Heterogeneous Knowledge Transfer in Video Emotion Recognition, Attribution and Summarization

Baohan Xu, Yanwei Fu, Yu-Gang Jiang, Boyang Li and Leonid Sigal

Abstract—Emotion is a key element in user-generated video. However, it is difficult to understand emotions conveyed in such videos due to the complex and unstructured nature of user-generated content and the sparsity of video frames expressing emotion. In this paper, for the first time, we propose a technique for transferring knowledge from heterogeneous external sources, including image and textual data, to facilitate three related tasks in understanding video emotion: emotion recognition, emotion attribution and emotion-oriented summarization. Specifically, our framework (1) learns a video encoding from an auxiliary emotional image dataset in order to improve supervised video emotion recognition, and (2) transfers knowledge from an auxiliary textual corpora for zero-shot recognition of emotion classes unseen during training. The proposed technique for knowledge transfer facilitates novel applications of emotion attribution and emotion-oriented summarization. A comprehensive set of experiments on multiple datasets demonstrate the effectiveness of our framework.

Index Terms—Video Emotion Recognition, Transfer Learning, Zero-Shot Learning, Summarization.

1 Introduction

Rapid development of mobile devices has led to an explosive growth of user-generated images and videos, which creates a demand for computational understanding of visual media content. In addition to recognition of objective content, such as objects and scenes, an important dimension of video content analysis is the understanding of emotional or affective content, i.e. estimating the emotional impact of the video on a viewer. Emotional content can strongly resonate with viewers and plays a crucial role in the video-watching experience. Some successes have been achieved with the use of deep-learning architectures trained for text at both sentence- and document-level [40] or image sentiment analysis [8]. However, understanding emotions from video, to a large extent, remains an unsolved problem .

Analysis of emotional content in video has many real-world applications. Video recommendation services can benefit from matching user interests with the emotions and interestingness [20], [21], [36] of video content, leading to improved user satisfaction. Better understanding of video emotions may enable socially appropriate advertising that is consistent with the main video's mood and help avoid inappropriateness such as placing a funny advertisement alongside a funeral video. Video summarization [68] and coding [60] can also benefit from understanding emotions, since an accurate summary should keep the emotional content conveyed by the original video.

- *Baohan Xu and Yu-Gang Jiang are with the School of Computer Science, Shanghai Key Lab of Intelligent Information Processing, Fudan University, Shanghai, China. Email:{bhxu14,ygj}@fudan.edu.cn.*
- *Yanwei Fu (corresponding author) is with the School of Data Science, Fudan University, Shanghai, China. Email:{yanweifu}@fudan.edu.cn*
- *Boyang Li and Leonid Sigal are with Disney Research. Email: {albert.li, lsigal}@disneyresearch.com.*

Unlike professionally produced videos, user-generated video content presents unique challenges for video understanding. Challenges arise from the diversity of the content, lack of structure, and typically poor production and editing quality (e.g., insufficient lighting). Analyzing the video emotional content in such videos is even more difficult, since (1) the complex spatio-temporal interactions between visual elements makes this intrinsically more complex than analysis of static images, and (2) the emotion is often expressed in only certain limited (sparse) keyframes or video clips.

To cope with these difficulties, we propose to employ heterogeneous knowledge extracted from external sources. In particular, we present an auxiliary Image Transfer Encoding (ITE) algorithm, which can leverage emotional information from auxiliary image data to aggregate frame-level features into a video-level emotion-sensitive representation. To demonstrate the power of this knowledge transfer technique, we tackle three inter-related tasks, namely emotion recognition, emotion attribution, and emotion-oriented summarization.

The first task of emotion recognition includes both supervised and zero-shot conditions. Zero-shot video emotion recognition aims to recognize emotion classes that are not seen during training. This task is motivated by recent cognitive theories [3], [4], [6], [45] that suggest human emotional experiences extend beyond the traditional "basic emotion" categories, such as Ekman's six emotions [13]. Rather, many cognitive processes cooperate closely to create rich emotional and affective experiences [23], [44], [54], such as ecstasy, nostalgia, or suspense. When operating in the real world, recognition systems trained with a small set of emotion labels will inevitably encounter emotion types that are not present in its training set. From large image and text corpora, we construct a semantic vector space where we can identify semantic relationships between visual representa-

tion and textual representation of emotions. Subsequently, we can exploit the semantic relationships to recognize emotions unknown to the system. To our best knowledge, our paper is the first to explore zero-shot emotion recognition.

For our second task, we define a novel problem, *video emotion attribution*, which aims to identify each frame's contribution to a video's overall emotion. This task is motivated by the observation that emotional videos, even those conveying strong emotions, typically contain many frames that are emotionally neutral. The neutral frames may serve important functions, such as setting up the context and introducing characters, but do not convey emotions themselves. Being able to detect emotional frames is a key component of video understanding and enables our third task.

Our third task is *emotion-oriented video summarization*. We argue that a good video summary should be succinct but also provide good coverage of the original video's emotion and information content. Hence our approach aims to balance emotion, information content, and length in providing an accurate video summary.

Contributions: We introduce a framework for transferring knowledge from heterogeneous sources, including image and text, for the understanding of emotions in videos. To the best of our knowledge, this is the first work on zero-shot emotion recognition achieved by applying knowledge learned from text sources to the video domain. We also propose the first definition and solution for the problems of emotion video attribution and emotion-oriented summarization. A user study demonstrates the advantages of emotion-oriented summaries created by the proposed technique, when compared to alternative methods that do not consider emotion. Our final contribution is the introduction of two new emotion-centric video datasets, VideoStory-P14 and YF-E6, to the research community[1].

2 Related Work

2.1 Psychological Theories of Emotion

It is a widely held view in psychology that emotions can be categorized into a number of static categories, each of which is associated with stereotypical facial expression(s), physiological measurements, behaviors, and external causes [12]. The most well-known model is probably Ekman's six pan-cultural basic emotions, including happiness, sadness, disgust, anger, fear, and surprise [13], [14]. However, the exact categories can vary from one model to another. Plutchik [62] added anticipation and trust to the list. Ortony, Clore and Collins's [58] model of emotion defined up to 22 emotions, including categories like hope, shame, and gratitude.

