ is the *a priori* probability of film-language, which is computed through shot transitions from official concert videos. A bigram language model is used to estimate the film-language with

$$P(\mathbf{s}) = P(s^1)\prod_{t=2}^{T} P(s^t\,|\,s^{t-1}) \quad (2)$$

where T denotes the total number of frames of high temporal resolution sequence. Thus, to yield the optimal shot sequence $\hat{\mathbf{s}}$ in (1), we are left to estimate $P(\mathbf{X},\mathbf{Y},\mathbf{Z}\,|\,\mathbf{s})$.

In practice, estimating $P(\mathbf{X},\mathbf{Y},\mathbf{Z}\,|\,\mathbf{s})$ is difficult or even unfeasible. An intuitive way to get around is to assume that the temporal resolutions $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$ are conditionally independent. However, such an assumption may lose the information of the correlation between temporal resolutions at different levels, which can be crucial for translating the visual storytelling of music. To tackle the issue, we adopt the method by Pan *et al*. [14], which was previously developed to estimate the joint probability of multi-sensory signals based on the principle of maximum entropy. It reduces the dimensionality of the joint data space into the practical range. Afterwards, the joint probability is estimated by combining the correlations (from mapped features) and the likelihoods (from original features) of multi-sensory features. We modify their approach to estimate the joint probability $P(\mathbf{X},\mathbf{Y},\mathbf{Z}\,|\,\mathbf{s})$ of multi-resolution feature representations. It follows from [14] that $P(\mathbf{X},\mathbf{Y},\mathbf{Z}\,|\,\mathbf{s})$ in (1) can be defined by

$$P(\mathbf{X},\mathbf{Y},\mathbf{Z}\,|\,\mathbf{s}) \overset{def}{=} \prod_{t=1}^{T} P(X^t\,|\,s^t)P(Y^{\lceil \frac{t}{2} \rceil}\,|\,s^t)P(Z^{\lceil \frac{t}{4} \rceil}\,|\,s^t) \times \frac{P(u^t, v^{\lceil \frac{t}{2} \rceil}, w^{\lceil \frac{t}{4} \rceil}\,|\,s^t)}{P(u^t\,|\,s^t)P(v^{\lceil \frac{t}{2} \rceil}\,|\,s^t)P(w^{\lceil \frac{t}{4} \rceil}\,|\,s^t)} \quad (3)$$

where $P(X^t\,|\,s^t)$, $P(Y^{\lceil \frac{t}{2} \rceil}\,|\,s^t)$, and $P(Z^{\lceil \frac{t}{4} \rceil}\,|\,s^t)$ are the likelihoods of music representations of high, middle, and low temporal resolutions, respectively. The symbol $\lceil\ \rceil$ represents the ceiling function. It is used to locate a frame of low or middle temporal resolution that corresponds to the high temporal

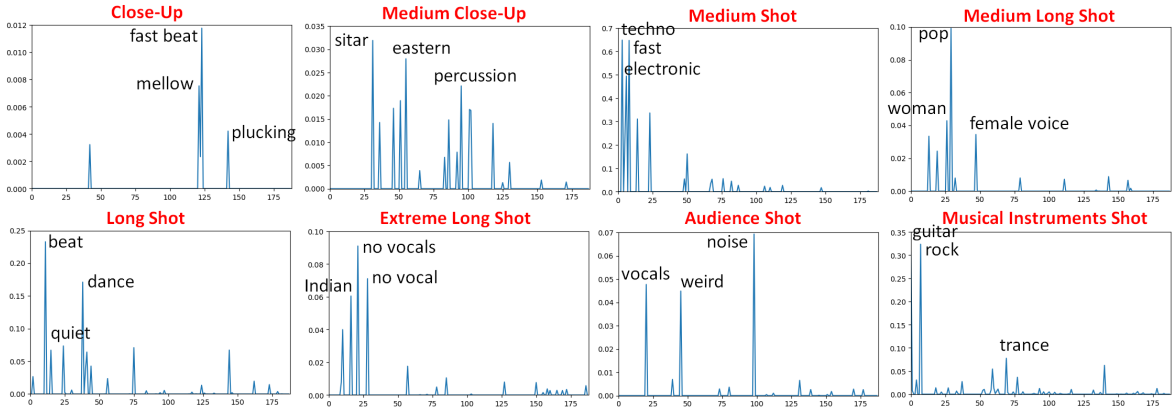

Fig. 3. The representative music tags in the high resolution unit statistics. Vertical and horizontal axes represent the average posterior probability (from JY-net) and the 188 music tags, respectively. For each shot type, the top-3 music tags are listed.

