DEEP-NET FUSION TO CLASSIFY SHOTS IN CONCERT VIDEOS

Wen-Li Wei ⋆, Jen-Chun Lin⋆, Tyng-Luh Liu⋆, Yi-Hsuan Yang†,
Hsin-Min Wang⋆, Hsiao-Rong Tyan‡, and Hong-Yuan Mark Liao⋆

⋆Institute of Information Science, Academia Sinica, Taiwan
†Research Center for Information Technology Innovation, Academia Sinica, Taiwan
‡Dept. of Information and Computer Engineering, Chung Yuan Christian Univ., Taiwan

ABSTRACT

Varying types of shots is a fundamental element in the language of film, commonly used by a visual storytelling director to convey the emotion, ideas, and art. To classify such types of shots from images, we present a new framework that facilitates the intriguing task by addressing two key issues. We first focus on learning more effective features by fusing the layer-wise outputs extracted from a deep convolutional neural network (CNN), pre-trained on a large-scale dataset for object recognition. We then introduce a probabilistic fusion model, termed as error weighted deep cross-correlation model (EW-Deep-CCM), to boost the classification accuracy. Specifically, the deep neural network-based cross-correlation model (Deep-CCM) is constructed to not only model the extracted feature hierarchies of CNN independently but also relate the statistical dependencies of paired features from different layers. Then, a Bayesian error weighting scheme for classifier combination is adopted to explore the contributions from individual Deep-CCM classifiers to enhance the accuracy of shot classification. We provide extensive experimental results on a dataset of live concert videos to demonstrate the advantage of the proposed EW-Deep-CCM over existing popular fusion approaches. The video demos can be found at https://sites.google.com/site/ewdeepccm2/demo.

Index Terms— Types of shots, convolutional neural networks, live concert, language of film

1. INTRODUCTION

With the prevalence of mobile devices, people can now easily film a live concert, and create video clips of specific performance. Popular websites such as YouTube or Vimeo have further boosted the phenomenon as data sharing becomes easy. Videos of this kind, recorded by audiences at different locations of the scene, provide those who could not attend the event the opportunity to enjoy the *same* performance. However, the viewing experience is usually unpleasant in that these videos are captured with no coordination, and incompleteness or redundancy happens always. To ensure pleasant viewing/listening experience, effective combination of the videos plus a smooth "decoration" process that would generate a single yet professional audio/visual stream is indispensable.

Video mashup, an emergent research topic in multimedia, can well satisfy the above-mentioned needs. A successful mashup process has to deal with all videos captured at different locations and convert them into a complete, non-overlapping, seamless, and high-quality outcome. Though a few attempts have been proposed to deal with video mashup in recent years [1, 2, 3], they actually pay little attention to the requirements of professional video editing. That is, these methods do not explicitly account for the visual storytelling of shots defined by the language of film. We note that in the language of film [4, 5], a type of shot is defined as how much a target subject and its surrounding area can be seen. Totally, six types of shots are defined, as described in Table 1. In film-making, an experienced director is good 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. 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. Motivated by the crucial relevance, we aim to classify the types of shots (defined by the language of film) for better portraying visual storytelling in a concert video, and plan to incorporate the technique in our upcoming effort for building an automatic mashup platform.

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

Types of Shots	Description
Close-Up (CU)	A Close-Up is used to show emotion on the subject's face.
Medium Close-Up (MCU)	A Medium Close-Up contains a person's head and shoulders completely.
Medium Shot (MS)	A Medium Shot contains a person from the waist to the top of the head.
Medium Long Shot (MLS)	A Medium Long Shot would contain a person from his/her knees to the top of the head.
Long Shot (LS)	A Long Shot would contain a person'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.