Nevertheless, more recent empirical findings and theories [3], [45] suggest emotional experiences are much more varied than previously assumed. It has been argued that the classical categories are only modal or stereotypical emotions, and large fuzzy areas exist in-between on the emotional landscape. Other theories [23], [44], [54] highlight the dynamics of emotion and the interactions between emotional processes and other cognitive processes. Together, the complex dynamics and interactions produce a rich set of emotional and affective experiences, and correspondingly rich natural language descriptions of those experiences, such as ecstasy, nostalgia, or suspense.

In order to cope with diverse emotional descriptions that may be practically difficult (or at least very costly) to label, in this paper, we investigate emotion recognition in a zero-shot setting (in addition to a traditional, supervised setting). After training, our recognition system is tested against emotional classes that do not appear in the training set. This zero-shot recognition task puts to test the system's ability to effectively utilize knowledge learned from heterogeneous sources in order to adapt to unseen emotional labels.

This paper is mostly concerned with recognizing emotion aroused from watching a video rather than recognizing facial expressions. Despite inherent subjectivity involved in emotional experiences and individual differences [28], there are likely modal responses that can be gathered from a reasonable and neutral audience. A number of recent works focused on recognizing the emotional impact of images and videos, as we review in the next section.

2.2 Automatic Emotion Analysis

In this section, we briefly review two relevant areas of research: recognition of emotional impact of images on viewers, and recognition of emotional impact from videos.

Recognizing emotional impact of still images on viewers. Machajdik and Hanbury [52] classified images into 8 affective categories: amusement, awe, contentment, excitement, anger, disgust, fear, and sadness. In addition to color, texture, and statistics about faces and skin area present in the image, they also made use of composition features such as the rule of the third and depth of field. Lu *et al.* [48] studied shape features along the dimensions of rounded-angular and simple-complex and their effects in arousing viewers' emotions. You *et al.* [79] designed a deep convolutional neural network (CNN) for visual sentiment analysis. After training on the entire training set, images on which the CNN performs poorly are stochastically removed. The remaining images were used to fine-tune the network. A few works [8], [76] also employed off-the-shelf CNN features.

Recognizing emotional impact of videos. A Large number of early works studied emotion in movies (e.g., [32], [38], [69]). Wang and Cheong [69] used an SVM with diverse audio-visual features to classify 2040 scenes in 36 Hollywood movies into 7 emotions. Jou *et al*. [37] worked on animated GIF files. Irie *et al.* [32] use Latent Dirichlet Allocation to extract audio-visual topics as mid-level features, which are combined with Hidden-Markov-like dynamic model. For a more comprehensive review, we refer reader to a survey [72].

SentiBank [5] contains a set of $1,553$ adjective-noun pairs, such as "beautiful flowers" and "sad eyes", and images exemplifying each pair. One linear SVM detector was trained for each pair. The best-performing $1,200$ detectors provide a mid-level representation for emotion recognition. Chen *et al.* [8] replaced the SVM detectors with deep convolutional neural networks. Jiang *et al.* [34] explored a large set of features and confirmed the effectiveness of mid-level

1. URL for downloading

representations like SentiBank. In this work, we transfer emotion information learned from the subset of the images in SentiBank [5] for the purpose of video emotion analysis.

An indirect approach for recognizing the emotional impact of a video is to recognize emotions exhibited by viewers of that video. This clever trick delegates the complex task of video understanding to human viewers, thereby simplifying the problem. McDuff *et al.* [55] analyzed facial expressions exhibited by viewers of video advertisements recorded with webcams. Histogram of Oriented Gradient (HOG) features were extracted based on 22 key points on the faces. Purchase intent is predicted based on the entire emotion trajectory over time. Kapoor *et al.* [39] used video, skin conductance, and pressure sensors on the chair and the mouse to predict frustration when a user interacted with an intelligent tutoring system. However, the success of this approach depends on the availability of a large number of human participants.

All previous work are limited as they aim to predict emotion or sentiment classes present in the training set. In this work, we utilize knowledge acquired from auxiliary images and text in order to identify emotion classes unseen in the training set. In addition, we also investigate related practical applications, previously unaddressed by the community, like emotion-oriented video attribution and summarization.

2.3 Zero-shot Learning

The tasks of identifying classes without any observed data is called zero-shot learning [43]; its main challenge is generalizing the recognition model to identify novel object categories without having access to any labelled instances. To cope with this challenge, existing approaches [18], [43] utilized semantic attributes describing the properties across object categories in order to transfer semantic knowledge from existing to novel object classes. However, these approaches require semantic attributes to be manually defined and annotated; the availability of annotation limits their scalability.

Recent work [15], [66] explored zero-shot learning with representation of words as points in a multi-dimensional vector space that is constructed from large-scale text corpora. The intuition underlying this lexical representation is the distributional hypothesis [30], which states that a word's meaning is captured by other words that co-occur with it. This representation has been demonstrated to exhibit generalization properties [56] and constructed vector space allows vector arithmetics Our own experiments corroborate the benefits of such a model. For example, when we add the vectors representing "surprise" and "sadness", we obtain a vector whose the nearest neighbor under cosine similarity is "disappointment". Adding vectors for "joy" and "trust" yields a vector whose nearest neighbor is "love".

In this paper, we explore zero-shot learning facilitated by a semantic vector space that affords reasoning over emotion categories with vector operations. The vector space and the operations capture knowledge learned from text corpora. As we build regressors that project image features into the text vector space, we are able to aggregate knowledge learned from both images and text and identify semantic relations between images and text. Subsequently, we can exploit the semantic relations to recognize emotions unknown to the system. This knowledge transfer framework is crucial to zero-shot emotion recognition. To the best of our knowledge, this is the first work on zero-shot emotion recognition.

2.4 Video Emotion Attribution

We define the novel task of *emotion attribution* as attributing the emotion of a video to its constituents such as frames or clips. The video emotion attribution problem is inspired by sentiment attribution in text [40]. Besides the difference in media (text *vs.* video), our attribution problem also considers multiple emotions whereas sentiment attribution considers only a binary classification (i.e., positive *vs.* negative).

2.5 Video Summarization

Video summarization has been studied for more than two decades. A complete review is beyond the scope of this paper and we refer readers to [68]. In broad strokes, we can classify work on video summarization into two major categories: approaches based on key frames [10], [16], [29], [46] and approaches based on video skims [17], [57], [71], [73]. Video summarization has been explored for various types of content, including professional videos like movies or news reports [57], [71], [73], egocentric videos [49], surveillance videos [16], [17], and, to a lesser extent, user-generated videos [26], [73].