resolution. For example, when t=1, the first frame of middle ($\lceil \frac{t}{2} \rceil$) and that of low ($\lceil \frac{t}{4} \rceil$) temporal resolutions are located to help translate the shot type of the first frame of high temporal resolution. Such correspondences are restricted within a multiple of three temporal resolutions (*i.e.*, 4-sec, 2-sec, and 1-sec per frame). The remaining in the right hand side of (3) is the correlation term of multiple temporal resolutions. Because the representations X, Y, and Z are continuous values, it is unfeasible to collect sufficient training data to construct statistical dependencies among the three, under the joint condition s. We use k-means clustering to construct a codebook and perform vector quantization to represent X, Y, and Z by their corresponding codeword, say, u, v, and w as in (3).

To further simplify the statistical dependency estimation, we assume that the dependencies among temporal resolutions are hierarchical (cf. Figure 2). Hence, (3) can be rewritten as

$$P(\mathbf{X},\mathbf{Y},\mathbf{Z}\,|\,\mathbf{s}) \approx \prod_{t=1}^{T} P(X^t\,|\,s^t)P(Y^{\lceil \frac{t}{2} \rceil}\,|\,s^t)P(Z^{\lceil \frac{t}{4} \rceil}\,|\,s^t) \times \frac{P(u^t\,|\,v^{\lceil \frac{t}{2} \rceil},s^t)P(v^{\lceil \frac{t}{2} \rceil}\,|\,w^{\lceil \frac{t}{4} \rceil},s^t)}{P(u^t\,|\,s^t)P(v^{\lceil \frac{t}{2} \rceil}\,|\,s^t)}. \quad (4)$$

Our implementation uses a dedicated RNN to separately model each of the three temporal resolutions and estimate their likelihood. We remark that the output by an RNN is a posterior probability. Hence, we approximate the first likelihood by $P(s^t\,|\,X^t)/P(s^t)$, the second by $P(s^t\,|\,Y^{\lceil \frac{t}{2} \rceil})/P(s^t)$, and the third by $P(s^t\,|\,Z^{\lceil \frac{t}{4} \rceil})/P(s^t)$.

With (2)-(4), we arrive at how $P(\mathbf{s}\,|\,\mathbf{X},\mathbf{Y},\mathbf{Z})$ in (1) is evaluated. Specifically, in the test phase, $P(\mathbf{s}\,|\,\mathbf{X},\mathbf{Y},\mathbf{Z})$, the posterior probability of each shot sequence $\mathbf{s}$, can be inferred by combining the outputs of RNNs, the correlation between music representations of three temporal resolutions, and the film-language. Then, the shot sequence $\mathbf{s}$ with maximum posterior probability is selected as the translation result $\hat{\mathbf{s}}$. In the training phase, each temporal resolution RNN (with LSTM units) is trained by using the back-propagation through time (BPTT) algorithm with the objective of softmax cross entropy. $P(u^t\,|\,v^{\lceil \frac{t}{2} \rceil},s^t)$, $P(v^{\lceil \frac{t}{2} \rceil}\,|\,w^{\lceil \frac{t}{4} \rceil},s^t)$, $P(u^t\,|\,s^t)$, and $P(v^{\lceil \frac{t}{2} \rceil}\,|\,s^t)$ in (4) are calculated by statistical co-occurrence dependencies over all the training data.

Table 2. The distribution of collected data.

