Knowledge-Augmented Multimodal Deep Regression Bayesian Networks for Emotion Video Tagging

Shangfei Wang, *Senior Member, IEEE*, Longfei Hao, and Qiang Ji, *Fellow, IEEE*

***Abstract*—The immanent dependencies between audio and visual modalities extracted from video content and the well-established film grammar (i.e., domain knowledge) are important for emotion video recognition and regression. However, these tools have yet to be exploited successfully. Therefore, we propose a multimodal deep regression Bayesian network (MMDRBN) to capture the relationship between audio and visual modalities for emotion video tagging. We then modify the structure of the MMDRBN to incorporate domain knowledge. A regression Bayesian network (RBN) is formed from one latent layer, one visible layer and directed links from the latent layer to the visible layer. RBN is able to fully represent the data, since it captures the dependencies not only among the visible variables but also among the latent variables given visible variables. For the MMDRBN, first, we learn several layers of RBNs using audio and visual modalities, and then stack these RBNs to form two deep networks. A joint representation is obtained from the top layers of the two deep networks, capturing the deep dependencies between audio and visual modalities. We also summarize the main audio and visual elements used by filmmakers to convey emotions and formulate them as semantical meaningful middle-level representation, i.e., attributes. Through these attributes, we construct the knowledge-augmented MMDRBN, which learns a hybrid middle-level video representation using video data and the summarized attributes. Experimental results of both emotion recognition and regression from videos on the LIRIS-ACCEDE database demonstrate that the proposed model can successfully capture the intrinsic connections between audio and visual modalities, and integrate the middle-level representation learning from video data and semantical attributes summarized from film grammar. Thus, it achieves superior performance on emotion video tagging compared to state-of-the-art methods.**

***Index Terms*—Regression Bayesian network, Multi-modal deep network, Domain knowledge, Emotion video tagging.**

Manuscript received June 18, 2018; revised February 28, 2019 and April 29, 2019; accepted August 7, 2019. Date of publication August 12, 2019; date of current version March 24, 2020. This work was supported in part by the project from Anhui Science and Technology Agency (1804a09020038) and in part by the National Science Foundation of China under Grants 917418129 and 61473270. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Benoit Huet. *(Corresponding author: Shangfei Wang.)*

S. Wang is with the Key Laboratory of Computing and Communication Software of Anhui Province, the School of Computer Science and Technology and the School of Data Science, University of Science and Technology of China, Hefei 230027, China (e-mail: sfwang@ustc.edu.cn).

L. Hao is with the School of Computer Science and Technology, University of Science and Technology of China, Hefei 230027, China (e-mail: hlf101@mai.ustc.edu.cn).

Q. Ji is with the Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180 USA (e-mail: qji@ecse.rpi.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2019.2934824

I. Introduction

EMOTION video tagging has attracted increasing attention in recent years due to the exponential growth of digital video repositories and the increase in user demand for videos consumed through the popular social networks. Mainstream works on emotion video tagging either detect the expected emotions from video content or recognize the induced emotions from audiences' spontaneous nonverbal responses while watching videos. Expected emotions are the emotions that a video should convey via audio and visual cues; induced emotions are the audiences' invoked emotions when they watch videos. This paper deals with expected emotions.

Video content consists of audio and visual elements. Both elements are utilized by video makers to convey emotions to audiences. For example, bright and vivid scenes may make people feel more positive, while dim scenes may invoke negative feelings. Loud noises may upset people, while soft music can be relaxing. The relations between audio and visual elements and their expressed emotions are referred to as domain knowledge. Audio and visual elements interact with each other to enhance the emotional atmosphere. For instance, a quick scene cut accompanied by urgent music may be used to convey excitement. Thoroughly exploring this domain knowledge and successfully integrating audio and visual elements should be beneficial for emotion video tagging.

Current emotion video tagging usually extracts several handcrafted audio and visual features to characterize the video content. Those handcrafted features represent the emotion-sensitive audio and visual elements to some extent, but cannot completely explore the domain knowledge. Several recent works have employed deep neural networks to learn feature representation from videos. Although the learned representations can take advantage of deep learning and the large scale of videos, they are entirely data-driven and do not consider domain knowledge.

After feature extraction, some works concatenate the extracted features as one feature vector, feeding them into either a classifier for emotion recognition or a regressor for emotion regression. We refer to this method as feature-level fusion. Obviously, the interactions between audio elements and visual elements cannot be thoroughly explored by simply linking all of the audio and visual features into one feature vector. Instead of feature-level fusion, some works employ decision-level fusion to integrate audio and visual elements for emotion video tagging. They first detect emotions from audio and visual elements separately, and then combine the recognition results

of the two modalities through certain fusion strategies. Either decision-level fusion cannot successfully model the interaction between audio signals and visual signals for emotion video tagging, since there exist very complex relations between audio content and visual content as well as the impact of video content on users' emotions [1]. To leverage the connections between audio and visual content for emotion video tagging, multimodal learning may be a better approach, since it jointly optimizes functions from multiple modalities and thus models the dependencies existed among multiple modalities to boost the performance.

We propose a new deep multimodal learning method, i.e., the multimodal deep regression Bayesian network (MMDRBN), to learn the joint representation of audio and visual modalities for emotion video tagging. Specifically, we stack several regression Bayesian networks (RBNs) for audio and visual modalities and then extract a joint representation through the stacked RBNs. Through minimizing the KL-divergence between the stacked deep network from two modalities and an inference network, we transform the MMDRBN into the inference network, which is used to predict affective scores from the video content.

We further extend the proposed multimodal deep model to a knowledge-augmented multimodal deep regression Bayesian network by constructing a joint representation from the visual modality, audio modality, and the well-established cinematography. We summarize the film grammar describing how filmmakers employ audio and visual elements to communicate emotions with the audience. Then, we define these audio and visual elements as attributes, and integrate them with the joint middle-level representations learned from video data. Finally, both the emotion labels and the attributes are used to further tune the parameters of the proposed knowledge-augmented multimodal deep network for emotion classification or regression from video content.

A previous paper proposing a multimodal deep regression Bayesian network appeared as [2]. Here, we extend the MMDRBN to a knowledge-augmented MMDRBN, which explores both the connections between audio and visual modalities as well as domain knowledge for emotion video tagging. Compared with the previous paper, the advantages of this paper are as follows: first, we have summarized the well-estimated film grammar and manually extracted these audio emotion-sensitive attributes and visual emotion-sensitive attributes as domain knowledge. Second, we have described the extended domain knowledge augmented multimodal deep regression Bayesian networks in the method section. Third, we have added experiments using domain knowledge augmented multimodal model on the LIRIS-ACCEDE database in the experiment section.

This paper is organized as follows. Section II provides an overview of the related work on emotion video tagging and multiview learning. Section III gives a problem statement. Section IV elaborates on the details of the proposed MMDRBN and proposed domain knowledge-augmented MMDRBN. Section V presents the experimental results of both classification and regression on the LIRIS-ACCEDE database. Section VI concludes our work.

II. Related Work

A. Emotion Video Tagging

Wang and Ji provided a comprehensive survey of current emotion video tagging in [3]. In this section, we briefly analyze how current works explore domain knowledge and the interaction between audio and visual elements for emotion classification and regression from video content.

Current emotion video tagging mainly explore domain knowledge by defining specific emotion-sensitive features to represent video content. For instance, Hanjalic and Xu [4] proposed to adopt the linear combination of the sound energy component, rhythm component and motion component as the arousal curve and valence curve. Grounded on psychology and cinematography, Wang *et al.* [5] explored a number of audio and visual cues. They addressed many audio features including Log-Frequency Power Coefficients (LFPC), low short-time energy ratio (LSTER), normalized octave energy bands, spectral flux, spectral roll-off and centroid, zero crossing rate (ZCR), Mel-frequency cepstral coefficients (MFCC), and music scale. For visual aspects, the visual excitement, lighting key, color energy and shot duration are explored.

These handcrafted features are inspired by psychological and cinematography research. Therefore, they can capture the emotion-discriminative audio and visual elements to some extent, but cannot fully explore the dependencies between expected emotions and visual-audio content. Recently, several works have adopted deep networks to learn representations for emotion video tagging. For example, Acar *et al.* [6] used convolutional neural networks (CNNs) to construct middle-level representations from MFCC and color values. The works employing deep networks leverage the power of deep multimodal networks and large-scale videos. However, they are driven by data and do not utilize domain knowledge.

To the best of our knowledge, only one work explicitly leverages the relations between audio-visual elements and expected emotions for emotion classification and regression from video content. Chen *et al.* [7] summarized the connections between emotions and three visual elements (i.e., motion, color and lighting), and then transferred these dependencies into loss constraints during the classifier learning process. Their work successfully takes advantages of domain knowledge to regularize the emotion classifiers from video content. However, they only considered visual elements and the probabilistic positive or negative correlations between these elements and valence or arousal. In addition to visual elements, audio elements are often used to enhance the emotional atmosphere in a movie. Furthermore, the dependencies between emotion and video content are much more complex than probabilistic positive or negative correlations. In this paper, we summarize the relations between visual elements and emotions as well as audio elements and emotions. Instead of simply modeling positive or negative correlations, we capture more complex and global dependencies through a latent regression Bayesian network.

Current emotion video tagging employ either feature-level fusion or decision-level fusion to handle interactions between audio and visual elements. For feature-level fusion, the extracted

audio and visual features are linked as one feature vector which is then used as the input of a classifier or regressor. Such fusion strategy can not thoroughly capture the complex interactions between audio and visual content. While decision-level fusion consists of two subsystems, which use either visual features or audio features and make the decision independently. The decision-level fusion strategy is then used to combine the decisions from subsystems to produce the final decision. Such fusion strategy ignores the synergy and the highly non-linear relationships between audio content and visual content [1]. Multimodal or multiview learning may be better approaches to leverage the connections among audio and visual content for emotion video tagging. However, to the best of our knowledge, little work thus far exploits multiview learning for emotion classification and regression from video content apart from Pang *et al.*'s [1]. Pang *et al.* [1] adopt the deep Boltzmann machine (DBM) to construct a multimodal DBM (MMDBM) from auditory, visual, and textual modalities for emotion classification and cross-modal retrieval. Specifically, the model first obtains middle-level representations from low-level auditory, visual, and textual features using DBM. The middle-level representations are combined to construct a joint representation. Then, the joint representations are fed into a logistic regression. Experimental results on web videos demonstrate that the MMDBM can effectively capture the non-linear and complex synergy among auditory, visual, and textual modalities in a joint space for better emotion tagging and retrieval. As a classical undirected graphic model, the restricted Boltzmann machine (RBM) promises that the hidden nodes are independent of each other given the neighboring visible layers. However, latent variables should explain the patterns of the input data cooperatively. This independence inevitably weakens the representation power of the MMDBM.