A diverse set of features have been proposed, including low-level features such as visual saliency [51] and motion cues [57], [59], mid-level information such as object trajectories [46], tag localization [71] and semantic recognition [73]. Dhall and Roland [11] considered smile/happy facial expressions. User's spontaneous reactions such as eye movement, blink and face expressions are measured by a Interest Meter system in [61] and used for video summarization. However, none of these approaches have considered video summarization based on more general video emotion content. Such video emotion is an important cue for finding the most "interesting" or "important" video highlights. For example, a good summary of a birthday party or a graduation ceremony should capture emotional moments in the event. Not considering the emotion dimension in the video summarization task risks losing these precious moments in the summary.

2.6 Multi-Instance Learning

The knowledge transfer approach adopted in this work is related to multi-instance learning (MIL), which has been extensively studied in the machine learning community. We hereby briefly review related techniques. MIL refers to recognition problems where each label is associated with a bag of instances, such as a bag of video frames, rather than one data instance per label in the traditional setting. The problem investigated in this work is intrinsically a multi-instance learning case as each video consists of many frame instances with possibly different emotions.

There are two main branches of MIL algorithms. The first branch attempts to enable "single-instance" supervised learning algorithms to be directly applicable to multi-instance feature bags. This branch includes most of early works on MIL [1], [63] such as miSVM [2], MIBoosting

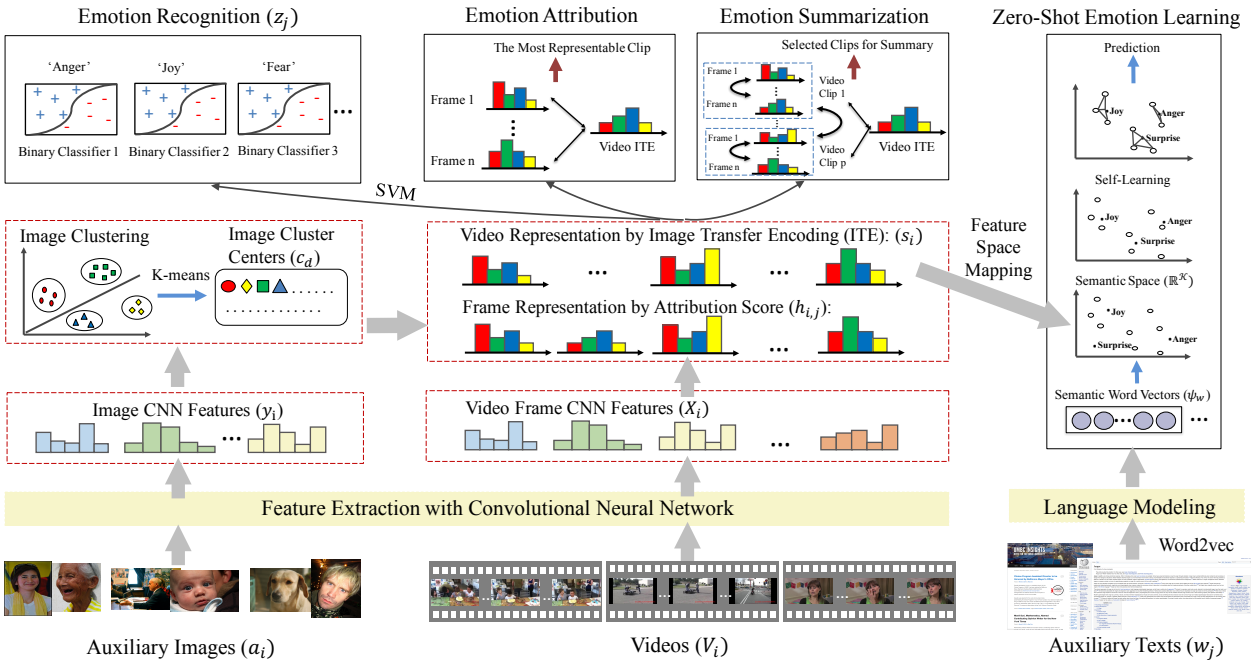

Fig. 1: **An overview of our framework.** Information from the auxiliary images (bottom left) is used to extract an emotion-centric dictionary from CNN-encoded image elements, which is subsequently used to encode video (bottom middle) and recognize emotion (top left). The same encoding is used for emotion attribution and summarization (top middle). Finally, information from a large text corpora is utilized for zero-shot recognition of emotions, as illustrated on the right.

[77], Citation-kNN [70], MI-Kernel [22], among others [24], [25]. These algorithms achieve satisfactory results in several applications [50], [75], but most of them can only handle small or moderate-sized data. In other words, they are computationally expensive and cannot be applied to deal with large-scale datasets.

The second branch of MIL adapt a multi-instance bag to a single data instance in the original instance space. Popular algorithms include constructive clustering based ensemble (CCE) [81], multi-instance learning based on the Fisher Vector representation (Mi-FV) [74] and multi-instance learning via embedded instance selection [9]. Inspired by these works, we encode the video frame bags into single-instance representations. It is worth noting our approach is different from existing MIL algorithms because (1) we perform the encoding process by using auxiliary image data, and demonstrate that transferring such knowledge is important for video emotion analysis; (2) our emotion recognition task is a multi-class multi-instance problem, while most previous MIL algorithms aimed at binary classification.

3 Approach

In this section, we start by presenting the problem formulation and common notations, and then discuss auxiliary image transfer encoding and the three problems we tackle: zero-shot recognition, video emotion attribution, and summarization. Figure 1 shows an overview of our framework.

3.1 Problem Setup

We define our training video dataset with n_{Tr} videos as:

$$Tr = \{(V_i, X_i, \boldsymbol{s}_i, z_i)\}_{i=1,\cdots,n_{Tr}}$$

where V_i denotes the i^{th} video, which is given an emotion label z_i and contains n_i frames $f_{i,1}, \ldots, f_{i,n_i}$. Each frame is described by the feature vector $\mathbf{x}_{i,j}$. As one video contains a set of features $X_i = \{\mathbf{x}_{i,j}\}_{j=1,\cdots,n_i}$ and a single emotion label, this is a typical multi-instance learning problem. Our MIL encoding process converts the bag of features X_i into a video-level feature vector $\boldsymbol{s}_i$ (see Section 3.2).