Data Type	# Concert	# Video	# Frame (H-resolution)	# Frame (M-resolution)	# Frame (L-Resolution)
Training	36	45	12,300	6,150	3,075
Validation	5	5	1,331	-	-
Testing	10	10	2,485	-	-

4. EXPERIMENTAL RESULTS

To demonstrate the effectiveness of MF-RNNs with film-language, we conduct experiments on a set of official concert videos downloaded from YouTube. 60 official concert videos[1], each of which belongs to a complete song, are collected from 51 live concerts, with a total of 452,848 annotated images (shot types). Since manually labeling 452,848 images to their corresponding shot type is very time-consuming, we use the newly-proposed shot classification method, namely, error weighted deep cross-correlation model (EW-Deep-CCM) [15] to automatically classify the eight shot types, as defined in this study. The shot (type) classification result is first regarded as the ground truth label for each image, and then used to generate the ground truth label for each temporal resolution unit. Specifically, for each low-, middle-, and high-resolution unit of music signal, the ground truth shot label is determined by the majority voting of shot labels of images that corresponding to each unit. Among the collected data, 45 official concert videos from 36 live concerts are included for training. We use the remaining 5 and 10 official concert videos from 5 and 10 live concerts for validation and testing, respectively. (See Table 2.)

We compare the performance of the proposed MF-RNNs (with film-language) with a conventional RNN approach that is trained with high-, middle-, and low-resolution unit, re-

[1]The video links are at https://sites.google.com/site/music2shots/dataset.

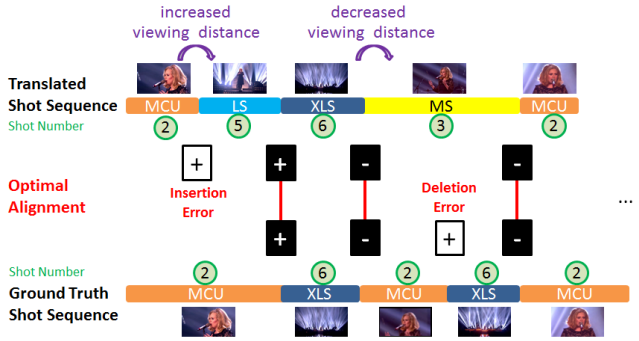

Fig. 4. Illustration of the dynamic programming-derived shot change trend-aligned path for a ground truth shot sequence and a translated shot sequence. The red line and black block denote the aligned shot change trend.

spectively, as well as a decision-level fusion approach [16]. Specifically, for high-resolution unit RNN (hRNN), features extracted from the output layer and the two fully-connected layers of JY-net are concatenated to form a combined feature vector of (188+512+512)-D, and then fed into the RNN for shot sequence translation. A similar procedure is used for middle-resolution unit RNN (mRNN) and low-resolution unit RNN (lRNN) translations. Since our goal is to obtain the shot sequence of high-resolution unit translation, the translation result of mRNN (or lRNN) needs up-sampling to match the number of frames of high-resolution unit. To this end, we simply repeat the middle-resolution unit (or low-resolution unit) translations locally such that the final output has the same number of frames as the high-resolution unit. For decision-level fusion, we use entropy as a decision metric [17] to measure the confidence from the outputs of three resolution unit RNNs, and the shot translation result of each frame of high-resolution unit is individually determined by one out of three resolution unit RNNs to prefer entropy. The RNN structure and parameter setting with respect to the hidden layers, neurons, learning rate, mini-batch size, the number of epochs for MF-RNNs with film-language and decision-level fusion are determined when their best performance is achieved for the validation set. That is, the respective meta-parameters of each method are optimally set for the validation error.

In the experiments, three metrics are used for performance evaluation, including the trend of shot change (TSC), the distance of shot types (DST), and the duration of shot (DS). For TSC measurement, similar to the word accuracy widely used in speech recognition [18], we define TSC accuracy as

$$TSC Accuracy = \frac{N - D - S - I}{N} \times 100\% \qquad (5)$$

where N is the total number of shot changes in the ground truth shot sequence; D, S, and I denote the deletion errors, substitution errors, and insertion errors of the shot change. Since the types of shots in the language-of-film are based on the viewing distance (from close view to far view, cf. Table 1), we define two trends of shot changes including increased viewing distance "+" and decreased viewing distance

Table 3. Average TSC, DST, and DS of the five approaches.

