Exploring Domain Knowledge for Affective Video Content Analyses

Tanfang Chen[†], Yaxin Wang[†], Shangfei Wang[*], Shiyu Chen
School of Computer Science and Technology
University of Science and Technology of China
Hefei, Anhui 230027
tanfang@mail.ustc.edu.cn,yaxin@mail.ustc.edu.cn,sfwang@ustc.edu.cn,sy1001@mail.ustc.edu.cn

ABSTRACT

The well-established film grammar is often used to change visual and audio elements of videos to invoke audiences' emotional experience. Such film grammar, referred to as domain knowledge, is crucial for affective video content analyses, but has not been thoroughly explored yet. In this paper, we propose a novel method to analyze video affective content through exploring domain knowledge. Specifically, take visual elements as an example, we first infer probabilistic dependencies between visual elements and emotions from the summarized film grammar. Then, we transfer the domain knowledge as constraints, and formulate affective video content analyses as a constrained optimization problem. Experiments on the LIRIS-ACCEDE database and the DEAP database demonstrate that the proposed affective content analyses method can successfully leverage well-established film grammar for better emotion classification from video content.

1 INTRODUCTION

Recent years have seen increasingly big amount of video data with the proliferation of mobile devices and the rapid development of online video service. Videos have become the medium for many people to communicate and to find entertainment in addition to sharing knowledge and information. Therefore, these exponentially growing video collections inevitably influence users' emotional states as they spread information and provide entertainment. In the background of this times, affective video content analyses have attracted increasing attentions.

Current study of affective video content analyses can be categorized into two groups: direct approaches and implicit approaches [31]. Direct affective video content analyses assign emotion tags to videos from the visual and audio features of videos, while implicit affective video content analyses infer videos' emotion tags from a user's spontaneous nonverbal response while watching the videos. This paper focuses on direct approaches of affective video content analyses.

Both cinematography and psychological research show that certain audio-visual cues are related to the affective content of a video. Take visual elements as an example, three visual elements, i.e., lighting, color and tempo, are often used to enhance users' emotional experience. High-key lighting is often used to generate the lighthearted and warm atmosphere, while the low-key lighting is used to create sad, frightening, or suspenseful scenes [39]. Color brightness is related to valence, while color saturation may influence arousal [28]. Higher tempo can be used to induce stress and excitement, and lower tempo can create a more relaxed and slow-paced scene [2, 19].

Inspired by cinematography and psychological research, most works of affective content analyses defined special audio and visual features. For example, Hanjalic and Xu [13] proposed motion component, rhythm component, and sound energy component for arousal curve modeling, as well as pitch-average component for valence curve modeling. Canini et al. [6] adopted a big amount of visual and audio features, including dominant color, color layout, scalable color, color structure, color codebook, color energy, lighting key I, lighting key II, saturation, motion dynamics, shot length, illuminant color, shot type transition rate, sound energy, low-energy ratio, zero-crossing rate, spectral rolloff, spectral centroid, spectral flux, Mel Frequency Cepstrum Coefficient (MFCC), sub-band distribution, beat histogram, and rhythmic strength. In addition to adopting hand-craft features to represent video content, several works explored deep learning to learn middle-level video representation. For example, Pang et al. [18] proposed to learn video representation from lower-level visual feature (i.e., DenseSIFT, GIST, HOG, LBP and SSIM), audio features (MFCC and AudioSix) and text features (word count vector) for multimodal affective video content analyses.

After feature extraction, both static and dynamic machine learning methods have been investigated to recognize emotions from video content. Static classifiers or regressors, such as support vector machine (SVM) [33], support vector regression (SVR) [6, 7, 9, 10, 40–42], multi-layer feed-forward neural networks (NNs) [32], Gaussian mixture models (GMMs) [37], and K-nearest neighbor (kNN) [36], capture the mapping between extracted features and emotion tags, ignoring the

[*]Dr. Shangfei Wang is the corresponding author.
[†]Tanfang Chen and Yaxin Wang are co-first authors.

MM'17, October 23–27, 2017, Mountain View, CA, USA.
© 2017 ACM. ISBN 978-1-4503-4906-2/17/10. . . $15.00
DOI: http://dx.doi.org/10.1145/3123266.3123352

dynamic aspects of affective video content. While dynamic classifiers, such as hidden Markov models (HMMs) [26, 34], dynamic Bayesian networks (DBNs) [3], and conditional random fields (CRFs) [35] can model the temporal dynamics. A comprehensive survey of affective video content analyses can be found in [31].

All these researches demonstrate the progress in affective video content analyses. However, most current works employ discriminative features and efficient classifiers for affective video content analyses, without explicitly exploring and leveraging domain knowledge for affective video content analyses. Therefore, in this paper, we propose a novel method to analyze affective video content through exploring domain knowledge. Both audio elements and visual elements are used by film makers to communicate emotions to audience. As a primary study to explore film grammar for affective video content analyses, this paper takes visual elements as an example to demonstrate the feasibility of the proposed affective video content analyses method enhanced through exploring domain knowledge. Specifically, we first infer probabilistic dependencies between visual elements and emotions from the summarized film grammar. Then we transfer the domain knowledge as constraints and formulate affective video content analyses as a constrained optimization problem. Experiments on two benchmark databases demonstrate the superiority of the proposed method.

2 DOMAIN KNOWLEDGE

Both audio elements and visual elements are used by film makers to communicate emotions to audiences. In this section, we introduce the dependencies between visual elements and emotions from the summarized film grammar. Specifically, three visual elements, i.e., lighting, color and tempo are discussed. We investigate how these visual elements affect audiences' emotion from the perspective of both film makers and audiences. Details are discussed in the following sections.