In addition, we define a test set with n_{Te} videos:

$$Te = \{(V_i, X_i, \boldsymbol{s}_i, z_i^\star)\}_{i=1,\cdots,n_{Te}}$$

where symbols are similarly defined except that $z_i^\star$ is an emotion label in the test set. The notational difference is due to the fact that in the zero-shot learning setting, no test labels exist in the training set. Let Z_{Tr} and Z_{Te} denote emotion labels in the training and test sets respectively, we have $Z_{Tr} \cap Z_{Te} = \varnothing$.

To enable knowledge transfer, we introduce a large-scale emotion-centric auxiliary image set and a text dataset. We denote the auxiliary image sentiment dataset as $A = \{(a_i, \boldsymbol{y}_i)\}_{i=1,\cdots,|A|}$ where $\boldsymbol{y}_i$ is the feature vector of an image a_i. The textual data are represented as a sequence of words $W = (w_0, \ldots, w_{|W|})$, $w_j \in \mathcal{V}$, where the vocabulary $\mathcal{V}$ is the set of unique words. We learn a $\mathcal{K}$-dimensional embedding $\boldsymbol{\psi}_w$ for each word $w \in \mathcal{V}$, as detailed in Section 3.3.

In this paper, we extract image features $\mathbf{x}_{i,j}$ and $\boldsymbol{y}_i$ with a deep Convolutional Neural Network (CNN) architecture, which was recently shown to greatly outperform traditional hand-crafted low-level features on several benchmark datasets, including MNIST and ImageNet [41]. Specifically, we retrain AlexNet [41] with all $2,600$ ImageNet classes and use the activation of the seventh layer ("fc7") as the feature vector for each frame.

3.2 Auxiliary Image Transfer Encoding (ITE)

3.2.1 The Encoding Scheme

We utilize emotion information from a large-scale emotional image dataset to encode each video into a video-level feature vector $\boldsymbol{s}_i$ using a Bag-of-Words (BoW) representation. We learn a dictionary by performing spherical k-means clustering [31] on the auxiliary images, which finds D spherical cluster centers $\boldsymbol{c}_1 \ldots, \boldsymbol{c}_D$. The similarity between a data point $\boldsymbol{x}_{i,j}$ and a cluster center $\boldsymbol{c}_d$ is cosine similarity:

$$\cos(\boldsymbol{x}_{i,j}, \boldsymbol{c}_d) = \frac{\boldsymbol{x}_{i,j}^\top \boldsymbol{c}_d}{\|\boldsymbol{x}_{i,j}\| \|\boldsymbol{c}_d\|}. \tag{1}$$

We use D cluster centers from a dictionary to encode a video into a D-dimensional BoW feature vector. Recall that a video V_i contains n_i frames and corresponding features $X_i = \{\mathbf{x}_{i,j}\}_{j=1,\cdots,n_i}$. For each frame, we identify its K nearest cluster centers. We can compute the assignment variables $\gamma_{i,j,d}$ as if we assign frame $f_{i,j}$ to the d^{th} cluster:

$$\gamma_{i,j,d} = \begin{cases} 1 & if\ \boldsymbol{c}_d \in K\text{-NN}\left(\mathbf{x}_{i,j}\right), \\ 0 & otherwise, \end{cases} \tag{2}$$

where K-NN$\left(\mathbf{x}_{i,j}\right)$ denotes the spherical K nearest neighbours to $\mathbf{x}_{i,j}$ from all cluster centers. The video-level encoding $\boldsymbol{s}_i$ is the accumulation of the frames; the d^{th} dimension of $\boldsymbol{s}_i$ is computed as:

$$s_{i,d} = \sum_{j=1}^{n_i} \gamma_{i,j,d} \cdot \cos\left(\mathbf{x}_{i,j}, \boldsymbol{c}_d\right). \tag{3}$$

3.2.2 Rationale for ITE

We utilize emotional information from a large-scale emotional image dataset to help encode the video content, i.e. ITE. This can be intuitively explained from the perspective of entropy. A dictionary built from the auxiliary emotion-related images can efficiently encode a video frame with emotion information as a sparse vector which concentrates on a few dimensions. In comparison, a frame without emotion information will likely be encoded less efficiently, producing a denser vector with small values in many dimensions. As a result, a non-emotional frame will have higher entropy than the emotional frame, and hence less impact on the resulted BoW representation.

Our encoding scheme ITE also differs from the standard BoW [64] and soft-weighting BoW [35], [80]. The standard BoW encodes local descriptors, such as SIFT and STIP, which requires a dictionary orders of magnitude greater than our frame set. Thus directly using standard BoW [64] on our problem will make the generated video-level features too sparse to be discriminative. Second, soft-weighting encoding, as a sophisticated and refined version of standard BoW, weight the significance of visual words with decaying weights on more cluster center bins. In contrast, due to the diverse nature of emotions (as explained in Section 2.1), one single video frame may equally evoke multiple emotions from viewers. We allow one feature vector $\mathbf{x}_{i,j}$ to equally contribute to multiple encoding bins (Eq. 3).

3.3 Zero-Shot Emotion Recognition

Canonical emotion theories [14] often define a fixed number of prototypical emotions. However, recent research [3], [45] highlights differences within each emotion category and argues that emotions are more diverse than previously imagined. This raises an interesting question: can we identify emotions that are not in our training set purely from their class labels? This is the zero-shot recognition problem.

To address this difficult challenge, we relate emotion class labels we have not seen before to the class labels we have seen. We learn a distributed representation for class label words from an auxiliary corpora of text containing emotional data, utilizing the linguistic intuition that words appear in similar contexts usually have similar meaning [30]. The distributed representations are embedded in a low-dimensional space $\mathbb{R}^{\mathcal{K}}$ in which emotion class labels can be related to each other.