We propose MMDRBN, a new multimodal learning method, to construct the high-level joint representation of audio and visual modalities for emotion classification and regression from video content. Although similar to current generative deep models in structure, the proposed MMDRBN is fundamentally different from these generative models since the dependencies among hidden nodes can be captured. Thus, the proposed MMDRBN can successfully capture the immanent dependencies between audio content and visual content and achieve better performance for emotion video tagging compared to state-of-the-art works.

Although Pang *et al.*'s method captures the complex and non-linear relations among different modalities in a joint space for better video emotion tagging and retrieval, it learns a multimodal middle-level video representation purely from training data. While convenient, such a learned video representation is sensitive to the quality and quantity of the training data. Moreover, the automatically constructed video representation is semantically meaningless. The well-established film grammar is often used to convey emotion through audio and visual elements and is a semantically meaningful middle-level representation that is crucial for emotion video tagging. This well-established knowledge can be exploited to help construct a hybrid middle-level video representation that can simultaneously leverage the available data and the existing domain knowledge. To this goal, we propose a knowledge-augmented multimodal deep network that combines the manually specified and semantically meaningful middle-level video attributes with automatically learned data-based middle-level video representation to produce a hybrid middle-level video representation for emotion video tagging. Specifically, we add a layer of domain knowledge at the joint representation layer. The domain knowledge aims to learn a better video representation and to bridge the gap between video representation and emotion video tagging.

B. Multiview Learning

Multiview data are readily available, since most features of data can be naturally or manually split into distinct feature sub-sets. For example, a video consists of audio and visual elements. Compared to single view learning, multiview learning can exploit the complementary information inherent in multiple views to learn more expressive representation and more powerful classifiers. Therefore, multiview learning has attracted increasing attention due to its promising potential in many applications. Comprehensive surveys on multiview learning can be found in [8], [9].

As stated in [9], correlation, consensus, and complementarity principles are adopted for multiview representation learning. Canonical correlation analysis (CCA) is a popular approach for correlation principles. It aims to find linear combinations of the two views with respect to maximum correlation with each other [10]. Kernel canonical correlation analysis (KCCA) extends CCA to the nonlinear change by leveraging the kernel method during CCA transformation [11]. The emergence of big data and the success of deep learning led to the proposal of deep canonical correlation analysis (DCCA) [12]. DCCA aims to learn feature representations of two maximally correlated views through employing deep neural networks (DNNs) as a nonlinear transformation. DCCA has proven more accurate than KCCA for nonlinear transformation tasks [13].

Multimodal deep Boltzmann machines (MMDBM) [14] and Multimodal deep belief networks (DBNs) [15] are typical models combining consensus and complementarity principles. They try to obtain a compact representation that best reconstructs the inputs and maximize the complementary information that exists in multiple views simultaneously. Both multimodal DBM and multimodal DBN consist of several layers of restricted Boltzmann machine (RBM). As an undirected graphical model, an RBM can effectively model global dependencies among visible units through the completely undirected links between the hidden layer and the visible layer as well as the assumption of independencies among hidden units given visible units. Introducing dependencies among hidden units will increase the models' ability to explain the patterns that inherent in the visible units. Instead of using undirected links like an RBM, RBN adopts directed links between hidden units and visible units to model not only the dependencies among the latent variables given visible variables but also the dependencies among visible variables, and thus better represents the patterns intrinsic in the visible units.

However, RBN inference is computationally intractable. General approximation algorithms can alleviate the problem but discard the dependencies among latent variables. To maintain

the dependencies, we propose an approach through Gibbs sampling. After learning several RBNs, we stack them to create an MMDRBN. The MMDRBN models the inherent connections between audio content and visual content. The MMDRBN is used to initialize an inference network through KL divergence; this inference network is a feed forward network for emotion classification or regression from video content.

Domain knowledge augments the proposed MMDRBN to further bridge the huge semantic gap between low-level features and video affect. Specifically, the audio and visual elements used to convey emotions to the audiences are aggregated into a domain knowledge layer. The layer bridges the gap between low-level features and high-level emotion semantic for video tagging.

Compared to related work, our contributions are as follows:

First, unlike the deep Boltzmann machine, which is an undirected graphic model with the independent assumption among latent nodes, the proposed deep regression Bayesian network is an directed graphic model, which is able to capture ont only the dependencies among the latent variables given the observation but also the dependencies among visible variables. Thus, the proposed knowledge-enhanced multimodal deep regression Bayesian network is more representative of the data.

Second, unlike most deep models which learn representations from data only, the proposed deep regression Bayesian network incorporates domain knowledge, i.e., the semantic middle-level video representation, into the deep learning. Therefore, the hybrid middle-level video representation is expected to be superior to both the manual and the data-based middle-level video representation.

Last, we are the first to summarize both the dependencies between visual elements and emotions as well as audio elements and emotions, and to propose a graphic model capturing this complex domain knowledge.

III. Problem Statement

The purpose of our work in this paper is to learn a multimodal network that considers the synergy between audio and visual data for emotion video tagging.

Let $\Omega = \{\boldsymbol{v}_n, \boldsymbol{a}_n, \boldsymbol{y}_n\}_{n=1}^N$ denote the training set, where N is the number of instances, $\boldsymbol{v}_n \in \mathbb{R}^{d_1}$ denotes the d_1 dimensional visual feature of the training instance, $\boldsymbol{a}_n \in \mathbb{R}^{d_2}$ denotes the d_2 dimensional audio feature of the training instance, and $\boldsymbol{Y} = \{y_n\}_{n=1}^N$ stores all ground truth emotion video tagging labels. Given the training set Ω, our goal is to train a multimodal network $\mathcal{N} : \{\mathbb{R}^{d_1}, \mathbb{R}^{d_2}\} \rightarrow \boldsymbol{y}$.

Domain knowledge enhances the multimodal network. Specifically, let $\Omega = \{\boldsymbol{v}_n, \boldsymbol{a}_n, \boldsymbol{at}_n, \boldsymbol{y}_n\}_{n=1}^N$ denote the training set, where $\boldsymbol{at}_n \in \mathbb{R}^{d_3}$ denotes the d_3 dimensional audio and visual domain knowledge of the training instance. Therefore, our purpose is to train a multimodal network enhanced by domain knowledge $\mathcal{N} : \{\mathbb{R}^{d_1}, \mathbb{R}^{d_2}, \mathbb{R}^{d_3}\} \rightarrow \boldsymbol{y}$.

IV. Proposed Approaches

A. Brief Introduction of RBN

The regression Bayesian network (RBN) [16] contains one visible layer, one latent layer, and completely directed links from hidden layer to visible layer, as shown in Fig. 1. These directed connections result in the "explaining away" effect, which introduces the independencies among the latent variables given the visible variables. To meet the requirements of our experimental data, we introduce both the Gaussian-Bernoulli RBN, which takes the continuous input, and the Bernoulli-Bernoulli RBN, which takes the binary input.

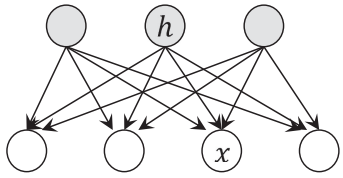


Fig. 1. The structure of a RBN.

Since the RBN is, in essence, a Bayesian network, the chain rule applies. Therefore, the joint probability of all visible and latent variables of a LRBN can be factorized into the product of prior probabilities for a latent variable h_j and conditional probabilities of a visible node v_i given all latent variables as shown in Eq. 1,

$$P(\boldsymbol{x}, \boldsymbol{h}) = \prod_{j=1}^{n_h} P(h_j) \prod_{i=1}^{n_d} P(x_i|\boldsymbol{h}) \,. \tag{1}$$

The prior probability for latent variable h_j is assumed to satisfy the Bernoulli distribution as shown in Eq. 2,

$$P(h_j) = \text{sigm}(d_j)^{h_j}(1 - \text{sigm}(d_j))^{1-h_j} \tag{2}$$

where $\text{sigm}(x) = 1/(1 + \exp(-x))$ and d_j is the bias of the hidden variable h_j.

The conditional probability of a visible variable x_i given all the latent variables $\boldsymbol{h}$ can be assumed as a linear Gaussian for continuous input as shown in Eq. (3) and Bernoulli distribution for binary input as shown in Eq. (4).

$$P\left(x_i|\boldsymbol{h}\right) \sim \mathcal{N}\left(\boldsymbol{w}_i^T\boldsymbol{h} + b_i,\, \sigma_i\right), \tag{3}$$

$$P(x_i|\boldsymbol{h}) = \sigma(\boldsymbol{w_i^T h} + b_i)^{x_i}(1 - \sigma(\boldsymbol{w_i^T h} + b_i))^{1-x_i} \tag{4}$$

where $\boldsymbol{w_i}$ is the weight between all of latent variables $\boldsymbol{h}$ and visible variable x_i, and b_i is the bias term for x_i. The mean of the Gaussian distribution is a linear combination of latent variables. The standard deviation is calculated from visible variables $\boldsymbol{x}$.

Therefore, when the visible variables are continuous, the RBN can be viewed as a mixed Gaussian model. When the visible variables are binary, the RBN can be viewed as a Bernoulli model.