To classify the types of shots from a concert video is a non-trivial task. A feasible approach needs to distinguish the differences between two different types of shots, which are mainly caused by varying the viewing distances (cf. Table 1). The difficulty can be further complicated by that the video quality of audience recordings in a live concert is often less satisfactory, due to the shakiness, blurriness, and poor illumination factors. To the best of our knowledge, research on language-of-film-based shot classification for concert videos has not been actively explored. In the past, the most relevant studies for shot classification were focused on sport videos and movies [6, 7, 8]. For example, Bagheri-Khaligh *et al*. [6] and Benini *et al*. [8] both propose to classify shot types into Close-Up, Medium

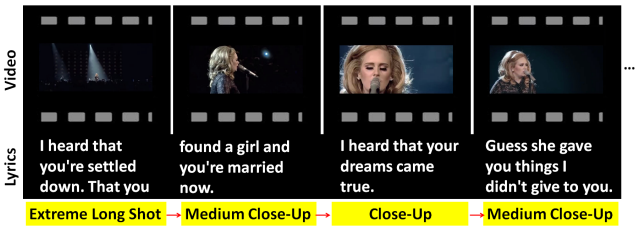

Fig. 1. An example of visual storytelling of an official concert video for the song "Someone Like You" by Adele live at Royal Albert Hall 2011.

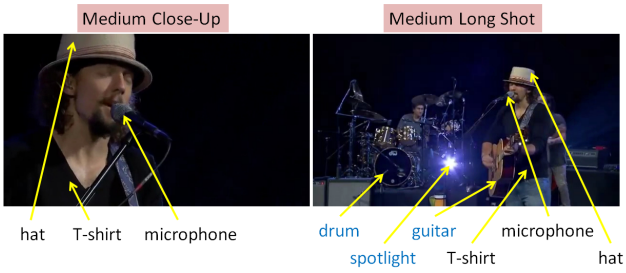

Fig. 2. Two images from an official concert video of the song "93 million miles" by Jason Mraz live at Hong Kong 2012. The left image is yielded by a MCU shot, and the right by a MLS.

shot, and Long shot in sport videos and movies, respectively. However, in their studies the three types of shots are not defined by the definitions in language of film. Also, for the purpose of visual storytelling, addressing only these three types of shots is generally not sufficient. In the technical aspect, their methods consider color distribution, motion activity maps, Hough lines and faces. Such features are all hand-crafted and may not be suitable for dealing with concert videos.

Inspired by the recent success in learning convolutional neural networks (CNNs) for object classification and feature representation, *e.g.*, [9, 10, 11, 12, 13], we first model a shot type as the composition of objects, as shown in Figure 2, and then construct a feature representation from the rich feature hierarchies of CNN to encode the shot types. To perform shot (type) classification, we introduce a novel probabilistic fusion model, termed as error weighted deep cross-correlation model (EW-Deep-CCM), by exploring the correlations between paired features, as well as their contributions to classification accuracy. (See Figure 3.) Finally, we note that besides the aforementioned six types of shots (cf. Table 1), we 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:

- This is the first work developed to classify the eight types of shots (six of them are defined by the language of film) as the basis for understanding visual storytelling in concert videos.
- We propose a novel probabilistic fusion model, EW-Deep-CCM, to significantly improve the classification accuracy for predicting the types of shots.

In the follows, we shall describe how the proposed EW-Deep-CCM framework learns a fusion deep-net model to classify the types of shots for images in detail.

2. IMAGE REPRESENTATIONS

Our idea builds upon the observation that deciding the shot of a given image from a live concert video can be casted as interpreting a composition of various objects in the scene. Take, for example, the two underlying shots discussed in Figure 2. In the left image, it comprises a hat, a T-shirt, and a microphone, while the corresponding shot is medium close-up. In contrast, the right image is taken by a medium long shot, and it includes a drum, a spotlight, a guitar, a hat, a T-shirt, and a microphone. The information indicates that a long shot tends to encompass more objects on the stage. Furthermore, even though the two types of shots could include a same object, say, hat, the size/location of the hat yielded by the two shot types are still quite different. Based on these observations, we thus model a shot (type) with a composition of objects, and use a CNN to better represent the constituent objects.

We use the 16-layer VGG-Net [12] to obtain the hierarchical features. The VGG-Net is trained on the ImageNet large-scale visual recognition challenge 2012 (ILSVRC-2012) dataset [11]. This collection includes 1.3 million images over 1,000 object categories. Along with the forward propagation in VGG-Net, we extract features from the output layer and the two fully-connected layers as the object representations for each input image, where the feature dimensions are 1000-D, 4096-D and 4096-D, respectively.