Following Mikilov *et al.* [56], we learn the distributed representation by predicting from each word its context words. Given a word w_t and its surrounding context words $(w_{t-M}, \cdots, w_{t-1}, w_{t+1}, \cdots, w_{t+M})$ within a window of size $2M$, we maximize the log likelihood of context words conditioned on w_t:

$$\max \sum_t \sum_{-M \le j \le M, j \neq 0} \log p(w_{t+j} | w_t) \tag{4}$$

We represent every unique word w in the vocabulary $\mathcal{V}$ as a vector $\boldsymbol{\psi}_w \in \mathbb{R}^{\mathcal{K}}$ and parameterize the above likelihood:

$$p(w_{t+j} | w_t) \propto \exp\left(\boldsymbol{\psi}_{w_{t+j}}^\top \boldsymbol{\psi}_{w_t}\right). \tag{5}$$

Directly optimizing Eq. 4 is intractable because computing the probability in Eq. 5 requires a summation over all words in the vocabulary. As an approximation, we use the negative sampling technique, which samples a few negative examples $w'_1, \cdots, w'_m$ that do not appear in the context window, and maximizes:

$$\sum_{-M \le j \le M, j \neq 0} \log \sigma\left(\boldsymbol{\psi}_{w_{t+j}}^\top \boldsymbol{\psi}_{w_t}\right) - \sum_{1 \le j \le m} \log \sigma\left(\boldsymbol{\psi}_{w'_j}^\top \boldsymbol{\psi}_{w_t}\right) \tag{6}$$

Training of the above model yields embeddings for each word in the vocabulary. We can then train a regressor $g(\cdot)$ from video-level features $\boldsymbol{s}_i$ to the embedding of its class label word (e.g., joy, sadness). In this work, we train a support vector regressor with a linear kernel for each dimension of the word vector.

However, regressors trained on the training set may generalize poorly to test classes that do not exist in the training set. This is mainly because the distribution of visual features in different classes differ. For example, videos of joy usually have positive frames with bright light and smiling faces, while a sad video would typically contain dark colors and people crying. Thus, the relation between video features and the class label's embedding may vary for different classes.

Algorithm 1 Pseudo-code describing the T1S algorithm.

Require:
- C : an auxiliary text dataset .
- Tr : training video set;
- Te : testing video set.
- $Z_{Tr} \cap Z_{Te} = \varnothing$.

1: Train the word2vec language model with the large-scale auxiliary text dataset ← Eq (4)
2: Project emotion words of training and test sets to embeddings $\boldsymbol{\psi}_{w_{Tr}}$ and $\boldsymbol{\psi}_{w_{Te}}$;
3: Train regressors from video features to class label's embeddings
4: Perform zero-shot emotion recognition ← Eq (7) and Eq (8).

To alleviate this generalization problem, we take inspiration from the Rocchio algorithm in information retrieval [18], [53], and use more relevant testing instances to update the query prototypes for better classification accuracy. We thus apply Transductive 1-Step Self-Training (T1S) to adjust the word vector of unseen emotion classes. Let Z_{Tr} and Z_{Te} denote emotion label words in the training and test sets respectively. For a test class $z_i^\star \in Z_{Te}$ that is previously unseen and its distributed representation $\boldsymbol{\psi}_{z_i^\star}$, we relate it to k nearest video neighbors in the test set. We compute a smoothed version $\bar{\boldsymbol{\psi}}_{z_i^\star}$:

$$\bar{\boldsymbol{\psi}}_{z_i^\star} = \frac{1}{K} \sum_{g(\boldsymbol{s}_k) \in K\text{-NN}\left(\boldsymbol{\psi}_{z_i^\star}\right), V_k \in Te} g\left(\boldsymbol{s}_k\right), \tag{7}$$

where K-NN$(\cdot)$ denotes the set of spherical K nearest neighbors in the semantic space. Eq. (7) aims to transductively ameliorate visual differences by averaging $\boldsymbol{\psi}_{z_i^\star}$ with nearest test instances. Here, to prevent semantic drift caused by self-training, we only perform self-training once.

After obtaining $\bar{\boldsymbol{\psi}}_{z_i^\star}$ for all unseen classes $z_i^\star \in Z_{Te}$, we can do nearest-neighbor classification in the vector space. Given a test video V_j and its video level features $\boldsymbol{s}_j$, its class label $z_j^\star$ can be estimated as:

$$\hat{z}_j^\star = \underset{z^\star \in Z_{Te}}{\operatorname{argmax}} \cos\left(g\left(\boldsymbol{s}_j\right), \bar{\boldsymbol{\psi}}_{z^\star}\right). \tag{8}$$

Compared with the zero-shot learning algorithm in [18], we skip the intermediate level of latent attributes and directly apply the 1-step self-training in the semantic word vector space. In addition, we use cosine similarity as the metric rather than the Euclidean distance since the semantic word vectors are intrinsically directional and cosine similarity is a better metric used in [19], [56]. The process is summarized in Algorithm 1.

3.4 Video Emotion Attribution

Emotion attribution aims to identify the contribution of each frame to the video's overall emotion. Emotion attribution can help find *video highlights* [29], which are defined as interesting or important events in the video. Generally, the concepts of "interesting" and "important" may be variable for different video domains and applications, such as the scoring of a goal in soccer videos, applause and cheering in talk-show videos, and exciting speech in presentation videos. Nevertheless, most of these "interesting" or "important" video events convey very strong video emotions, thus providing important signal for highlighting the core parts of the whole video.

Formally, for a video V_i containing a sequence of frames $f_{i,1}, \ldots, f_{i,n_i}$, we want to find the frames that substantially contribute to the overall video emotion. Using the ITE technique described in Section 3.2, we can encode a frame $f_{i,j}$ as its similarity to D cluster centers:

$$\boldsymbol{h}_{i,j} = \left[\cdots, \gamma_{i,j,d} \cdot \cos(\mathbf{x}_{i,j}, \boldsymbol{c}_d), \cdots\right]_{1 \le d \le D} \tag{9}$$

where $\gamma_{i,j,d}$ is defined in Eq (2). The vector $\boldsymbol{h}_{i,j}$ uses the auxiliary image dataset A to evaluate the emotions in the j^{th} frame. Note that in this case $\sum_j \boldsymbol{h}_{i,j} = \boldsymbol{s}_i$.

The video emotion attribution can then be formulated as measuring the similarity between the video-level emotion vector and the frame-level vectors. Specifically, the attribution score of the j^{th} video frame is computed as the cosine similarity between the video-level feature $\boldsymbol{s}_i$, and frame-level feature $\boldsymbol{h}_{i,j}$. We thus find the frame that contributes the most to the overall emotion of V_i by

$$\underset{j \in [1,\ldots,n_i]}{\operatorname{argmax}} \cos\left(\boldsymbol{s}_i, \boldsymbol{h}_{i,j}\right). \tag{10}$$

The emotion attribution procedure can also be extended to a list of pre-partitioned video clips $\{E_1, \ldots E_P\}$ by using clip-level feature $\boldsymbol{h}_p$:

$$\boldsymbol{h}_p = \left[\cdots, \sum_{f_j \in E_p} \nu_{i,j,d} \cdot \cos(\mathbf{x}_{i,j}, \boldsymbol{c}_d), \cdots\right]_{1 \le d \le D} \tag{11}$$

3.5 Emotion-Oriented Video Summarization

One important problem that is enabled by emotion attribution is video summarization. Leveraging our proposed technique, here we present a video summarization method for preserving the emotional content in a video and balancing it against information coverage.