Given continuous visible variables, the joint probability of RBN, Eq. (1), can be rewritten by combining Eq. (2) and Eq. (3):

$$\begin{aligned} P_\Theta(\boldsymbol{x}, \boldsymbol{h}) &= \prod_j \frac{\exp(d_j h_j)}{1 + \exp(d_j)} \prod_i \mathcal{N}(x_i : \boldsymbol{w}_i^T\boldsymbol{h} + b_i,\, \sigma_i) \\ &= \frac{\exp(-\psi_\Theta(\boldsymbol{x}, \boldsymbol{h}))}{(2\pi)^{n_d/2} \prod_i \sigma_i \prod_j (1 + \exp(d_j))} \end{aligned} \tag{5}$$

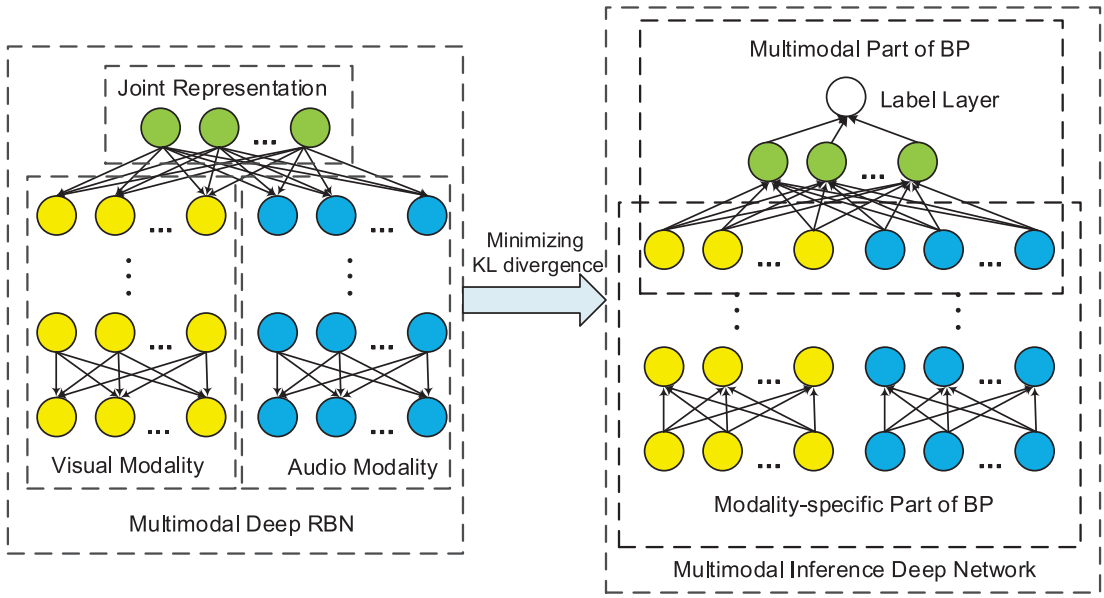


Fig. 2. The framework of the Proposed Multimodal Deep Regression Bayesian Networks for Emotion Video Tagging.

where $\Theta = \{\boldsymbol{W}, \boldsymbol{\sigma}, \boldsymbol{b}, \boldsymbol{d}\}$, and

$$\psi_\Theta(\boldsymbol{x}, \boldsymbol{h}) = \sum_i \frac{(x_i - b_i)^2}{2\sigma_i^2} - \sum_i \frac{x_i - b_i}{\sigma_i^2} \boldsymbol{w}_i^T \boldsymbol{h} + \sum_i \frac{1}{2\sigma_i^2} (\boldsymbol{w}_i^T \boldsymbol{h})^2 - \boldsymbol{d}^T \boldsymbol{h}, \tag{6}$$

Given binary visible variables, the joint probability of RBN, Eq. (1), can be rewritten by combining Eq. (2) and Eq. (4):

$$P_\Theta(\boldsymbol{x}, \boldsymbol{h}) = \prod_j \frac{\exp(d_j h_j)}{1 + \exp(d_j)} \prod_i \frac{\exp((\boldsymbol{w}_i^T \boldsymbol{h} + b_i) x_i)}{1 + \exp(\boldsymbol{w}_i^T \boldsymbol{h} + \boldsymbol{b})} = \frac{\exp(-\psi_\Theta(\boldsymbol{x}, \boldsymbol{h}))}{\prod_j (1 + \exp(d_j))}, \tag{7}$$

where $\Theta = \{\boldsymbol{W}, \boldsymbol{b}, \boldsymbol{d}\}$, and

$$\psi_\Theta(\boldsymbol{x}, \boldsymbol{h}) = -\sum_i (\boldsymbol{w}_i^T \boldsymbol{h} + b_i) x_i - \sum_j d_j h_j + \sum_i \log(1 + \exp(w_i^T \boldsymbol{h} + b_i)). \tag{8}$$

There are two primary differences between the joint probability of RBNs and the joint probability of RBMs. First, the RBN adopts directed connections rather than undirected connections. Thus, the mixed Gaussian RBN model has the extra term $\sum_i \frac{1}{2\sigma_i^2} (\boldsymbol{w}_i^T \boldsymbol{h})^2$, which the Gaussian-Bernoulli RBM does not have. The Bernoulli RBN model has the extra term $\sum_i (1 + \exp(\boldsymbol{w}_i^T \boldsymbol{h} + b_i))$, which the Bernoulli-Bernoulli RBM lacks. According to the formulation, the dependencies among latent variables are modeled faithfully by the extra term. Sequentially, the model should better capture the relationships embedding in the data. Second, unlike the partition function of the RBM, which is computationally intractable, the normalized term of the RBN can be calculated easily.

B. Multimodal Deep Regression Bayesian Networks

Fig. 2 shows the framework of the proposed multimodal deep regression Bayesian network (MMDRBN), which contains two deep stacked RBNs constructed for audio and visual modalities, respectively. Since audio and visual features are continuous, the bottom two modalities are mixed Gaussian RBNs, while the others are Bernoulli-Bernoulli RBNs. First, the two deep stacked RBNs are layer-wisely trained. The hidden layer of the lower RBN is employed as the visible layer of its upper RBN. After obtaining the trained deep networks of the two modalities, a joint representation layer is added to the top layer. Specifically, the visible variables of the joint representation layer are a link of the top layers of the audio modality network and the visual modality network. The latent variables of the joint representation layer can be viewed as a joint representation of the audio and visual modalities. After layer-wise learning, we obtain an MMDRBN, as shown in the left part of Fig. 2.

For the MMDRBN, exact inference is intractable due to the directed connections. Therefore, we transform the MMDRBN into an inference network by minimizing the KL divergence between the MMDRBN and the inference network, as shown in the right side of Fig. 2. Then, a label layer is added on the top of the inference network. Given the label layer, the inference network can be fine-tuned through a back-propagation (BP) algorithm. After fine-tuning, the inference network is utilized to predict discrete emotion labels or estimate continuous emotion scores from video content. The methods for learning and fine-tuning these networks will be discussed in detail in the following three sections.

1) A Multimodal Generative Network Learning: In this subsection, we first propose an efficient parameter learning method for the RBN based on stochastic approximation procedure (SAP), and then briefly explain how to stack RBNs. As the mixed Gaussian RBN and Bernoulli-Bernoulli RBN are similar, we just

use mixture Gaussian RBN as an example to illustrate the SAP algorithm.

Considering Eq. (5), maximizing marginal log-likelihood can be used to learn model parameters Θ. Let $\mathscr{D} = \{\boldsymbol{x}^{(m)}\}_{m=1}^{M}$ denote the training data set, the target of proposed model can be shown as Eq. (9):

$$\begin{aligned}\mathscr{L}(\mathscr{D};\Theta) &= \sum_m \log P_\Theta(\boldsymbol{x}^{(m)}) \\ &= \sum_m \log\left(\sum_{\boldsymbol{h}} P_\Theta(\boldsymbol{x}^{(m)},\boldsymbol{h})\right). \end{aligned} \tag{9}$$

Parameters are typically computed by gradient ascent to maximize the object function. The parameter gradient is shown in Eq. (10):

$$\nabla_\theta \mathscr{L}(\mathscr{D};\Theta) = \sum_m \sum_h P_\Theta(\boldsymbol{h}|\boldsymbol{x}^{(m)}) \frac{\partial - E_\Theta(\boldsymbol{x}^{(m)},\boldsymbol{h})}{\partial\theta}. \tag{10}$$

According to Eq. (10), it is intractable to compute the exact gradient. Due to the following two reasons, the posterior probability $P_\Theta(h|x^{(m)})$ cannot be calculated. First, it is intractable to computed because of the top-down connections of RBN. Second, the exact gradient requires exponential summations of all possible latent variables $\boldsymbol{h}$.

Variational approximation algorithms, such as the mean field algorithm [17], can be used to solve the above intractable inference. However, such approximations inevitably introduce a gap between the true posterior probability and corresponding approximate, since the relationships will be discarded in the approximate distribution.

The true posterior probability can be estimated through sampling. In the work, we try to remain the dependencies by directly sampling. Specifically, Gibbs sampling is employed to draw samples for the hidden variables. One latent variable is sampled under fixing all the other variables. In this way, dependencies can be preserved to some degrees.

Since the specific form of $P(\boldsymbol{h}|\boldsymbol{x})$ is not available, it is intractable to draw samples exactly from $P(\boldsymbol{h}|\boldsymbol{x})$ through Gibbs sampling. To address the problem, we utilize some approximations during the sampling:

$$P(\boldsymbol{h}|\boldsymbol{x}) = \prod_j P(h_j|h_1,\ldots,h_{j-1},\boldsymbol{x}) \approx \prod_j P(h_j|\boldsymbol{h}_{-j},\boldsymbol{x}). \tag{11}$$

where $h_{-j} = \{h_1,\ldots;,h_{j-1},h_{j+1},\ldots,h_{n_h}\}$. Specifically, we update hidden variable h_j through Eq. (12).

$$h_j^t \sim P(h_j|\boldsymbol{x},\boldsymbol{h}_{-j}^{t-1}). \tag{12}$$

Given $\boldsymbol{x}$ and $\boldsymbol{h}_{-j}$,

$$P(h_j = 1|\boldsymbol{x},\boldsymbol{h}_{-j}^{t-1}) = \frac{P(h_j=1,\boldsymbol{h}_{-j}^{t-1},\boldsymbol{x})}{P(h_j=1,\boldsymbol{h}_{-j}^{t-1},\boldsymbol{x} + P(h_j=0,\boldsymbol{h}_{-j}^{t-1},\boldsymbol{x})} \tag{13}$$

According to Eq. (5) or Eq. (7), the probability can be calculated and the hidden variable h_j will be updated through sampling. In this way, one latent variable is sampled under fixing all the other variables. The sampling procedure goes over for several iterations until convergence, and then a sample is collected and used to update parameters.