3. EW-DEEP-CCM CLASSIFICATION

The proposed EW-Deep-CCM is a deep-net extension to a previous audio-visual emotion recognition model called error weighted semi-coupled (EWSC) HMM [14]. Learning the EWSC-HMM considers not only the temporal relationship between audio and visual streams but also the contributions of different audio-visual feature pairs for obtaining a better emotion recognition result. To model and classify the shot type of an image, EW-Deep-CCM is constructed to explore the relationship between paired shot representations, *i.e.*, derived from the hidden layers of a deep neural network (DNN) given the object representation input, as well as their contributions for classifying different shot types as shown in Figure 3.

Given a deep CNN, we use O_i^{out} and O_j^{fc} to represent features extracted from the ith output layer and the jth fully-connected layer, and $\Lambda_{ij} = (\Lambda_i^{\text{out}}, \Lambda_j^{\text{fc}})$ for the resulting feedforward DNN-based cross-correlation model (Deep-CCM) classifier. Depending on the CNN architecture, we could have two sets of object representations, $\mathbf{O}^{\text{out}} = \{O_i^{\text{out}}\}_{i=1}^{C}$ and $\mathbf{O}^{\text{fc}} = \{O_j^{\text{fc}}\}_{j=1}^{D}$, and $C \times D$ Deep-CCM classifiers. For simplicity, we write $\mathbf{O} = (\mathbf{O}^{\text{out}}, \mathbf{O}^{\text{fc}})$ and $O_{ij} = (O_i^{\text{out}}, O_j^{\text{fc}})$. With the VGG-Net [12], we have $C = 1$ and $D = 2$.

Based on the feature representation $\mathbf{O}$ and the Deep-CCM classifiers $\{\Lambda_{ij}\}$, the task to decide the shot type w of an image out of totally K classes can be casted as estimating the following posterior probability:

$$\begin{aligned} P(w|\mathbf{O}) &= \sum_{i=1}^{C}\sum_{j=1}^{D} P(w, \Lambda_{ij}|O_{ij}) \\ &= \sum_{i=1}^{C}\sum_{j=1}^{D} P(w|O_{ij}, \Lambda_{ij}) P(\Lambda_{ij}|O_{ij}) \end{aligned} \tag{1}$$

where $P(w|O_{ij}, \Lambda_{ij})$ is the probability of w given by the classifier Λ_{ij}, together with the paired input from O_{ij}. In addition, $P(\Lambda_{ij}|O_{ij})$ can be considered an empirical weight assigned to the Deep-CCM classifier, representing the confidence of the decision of Λ_{ij}. Following [14], we estimate the weight from the confusion matrix of Λ_{ij}.