2.1 Lighting

In the film makers' perspective, lighting has great power to establish the mood of a scene and can greatly affect the emotions of the audiences [30]. Generally, two aesthetic lighting techniques called *high-key lighting* and *low-key lighting* are frequently used. *High-key lighting* is a flat lighting deemphasizing the light/dark contrast whereas *low-key lighting* is characterized by the contrast between light and shadow areas [5, 39]. As mentioned in [39], *high-key lighting* is often designed to create the lighthearted and warm atmosphere, which invokes high valence and low arousal mood from the audiences. On the contrary, the *low-key lighting* is often adopted to create sad, frightening, or suspenseful scenes, and this invokes low valence and high arousal mood from the audiences.

From the perspective of the audiences, the perceived lighting can fully affect their feelings. Darkness heightens mystery and intrigue [14]. In general, darkness means less information and uncertainty. While losing in the fully darkness, people will intuitively feel frightened and great potential for the danger [14]. In this vulnerable position, the fear and anxiety increase and the audiences feel high arousal and low valence. While staying in high lighting, people know the around exactly and feel relaxed. In this bright position, people feel high valence and low arousal.

The differences between high-key lighting and low-key lighting are mainly determined by two factors: 1) the general level of light and 2) the proportion of shadow area [30]. In this paper, to accurately quantify the general level of light and the proportion of shadow area, the lighting key is formulated as below:

$$lighting\ key = Med_l * Pro_s \tag{1}$$

where Med_l is the median lighting, which represents the general level of light, and Pro_s is the proportion of pixels whose lightness fall below a certain shadow threshold and represents the proportion of shadow area. This threshold is determined to be 0.18 according to [30].

After extracting lighting key features, we binarize the key lighting of the video. The median lighting key is used as the threshold. We determine the video frames as high-key lighting if its lighting key is higher than the median lighting key, while low-key lighting is assigned to the video frames if its lighting key is lower than the median lighting key.

In all, from the film makers' design and the audience's psychological response, high-key lighting videos have more chances to invoke high valence and low arousal mood from audiences, while the low-key lighting videos are more likely to invoke the low valence and high arousal mood from the audiences.

2.2 Color

In the cinematographic perspective, color is the most important visual element for film presentation. Generally, colors are categorized into two groups: warmer colors and cooler colors. The cooler colors, which contain green, cyan, blue, and magenta, are less bold and provocative [14]. By creating the scene with cooler colors, the film makers intend to present a scene of calm and turning inward. On the contrary, the warmer colors, which include red, orange, and yellow are often used to present a scene of energy, life, and outward tendencies [14]. Thus, warmer colors are mainly adopted for invoking high valence and high arousal from the audiences, whereas cooler colors are used to invoke low valence and low arousal. From the perspective of the audiences, studies on colors also show that valence is strongly correlated to brightness while arousal is strongly correlated to saturation [28].

In this paper, we introduce *color energy* [30] for measuring the joint valence-arousal quality of a scene arising from the color composition. Color energy is defined as the product of the raw energy and color contrast:

$$color\ energy = \sum_i \sum_j p(c_i) * p(c_j) * d(c_i, c_j) \sum_k^M E(h_k) s_k v_k \tag{2}$$

Table 1: The dependencies between three visual elements, i.e., lighting key, color energy and ASD and emotions. Note that the ✓demonstrates great dependencies between emotion and the visual elements. For example, high-key lighting and high valence are marked ✓since the occurrence probability of high valence can be improved by the high-key lighting. Low color energy and high valence are not marked ✓since such dependency do not exist. Details are discussed in *Sec. 2* .

	high valence	low valence	high arousal	low arousal
high-key lighting	✓			✓
low-key lighting		✓	✓	
high color energy	✓		✓	
low color energy		✓		✓
long ASD	✓			✓
short ASD		✓	✓	

where c is a histogram bin indexed by i, j to iterate every single bin index in the HLS histogram of an image, $p(\cdot)$ is the histogram probability, $d(c_i, c_j)$ is the L2-norm in HLS space, s_k and v_k are the saturation and lightness values of k^{th} pixel respectively, and $E(h_k)$ is the energy of the hue of k^{th} pixel, assigned a range be-tween $[0.75 - 1.25]$, depending on its angular distance to blue and red respectively. M is the total number of pixels.

After obtaining the color energy, we adopt the median color energy as the threshold and classify the video clips into high color energy and low color energy. Specifically, video clips whose color energy are above the median are assigned as *high color energy* while video clips whose color energy are below the median are assigned as *low color energy.*

In all, considering the goal of the film makers and the audiences' psychological response to the colors, the audiences are more likely to feel high valence and high arousal after watching high color energy videos. The audiences tend to feel low valence and low arousal after watching low color energy videos.

2.3 Tempo

In the film makers' perspective, the editing effects (e.g. cuts) are frequently used to affect the audience's perceived passage of time. The cuts define the shot length [39]. As each shot conveys an event, the film makers can heighten arousal and intensify a scene by increasing the event density via rapid shot changes [39]. Generally, shorter shots generate greater excitement and longer shots bring relaxation [24]. To the audience, they feel dynamic and breathtaking excitement while watching rapid shot changes [2, 20]. Thus the short shot length videos induce high arousal and low valence from the audiences, while the long shot length videos induce low arousal and high valence from the audiences.

In this paper, we introduce *average shot duration (ASD)* [32] to measure the pace of a sequence in a movie clip. We first compute the distance between adjacent frames according to the distribution of color and light values. Then we determine the shot boundary by comparing difference between consecutive frames. A threshold is calculated for every 100 frames according to [32]. We compute the *ASD* by averaging the shot durations.

After extracting ASD, we adopt the median *ASD* as the threshold and categorize the video clips into two groups: long ASD and short ASD. We assign video clips as long ASD if its ASD are above the median, while the video clips whose ASD are below the median are assigned as short ASD.

From the film makers' design and the audience's psychological response, audiences tend to feel high valence and low arousal while watching long ASD videos. On the contrary, audiences tend to feel low valence and high arousal while watching short ASD videos.

In conclusion, the dependencies between emotions and visual elements discussed above are shown in *Table 1.*

3 METHODOLOGY

Given the domain knowledge, in this section, we introduce the method to exploit them to train classifiers.