To address the exponential summations, Markov Chain Monte Carlo methods are typically adopted. An intuitive estimation is shown in Eq. (14).

$$\nabla_\theta \mathscr{L}(\mathscr{D};\Theta) \approx \frac{1}{n}\sum_m \sum_s \frac{\partial - E_\Theta(\boldsymbol{x}^{(m)},\boldsymbol{h}^{(m,s)})}{\partial\theta}, \tag{14}$$

where $\boldsymbol{h}^{(m,1)},\ldots,\boldsymbol{h}^{(m,n)}$ are n samples from $P(\boldsymbol{h}|\boldsymbol{x}^{(m)})$. In this work, we avoid multiple Gibbs chains by employing the stochastic approximation procedure (SAP) framework [18], so only one sample of the latent variables is used to estimate the gradient.

Under some mild assumptions, the SAP is guaranteed to converge to a local optimum [19] if the learning rate γ_t satisfies Eq. (15),

$$\begin{aligned}\sum_{t=1}^{\infty}\gamma_t &= \infty, \\ \sum_{t=1}^{\infty}\gamma_t^2 &< \infty. \end{aligned} \tag{15}$$

Then, the gradient is calculated shown in Eq. (16).

$$\nabla_\theta \mathscr{L}(\mathscr{D};\Theta) \approx \sum_m \frac{\partial - E_\Theta(\boldsymbol{x}^{(m)},\boldsymbol{h}^{(m)})}{\partial\theta}, \tag{16}$$

Since the energy function is merely a linear function of the parameters, the derivate has a simple formulation. Due to $\Theta = \{\boldsymbol{W},\boldsymbol{\sigma},\boldsymbol{b},\boldsymbol{d}\}$, the four parts of Θ can be calculated as following:

$$\frac{\partial - E_\Theta(\boldsymbol{x}^{(m)},\boldsymbol{h}^{(m)})}{\partial w_{ij}} = \frac{h_j^{(m)}(x_i^{(m)} - b_i - \boldsymbol{w}_i^T\boldsymbol{h}^{(m)})}{\sigma_i^2}. \tag{17}$$

$$\begin{aligned}\frac{\partial - E_\Theta(\boldsymbol{x}^{(m)},\boldsymbol{h}^{(m)})}{\partial\sigma_i^2} &= -\sum_i \frac{(x_i^{(m)} - b_i)^2}{\sigma_i^3} \\ &+ \sum_i \frac{2(x_i^{(m)} - b_i)}{\sigma_i^3}\boldsymbol{w}_i^T\boldsymbol{h} - \sum_i \frac{1}{\sigma_i^3}(\boldsymbol{w}_i^2\boldsymbol{h})^2 \end{aligned} \tag{18}$$

$$\frac{\partial - E_\Theta(\boldsymbol{x}^{(m)},\boldsymbol{h}^{(m)})}{\partial b_i} = \frac{x_i^{(m)} - b_i - w_i^T h}{\sigma_i^2} \tag{19}$$

$$\frac{\partial - E_\Theta(\boldsymbol{x}^{(m)},\boldsymbol{h}^{(m)})}{\partial d_j} = h_j^{(m)} \tag{20}$$

To make the learning phase fast and enable our method to scale up to large dataset in practice, the mini-batch stochastic gradient ascent algorithm is adopted, which calculates the gradient through a random mini-batch of training data. In the work, we go over the training set several epochs until convergence is reached.

Detailed learning algorithm of RBN is summarized in Algorithm 1.

During layer-wise training, each RBN is pre-trained using the above learning algorithm, and the hidden layer of the lower RBN is used as the visible layer of its upper RBN. We train two stacked RBNs - one for the visual modality and one for the audio modality. After that, we stack a joint representation layer upon

Algorithm 1: Parameter Learning for an RBN [16]

Input training data $\mathscr{D} = \{\boldsymbol{x}^{(m)}\}_{m=1}^{M}$;
Output optimal parameters $\Theta = \{\boldsymbol{W}, \boldsymbol{\sigma}, \boldsymbol{b}, \boldsymbol{d}\}$.

```
the parameters of RBN Θ are initialized randomly;
at first, hidden variables are generated by Gibbs sampling;
while learned parameters not optimal, do
    a mini-batch of training data x are randomly selected from 𝒟;
    for the mini-batch of training data, corresponding hidden variables are sampled by Gibbs sampling, h^(t) ~ P(h|x, h^(t−1));
    the gradient is calculated by Algorithm 1;
    Update the parameters, θ_t = θ_{t−1} + γ_t ∇_θ 𝓛(x) .
end while
return Θ = {W, σ, b, d}.
```

the top layer of the two stacked RBNs. Specifically, the visible variables of the joint representation layer are a concatenation of the top layers of the audio and visual modality networks.

The latent variables of the joint representation layer are considered as the joint representation from two different modalities. The multimodal layer is also pre-trained using the above learning algorithm. After stacking the trained RBNs, we obtain an MMDRBN as shown in the left part of Fig. 2.

2) An Inference Network Learning: Due to the edge direction, it is intractable to compute exact inference for the RBN model. Therefore, we learn an inference network as the basis of the MMDRBN. To approximate the inference of the RBNs, another distribution Q_Φ is introduced. To close to $P_\Theta(h|x)$, we minimize the KL divergence between them. Φ is the parameters set of Q and P_Θ is the RBN learned with our proposed method. The equation is shown in Eq. (21):

$$KL(Q_\Phi(h|x)||P_\Theta(h|x)) = \sum_h Q_\Phi(h|x) log \frac{Q_\Phi(h|x)}{P_\Theta(h|x)}. \quad (21)$$

Distribution Q is adopted to close to the intractable exact inference $P_\Theta(h|x)$ through minimizing KL divergence. We optimize the KL divergence by gradient descent, and the gradient is calculated as follows:

$$\begin{aligned}
&\frac{\partial KL(Q_\Phi(h|x)||P_\Theta(h|x))}{\Phi} \\
&= \sum_h \frac{\partial Q_\Phi(h|x)}{\Phi} log Q_\Phi + \sum_h \frac{\partial Q_\Phi(h|x)}{\Phi} \\
&\quad - \sum_h log P_\Theta(x,h) \frac{\partial Q_\Phi(h|x)}{\Phi} \\
&= E((log P_\Theta(x,h) - log Q_\Phi(h|x)) \\
&\quad \times \frac{\partial log Q_\Phi(h|x)}{\partial \Phi}. \quad (22)
\end{aligned}$$

Since it is time consuming to calculate the expectation in Eq. (22), a sampling-based method is used to estimate the expectation. After sampling n samples from the distribution $Q_\Phi(h|x)$, the approximation of the expectation can be calculated as Eq. (23):

$$\begin{aligned}
&\frac{\partial KL(Q_\Phi(h|x)||P_\Theta(h|x))}{\Phi} \\
&= \frac{1}{n} \sum_{i=1}^{n} ((log P_\Theta(x, h^{(i)}) - log Q_\Phi(h^{(i)}|x)) \\
&\quad \times \frac{\partial log Q_\Phi(h^{(i)}|x)}{\partial \Phi}. \quad (23)
\end{aligned}$$

The mini-batch stochastic gradient ascent algorithm is also employed here.

3) Discriminative Fine Turning and Inference: After initializing an inference network from an MMDRBN, a layer of labels is added on the top of the inference network. Given the label layer, we can fine-tune the inference network through a back-propagation algorithm. Since it is a multimodal inference network, there is a little difference between the adopted back-propagation and the standard back-propagation algorithm (BP). There are two steps for fine-tuning the parameters. In first step, the BP is applied to the multimodal part of Fig. 2. The top layers of the audio and visual modality networks are treated as a individual layer in this step. After fine-tuning the multimodal part, the errors propagate to the top layers of the audio and visual modality networks. In second step, the back-propagated errors are collected from the top layer of the two modalities. According to the errors, the remaining parts of the two networks are fine-tuned through the BP. The final inference network can be used to predict discrete or continuous emotion scores from video content through a forward propagation algorithm.

C. Knowledge-Augmented Multimodal Deep Regression Bayesian Networks

Since the joint representation learned from the audio and visual modalities may be semantically meaningless, domain knowledge should be introduced to bridge the semantic gap. Therefore, we extend the proposed MMDRBN to a knowledge-augmented MMDRBN, as shown in Fig. 3.