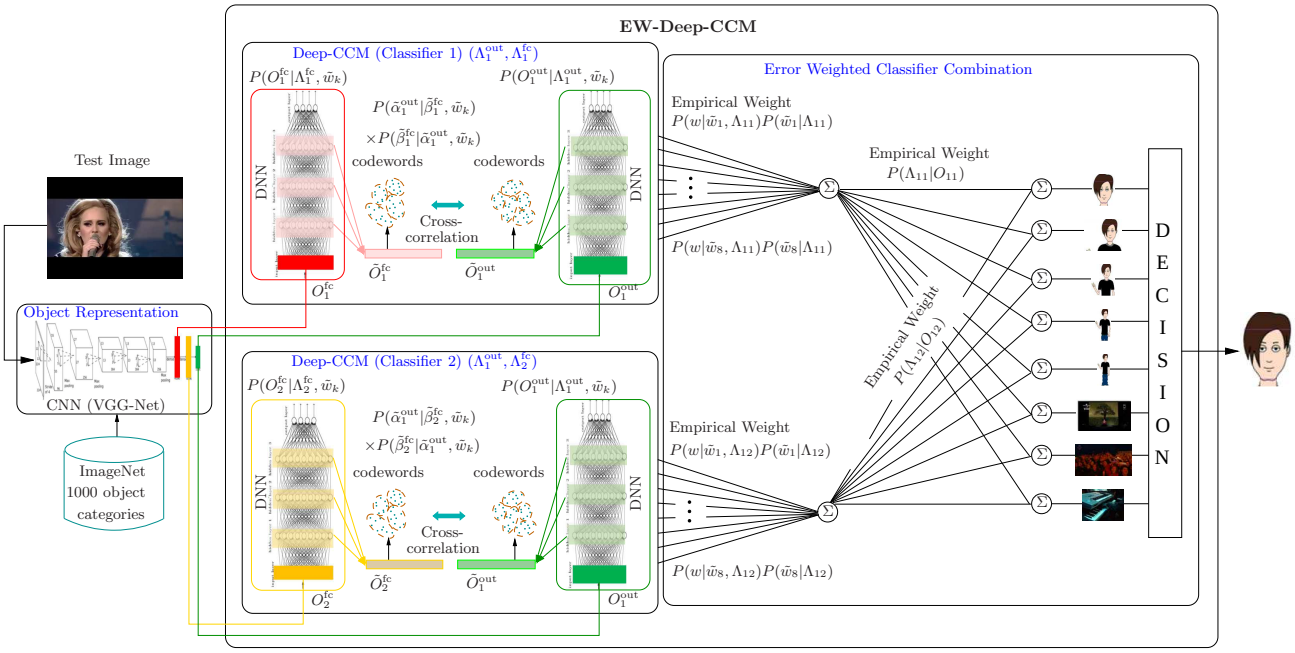

Fig. 3. Illustration of the proposed EW-Deep-CCM framework for shot classification.

Observe that the classification of w can be made by combining the individual predictions from each Λ_{ij}. Let $\tilde{w}$ be the prediction of a shot type by an individual Λ_{ij}. $P(w|O_{ij}, \Lambda_{ij})$ in (1) can then be further decomposed by

$$\begin{aligned} P(w|O_{ij}, \Lambda_{ij}) &= \sum_{k=1}^{K} P(w, \tilde{w} = k|O_{ij}, \Lambda_{ij}) \\ &= \sum_{k=1}^{K} P(w|O_{ij}, \Lambda_{ij}, \tilde{w}_k) P(\tilde{w}_k|O_{ij}, \Lambda_{ij}) \\ &\approx \sum_{k=1}^{K} P(w|\Lambda_{ij}, \tilde{w}_k) P(\tilde{w}_k|O_{ij}, \Lambda_{ij}) \\ &\propto \sum_{k=1}^{K} P(w|\Lambda_{ij}, \tilde{w}_k) P(\tilde{w}_k|\Lambda_{ij}) P(O_{ij}|\Lambda_{ij}, \tilde{w}_k) \end{aligned} \tag{2}$$

where $K = 8$ and the approximation assumes the independence of the shot prediction w and O_{ij}, given that we already know the individual shot prediction $\tilde{w}_k$ by Λ_{ij}. Thus, to yield $P(w|\mathbf{O})$ in (1), we are left to estimate $P(O_{ij}|\Lambda_{ij}, \tilde{w}_k)$ in that both $P(w|\Lambda_{ij}, \tilde{w}_k)$ and $P(\tilde{w}_k|\Lambda_{ij})$ can be computed from the the confusion matrix of Λ_{ij}, as in [14].

Approximating $P(O_{ij}|\Lambda_{ij}, \tilde{w}_k)$ can be achieved by exploring the co-occurrence dependencies between paired features, which not only individually model the extracted features of object representations from the output and fully-connected layers, but also construct the statistical dependencies among shot representations. We have

$$\begin{aligned} P(O_{ij}|\Lambda_{ij}, \tilde{w}_k) &\approx P(O_{ij}|\Lambda_i^{\text{out}}, \tilde{w}_k) P(O_{ij}|\Lambda_j^{\text{fc}}, \tilde{w}_k) \\ &= P(O_i^{\text{out}}|\Lambda_i^{\text{out}}, \tilde{w}_k) P(O_j^{\text{fc}}|O_i^{\text{out}}, \Lambda_i^{\text{out}}, \tilde{w}_k) \\ &\times P(O_i^{\text{out}}|O_j^{\text{fc}}, \Lambda_j^{\text{fc}}, \tilde{w}_k) P(O_j^{\text{fc}}|\Lambda_j^{\text{fc}}, \tilde{w}_k) \end{aligned} \tag{3}$$

where $P(O_i^{\text{out}}|\Lambda_i^{\text{out}}, \tilde{w}_k)$ is the likelihood of the feature of the ith output layer, and $P(O_j^{\text{fc}}|O_i^{\text{out}}, \Lambda_i^{\text{out}}, \tilde{w}_k)$ is the probability of the co-occurrence dependency between features of the ith output layer and the jth fully-connected layer. The remaining in the right hand side of (3) can be explained analogously.