3.1 Problem Statement

Let $S = \{(x_i, h_i, y_i)\,|i = 1, ..., N\}$ denote training samples where x_i represents D-dimensional features from observation, $h_i = (h_i^l, h_i^c, h_i^m) \in \{0,1\}^3$ represents the binarized lighting key, color energy and ASD respectively, $y_i \in \{y_i^v, y_i^a | y_i^v, y_i^a \in \{-1, 1\}\}$ represents arousal and valence label respectively, and N is the number of training samples. The goal of the *emotion tagging* task is to learn classifier $f(x, \omega)$ according to Eq. 3:

$$\underset{w}{min} \;\; \alpha \sum_{i=1}^{N} L(f(x_i, w), y_i) + \beta \sum_{i=1}^{N} L(f(x_i, w), \Delta(h_i, y_i)) + R(w) \tag{3}$$

where α and β are the coefficients, and $\Delta(h_i, y_i)$ indicates the dependencies between visual elements h and the emotion label y. The first term represents the loss function over training samples. The second term represents the regularization term reflecting domain knowledge. The last term represents the regularization term of the weights. For the first term, any loss function can be used. For the second term, any domain knowledge, i.e, the relations between any visual or audio elements and emotions, can be exploited to build better emotion classifiers from videos. In this paper, domain knowledge of three visual elements, i.e., lighting key, color energy and ASD are discussed.

3.2 Representation of Domain Knowledge

Lighting-based domain knowledge In the valence space, high-key lighting videos have more chances to invoke high valence mood from audiences, while the low-key lighting videos have more likely to invoke the low valence from the audiences. Thus we infer the probabilistic dependencies between lighting and valence emotion as:

$$\begin{aligned} p(\hat{y}^v = 1|h^l = 1) \geq p(\hat{y}^v = -1|h^l = 1) \\ p(\hat{y}^v = -1|h^l = 0) \geq p(\hat{y}^v = 1|h^l = 0) \end{aligned} \tag{4}$$

where $p(\hat{y}^v = 1|h^l = 1)$ and $p(\hat{y}^v = -1|h^l = 1)$ indicate the probabilities of high valence and low valence given high-key lighting. $p(\hat{y}^v = -1|h^l = 0)$ and $p(\hat{y}^v = 1|h^l = 0)$ indicate the probabilities of low valence and high valence given low-key lighting. In this paper, we adopt Relu function to penalize the samples violating the domain knowledge. The corresponding penalty $\ell_i^{lv}(x_i, h_i, \hat{y}_i)$ from the lighting-based domain knowledge according to Eq. 4 is encoded as below:

$$\begin{aligned} \ell_i^{lv}(x_i, h_i, \hat{y}_i) &= h_i^l * [p(\hat{y}_i^v = -1|h_i^l = 1) - p(\hat{y}_i^v = 1|h_i^l = 1)]_+ \\ &+ (1 - h_i^l) * [p(\hat{y}_i^v = 1|h_i^l = 0) - p(\hat{y}_i^v = -1|h_i^l = 0)]_+ \\ &= h_i^l * [1 - 2 * p(\hat{y}_i^v = 1|h_i^l = 1)]_+ \\ &+ (1 - h_i^l)[2 * p(\hat{y}_i^v = 1|h_i^l = 0) - 1]_+ \end{aligned} \tag{5}$$

where $[\cdot]_+ = max(\cdot, 0)$.

In the arousal space, high-key lighting videos have more chances to invoke low arousal mood from audiences, while the low-key lighting videos are more likely to invoke high arousal mood from the audiences. We infer the probabilistic dependencies between lighting and arousal emotion as:

$$\begin{aligned} p(\hat{y}^a = 1|h^l = 1) \leq p(\hat{y}^a = -1|h^l = 1) \\ p(\hat{y}^a = -1|h^l = 0) \leq p(\hat{y}^a = 1|h^l = 0) \end{aligned} \tag{6}$$

Thus the corresponding constraint $\ell_i^{la}(x_i, h_i, \hat{y}_i)$ for arousal according to Eq. 6 is encoded as below:

$$\begin{aligned} \ell_i^{la}(x_i, h_i, \hat{y}_i) &= h_i^l * [p(\hat{y}_i^a = 1|h_i^l = 1) - p(\hat{y}_i^a = -1|h_i^l = 1)]_+ \\ &+ (1 - h_i^l) * [p(\hat{y}_i^a = -1|h_i^l = 0) - p(\hat{y}_i^a = 1|h_i^l = 0)]_+ \\ &= h_i^l * [2 * p(\hat{y}_i^a = 1|h_i^l = 1) - 1]_+ \\ &+ (1 - h_i^l)[1 - 2 * p(\hat{y}_i^a = 1|h_i^l = 0)]_+ \end{aligned} \tag{7}$$

Color-based domain knowledge In the valence space, the audiences are more likely to feel high valence after watching high color energy videos. The audiences tend to feel low valence after watching low color energy videos. Thus we infer the probabilistic dependencies between color and valence emotion as:

$$\begin{aligned} p(\hat{y}^v = 1|h^c = 1) \geq p(\hat{y}^v = -1|h^c = 1) \\ p(\hat{y}^v = -1|h^c = 0) \geq p(\hat{y}^v = 1|h^c = 0) \end{aligned} \tag{8}$$

Thus the corresponding penalty $\ell_i^{cv}(x_i, h_i, \hat{y}_i)$ for valence according to Eq. 8 is encoded as below:

$$\begin{aligned} \ell_i^{cv}(x_i, h_i, \hat{y}_i) &= h_i^c * [p(\hat{y}_i^v = -1|h_i^c = 1) - p(\hat{y}_i^v = 1|h_i^c = 1)]_+ \\ &+ (1 - h_i^c) * [p(\hat{y}_i^v = 1|h_i^c = 0) - p(\hat{y}_i^v = -1|h_i^c = 0)]_+ \\ &= h_i^c * [1 - 2 * p(\hat{y}_i^v = 1|h_i^c = 1)]_+ \\ &+ (1 - h_i^c)[2 * p(\hat{y}_i^v = 1|h_i^c = 0) - 1]_+ \end{aligned} \tag{9}$$