Like the proposed MMDRBN, the knowledge-augmented MMDRBN contains of two stacked RBNs, one for the visual modality and one for the audio modality. The two stacked RBNs are layerwisely trained and stacked using the algorithm proposed in Section IV-B1. After obtaining a network of two modalities, a multimodal layer and domain knowledge layer are added upon the top layer. Specifically, the visible variables of the multimodal layer are a link of the top layers of the audio and visual modality networks. The latent variables are the joint representation of the two modalities and attributes from the summarized domain knowledge. Since the RBN can model the dependencies between visible variables and latent variables as well as the dependencies among latent variables, there are relations not only between attributes and the multimodal layer, but also between attributes and joint representation. The attributes function as domain knowledge, leading to semantically meaningful joint

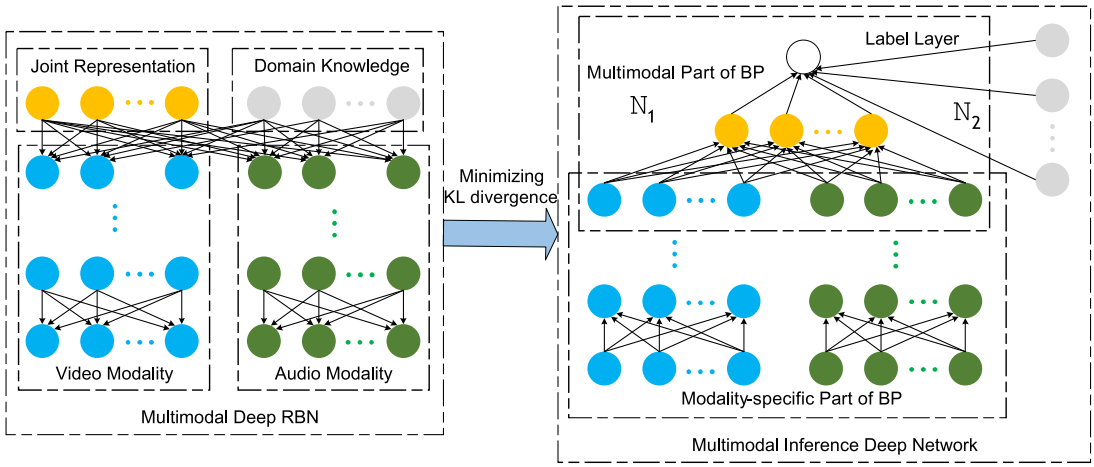


Fig. 3. The framework of the proposed Knowledge-Augmented Multimodal Deep Regression Bayesian Networks for Emotion Video Tagging.

representation. After stacking the trained RBNs, we obtain a domain knowledge-augmented MMDRBN as shown on the left side of Fig. 3. Since exact inference is intractable, we transfer the domain knowledge-augmented MMDRBN into an inference network by minimizing the KL divergence, as shown on the right side of Fig. 3. We then add a label layer for the multimodal and domain knowledge parts. The label layer is utilized to fine-tune the two networks through the back propagation algorithm.

Compared to the MMDRBN, this model has an extra layer representing domain knowledge, which can further capture the dependencies among both modalities to use domain knowledge for superior emotion video tagging.

1) Summarized Domain Knowledge: Both audio and visual elements are employed by filmmakers to communicate emotions to audiences. These elements, also called attributes, capture film production rules and bridge the gap between the high-level affective semantics of movie content and its low-level audio and visual features.

Following Chen *et al.*'s [20] work, we focus on three visual video elements: color, lighting, and tempo. Color is an crucial film element that can be changed to affect the audiences' emotion. Generally, colors are categorized into two groups: cool colors and warm colors. The cool colors are less bold and provocative than warm colors. By creating a scene with cool colors, the filmmakers intend to present a scene of calmness and introspection. On the contrary, the warm colors are typically utilized to present a scene of energy, life, and outward tendencies. We use color energy [5] to measure the color composition, and describe a video's color as either high color energy or low color energy.

Lighting is another powerful video element that can be manipulated to establish certain moods for the viewers. Lighting can be categorized as high key or low key. High key is typically employed to create the lighthearted and warm atmosphere typical of joyous scenes, while the latter generates sad, surprising, frightening, or suspenseful scenes with dim lights, shadow play, and predominantly dark backgrounds. Based on this, we propose to use lighting key [21] to measure lighting, describing a video's lighting as either high key or low key.

Finally, tempo is a measure of video dynamics. Tempo has significant power to affect emotional intensity. For example, a high tempo of action can induce stress and excitement, while a slow tempo creates a more relaxed and slower-paced scene. In this paper, average shot duration [22] is used to measure the pace of a sequence in a movie clip. The movie's tempo is labelled high or low, depending on whether or not the average shot duration of the clip is longer than the median value.

In addition to the visual elements, a video's emotion tagging can be characterized by certain audio elements. Scherer [23] summarized several emotion-sensitive audio elements including fundamental frequency (F0), formants, intensity, spectral parameters, and duration, as shown in Table I. For example, a high F1 formant mean is often related to negative emotions such as disgust and sadness, while low F1 formant means typically represent happiness. Happiness is produced with wider first formant bandwidth, while disgust and sadness are produced with narrow first formant bandwidth. Energy concentrates in high-frequency regions for high arousal emotions like anger, and in low-frequency regions for low arousal emotions such as sadness. The audio elements summarized by Scherer [23] are adopted as audio attributes in the proposed knowledge-augmented MMDRBN.

2) Learning and Inference: After summarizing the well-established film grammar and manually defining the emotion-sensitive attributes, we use these attributes as the input of the domain knowledge layer as shown in Fig. 3.

The learning process of the knowledge-augmented MMDRBN consists of three steps: learning a multimodal generative network, learning an inference network as well as discriminative fine-tuning and inference.

The learning process of generative network for the domain knowledge-augmented MMDRBN includes layer-wised training of several RBN according to Algorithm 1. All the RBNs use the same learning algorithms, except for the top RBN. The latent nodes of the top RBN are joint representation and domain knowledge rather than joint representation only. Let $\boldsymbol{h} = \{\boldsymbol{h_j}, \boldsymbol{h_d}\}$, where $\boldsymbol{h}$ represents hidden nodes consisting of joint representation $\boldsymbol{h_j}$ and domain knowledge $\boldsymbol{h_d}$. Since $\boldsymbol{h_d}$ is known, the

TABLE I
THE DEPENDENCIES BETWEEN EMOTION VIDEO TAGGING AND AUDIO ATTRIBUTES [23]

	Enjoyment/ Happiness	Elation/ Joy	Displeasure/ Disgust	Contempt/ Scorn	Sadness/ Dejection	Grief/ Desperation	Anxiety/ Worry	Fear/ Terror	Irritation/ Cold anger	Rage/ Hot anger	Boredom/ Indifference	Shame/ Guilt
Fundamental frequency												
Shift regularity		↓						↓		↓	↑	
Perturbation	↓	↑			↑	↑		↑		↑		
Variability	↓	↑			↓	↑		↑↑	↓	↑↑		
Range	↓	↑	↑	↓↑	↓↑	↑	↑	↑↑	↓↑	↓↑	↓	↑
Contour	↓	↑			↓	↑	↑	↑↑	↓			↑
Formants												
Formant precision		↑	↑	↑	↓	↑	↑	↑	↑	↑		↑
Second formant Mean			↓	↓	↓	↓	↓	↓	↓	↓	↓	↓
First formant Mean	↓		↓	↑	↑	↑	↑	↑	↑	↑	↑	↑
First formant Bandwidth	↑	↓↑	↓↓	↓	↓↑	↓↓	↓	↓↓	↓↓	↓↓	↓	↓
Intensity												
Variability	↓	↑			↓			↑		↑		
Range	↓	↑			↓			↑	↑	↑		
Mean	↑	↑	↑	↑↑	↓↓	↑		↑	↑	↑↑	↓↑	
Spectral parameters												
Spectral noise					↑							
High-frequency energy	↓	↓↑	↑	↑	↓↑	↑↑	↑	↑↑	↑↑	↑↑	↓↑	↑
Frequency range	↑	↑	↑	↑↑	↑	↑↑		↑↑	↑	↑	↑	
Duration												
Transition time	↑	↓			↑	↓		↓		↓		
Speech rate	↓	↑			↓	↑		↑↑		↑		

Note: ↑ represents increase; ↓ represents decrease. Double symbols refer to enhanced chagnes. Two opposite arrows indicate that the antecedent audio elements exert oppoising influence.

Gibbs sampling shown in Eq. (12) should be modified as follows:

$$\boldsymbol{h_j}_k^t \sim P(h_{j\,k}|\boldsymbol{x}, \boldsymbol{h_j}_{-k}^{t-1}, \boldsymbol{h_d})\,. \tag{24}$$

The learning process of the inference network for the knowledge-augmented MMDRBN is as the same as that for MMDRBN, described in Section IV-B.

After learning the inference network, the connections of the domain knowledge-augmented MMDRBN become bottom-up rather than top-down. Since the value of domain knowledge is known, the connections pointing to it can be removed.

During the last step, i.e., discriminative fine turning and inference, the domain knowledge is used to create an extra two-layer network, $\mathcal{N}_2$, as shown on the right side of Fig. 3. Let $\mathcal{N}_1$ represents the predicted emotion labels from visual and auditory content and $\mathcal{N}_2$ represents the predicted emotion labels from domain knowledge. γ is a tradeoff between $\mathcal{N}_1$ and $\mathcal{N}_2$. The output of domain knowledge-augmented MMDRBN $\mathcal{N}$ can be obtained as follows:

$$\mathcal{N} = \gamma\mathcal{N}_1 * (1-\gamma)\mathcal{N}_2\,. \tag{25}$$

Similar as Section IV-B3, back-propagation is adopted for discriminative fine tuning.

V. EXPERIMENTS

A. Databases

Several benchmark databases can be used for emotion video tagging, including the LIRIS-ACCEDE database [24], the MAHNOB-HCI database [25], and the Database for Emotion Analysis using Physiological Signals (DEAP) [26]. They contains 9,800 film excerpts, 20 emotional videos, and 120 one-minute music video excerpts respectively. Since the proposed model requires as many samples as possible for better performance and to avoid over-fitting, the LIRIS-ACCEDE database is adopted for the evaluation.

The video clips from the LIRIS-ACCEDE database are relative assessed along the induced arousal and valence axes ranging from 0 to 9,799. Based on these ranks of valence and arousal, MediaEval 2015 [27] propose classification task, and MediaEval 2016 [28] and MediaEval 2017 [29] propose regression tasks. MediaEval 2015 uniformly rescales the ranks to a more common $[-1, 1]$ range, and then assigns valence and arousal scores as -1, 0, or 1 corresponding to three ranges $[-1, -0.15]$, $[-0.15, 0.15]$ and $[0.15, 1]$. MediaEval 2016 provides the absolute affective scores, which is estimated from initial ranks using Gaussian regression models, as the ground truth for the regression task. MediaEval 2017 [29] proposes to predict valence and arousal on long movies, i.e. consecutive 10-second segments sliding over the whole movie with a shift of 5 seconds.