It is possible to better model statistical dependencies in (3) for classifying shot types. In particular, in the stage of testing, we perform forward propagation through Λ_i^{out} and Λ_j^{fc} to map the object representations, O_i^{out} and O_j^{fc}, into the shot representations, $\tilde{O}_i^{\text{out}}$ and $\tilde{O}_j^{\text{fc}}$, by respectively concatenating the features of the hidden layers. (See Figure 3.) It implies $P(O_j^{\text{fc}}|O_i^{\text{out}}, \Lambda_i^{\text{out}}, \tilde{w}_k)$ and $P(O_i^{\text{out}}|O_j^{\text{fc}}, \Lambda_j^{\text{fc}}, \tilde{w}_k)$ in (3) are approximated by $P(\tilde{O}_j^{\text{fc}}|\tilde{O}_i^{\text{out}}, \tilde{w}_k)$ and $P(\tilde{O}_i^{\text{out}}|\tilde{O}_j^{\text{fc}}, \tilde{w}_k)$, respectively. In addition, because $\tilde{O}_i^{\text{out}}$ and $\tilde{O}_j^{\text{fc}}$ are continuous values, it is unfeasible to collect sufficient amount of training data to construct statistical dependencies between the two, under the joint condition $\tilde{w}_k$. We use k-means clustering to construct a codebook and perform vector quantization to represent $\tilde{O}_i^{\text{out}}$ and $\tilde{O}_j^{\text{fc}}$ by their corresponding codeword, say, $\tilde{\alpha}_i^{\text{out}}$ and $\tilde{\beta}_j^{\text{fc}}$. Hence, (3) can be rewritten as

$$\begin{aligned} P(O_{ij}|\Lambda_{ij}, \tilde{w}_k) \approx\ & P(O_i^{\text{out}}|\Lambda_i^{\text{out}}, \tilde{w}_k) P(\tilde{\beta}_j^{\text{fc}}|\tilde{\alpha}_i^{\text{out}}, \tilde{w}_k) \\ & \times P(\tilde{\alpha}_i^{\text{out}}|\tilde{\beta}_j^{\text{fc}}, \tilde{w}_k) P(O_j^{\text{fc}}|\Lambda_j^{\text{fc}}, \tilde{w}_k). \end{aligned} \tag{4}$$

We remark that the output by a DNN is a posterior probability, but $P(O_i^{\text{out}}|\Lambda_i^{\text{out}}, \tilde{w}_k)$ and $P(O_j^{\text{fc}}|\Lambda_j^{\text{fc}}, \tilde{w}_k)$ in (4) are likelihoods. In our implementation, we approximate the first likelihood by $P(\tilde{w}_k, \Lambda_i^{\text{out}}|O_i^{\text{out}})/P(\Lambda_i^{\text{out}}, \tilde{w}_k)$ and the second by $P(\tilde{w}_k, \Lambda_j^{\text{fc}}|O_j^{\text{fc}})/P(\Lambda_j^{\text{fc}}, \tilde{w}_k)$.

With (2)-(4), we arrive at how $P(w|\mathbf{O})$ in (1) is evaluated. Specifically, in the test phase, $P(w|\mathbf{O})$, the posterior probability of each shot type w, can be inferred from every predicted shot type $\tilde{w}_k$ by combining the outputs of Deep-CCM classifiers and empirical weights in the paired object representations $O_{ij} = (O_i^{\text{out}}, O_j^{\text{fc}})$. Then, the shot type w with maximum posterior probability is selected as the classification result.

In the training phase, the feedforward DNNs Λ_i^{out} and Λ_j^{fc} are trained separately using the back-propagation algorithm with the objective of softmax cross entropy. $P(\tilde{\beta}_j^{\text{fc}}|\tilde{\alpha}_i^{\text{out}}, \tilde{w}_k)$ and $P(\tilde{\alpha}_i^{\text{out}}|\tilde{\beta}_j^{\text{fc}}, \tilde{w}_k)$, the terms of cross-correlation are calculated by

Fig. 4. The eight types of shots and examples.