In the arousal space, the audiences are more likely to feel high arousal after watching high color energy videos. The audiences tend to feel low arousal after watching low color energy videos. Thus we infer the probabilistic dependencies between color and arousal emotions as:

$$\begin{aligned} p(\hat{y}^a = 1|h^c = 1) \geq p(\hat{y}^a = -1|h^c = 1) \\ p(\hat{y}^a = -1|h^c = 0) \geq p(\hat{y}^a = 1|h^c = 0) \end{aligned} \tag{10}$$

Thus the corresponding constraint $\ell_i^{ca}(x_i, h_i, \hat{y}_i)$ for arousal according to Eq. 10 is encoded as below:

$$\begin{aligned} \ell_i^{ca}(x_i, h_i, \hat{y}_i) &= h_i^c * [p(\hat{y}_i^a = -1|h_i^c = 1) - p(\hat{y}_i^a = 1|h_i^c = 1)]_+ \\ &+ (1 - h_i^c) * [p(\hat{y}_i^a = 1|h_i^c = 0) - p(\hat{y}_i^a = -1|h_i^c = 0)]_+ \\ &= h_i^c * [1 - 2 * p(\hat{y}_i^a = 1|h_i^c = 1)]_+ \\ &+ (1 - h_i^c)[2 * p(\hat{y}_i^a = 1|h_i^c = 0) - 1]_+ \end{aligned} \tag{11}$$

Tempo-based domain knowledge In the valence space, audiences tend to feel high valence while watching long ASD videos. On the contrary, audiences tend to feel low valence while watching short ASD videos. Thus we infer the probabilistic dependencies between lighting and emotions as:

$$\begin{aligned} p(\hat{y}^v = 1|h^m = 1) \geq p(\hat{y}^v = -1|h^m = 1) \\ p(\hat{y}^v = -1|h^m = 0) \geq p(\hat{y}^v = 1|h^m = 0) \end{aligned} \tag{12}$$

Thus the corresponding constraint $\ell_i^{mv}(x_i, h_i, \hat{y}_i)$ for valence according to Eq. 12 is encoded as below:

$$\begin{aligned} \ell_i^{mv}(x_i, h_i, \hat{y}_i) &= h_i^m * [p(\hat{y}_i^v = -1|h_i^m = 1) - p(\hat{y}_i^v = 1|h_i^m = 1)]_+ \\ &+ (1 - h_i^m) * [p(\hat{y}_i^v = 1|h_i^m = 0) - p(\hat{y}_i^v = -1|h_i^m = 0)]_+ \\ &= h_i^m * [1 - 2 * p(\hat{y}_i^v = 1|h_i^m = 1)]_+ \\ &+ (1 - h_i^m)[2 * p(\hat{y}_i^v = 1|h_i^m = 0) - 1]_+ \end{aligned} \tag{13}$$

In the arousal space, audiences tend to feel low arousal while watching long ASD videos. On the contrary, audiences tend to feel high arousal while watching short ASD videos. We infer the probabilistic dependencies between lighting and emotions as:

$$\begin{aligned} p(\hat{y}^a = 1|h^m = 1) \leq p(\hat{y}^a = -1|h^m = 1) \\ p(\hat{y}^a = -1|h^m = 0) \leq p(\hat{y}^a = 1|h^m = 0) \end{aligned} \tag{14}$$

Thus the corresponding constraint $\ell_i^{ma}(x_i, h_i, \hat{y}_i)$ for arousal according to Eq. 14 is encoded as below:

$$\begin{aligned} \ell_i^{ma}(x_i, h_i, \hat{y}_i) &= h_i^m * [p(\hat{y}_i^a = 1|h_i^m = 1) - p(\hat{y}_i^a = -1|h_i^m = 1)]_+ \\ &+ (1 - h_i^m) * [p(\hat{y}_i^a = -1|h_i^m = 0) - p(\hat{y}_i^a = 1|h_i^m = 0)]_+ \\ &= h_i^m * [2 * p(\hat{y}_i^a = 1|h_i^m = 1) - 1]_+ \\ &+ (1 - h_i^m)[1 - 2 * p(\hat{y}_i^a = 1|h_i^m = 0)]_+ \end{aligned} \tag{15}$$

3.3 Proposed Approaches

We now introduce the proposed method to learn video emotion tagging classifier subject to these domain knowledge. In this paper, we adopt hinge loss as our loss function as below:

$$\ell_i(x_i, y_i) = [1 - y^{\{v,a\}} f(x, w)]_+ \tag{16}$$

We propose to learn classifier by exploiting domain knowledge as below:

$$F^{\{v,a\}} = \alpha \sum_{i=1}^{N} \ell_i(x_i, y_i) + \beta^l \sum_{i=1}^{N} \ell_i^{\{lv,la\}}(x_i, h_i^l, \hat{y}_i) + \beta^c \sum_{i=1}^{N} \ell_i^{\{cv,ca\}}(x_i, h_i^c, \hat{y}_i) + \beta^m \sum_{i=1}^{N} \ell_i^{\{mv,ma\}}(x_i, h_i^m, \hat{y}_i) + \frac{1}{2} w^T w \tag{17}$$

where w is the parameter of the classifier, α, β^l, β^c, and β^m are coefficients. The first term is the hinge loss of training samples. The second term is the penalty causing by violating the lighting-based domain knowledge as Eq. 5 and Eq. 7. The third term is the penalty causing by violating the color-based domain knowledge as Eq. 9 and Eq. 11. The fourth term is the penalty causing by violating the tempo-based domain knowledge as Eq. 13 and Eq. 15. The last term is the regularization on parameters of classifier.