B. Experimental Conditions

In our experiments, the classification task proposed by MediaEval 2015, the regression task proposed by MediaEval 2016 and MediaEval 2017 are considered.

For the classification task, the features proposed in [24] are adopted in our work; a set of 17 features is employed for valence and a set of 12 features is used for arousal. We also augment the features with 31 dimensional audio features and three visual features adopted in [30]. All features are normalized prior to

TABLE II
EXPERIMENTAL RESULTS OF EMOTION TAGGING FROM VIDEO CONTENT ON THE LIRIS-ACCEDE

	MediaEval 2015		MediaEval 2016				MediaEval 2017			
	Valence	Arousal	Valence		Arousal		Valence		Arousal	
	Acc	Acc	MSE	PCC	MSE	PCC	MSE	PCC	MSE	PCC
LCCA	42.16	63.33	0.416	0.199	0.923	0.226	0.141	0.214	0.126	0.095
KCCA	42.95	63.32	0.402	0.211	0.891	0.245	0.138	0.221	0.126	0.040
DCCA	41.77	63.58	0.393	0.241	0.882	0.256	0.135	0.243	0.125	0.116
DCCAE	43.38	63.74	0.414	0.231	0.879	0.271	0.131	0.251	0.122	0.163
MMDBM	41.38	63.75	0.388	0.186	0.921	0.202	0.127	0.282	0.110	0.285
Early fusion	43.74	63.85	0.344	0.331	0.834	0.312	0.120	0.297	0.107	0.306
Late fusion	42.22	63.76	0.362	0.260	0.816	0.323	0.122	0.304	0.104	0.298
none	44.26	64.30	0.332	0.387	0.766	0.416	0.110	0.351	0.103	0.315
audio	45.80	64.80	0.325	0.429	0.721	0.451	0.109	0.366	0.100	0.329
visual	44.59	64.48	0.323	0.386	0.738	0.452	0.111	0.359	0.102	0.314
audio+visual	**46.73**	**65.10**	**0.303**	**0.450**	**0.713**	**0.470**	**0.103**	**0.375**	**0.099**	**0.330**

the classification task. For the regression task, we do not use the slope of the power spectrum and audio flatness envelope from the arousal feature set following the baseline features provided by MediaEval 2016. On the MediaEval 2017 dataset, we adopt the features provided by the organizers for audio and visual modalities. Specifically, for audio modality, 1582 features provided by the organizers are normalized. And then, we reduce the dimensionality of the features by 99% of variation via principal component analysis (PCA). For visual modality, all features are concatenated and the mean value and standard deviation value of corresponding segments are calculated. And then, PCA is adopted to reduce feature dimensionality by keep 99% variation. As to domain knowledge, the features summarized in Section IV-C1 are extracted.

Ten-fold cross-validation is used on the MediaEval 2015 and MediaEval 2016. For MediaEval 2017, we train our models on the DevSet and test it on the TestSet provided by the organizers. The deep RBNs for audio modality and visual modality each consist of two layers of RBNs. For the visual modality, the number of nodes of the first hidden layer and that of the second hidden layer are set to eight and 18 respectively. For the audio modality, the number of nodes of the first hidden layer and that of the second hidden layer are set to 30 and 18 respectively. The dimension of the joint representation layer is 18. We utilize accuracy as the metric for the classification task, and Pearson correlation coefficient (PCC) and mean squared error (MSE) as the metrics for the regression task.

Under the above data and experimental settings, we designed several experiments to demonstrate the effectiveness of our method from many aspects.

To demonstrate the superiority of the proposed method over other multimodal methods, our method is compared to the results of MMDBM, LCCA, KCCA, DCCA and DCCAE under the same experimental conditions. To demonstrate the superiority of the multimodal process in our method, the proposed model is compared with the early fusion strategy and late fusion strategy. For the early fusion strategy, the two modalities are concatenate into one vector and trained for the concatenate feature vectors. For the late fusion strategy, two individual inference networks are learned through the proposed method. The two inference networks can analyze the emotion video content separately. And then, we combined the output of the two individual networks to predicted the emotion label. We also demonstrate that the joint representation of the abstracted features is superior to simply concatenate features, which is clarified in [14]. To demonstrate the advantage of domain knowledge in our method, we compare MMDRBN (**none**), MMDRBN augmented by audio domain knowledge only (**audio**), MMDRBN augmented by visual domain knowledge only (**visual**), and MMDRBN augmented by both audio and visual domain knowledge (**audio+visual**). The results are shown in Table II. We also compare our classification and regression results to the state-of-the-art works [20], [31]–[34] in Table III, [35]–[39] in Table IV and [40]–[44] in Table V.

TABLE III
COMPARISON WITH RELATED WORK ON MEDIAEVAL 2015

	Valence	Arousal
MIC-TJU [31]	41.95	55.93
NII-UIT [34]	42.96	55.91
ICL-TUM-PASSAU [32]	41.48	55.72
Fudan-Huawei [33]	41.80	48.80
TCS-ILAB [45]	35.66	48.95
UMons [46]	37.28	52.44
RFA [47]	33.03	45.04
KIT [48]	38.50	51.90
Chen *et al.* [20]	43.18	60.88
audio+visual	**46.73**	**65.10**

TABLE IV
COMPARISON WITH RELATED WORK ON MEDIAEVAL 2016

	Valence		Arousal	
	MSE	PCC	MSE	PCC
RUC [35](run 1)	0.218	0.312	1.479	0.405
THU-HCSI [36]	**0.214**	0.296	1.531	0.267
AUTH-SGP [37]	/	0.290	/	0.303
BUL [38]	0.231	0.149	1.413	0.271
Liu *et al.* [39]	0.240	0.102	1.185	0.159
audio+visual	0.303	**0.450**	**0.713**	**0.470**

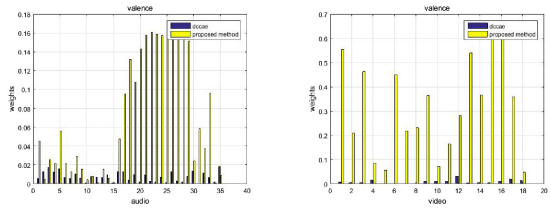
C. Experimental Results and Analysis of MMDRBN

From Table II, we obtain several observations.

First, the proposed MMDRBN shows good performance on the three datasets. Specifically, on the MediaEval 2015 dataset, the accuracy is 44.26% for valence and 64.30% for arousal. On the MediaEval 2016 dataset, the MSE is 0.332 and

TABLE V
COMPARISON WITH RELATED WORK ON MEDIAEVAL 2017

	Valence		Arousal	
	MSE	PCC	MSE	PCC
BOUN-NKU (run 5) [40]	0.188	0.090	0.113	0.219
HKBU (run 1) [41]	0.191	-0.151	0.126	0.045
MIC-TJU (run 3) [42]	0.213	0.153	0.140	0.082
THUHCSI (run 3) [43]	0.183	0.371	0.117	0.321
THNJ-CS (run 2) [44]	**0.046**	0.006	0.113	0.215
audio+visual	0.103	**0.375**	**0.099**	**0.330**



(a) (b)

Fig. 4. Visualization of parameters for classification task.

the PCC is 0.387 in the valence space. For arousal, the proposed MMDRBN achieves 0.766 for MSE and 0.416 for PCC. On the MediaEval 2017, the MSE is 0.110 and the PCC is 0.351 in the valence space. For arousal, the proposed MMDRBN achieves 0.103 for MSE and 0.315 for PCC. The results on arousal are higher than that of valence for the three datasets. This may demonstrate that the inner pattern of arousal data is more helpful for recognition.

Second, the proposed MMDRBN outperforms the related multimodal methods for the three datasets. LCCA and KCCA aim to construct the common space from different modalities. They directly construct a representation from the input modalities, without the layer-wise feature abstraction. Therefore, their input features contain modality-specific information, making it harder to discover relationships across modalities [14]. While our proposed model eliminates the modality-specific information through deep networks, and then constructs a joint representation of two modalities. This may be the reason why our model outperforms LCCA, and KCCA. DCCA, DCCAE, and MMDBM all consist of several RBMs. DCCA and DCCAE combine the RBM-based deep neural network and CCA method. MMDBM is constructed with DBMs of several modalities. Our experimental results are also superior to those of DCCA, DCCAE, and MMDBM, since the RBN is able to capture more dependencies than the RBM can, as discussed in Section IV. For an in-depth analysis of the comparison results, we visualize and analyze the learned features of our model and DCCAE which perform best among LCCA, KCCA and DCCA. For example, in the classification task, we visualize the parameters of the valence model of the bottom layers for the LRBN model and DAACE model in Fig. 4. In the visual figure, parameters of our model focus on colorfulness, fades, and number of scene cuts, as these features are closely related to emotion according to [5]. In the audio figure, we obtain larger parameters on MFCC-related features, which are commonly used features in affective computing.

Third, the proposed MMDRBN obtains better performance than the early fusion and late fusion methods. Both early fusion and late fusion methods are based on RBNs and are learned using Algorithm 1 proposed in Section IV. This comparison further demonstrates the important role of joint representation in a multimodal method. The joint representation can capture the relations between the features of different modalities with fewer differences in modal concepts that are found when merging the raw features directly. From the comparison results, we can infer that representing multimodal data in the same output space is one of the most important aspects for boosting model performance.

D. Experimental Results and Analysis of Knowledge-Augmented MMDRBN

From Table II, we find that: compared to different domain knowledge configurations, the MMDRBN augmented by both audio and visual domain knowledge achieves the highest results. The four methods share parameters. This demonstrates that the proposed model, which leverages more domain knowledge, can model more inherent connections between video and emotions. Additionally, MMDRBN augmented by audio domain knowledge only outperforms MMDRBN augmented by visual domain knowledge only. It is reasonable, since there are about five times as many audio attributes as there are visual attributes.