Metrics \ Method	Ground Truth	hRNN	mRNN	lRNN	Decision-Level Fusion	Ours
Avg. TSC (%)	-	66.09	61.25	55.17	60.77	69.23
Avg. DST	-	1.42	1.33	1.48	1.47	1.32
Avg. DS (sec.)	4.59	4.09	6.43	8.01	3.80	4.59

"-" to measure the TSC accuracy between the ground truth and the translated shot sequences. For example, as shown in Figure 4, the type of shot from MCU to LS is categorized into the trend of increased viewing distance "+", while from XLS to MS is categorized into the trend of decreased viewing distance "-". Afterwards, a dynamic programming algorithm is employed to find an optimal alignment of shot change trends between the ground truth shot sequence and the translated shot sequence [18], to obtain the D, S, and I for TSC accuracy estimation. To further validate MF-RNNs with film-language, we extend the performance evaluation by measuring DST. To this end, we number the eight types of shots as one to eight according to the viewing distance. And, the numbered shots (see Figure 4) between aligned shot trends are estimated one-by-one by setting DST as the absolute error measurement. Finally, DS is considered for estimating the time length (sec.) in each ground truth (or translated) shot. We report the average TSC, DST, and DS over the testing set.

Table 3 shows the performance of the five approaches for all the mentioned metrics. In comparing the RNNs of the three temporal resolutions, hRNN is superior to mRNN and lRNN in almost all the three metrics. This could be due to that the use of an up-sampling technique in mRNN and lRNN could lose accuracy locally in translating the high-resolution unit. Regarding the decision-level fusion, it is evident from Table 3 that the fusion improves mRNN or lRNN by combining the three resolution unit RNNs. However, the performance is still worse than hRNN. This may be because the shot type estimated from long-range temporal unit (*i.e.*, middle- and low-resolution units) may not accurately reflect the behavior of the detailed local (high-resolution) temporal unit. Thus, the shot estimated from mRNN or lRNN, even with high confidence in entropy estimation, does not imply that it can be directly regarded as the result of high-resolution unit. Furthermore, such a hard-decision fusion approach may lose the correlation among three temporal resolution units, which may be crucial in music-to-shot sequence translations. Based on the above analyses, MF-RNNs with film-language is designed to not only individually model the extracted music representations of three temporal resolution units but also construct the statistical dependencies between them as well as integrating with a film-language model to improve the translation performance. Compared with the decision-level fusion and hRNN approaches, the results support that MF-RNNs with film-language can capture longer temporal-dependencies by appropriately introducing the middle- and low-resolution units in a soft decision way, and achieves the best overall performance over all three metrics.

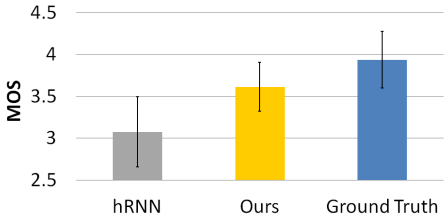

Fig. 5. Results of the subjective evaluation.

Subjective evaluation[2] in terms of 5-point mean opinion score (MOS) is conducted on five concert video sets. Each concert video set includes the original official concert video (the ground truth) and the concert videos generated by using shot (type) sequences that are translated from the hRNN and the proposed MF-RNNs with film-language, respectively. The three concert videos are provided in a random order. After viewing each concert video, the subject is asked to rate a MOS for the indicator, "whether the visual storytelling of shots matches the music?". Each concert video is evaluated by 16 subjects (recruited from the authors laboratory and university). The average MOS over all concert videos and subjects is shown in Figure 5. It is clear that the proposed MF-RNNs with film-language outperforms hRNN. The results reveal that modeling the relation among low-, middle-, and high-resolution units and considering the film-language can indeed generate more attractive visual storytelling (shots) that enhances subjects' viewing and listening experiences. The results also show that the MOS of the concert videos generated by the proposed MF-RNNs with film-language is quite close to that of the ground truth concert videos, which is really encouraging. For more details about the experiments, we also provide video demos at https://sites.google.com/site/music2shots/demo.