In this paper, we use $f(x, w) = w \cdot \phi(x)$ as our score function where $\phi(x)$ maps the feature space into the kernel space. By applying sigmoid function, the probabilistic dependencies between visual elements and emotion labels are represented as:

$$\begin{aligned} p(\hat{y} = 1|h) &= \sigma(f(x, w)) \\ p(\hat{y} = -1|h) &= 1 - \sigma(f(x, w)) \end{aligned} \tag{18}$$

where $\sigma(x) = \frac{1}{1+e^{-x}}$.

Now our goal is to minimize F and obtain the weight w. We adopt the stochastic gradient descent (SGD) to solve the problem.

$$w^{(t+1)} = w^{(t)} - \eta^{(t)} \frac{\partial F^{\{v,a\}}}{\partial w} \tag{19}$$

where t and η indicate the number of iterations and the learning rate respectively.

The gradient of loss function to the weight can be computed as below:

$$\frac{\partial F^{\{v,a\}}}{\partial w} = \alpha \sum_{i=1}^{N} \frac{\partial \ell_i(x_i, y_i)}{\partial w} + \beta^l \sum_{i=1}^{N} \frac{\partial \ell_i^{\{lv,la\}}(x_i, h_i^l, \hat{y}_i)}{\partial w} + \beta^c \sum_{i=1}^{N} \frac{\partial \ell_i^{\{cv,ca\}}(x_i, h_i^c, \hat{y}_i)}{\partial w} + \beta^m \sum_{i=1}^{N} \frac{\partial \ell_i^{\{mv,ma\}}(x_i, h_i^m, \hat{y}_i)}{\partial w} + w \tag{20}$$

where the specific gradient of loss function to the weight can be computed as below:

$$\frac{\partial \ell_i(x_i, y_i)}{\partial w} = \begin{cases} -y_i \phi(x_i), & if \quad 1 - y_i f(x_i, w) \geq 0 \\ 0, & otherwise \end{cases} \tag{21}$$

$$\frac{\partial \ell_i^{lv}(x_i, h_i, \hat{y}_i)}{\partial w} = \begin{cases} -2\sigma(f(x_i, w))[1 - \sigma(f(x_i, w))]\phi(x_i), \\ if \quad h_i^l = 1 \; and \; 1 - 2\sigma(f(x_i, w)) \geq 0 \\ 2\sigma(f(x_i, w))[1 - \sigma(f(x_i, w))]\phi(x_i), \\ if \quad h_i^l = 0 \; and \; 2\sigma(f(x_i, w)) - 1 \geq 0 \\ 0 \quad otherwise \end{cases} \tag{22}$$

Gradients of ℓ_i^{la}, ℓ_i^{cv}, ℓ_i^{ca}, ℓ_i^{mv}, and ℓ_i^{ma} can be computed as Eq. 22 similarly.

The detailed learning algorithm is shown in Algorithm 1.

Algorithm 1 Training algorithm for the proposed method

```
Input:
    training samples $(x_i, h_i, y_i)$,
    coefficient $\alpha$, $\beta^l$, $\beta^c$, and $\beta^m$, learning rate $\eta$
Output:
    optimized parameter $w$
Randomly initialize $w$;
repeat
    for each training sample $(x_i, h_i, y_i)$ do
        Calculate the probabilistic dependencies $p(\hat{y} = 1|h)$ and $p(\hat{y} = -1|h)$ as Eq. 18;
        Calculate the specific gradient as Eq. 21 and 22;
    end for
    Calculate $\frac{\partial F^{\{v,a\}}}{\partial w}$ as Eq. 20;
    $w \leftarrow w - \eta(\frac{\partial F^{\{v,a\}}}{\partial w})$;
until Converges
Return $w$.
```

After learning, the proposed approach can infer the affective label for testing samples according to Eq. 23.

$$\hat{y}^{\{a,v\}} = sign(f(x, w)) = \begin{cases} 1, & if \quad f(x, w) >= 0 \\ -1, & if \quad f(x, w) < 0 \end{cases} \tag{23}$$

where $f(x, w)$ is our score function.

4 EXPERIMENTS

4.1 Experimental Condition

To demonstrate the effectiveness of the proposed method, two benchmark video clip databases are used, i.e., the LIRIS-ACCEDE database [4] and the Database for Emotion Analysis using Physiological signals (DEAP) [25].

The LIRIS-ACCEDE database is now the largest video database for video content analysis, which consists of 9800 video excerpts, extracted from 160 feature films and short films. Affective annotation rankings along the induced arousal and valence axis initially ranging from 0 to 9,799 are achieved using crowdsourcing through a pairwise video comparison protocol. Based on these valence and arousal ranks, MediaEval 2015 [23] proposed classification tasks on the LIRIS-ACCEDE database in which the ranks are re-scaled uniformly to a more common $[-1, 1]$ range. Then valence or arousal scores are assigned with -1, 0, 1 corresponding to three ranges [-1, -0.15), [-0.15,0.15] and (0.15, 1].

Table 2: Affective video content analyses results on the LIRIS-ACCEDE database and the DEAP database.

Method	LIRIS-ACCEDE		DEAP	
	valence	arousal	valence	arousal
none	36.03/.2992	43.51/.3560	63.16/.6303	68.42/.6607
lighting	39.29/.3792	47.14/.2983	76.32/.7617	71.05/.6841
color	38.59/.3824	48.56/.3769	68.42/.6842	73.68/.7173
tempo	38.47/.3811	47.35/.3716	73.68/.7339	71.05/.6841
lighting+color	42.06/.4093	57.64/.3772	78.95/.7841	76.32/.7415
lighting+tempo	40.12/.3968	56.44/.3771	76.32/.7630	78.95/.7564
color+tempo	40.98/.4015	51.31/**.3888**	81.58/.8146	81.58/.7914
lighting+color+tempo	**43.18/.4209**	**60.88**/.3702	**84.21/.8417**	**84.21/.8173**

”-/-” refers to accuracy and F1 score respectively.