Table 2. The distribution of collected data.

Data Type	Data Source	# Concert	#Video	# Frame
Training	Official	14	17	22,292
Test	Official	3	5	11,247
Test	Audience	1	9	4,352

statistical co-occurrence dependencies over all the training data. The empirical weights $P(\Lambda_{ij}|O_{ij})$ in (1), $P(w|\Lambda_{ij}, \tilde{w}_k)$ and $P(\tilde{w}_k|\Lambda_{ij})$ in (2) are calculated through a confusion matrix [14] over all the training data based on the Deep-CCM classifiers.

4. EXPERIMENTAL RESULTS

To demonstrate the effectiveness of EW-Deep-CCM, we conduct experiments on a set of concert videos downloaded from YouTube. 31 video clips (including official and audience recordings) are collected from 18 live concerts, with a total of 37,891 annotated images as in Table 2. Each image is specified with an appropriate shot type based on the definition of the language of film [4, 5]. Our experiments aim to classify totally eight shot types as listed in Figure 4. Among the collected images, 22,292 (official recording) image frames from 14 live concerts are included for training. We use 11,247 (official recording) and 4,352 (audience recording) image frames from the remaining 4 live concerts for testing (see Table 2).

We compare the performance of the proposed EW-Deep-CCM with those of two popular approaches to data fusion, namely, early fusion and late fusion [15]. For early fusion, features extracted from the output layer and the two fully-connected layers of VGG-Net are concatenated to construct a combined feature vector (1000+4096+4096), and then fed into a DNN classifier for shot (type) classification. For late fusion, features extracted from the three layers are first modeled by the corresponding DNN classifier. The preliminary shot classification outputs from each DNN classifier are then concatenated (8+8+8 shot type outputs) and fed into the final DNN classifier for decision. The structure and parameter setting of DNN with respect to the hidden layers, neurons, learning rate, mini-batch size, the number of epochs for EW-Deep-CCM, early fusion, and late fusion are determined when their best classification accuracy is achieved in the experiments. Finally, taking account of that the number of testing data of the eight shot types are imbalanced, we report both the weighted average (WA) and unweighted average (UA, better reflecting the imbalances among classes) classification rates in the experiments.

We first evaluate the performance of each layer-wise feature representation for shot classification. In Table 3, both UA and WA accuracy rates indicate that object representations of different layers

Table 3. UA/WA classification accuracy (%).

Data	DNN-o1	DNN-fc2	DNN-fc1	Early	Late	Ours
Official (UA)	63.53	70.36	74.31	76.14	78.54	83.32
Official (WA)	58.67	72.80	78.77	74.62	79.10	81.25
Audience (UA)	50.65	62.84	53.47	59.00	64.64	77.50
Audience (WA)	50.62	62.78	66.06	64.09	67.37	73.21

Table 4. Layer-wise classification accuracy (%, official recordings).

Shot Types	CU	MCU	MS	MLS	LS	XLS	ADS	MIS
DNN-o1	56.33	64.48	54.33	33.94	79.90	62.21	92.31	64.52
DNN-fc2	82.68	91.99	48.81	13.97	60.78	75.50	92.31	96.87
DNN-fc1	68.57	91.71	92.30	31.21	39.71	86.32	100.0	84.71

(denoted as o1, fc1 and fc2) in VGG-Net have distinct contributions to shot classification. Table 4 further shows the classification outcomes of eight shot types for three layer-wise feature representations in official recordings. Specifically, DNN-fc1 achieves better classification results for MS and XLS shot types, and DNN-fc2 for CU and MIS shot types, while DNN-o1 yields better accuracy for MLS and LS shot types. The finding inspires our use of a fusion approach to integrating the representation power from different layers. Regarding the fusion comparison, it is evident from Table 3 that late fusion outperforms early fusion in our experiments. This could be due to that the use of a high-dimensional feature vector in early fusion is prone to trap itself into the problems of data sparseness and overfitting. Still, the UA classification rate of early fusion is mostly better than that of each individual layer-wise representation, and that indeed suggests the advantage of model fusion. On the other hand, although late fusion does not concatenate the features and can thus avoid the data sparseness and overfitting problems, the assumption of conditional independence among the three layer-wise representations is practical, but not appropriate. Based on the above analyses, EW-Deep-CCM is designed to not only individually model the extracted feature hierarchies of the VGG-Net but also construct the statistical dependencies among paired features as well as exploring their contributions to improve the classification accuracy. Among the six approaches in Table 3, the proposed EW-Deep-CCM achieves the best UA and WA classification rates. The Chi-squared test statistic [16] performed on two-sided tests with a significance level of 0.05 further justifies that the differences of performances among the six approaches are statistically significant with $\chi^2(5) = 1984.169, P < 0.0001$ and $\chi^2(5) = 530.749, P < 0.0001$ for official and audience recordings, respectively. Finally, despite the less-satisfactory quality of audience recordings, EW-Deep-CCM can still yield an UA accuracy rate close to 80%, not far behind that for official recordings.