We summarize a video by extracting a number of key frames from it. Let U_i denote this set of key frames for video V_i, we select highest scored frames according to the following:

$$U_i = \underset{f_j \in V_i}{\operatorname{argmax}} \cos\left(\boldsymbol{s}_i, \boldsymbol{h}_{i,j}\right) + \lambda \sum_{f_k \in V_i} \cos\left(\mathbf{x}_{i,j}, \mathbf{x}_{i,k}\right), \tag{12}$$

where λ is a weight parameter, and the second term $\cos\left(\mathbf{x}_{i,j}, \mathbf{x}_{i,k}\right)$ rewards key frames that are the most similar to other frames in the same video, which means that the selected frames are representative of the entire video. Note that (1) the cosine similarity in the first term is computed using ITE, while in the second term the similarity is defined directly in the feature space; (2) We empirically set $\lambda = 1$ to equally consider both emotion content and representativeness of the video.

Comparing with previous work [16], [17], [29], [68], [73], Eq (12) considers the summary of both video highlights (by the first term for emotion attribution) and information coverage (by the second term for eliminating redundancy and selecting information-centric frames/clips). Thus our method can produce a condensed, succinct and emotion-rich summary which can facilitate the browsing, retrieval and storage of the original video content. Particularly, our summary results are more emotionally interpretable due to the emotion attribution.

	P_0	P_e	Kappa
YouTube-24	0.74	0.31	0.62
YF-E6	0.82	0.33	0.73

TABLE 1: Cohen's kappa scores for the annotations of the two newly annotated datasets. P_0 is the relative observed agreement among annotators. P_e is the hypothetical probability of chance agreement. For YouTube-24, the annotators were tasked with classifying each video in one of the 8 emotional categories further into 3 additional sub-categories. This is the reason P_e is relatively high, as chance agreement is only among the 3 sub-categories.

4 EXPERIMENTS

4.1 Datasets and Settings

We utilize three video emotion datasets for evaluation. Among them, the VideoStory-P14 and YF-E6 datasets are introduced by us and are made available to the community.

YouTube emotion datasets [34]. The YouTube dataset contains 1,101 videos annotated with 8 basic emotions from the Plutchik's Wheel of Emotions. To facilitate the zero-shot emotion recognition task, we re-annotate the videos into 24 emotions, by adding 3 variations to each basic emotion according to Plutchik's definition. For example, we split the *joy* class into *ecstasy*, *joy* and *serenity* according to arousal. We use the short-hand **YouTube-8** and **YouTube-24** for the original and re-annotated datasets respectively. To annotate the **YouTube-24** dataset, each video was labeled by 5 annotators using majority vote with a high Cohen's kappa score[2] [65] of 0.62. Table 1 shows the details.

We hereby report statistics for emotions included in the datasets. YouTube-8 dataset contains 101 videos labeled as *anger*, 101 as *anticipation*, 115 *disgust*, 167 *fear*, 180 *joy*, 101 *sadness*, 236 *surprise* and 100 *trust*. YouTube-24 has 36 videos labeled as *anger*, 33 *annoyance*, 32 *rage*, 44 *anticipation*, 32 *interest*, 25 *vigilance*, 42 *boredom*, 64 *disgust*, 9 *loathing*, 12 *apprehension*, 79 *fear*, 76 *terror*, 23 *ecstacy*, 76 *joy*, 81 *serenity*, 27 *grief*, 11 *pensiveness*, 63 *sadness*, 29 *amazement*, 59 *distraction*, 148 *surprise*, 39 *acceptance*, 26 *admiration*, 35 *trust*.

YouTube/Flickr-EkmanSix (YF-E6) dataset. As discussed in the related work section, Ekman [14] found a high agreement of emotions across cultures and proposed 6 basic emotion types. We collect the YF-E6 emotion dataset using the 6 basic emotion type as keywords on social video-sharing websites including YouTube and Flickr, leading to a total of 3000 videos. The dataset is labeled through crowdsourcing by 10 different annotators (5 males and 5 females), whose age ranged from 22 to 45. Annotators were given detailed definition for each emotion before performing the task. Every video is manually labeled by all the annotators. A video is excluded from the final dataset when over half of annotations are inconsistent with the initial search keyword. Due to high agreement of Ekman emotions, we observe very high consistency of the annotations: 85% videos were given the same label by 7 or more annotators with the high kappa score 0.73 (in Table 1). The final dataset comprises $1,637$ videos across the 6 emotion classes, with an average duration of 112 seconds. Specifically, the YF-E6 dataset contains 225 *anger*, 239 *disgust*, 287 *joy*, 221 *sadness*, and 360 *surprise* videos.

The VideoStory-P14 dataset. The VideoStory-P14 dataset is derived from the recently proposed VideoStory dataset [27]. We use all the keywords of the Plutchik's Wheel of Emotions [62] to query the VideoStory dataset in terms of its video captions. Emotion keywords are matched against all the words in the video's caption. This leads to a set of 626 videos belonging to 14 emotion classes. The dataset contains 83 videos labeled as *anger*, 30 as *annoy*, 27 *aggressive*, 119 *rage*, 28 *interest*, 14 *disgust*, 29 *distract*, 16 *fear*, 23 *terror*, 67 *love*, 80 *joy*, 81 *surprise*, 11 *submission*, 18 *trust*.

Auxiliary emotional image and text datasets. From the Flickr image dataset [5], we select as the auxiliary image data a subset of $110K$ images of Adjective-Noun Pairs (ANPs) that have top ranks with respect to the emotions (see Table 2 in [5]). These images are clustered into $2,000$ clusters (i.e. $D = 2000$ in Eq (1)). As shown in [56], the large-scale text data can greatly benefit the trained language model. We train the Skip-gram model (Eq 5) on a large-scale text corpora, which includes around 7 billion words from the UMBC WebBase (3 billion words), the latest Wikipedia articles (3 billion words) and some other documents (1 billion words). The trained model contains roughly 4 million unique words, bi-gram and tri-gram phrases (i.e., $|\mathcal{V}| \approx 4$ million). Most of the documents are formal texts which have clear definitions, descriptions and usage of the emotion and sentiment related words.