The DEAP database is collected from 32 participants while watching 40 stimulating music video clips. The self-assessment evaluation of users’ induced emotions after watching are reported in 9-point rating scales for valence and arousal. Since two videos (experiment ID: 17 and 18) cannot be downloaded due to the copyright issues, 38 videos are involved in our experiments. The label category of a video is determined by the average rating of the viewer. We first average all the evaluations of a video as the average rating of the video $Score_v$. Then we average $Score_v$ of all videos and choose this mean value as the threshold. If the average rating of the video $Score_v$ is larger than the threshold, the positive valence or high arousal is assigned to the video. Otherwise, the video is regarded as negative valence or low arousal.

For the LIRIS-ACCEDE database, we use the features provided by [4] for experiments, including alpha, audio asymmetry, audio asymmetry envelop, audio frequency centroid, colorfulness, color contrast, compositional balance, length of scene cuts. depth of field, audio energy, entropy complexity, audio flatness, audio flatness envelop, global activity, hue count, lightning, number of max salient pixels, median lightness, number of fades per frame, number of scene cuts per frame, normalized number of white frames, disparity of salient pixels, spatial edge distribution area, standard deviation of local maxima, spectral roll-off, standard deviation of the wavelet coefficients of the audio signal, and Zero Crossing Rate (ZCR). We also extract color energy [30], lighting key [39] and average shot duration (ASD) [32].

For the DEAP database, we extract visual features, including color energy, lighting key and ASD. We also extract audio features, including Mel-frequency Cepstrum Coefficients (M-FCC) and spectrum flux[17] for the DEAP database.

To investigate the effect of domain knowledge, we compare the following eight methods: affective video content analyses from features only ignoring constraints from domain knowledge (**none**), i.e. the objective function only consists of the first and last terms of Eq. 17, affective video content analyses through exploring lighting-based domain knowledge (**lighting**), affective video content analyses through exploring color-based domain knowledge (**color**), affective video content analyses through exploring tempo-based domain knowledge (**tempo**), affective video content analyses enhanced by both lighting-based and color-based domain knowledge (**lighting+color**), affective video content analyses enhanced by both lighting-based and tempo-based domain knowledge (**lighting+tempo**), affective video content analyses enhanced by both color-based and tempo-based domain knowledge (**color+tempo**), and the proposed method with all three domain knowledge (**lighting+color+tempo**).

For experiments, 10-fold cross-validation protocol is adopted on the LIRIS-ACCEDE database and leave-one-video-out cross-validation is adopted on the DEAP database. During model training, we first initialize the weights to small random number, then we conduct model selection with grid search, by choosing the hyper parameter α, β_l, β_c, and β_m ranging from $\{1, 10, 20, 50\}$ for simplicity. For each method, we monitor the objective cost on the validation set and the hyper parameters with the smallest objective cost are chosen.

For evaluation, we use the accuracy and averaged F1 score.

4.2 Experimental Results of Affective Video Content Analyses

Affective video content analyses results on the LIRIS-ACCEDE database and the DEAP database are shown in *Table 2*. From the table, we find follows:

First, the proposed affective video content analyses enhanced by three visual domain knowledge performs best among the eight methods with the highest accuracy and F1 scores in most cases. Specifically, compared with affective video content analyses from features only ignoring domain knowledge, the proposed method increases accuracy of 7.15% and 17.37% and F1 score of 0.1217 and 0.0142 for valence and arousal respectively on the LIRIS-ACCEDE database. On the DEAP database, the proposed method achieves 21.05% and 15.79% improvements of accuracy and 0.2114 and 0.1566 improvements of F1 score for valence and arousal respectively. Video content analyses from features only leverages the extracted features to describe important visual and audio elements in videos, and maps features to emotion labels through classifier learning. Its learning process is totally data-driven, ignoring well-established film grammar. While the

Table 3: Comparison with MediaEval 2015 related works on the LIRIS-ACCEDE database

method	valence	arousal
MIC-TJU[38]	41.95	55.93
NII-UIT[15]	42.96	55.91
ICL-TUM-PASSAU[27]	41.48	55.72
Fudan-Huawei[11]	41.80	48.80
TCS-ILAB[8]	35.66	48.95
UMons[21]	37.28	52.44
RFA[16]	33.03	45.04
KIT[29]	38.50	51.90
Ours	**43.18**	**60.88**

proposed method successfully captures domain knowledge as constraints during training. Therefore, the proposed method explores both domain knowledge and training data to obtain better emotion classifiers from video content, and thus achieves better performance.

Second, the methods leveraging more domain knowledge have better performance than that leveraging less domain knowledge. Specifically, the methods exploring two domain knowledge outperform the methods exploring one domain knowledge. Lighting, color and tempo describe the video from different aspects. Their effects on affective video content are complementary. Thus, the methods leveraging more domain knowledge can capture more dependencies between visual elements and emotion, and result in better performance.

Third, compared the three methods which employ one domain knowledge, their performances are comparable. It may indicate that lighting, color and tempo have the similar importance in affective video content analyses.

4.3 Comparison with Related Work

To demonstrate the effectiveness of the proposed method, we compare the proposed method to the related works. On the LIRIS-ACCEDE database, the MediaEval proposed affective content analyses tasks. Since MediaEval 2016 proposed regression tasks on the LIRIS-ACCEDE database, and in this paper we explore the classification problem of affective content analyses, we do not compare with MediaEval 2016 [12]. Instead we make comparison with MediaEval 2015 [23], which proposed classification tasks on the LIRIS-ACCEDE database. On the DEAP database, we compare the proposed method to Chen et al.'s work [22] and Acar et al.'work [1].

Table 3 shows the comparison results with the works published in MediaEval 2015 in terms of accuracy. Considering that the features we used are simplest and the results we achieved are highest among all the related works, this indicates the effectiveness of the proposed method for affective video content analyses. Unlike the related works, which adopts extracted features to map emotion labels directly, the proposed method utilizes domain knowledge, i.e. dependencies between the visual elements (i.e., lighting key, color energy, and ASD) and emotions. Thus, the proposed method is superior to state of the art.