5. CONCLUSIONS

We have introduced a novel probabilistic fusion framework, named as error weighted deep cross-correlation model (EW-Deep-CCM), to perform shot classification for concert videos. Our experiments on both official and audience recordings have demonstrated that EW-Deep-CCM outperforms the current popular fusion strategies, and can offer satisfactory shot-type classification results. Leveraging with these promising outcomes, our future work along this line would focus on addressing the challenging issues of learning visual storytelling of professionally edited videos, which is crucial in developing automatic techniques for video mashup.

Acknowledgment: This work was supported in part by MOST grants 105-2221-E-001-018-MY3 and 105-2221-E-001-027-MY2.

6. REFERENCES

[1] Prarthana Shrestha, Peter HN de With, Hans Weda, Mauro Barbieri, and Emile HL Aarts, "Automatic mashup generation from multiple-camera concert recordings," in *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 2010, pp. 541–550.

[2] Mukesh Kumar Saini, Raghudeep Gadde, Shuicheng Yan, and Wei Tsang Ooi, "Movimash: online mobile video mashup," in *Proceedings of the 20th ACM international conference on Multimedia*. ACM, 2012, pp. 139–148.

[3] Yue Wu, Tao Mei, Ying-Qing Xu, Nenghai Yu, and Shipeng Li, "Movieup: Automatic mobile video mashup," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 25, no. 12, pp. 1941–1954, 2015.

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

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

[6] Ali Bagheri-Khaligh, Ramin Raziperchikolaei, and M Ebrahimi Moghaddam, "A new method for shot classification in soccer sports video based on svm classifier," in *Image Analysis and Interpretation (SSIAI), 2012 IEEE Southwest Symposium on*. IEEE, 2012, pp. 109–112.

[7] Ling-Yu Duan, Min Xu, Qi Tian, Chang-Sheng Xu, and Jesse S Jin, "A unified framework for semantic shot classification in sports video," *IEEE Transactions on Multimedia*, vol. 7, no. 6, pp. 1066–1083, 2005.

[8] Sergio Benini, Luca Canini, and Riccardo Leonardi, "Estimating cinematographic scene depth in movie shots," in *Multimedia and Expo (ICME), 2010 IEEE International Conference on*. IEEE, 2010, pp. 855–860.

[9] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[10] Li-Jia Li, Hao Su, Li Fei-Fei, and Eric P Xing, "Object bank: A high-level image representation for scene classification & semantic feature sparsification," in *Advances in neural information processing systems*, 2010, pp. 1378–1386.

[11] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al., "ImageNet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[12] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[13] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva, "Learning deep features for scene recognition using places database," in *Advances in neural information processing systems*, 2014, pp. 487–495.

[14] Jen-Chun Lin, Chung-Hsien Wu, and Wen-Li Wei, "Error weighted semi-coupled hidden markov model for audio-visual emotion recognition," *Multimedia, IEEE Transactions on*, vol. 14, no. 1, pp. 142–156, 2012.

[15] Chung-Hsien Wu, Jen-Chun Lin, and Wen-Li Wei, "Survey on audiovisual emotion recognition: databases, features, and data fusion strategies," *APSIPA Transactions on Signal and Information Processing*, vol. 3, pp. e12, 2014.

[16] Harris Cooper, Larry V Hedges, and Jeffrey C Valentine, *The handbook of research synthesis and meta-analysis*, Russell Sage Foundation, 2009.