To further validate the effectiveness of the proposed models in learning data representation, we employ t-SNE to represent the learned middle-level representations. For example, for valence classification on MediaEval 2015, Fig. 5 shows the t-SNE embedding of raw video data, the t-SNE embedding of the learned joint representations using a multimodal deep regression Bayesian network without exploring domain knowledge, and the t-SNE embedding of the learned joint representations using the proposed knowledge-augmented multimodal deep regression Bayesian network. As shown in Fig. 5(a), the videos with different valence labels tend to distribute uniformly, and the centers of three categories are very close to each other. This demonstrates that it is difficult to classify videos using the raw input features. In Fig. 5(b), videos with different valence labels distribute slightly differently, and the sample centers are separated. This demonstrates that, compared to the raw features, the learned joint representation from video data is more discriminative for emotion video tagging, since it can leverage the inherent relations between the audio and visual modalities. In Fig. 5(c), the videos with different valence labels are distributed differently. The sample centers are farther apart, especially the negative videos (in red points). This confirms that the learned joint representation from both video data and film grammar is superior to either the raw features or the data-based middle-level video representation for emotion video tagging. Since emotion video tagging is challenging for the ACCEDE-LIRIS database, three valence classes do not completely distribute separately in these figures, demonstrating the challenges of emotion video tagging.

To analyze the contributions of the domain knowledge, we conduct an ablation study by adjusting tradeoff γ from 0 to 1 in Eq. (25). Fig. 6 shows the experimental results on four tasks.

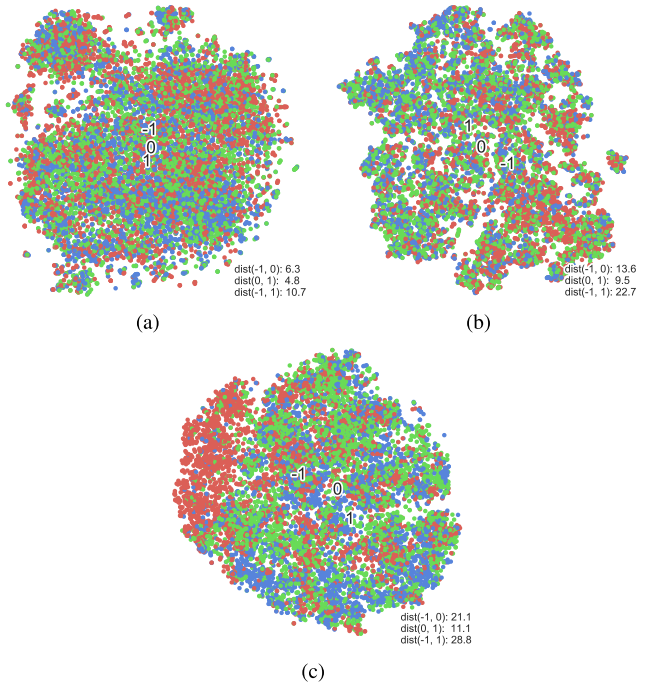


Fig. 5. (a): A t-SNE embedding of video features on the LRIRS-ACCEDE database; (b): A t-SNE embedding of joint representation learned from data only; (c) A t-SNE embedding of joint representation learned from both data and domain knowledge. -1, 0, and -1 in the figures represent the sample centers of negative videos, neutral videos and positive videos. dist means distance between sample centers.

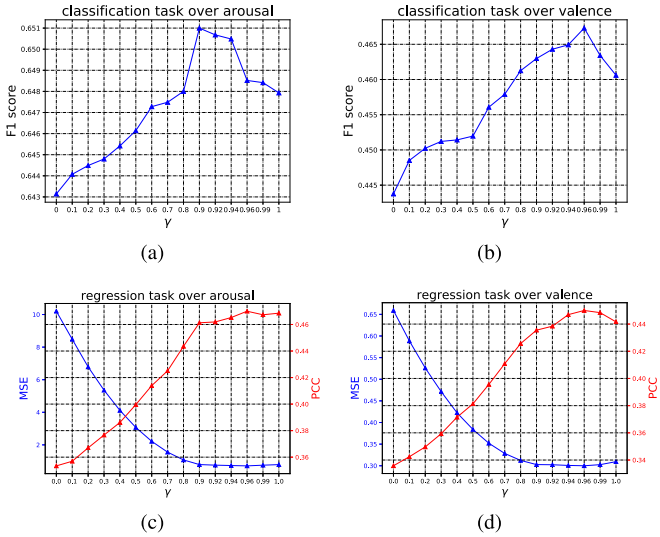


Fig. 6. Experimental results varying the parameter γ in the range from 0 to 1.

From the four figures, we can find that when given an appropriate γ, our method considering both the multimodal network and domain knowledge outperforms the methods that only consider the multimodal network or only use domain knowledge. In addition, the multimodal network plays a more important role in our method. This is reasonable, considering that there are fewer inputs and a simpler structure.

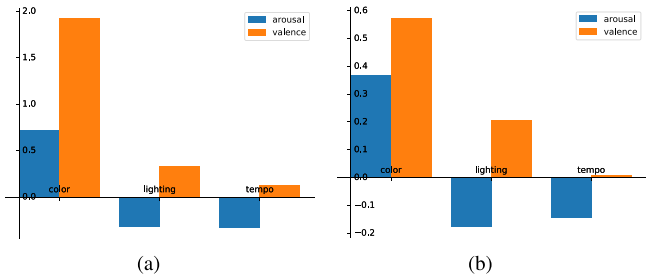


Fig. 7. Visualization of parameters for visual domain knowledge on the (a) MediaEval 2016 and (b) MediaEval 2017 datasets.

To further analyze the dependencies between emotion and domain knowledge, we have visualized the learned parameters for three visual domain knowledge on the MediaEval 2016 and MediaEval 2017 datasets as examples. From Fig. 7, we can find the following observations: first, among the three visual elements, the learned weight of color is the largest for both valence and arousal on two datasets. It may indicate that color has the most significant influence on emotions. It is reasonable, since color is a crucial film element that can be changed to affect the audiences' emotions. Second, compared with valence, tempo has more influence on arousal since the learned absolute weight for arousal is larger than that for valence on both datasets. It is expected, since tempo is a measure of video dynamics, and thus has power to affect emotional intensity, which is related to arousal. In conclusion, the different correlations between emotion tag and domain knowledge may be the root cause of different γ.

E. Comparison to Related Works

Table III lists the results of all works published in MediaEval 2015 [27]. Because the experimental settings and used features differ, we compare the highest results for reference only. Considering that our features are the simplest among all the related works, our high results demonstrate the strong ability of the proposed method to capture the dependencies among data. In Chen *et al.*'s work [20], only visual domain knowledge was adopted and audio domain knowledge was ignored. For the proposed method, the domain knowledge is used to help model the joint representation for audio and visual modalities and to bridge the semantic gap between representation and emotion video tagging. Thus, the proposed model achieves better performance.

Table IV lists the results of all works published in MediaEval 2016 [28]. Just as the above comparisons, this information should only be used for reference due to differences in experimental data and settings. Our model outperforms all of the listed works except for RUC [35], which achieves the best overall experimental results among all the participants of MediaEval 2016. Our results are competitive, though the MSE of valence is poorer than the results achieved by other methods.

Table V lists the results of all works published in MediaEval 2017 [29]. Because their used features are different from ours, we compare with their highest results for reference only. For arousal prediction, the proposed method outperforms all the

listed work. For valence prediction, the proposed method performs better than all the listed work expect for TCNJ-CS [44], who has a better MSE but a poorer PCC. Therefore, our results are still competitive.

Considering the results of three tasks, our model not only achieves good results for video tagging, but also has an excellent generalization ability.

VI. Conclusion

This paper proposes a knowledge-augmented multimodal deep regression Bayesian network to explore both the relationships between the audio and visual modalities and the domain knowledge for emotion video tagging. First, the well-established film grammar is investigated and the emotion-sensitive attributes are defined. Then, a knowledge-augmented multimodal deep regression Bayesian network is constructed to learn the middle-level representation from both video data and the emotion-sensitive attributes. Efficient learning and inference algorithms are also developed. Experimental results and comparisons on the LIRIS-ACCEDE database show the superiority of the proposed method. Although the proposed method successfully leverages the well-established film grammar for emotion tagging from videos produced by professionals, it may be not suitable to detect emotions from videos generated by ordinary users, who usually do not have the expertise to apply various film grammars to induce viewers' emotions. How to define the semantically meaning middle-level representation for user-generated videos is worthy of further study.

References

[1] L. Pang, S. Zhu, and C.-W. Ngo, "Deep multimodal learning for affective analysis and retrieval," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 2008–2020, Nov. 2015.

[2] Q. Gan, S. Wang, L. Hao, and Q. Ji, "A multimodal deep regression Bayesian network for affective video content analyses," in *Proc. IEEE Int. Conf. Comput. Vision*, 2017, pp. 5123–5132.

[3] S. Wang and Q. Ji, "Video affective content analysis: A survey of state-of-the-art methods," *IEEE Trans. Affect. Comput.*, vol. 6, no. 4, pp. 410–430, Oct.–Dec. 2015.

[4] A. Hanjalic and L.-Q. Xu, "Affective video content representation and modeling," *IEEE Trans. Multimedia*, vol. 7, no. 1, pp. 143–154, Feb. 2005.

[5] H. L. Wang and L.-F. Cheong, "Affective understanding in film," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 6, pp. 689–704, Jun. 2006.

[6] E. Acar, F. Hopfgartner, and S. Albayrak, "Understanding affective content of music videos through learned representations," in *Proc. Int. Conf. Multimedia Modeling*, Springer, 2014, pp. 303–314.

[7] S. Chen *et al.*, "Implicit hybrid video emotion tagging by integrating video content and users' multiple physiological responses," in *Proc. IEEE 23rd Int. Conf. Pattern Recognit.*, 2016, pp. 295–300.

[8] S. Sun, "A survey of multi-view machine learning," *Neural Comput. Appl.*, vol. 23, no. 7–8, pp. 2031–2038, 2013.

[9] Y. Li, M. Yang, and Z. Zhang, "Multi-view representation learning: A survey from shallow methods to deep methods," 2016, *arXiv:1610.01206*.