Table 4: Comparison with related works on the DEAP database

method	accuracy
Chen et al. [22] (valence)	73.68
Chen et al. [22] (arousal)	81.58
Acar et al. [1]	81.08
Ours (valence)	84.21
Ours (arousal)	84.21

Table 4 shows the comparison results with related works on the DEAP database. Since the used features and experimental settings are different, the comparison results are listed for reference only. Chen et al. [22] proposed an implicit hybrid video emotion tagging with the help of user's spontaneous nonverbal response while watching the videos, while the proposed method does not use the users' physiological responses. Acar et al. [1] adopted VA-based classification schemes, while the proposed method is designed for valence and arousal respectively. Compared with the two works, the proposed method has best performance. The good performance of the proposed method further demonstrates its superiority of related works.

Taking the performance on the LIRIS-ACCEDE database and the DEAP database into account, the proposed method has a excellent generalization ability for affective video content analyses. This demonstrates the effectiveness of the proposed method.

5 CONCLUSIONS

In this paper, we propose a novel method to analyze affective video content through exploring domain knowledge. We first investigate the probabilistic dependencies between emotions and visual elements, i.e., lighting, color and tempo. Then we transfer such probabilistic dependencies as the domain knowledge constraints for affective video analyses. The experimental results on the LIRIS-ACCEDE database and the DEAP database demonstrate the importance of the domain knowledge. This further demonstrates the superiority of the proposed method to the state of the art.

Both audio elements and visual elements are used by film makers to communicate emotions to audiences. As a primary study to explore film grammar for affective video content analyses, this paper focuses on visual elements. In the future work, we plan to explore the dependencies among audio elements and emotions for affective video content analyses.

ACKNOWLEDGMENTS

This work has been supported by the National Science Foundation of China (Grant No. 61473270, 61228304, 61175037), and the project from Anhui Science and Technology Agency (1508085SMF223).

REFERENCES

[1] Esra Acar, Frank Hopfgartner, and Sahin Albayrak. 2016. A comprehensive study on mid-level representation and ensemble learning for emotional analysis of video material. *Multimedia Tools and Applications* (2016), 1–29.

[2] Brett Adams, Chitra Dorai, and Svetha Venkatesh. 2002. Toward automatic extraction of expressive elements from motion pictures: tempo. *IEEE Transactions on Multimedia* 4, 4 (2002), 472–481.

[3] Sutjipto Arifin and Peter YK Cheung. 2007. A Novel Probabilistic Approach to Modeling the Pleasure-Arousal-Dominance Content of the Video based on" Working Memory". In *Semantic Computing, 2007. ICSC 2007. International Conference on*. IEEE, 147–154.

[4] Yoann Baveye, Emmanuel Dellandrea, Christel Chamaret, and Liming Chen. 2015. Liris-accede: A video database for affective content analysis. *IEEE Transactions on Affective Computing* 6, 1 (2015), 43–55.

[5] David Bordwell, Kristin Thompson, and Jeremy Ashton. 1997. *Film art: An introduction*. Vol. 7. McGraw-Hill New York.

[6] Luca Canini, Sergio Benini, and Riccardo Leonardi. 2013. Affective recommendation of movies based on selected connotative features. *IEEE Transactions on Circuits and Systems for Video Technology* 23, 4 (2013), 636–647.

[7] Luca Canini, Sergio Benini, Pierangelo Migliorati, and Riccardo Leonardi. 2009. Emotional identity of movies. In *Image Processing (ICIP), 2009 16th IEEE International Conference on*. IEEE, 1821–1824.

[8] Rupayan Chakraborty, Avinash Kumar Maurya, Meghna Pandharipande, Ehtesham Hassan, Hiranmay Ghosh, and Sunil Kumar Kopparapu. 2015. TCS-ILAB-MediaEval 2015: Affective Impact of Movies and Violent Scene Detection. In *MediaEval*.

[9] Yue Cui, Jesse S Jin, Shiliang Zhang, Suhuai Luo, and Qi Tian. 2010. Music video affective understanding using feature importance analysis. In *Proceedings of the ACM International Conference on Image and Video Retrieval*. ACM, 213–219.

[10] Yue Cui, Suhuai Luo, Qi Tian, Shiliang Zhang, Yu Peng, Lei Jiang, and Jesse S Jin. 2013. Mutual information-based emotion recognition. In *The Era of Interactive Media*. Springer, 471–479.

[11] Qi Dai, Rui-Wei Zhao, Zuxuan Wu, Xi Wang, Zichen Gu, Wenhai Wu, and Yu-Gang Jiang. 2015. Fudan-Huawei at MediaEval 2015: Detecting Violent Scenes and Affective Impact in Movies with Deep Learning. In *MediaEval*.

[12] Emmanuel Dellandréa, Liming Chen, Yoann Baveye, Mats Sjöberg, Christel Chamaret, and ECD Lyon. 2016. The mediaeval 2016 emotional impact of movies task. In *Proc. of the MediaEval 2016 Workshop, Hilversum, Netherlands*.

[13] Alan Hanjalic and Li-Qun Xu. 2005. Affective video content representation and modeling. *IEEE transactions on multimedia* 7, 1 (2005), 143–154.

[14] Greg Keast. 2014. *Shot Psychology: The Filmmaker's Guide For Enhancing Emotion And Meaning*. Kahala Press.

[15] Vu Lam, Sang Phan Le, Duy-Dinh Le, Shin'ichi Satoh, and Duc Anh Duong. 2015. NII-UIT at MediaEval 2015 Affective Impact of Movies Task. In *MediaEval*.

[16] Ionut Mironica, Bogdan Ionescu, Mats Sjöberg, Markus Schedl, and Marcin Skowron. 2015. RFA at MediaEval 2015 Affective Impact of Movies Task: A Multimodal Approach. In *MediaEval*.