Experimental settings. Each video is uniformly sampled at 5 frame increments for feature extraction to reduce the computational cost. The dimension of the real-valued semantic vectors $\boldsymbol{\psi}_w$ (Eq (5)) is set to 500 to balance computational cost of training $\boldsymbol{\psi}_w$ from large-scale text corpora and the effectiveness of the syntactic and semantic regularities of representations [56]. Our AlexNet CNN model is trained by ourselves using $2,600$ ImageNet classes with the Caffe toolkit [33], and we use the $4,096$-dimensional activations of the 7th fully-connected layer after the Rectified Linear Units (*i.e.* $fc7$) as features. The number of nearest neighbors in Eq (2) is empirically set to 10% of the image clusters (i.e. $K = D/10$), which balances the computational cost with a good representation in Eq (3).

4.2 Supervised Emotion Recognition

To illustrate benefits of our ITE encoding scheme, we first perform supervised emotion recognition with a support vector machines (SVM) classifier with chi-square kernel. We compare our method with the following alternative baselines.

MaxP [47]. The instance-level classifiers are trained using the labels inherited from their corresponding bags. These classifiers can be used to predict instance labels of testing videos. The final bag labels are produced by majority vote of instance labels. This method is a variant of the Key Instance Detection (KID) [47] in multi-class multi-instance setting.

2. Cohen's kappa coefficient is a metric for measuring the inter-rater agreement for qualitative items.

AvgP [78]. We average the frame-level image features of one video as video-level feature descriptions for classification. For the i^{th} video, its average pooling feature is computed as $\frac{1}{n_i}\sum_{j=1}^{n_i}\mathbf{x}_{i,j}$. The average pooling is the standard approach of aggregating frame-level features into video-level descriptions as mentioned in [78].

Mi-FV [74]. MIL bags of training videos are mapped into a new bag-level Fisher Vector representation. Mi-FV is able handle large-scale MIL data efficiently.

CCE [81]. The instances of all training bags are clustered into b groups, and each bag is re-represented by b binary features, where the value of the i^{th} feature is 1 if the concerned bag has instances falling into the i^{th} group and 0 otherwise. This is essentially a simplified version of our ITE method encoded by training instances only.

The linear kernel is used for Mi-FV and MaxP due to the large number of samples/dimensions, and the Chi-square kernel[3] is used for others. A binary two-class SVM model is trained for each emotion class separately. The key parameters are selected by 3-fold cross-validation.

The experimental results are shown in Figure 2, which clearly demonstrate that our ITE method significantly outperforms the four alternatives on all three datasets. This validates the effectiveness of our method in generating a better video-level feature representation based on the auxiliary images. In particular, the improvement of ITE over CCE and Mi-FV verifies that the knowledge transferred from the auxiliary emotional image dataset is probably more critical than that existing in the training video frames. This supports our argument that most of the frames of these videos have no direct relation to the emotions expressed by the videos, and underscores the importance of knowledge transfer. Particularly, we use the same training/testing split as [34] on **YouTube-8** dataset. The AvgP result 40.1% is comparable with the result 41.9% in [34] with the method of combining different types of hand-crafted visual features with the state-of-the-art multi-kernel strategy. This validates that the performance of the deep features we use can match that of multi-kernel combination of hand-crafted features.

Our ITE results show 1.2, 3.2, 4.6, 0.8 absolute percentage points improvement over AvgP on VideoStory-P14, YF-E6, Youtube-8 and YouTube-24 video dataset. This further validates the effectiveness of our method. In particular, we found that (1) a portion of videos in VideoStory-P14 dataset are surveillance videos (from VideoStory dataset), which have very poor visual quality and thus make all the methods fail. This explains the slightly lower improvement margin of ITE over AvgP on VideoStory-P14 dataset. (2) Also note that Youtube-24 is a re-annotated version of Youtube-8 and thus it is even harder because in contains larger number of classes and fewer training instances per class. While it is difficult for all the methods to classify emotions, our ITE results are still the best among the competitors.

One should notice that CCE has the worst performance. CCE re-encodes the multi-instances into *binary* representations by ensemble clustering. Such representations may have better performance than the hand-crafted features used in [81], but they cannot beat the recently proposed deep features, which have been shown to be able to extract higher level information [41]. In other words, the re-encoding process of CCE loses discriminative information gained from the deep features, and is therefore unsuited for the task.

3. The RBF kernel is also evaluated but shows slightly lower performance than that of the Chi-square kernel.

Methods	ITE		AvgP	
Features	*fc7*	*fc6*	*fc7*	*fc6*
YouTube-8	43.8	**45.6**	41.1	42.0
YF-E6	**50.9**	49.4	48.4	48.7

TABLE 2: Layer-by-layer analysis for supervised learning. Results obtained from convolutional layers *conv5* and *conv4* are $22.5 \pm 2\%$, which are significant lower than the above.

In addition, Mi-FV and MaxP have similar performance: MaxP is slightly better on VideoStory-P14, YF-E6 and Mi-FV is slightly better on YouTube. However, the results of Mi-FV and MaxP are much worse than those of AvgP. These differences can be explained by the different choices of kernels. We validate that the AvgP with linear SVM classifier has similar performance (with a variance of 2%) as MaxP and Mi-FV. Nevertheless, due to high dimensions of Fisher Vectors and large amount of training instances in MaxP, nonlinear kernels will introduce prohibitive computational cost. Thus, in subsequent experiments, we use AvgP as the main alternative baseline to ITE, since other alternatives do not demonstrate competitive advantages. We illustrate the confusion matrix in Fig. 3. The matrices of ITE shows clear diagonal structure and the results are better than AvgP in most of classes.

Some qualitative results of supervised emotion prediction are shown in Figure 5. In the successful cases, test videos share visual characteristics with auxiliary image dataset, such as *bright light* and *smiling faces* in the "joy" category. The "anger" videos are wrongly classified as "fear". Comparing with "anger", the "fear" category is more highly correlated with *dark lightning* and *screaming faces* which are visually dominant in the failure case. The video wrongly labeled as "joy" has festive colors which resemble a Christmas tree.