[10] T. W. Anderson, T. W. Anderson, T. W. Anderson, and T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*. vol. 2, New York, NY, USA: Wiley, 1958.

[11] S. Akaho, "A Kernel method for canonical correlation analysis," 2006, *arXiv:cs/0609071*.

[12] W. Wang, R. Arora, K. Livescu, and J. Bilmes, "On deep multi-view representation learning," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1083–1092.

[13] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proc. 30th Int. Conf. Mach. Learn.*, 2013, pp. 1247–1255.

[14] N. Srivastava and R. R. Salakhutdinov, "Multimodal learning with deep Boltzmann machines," in *Proc. Advances Neural Inf. Process. Syst.*, 2012, pp. 2222–2230.

[15] J. Ngiam *et al.*, "Multimodal deep learning," in *Proc. 28th Int. Conf. Mach. Learn. (ICML-11)*, 2011, pp. 689–696.

[16] Q. Gan, S. Nie, S. Wang, and Q. Ji, "Differentiating between posed and spontaneous expressions with latent regression Bayesian network," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 4039–4045.

[17] L. K. Saul, T. Jaakkola, and M. I. Jordan, "Mean field theory for sigmoid belief networks," *J. Artif. Intell. Res.*, vol. 4, pp. 61–76, 1996.

[18] H. Robbins and S. Monro, "A stochastic approximation method," in *Herbert Robbins Selected Papers*. New York, NY, USA: Springer, 1985, pp. 102–109.

[19] A. L. Yuille, "The convergence of contrastive divergences," in *Proc. Advances Neural Inf. Process. Syst.*, 2005, pp. 1593–1600.

[20] T. Chen, Y. Wang, S. Wang, and S. Chen, "Exploring domain knowledge for affective video content analyses," in *Proc. ACM Multimedia Conf.*, 2017, pp. 769–776.

[21] H. Zettl, *Sight, Sound, Motion: Applied Media Aesthetics.* Boston, MA, USA: Cengage Learning, 2013.

[22] S. C. Watanapa, B. Thipakorn, and N. Charoenkitkarn, "A sieving ANN for emotion-based movie clip classification," *IEICE Trans. Inf. Syst.*, vol. 91, no. 5, pp. 1562–1572, 2008.

[23] K. R. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech Commun.*, vol. 40, no. 1–2, pp. 227–256, 2003.

[24] Y. Baveye, E. Dellandrea, C. Chamaret, and L. Chen, "LIRIS-ACCEDE: A video database for affective content analysis," *IEEE Trans. Affect. Comput.*, vol. 6, no. 1, pp. 43–55, Jan.–Mar. 2015.

[25] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 42–55, Jan.–Mar. 2012.

[26] S. Koelstra *et al.*, "DEAP: A database for emotion analysis; using physiological signals," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 18–31, Jan.–Mar. 2012.

[27] M. Sjöberg *et al.*, "The MediaEval 2015 affective impact of movies task," presented at the MediaEval 2015 Workshop, Wurzen, Germany, Sept. 14–15, 2015.

[28] E. Dellandréa, L. Chen, Y. Baveye, M. Sjöberg, and C. Chamaret, "The MediaEval 2016 emotional impact of movies task," presented at the MediaEval 2016 Workshop, Hilversum, The Netherlands, Oct. 20–21, 2016.

[29] E. Dellandréa, M. Huigsloot, L. Chen, Y. Baveye, and M. Sjöberg, "The MediaEval 2017 b4emotional impact of movies task," presented at the MediaEval'17 Workshop, Dublin, Ireland, Sep. 13–15, 2017.

[30] S. Chen *et al.*, "Implicit hybrid video emotion tagging by integrating video content and users' multiple physiological responses," in *Proc. Int. Conf. Pattern Recognit.*, 2016, pp. 295–300.

[31] Y. Yi, H. Wang, B. Zhang, and J. Yu, "MIC-TJU in MediaEval 2015 affective impact of movies task," presented at the MediaEval 2015 Workshop, Wurzen, Germany, Sep. 14–15, 2015.

[32] G. Trigeorgis *et al.*, "The ICL-TUM-PASSAU approach for the MediaEval 2015 "affective impact of movies" task," presented at the MediaEval 2015 Workshop, vol. 2015, Wurzen, Germany, Sep. 14–15, 2015.

[33] Q. Dai *et al.*, "Fudan-Huawei at MediaEval 2015: Detecting violent scenes and affective impact in movies with deep learning," presented at the MediaEval 2015 Workshop, Wurzen, Germany, Sep. 14–15, 2015.

[34] V. Lam, S. Phan, D.-D. Le, S. Satoh, and D. A. Duong, "NII-UIT at MediaEval 2015 affective impact of movies task," presented at the MediaEval 2015 Workshop, Wurzen, Germany, Sep. 14–15, 2015.

[35] S. Chen and Q. Jin, "RUC at MediaEval 2016 emotional impact of movies task: Fusion of multimodal features," in presented at the MediaEval 2016 Workshop, Hilversum, The Netherlands, Oct. 20–21, 2016.

[36] Y. Ma, Z. Ye, and M. Xu, "THU-HCSI at MediaEval 2016: Emotional impact of movies task," presented at the MediaEval 2016 Workshop, Hilversum, The Netherlands, Oct. 20–21, 2016.

[37] T. Anastasia and H. Leontios, "AUTH-SGP in MediaEval 2016 emotional impact of movies task," presented at the MediaEval 2016 Workshop, vol. 1739, Hilversum, The Netherlands, Oct. 20–21, 2016.

[38] A. Jan, Y. F. B. A. Gaus, H. Meng, and F. Zhang, "BUL in MediaEval 2016 emotional impact of movies task," presented at the MediaEval 2016 Workshop, Hilversum, The Netherlands, Oct. 20–21, 2016.

[39] Y. Liu, Z. Gu, Y. Zhang, and Y. Liu, "Mining emotional features of movies," presented at the MediaEval 2016 Workshop, Hilversum, The Netherlands, Oct. 20–21, 2016.

[40] N. Karslioglu, Y. Timar, A. A. Salah, and H. Kaya, "BOUN-NKU in MediaEval emotional impact of movies task," presented at the MediaEval'17 Workshop, Dublin, Ireland, Sep. 13–15, 2017.
[41] Y. Liu, Z. Gu, and T. H. Ko, "HKBU at MediaEval 2017 emotional impact of movies task," presented at the MediaEval'17 Workshop, Dublin, Ireland, Sep. 13–15, 2017.
[42] Y. Yi, H. Wang, and J. Wei, "Mic-tju in MediaEval 2017 emotional impact of movies task," presented at the MediaEval'17 Workshop, Dublin, Ireland, Sep. 13–15, 2017.
[43] Z. Jin, Y. Yao, Y. Ma, and M. Xu, "THUHCSI in MediaEval 2017 emotional impact of movies task," presented at the MediaEval'17, Dublin, Ireland, Sep. 13–17, 2017.
[44] S. Yoon, "TCNJ-CS @ MediaEval 2017 emotional impact of movie task," presented at the MediaEval'17 Workshop, Dublin, Ireland, Sep. 13–17, 2017.
[45] R. Chakraborty *et al.*, "TCS-ILAB-MediaEval 2015: Affective impact of movies and violent scene detection," presented at the MediaEval 2015 Workshop, Wurzen, Germany, Sep. 14–15, 2015.
[46] O. Seddati *et al.*, "UMons at MediaEval 2015 affective impact of movies task including violent scenes detection," presented at the MediaEval 2015 Workshop, Wurzen, Germany, Sep. 14–15, 2015.
[47] I. Mironica, B. Ionescu, M. Sjöberg, M. Schedl, and M. Skowron, "RFA at MediaEval 2015 affective impact of movies task: A multimodal approach," presented at the MediaEval 2015 Workshop, Wurzen, Germany, Sep. 14–15, 2015.
[48] P. Marin Vlastelica, S. Hayrapetyan, M. Tapaswi, and R. Stiefelhagen, "KIT at MediaEval 2015-evaluating visual cues for affective impact of movies task," presented at the MediaEval 2015 Workshop, Wurzen, Germany, Sep. 14–15, 2015.

Shangfei Wang (SM'15) received the BS degree in electronic engineering from Anhui University, Hefei, Anhui, China, in 1996. She received the MS degree in circuits and systems, and the PhD degree in signal and information processing from University of Science and Technology of China (USTC), Hefei, Anhui, China, in 1999 and 2002. From 2004 to 2005, she was a Postdoctoral Research Fellow with Kyushu University, Japan. Between 2011 and 2012, she was a Visiting Scholar at Rensselaer Polytechnic Institute in Troy, NY, USA. She is currently a Professor with the School of Computer Science and Technology, USTC. Her research interests include affective computing and probabilistic graphical models. She has authored or coauthored over 90 publications. She is a member of the ACM.

Longfei Hao received the B.S. degree in computer science from Anhui University, Hefei, Anhui, China, in 2016. He is currently working toward the M.S. degree in computer science at the University of Science and Technology of China, Hefei, China. His research interest is in affective computing.

Qiang Ji (F'15) received the Ph.D. degree in electrical engineering from the University of Washington, Seattle, WA, USA. From 2009 to 2010, he was the Program Director of the National Science Foundation (NSF), Arlington, VA, USA, where he managed NSF's computer vision and machine learning programs. He also held teaching and research positions at the Beckman Institute, University of Illinois at Urbana–Champaign, Urbana, IL, USA, the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA, the Department of Computer Science, University of Nevada at Reno, Reno, NV, USA, and the Air Force Research Laboratory, Rome, NY, USA. He is currently a Professor with the Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, where he is also the Director of the Intelligent Systems Laboratory. His research interests include computer vision, probabilistic graphical models, and machine learning and their applications in various fields.

He has authored or coauthored over 230 papers in peer-reviewed journals and conferences. He is a fellow the IAPR. He is a program committee member of numerous international conferences/workshops. He received multiple awards for his work. He has served as the general chair, the program chair, and the technical area chair for numerous international conferences/workshops. He is currently an editor of several his research-related IEEE and international journals.