[17] Sirko Molau, Michael Pitz, Ralf Schluter, and Hermann Ney. 2001. Computing mel-frequency cepstral coefficients on the power spectrum. In *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, Vol. 1. IEEE, 73–76.

[18] Lei Pang, Shiai Zhu, and Chong-Wah Ngo. 2015. Deep multimodal learning for affective analysis and retrieval. *IEEE Transactions on Multimedia* 17, 11 (2015), 2008–2020.

[19] Zeeshan Rasheed, Yaser Sheikh, and Mubarak Shah. 2005. On the use of computable features for film classification. *IEEE Transactions on Circuits and Systems for Video Technology* 15, 1 (2005), 52–64.

[20] Zeeshan Rasheed, Yaser Sheikh, and Mubarak Shah. 2005. On the use of computable features for film classification. *IEEE Transactions on Circuits and Systems for Video Technology* 15, 1 (2005), 52–64.

[21] Omar Seddati, Emre Kulah, Gueorgui Pironkov, Stéphane Dupont, Saïd Mahmoudi, and Thierry Dutoit. 2015. UMons at MediaEval 2015 Affective Impact of Movies Task including Violent Scenes Detection. In *MediaEval*.

[22] Chen Shiyu, Wang Shangfei, Wu Chongliang, Gao Zhen, Shi Xiaoxiao, and Ji Qiang. 2016. Implicit Hybrid Video Emotion Tagging by Integrating Video Content and Users Multiple Physiological Responses. In *International Conference on Pattern Recognition*.

[23] Mats Sjöberg, Yoann Baveye, Hanli Wang, Vu Lam Quang, Bogdan Ionescu, Emmanuel Dellandréa, Markus Schedl, Claire-Hélène Demarty, and Liming Chen. 2015. The MediaEval 2015 Affective Impact of Movies Task.. In *MediaEval*.

[24] Greg M Smith. 2003. *Film structure and the emotion system*. Cambridge University Press.

[25] Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic. 2012. A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing* 3, 1 (2012), 42–55.

[26] Kai Sun and Junqing Yu. 2007. Video affective content representation and recognition using video affective tree and hidden markov models. In *International Conference on Affective Computing and Intelligent Interaction*. Springer, 594–605.

[27] George Trigeorgis, Eduardo Coutinho, Fabien Ringeval, Erik Marchi, Stefanos Zafeiriou, and Björn W Schuller. 2015. The ICL-TUM-PASSAU Approach for the MediaEval 2015" Affective Impact of Movies" Task. In *MediaEval*.

[28] Patricia Valdez and Albert Mehrabian. 1994. Effects of color on emotions. *Journal of experimental psychology: General* 123, 4 (1994), 394.

[29] Marin Vlastelica, Sergey Hayrapetyan, Makarand Tapaswi, and Rainer Stiefelhagen. 2015. KIT at MediaEval 2015-evaluating visual cues for affective impact of movies task. (2015).

[30] Hee Lin Wang and Loong-Fah Cheong. 2006. Affective understanding in film. *IEEE Transactions on circuits and systems for video technology* 16, 6 (2006), 689–704.

[31] Shangfei Wang and Qiang Ji. 2015. Video affective content analysis: a survey of state-of-the-art methods. *IEEE Transactions on Affective Computing* 6, 4 (2015), 410–430.

[32] Saowaluk C Watanapa, Bundit Thipakorn, and Nipon Charoenkitkarn. 2008. A Sieving ANN for Emotion-Based Movie Clip Classification. *IEICE Transactions on Information and Systems* 91, 5 (2008), 1562–1572.

[33] Cheng-Yu Wei, Nevenka Dimitrova, and Shih-Fu Chang. 2004. Color-mood analysis of films based on syntactic and psychological models. In *Multimedia and Expo, 2004. ICME'04. 2004 IEEE International Conference on*, Vol. 2. IEEE, 831–834.

[34] Min Xu, Jesse S Jin, Suhuai Luo, and Lingyu Duan. 2008. Hierarchical movie affective content analysis based on arousal and valence features. In *Proceedings of the 16th ACM international conference on Multimedia*. ACM, 677–680.

[35] Min Xu, Changsheng Xu, Xiangjian He, Jesse S Jin, Suhuai Luo, and Yong Rui. 2013. Hierarchical affective content analysis in arousal and valence dimensions. *Signal Processing* 93, 8 (2013), 2140–2150.

[36] Ashkan Yazdani, Krista Kappeler, and Touradj Ebrahimi. 2011. Affective content analysis of music video clips. In *Proceedings of the 1st international ACM workshop on Music information retrieval with user-centered and multimodal strategies*. ACM, 7–12.

[37] Ashkan Yazdani, Evangelos Skodras, Nikolaos Fakotakis, and Touradj Ebrahimi. 2013. Multimedia content analysis for emotional characterization of music video clips. *EURASIP Journal on Image and Video Processing* 2013, 1 (2013), 26.

[38] Yun Yi, Hanli Wang, Bowen Zhang, and Jian Yu. 2015. MIC-TJU in MediaEval 2015 Affective Impact of Movies Task. In *MediaEval*.

[39] Herbert Zettl. 2013. *Sight, sound, motion: Applied media aesthetics*. Cengage Learning.

[40] Shiliang Zhang, Qingming Huang, Shuqiang Jiang, Wen Gao, and Qi Tian. 2010. Affective visualization and retrieval for music video. *IEEE Transactions on Multimedia* 12, 6 (2010), 510–522.

[41] Shiliang Zhang, Qi Tian, Qingming Huang, Wen Gao, and Shipeng Li. 2009. Utilizing affective analysis for efficient movie browsing. In *Image Processing (ICIP), 2009 16th IEEE International Conference on*. IEEE, 1853–1856.

[42] Shiliang Zhang, Qi Tian, Shuqiang Jiang, Qingming Huang, and Wen Gao. 2008. Affective MTV analysis based on arousal and valence features. In *Multimedia and Expo, 2008 IEEE International Conference on*. IEEE, 1369–1372.