5. CONCLUSIONS

We have introduced a novel probabilistic-based fusion approach, named as multi-resolution fused recurrent neural networks (MF-RNNs) with film-language, to perform music-to-shot sequence translation for concert videos. Our experiments on both objective and subjective evaluations have demonstrated that MF-RNNs with film-language outperforms the conventional RNN approaches and a popular fusion strategy, and can offer satisfactory music-to-shot sequence translation results. Leveraging with these promising outcomes, our future work along this line would focus on addressing the challenging issues of enhancing video and audio qualities, which is crucial in developing a high quality video mashup system.

Acknowledgment: This work was supported in part by MOST grants 107-2634-F-001-003 and 107-2634-F-001-002.

[2]MOS results and videos are at https://sites.google.com/site/music2shots/mos.

6. REFERENCES

[1] P. Shrestha, H. Weda, M. Barbieri, E. HL Aarts, et al., "Automatic mashup generation from multiple-camera concert recordings," in *Proc. ACM Multimedia*, 2010, pp. 541–550.

[2] M. K. Saini, R. Gadde, S.c. Yan, and W.T. Ooi, "MoViMash: online mobile video mashup," in *Proc. ACM Multimedia*, 2012, pp. 139–148.

[3] Y. Wu, T. Mei, Y.-Q. Xu, N. Yu, and S.P. Li, "MoVieUp: Automatic mobile video mashup," *IEEE TCSVT*, vol. 25, no. 12, pp. 1941–1954, 2015.

[4] D. Andrews, *Digital overdrive: Communications & multimedia technology*, Digital Overdrive, 2011.

[5] G. Mercado, *The filmmaker's eye: Learning (and breaking) the rules of cinematic composition*, Taylor & Francis, 2010.

[6] P. N. Juslin and D. Västfjäll, "Emotional responses to music: the need to consider underlying mechanisms," *Behavioral and Brain Sciences*, vol. 31, no. 5, pp. 559–621, 2008.

[7] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and Saenko K., "Sequence to sequence - video to text," in *Proc. ICCV*, 2015, pp. 4534–4542.

[8] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. CVPR*, 2015, pp. 2625–2634.

[9] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. Mooney, and Saenko K., "Translating videos to natural language using deep recurrent neural networks," in *Proc. NAACL*, 2015, pp. 1495–1504.

[10] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[11] J. Hart, *The Art of the Storyboard: A Filmmakers Introduction*, Elsevier, 2008.

[12] J.-Y. Liu and Y.-H. Yang, "Event localization in music auto-tagging," in *Proc. ACM Multimedia*, 2016, pp. 1048–1057.

[13] E. Law, K. West, M. Mandel, M. Bay, and J. S. Downie, "Evaluation of algorithms using games : The case of music tagging," in *Proc. ISMIR*, 2009, pp. 387–392.

[14] H. Pan, Liang Z.-P., and T. S. Huang, "Estimation of the joint probability of multisensory signals," *Pattern Recognition Letters*, vol. 22, no. 13, pp. 1431–1437, 2001.

[15] W.-L. Wei, J.-C. Lin, T.-L. Liu, Y.-H. Yang, H.-M. Wang, H.-R. Tyan, and H.-Y. M. Liao, "Deep-net fusion to classify shots in concert videos," in *Proc. ICASSP*, 2017, pp. 1383–1387.

[16] C.-H. Wu, J.-C. Lin, and W.-L. Wei, "Survey on audiovisual emotion recognition: databases, features, and data fusion strategies," *APSIPA TSIP*, vol. 3, no. e12, pp. 1–18, 2014.

[17] S. Teerapittayanon, B. McDanel, and H.T. Kung, "BranchyNet: Fast inference via early exiting from deep neural networks," in *Proc. ICPR*, 2016, pp. 2464–2469.

[18] S. J. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book Version 3.4*, Cambridge Univ. Press, 2006.