4.2.1 Hyper-Parameters and Deep Network Configurations

We conduct further experiments to investigate whether our ITE technique can maintain its advantage over baselines under different hyper-parameter settings, CNN configurations, and with additional audio features. Experimental results show that ITE consistently outperforms the baselines and suggest ITE's advantage is robust under many different conditions. For simplicity, our ablation studies cover the Youtube-8 and YF-E6 datasets in supervised setting.

Layer-by-layer Analysis. The CNN we adopt in this paper, AlexNet, contains 5 convolution layers and 2 fully connected layers. Although it is generally acknowledged that lower layers preserve more local information and higher layers contain more global information, it remains unclear which layers are the most conducive to emotion detection. In this experiment, we employ the outputs from the last and the second last convolutional layers (denoted as *conv5* and *conv4* respectively) as well as the first and second

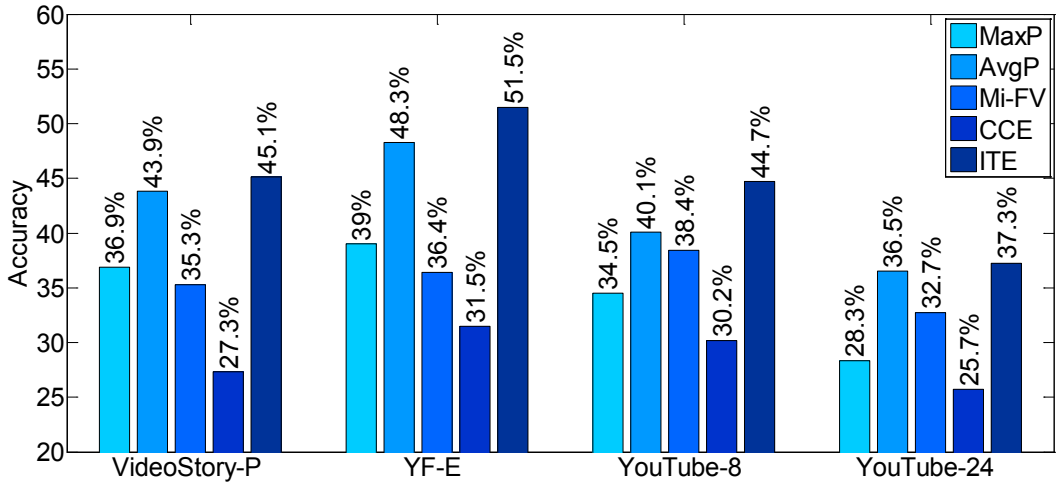

Fig. 2: Average accuracy of supervised emotion recognition.

	denseSIFT	MFCC	ITE($fc7$)	[ITE($fc7$), denseSIFT]	[ITE($fc7$), MFCC]	[ITE($fc7$), denseSIFT, MFCC]
YouTube-8	35.6	44.0	43.8	43.8	**52.6**	46.7
YF-E6	38.6	39.0	50.9	48.8	**51.2**	50.4

TABLE 3: Concatenated results of hand-crafted feature and deep features. ITE is computed from $fc7$.

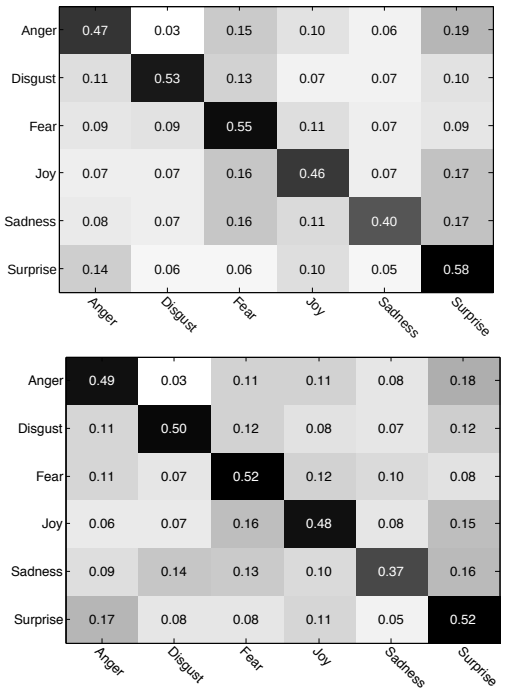

Fig. 3: Confusion matrices for supervised learning on the YF-E6 dataset using our ITE encoding (top) and the AvgP method (bottom).

fully connected layers (*fc6* and *fc7* respectively) as candidate features for video frames. We compare the ITE method and the AvgP method. Table 2 shows the results.

We observe that features from *fc6* perform better for YouTube-8 and features from *fc7* perform only slightly better for YF-E6. In both cases, the ITE technique outperforms the AvgP technique, suggesting our encoding mechanism is quite general and is not tied to a particular layer in the network. Features extracted from the fully connected layers significantly outperform those from convolutional layers (which is $22.5 \pm 2\%$), which suggests that features of higher layers contain more semantic information that is beneficial for video emotion understanding task. For the rest of the paper, we use features from *fc7*.

Complementarity of CNN and hand-crafted features. Table 3 reports results of ITE encoding concatenated with hand-crafted features. We find that (1) results of concatenation with the visual features (denseSIFT) are comparable to those of raw ITE on two dataset. This shows that visual hand-crafted features does not add much to the automatically learned features from the deep neural network. It also demonstrates the ITE outperforms the traditional hand-crafted features. (2) Using audio features (MFCC) alone can achieve very high accuracy for video emotion recognition, indicating that the audio track offers significant utility in video emotion recognition; (3) The hand-crafted audio features (MFCC) are highly complementary to visual features, since they represent the signal from a different modality. The use of multi-modal information in emotion recognition, especially in the zero-shot setting, may merit further investigation. (4) Concatenating all features leads to worse results than that of $[ITE(fc7), \text{MFCC}]$ due to the increased dimensionality that results from adding hand-crafted visual features, which do not contribute to performance.

We test different deep architectures and find that VGG-16/VGG-19/AlexNet $>$ GoogLeNet. While previous experiments showed satisfactory results on emotion analysis task by using AlexNet architecture, we want to compare with different architectures to better understand deep features. VGG-16 and VGG-19 [7] and GoogLeNet-22 [67] achieved the state-of-the-art for image classification on

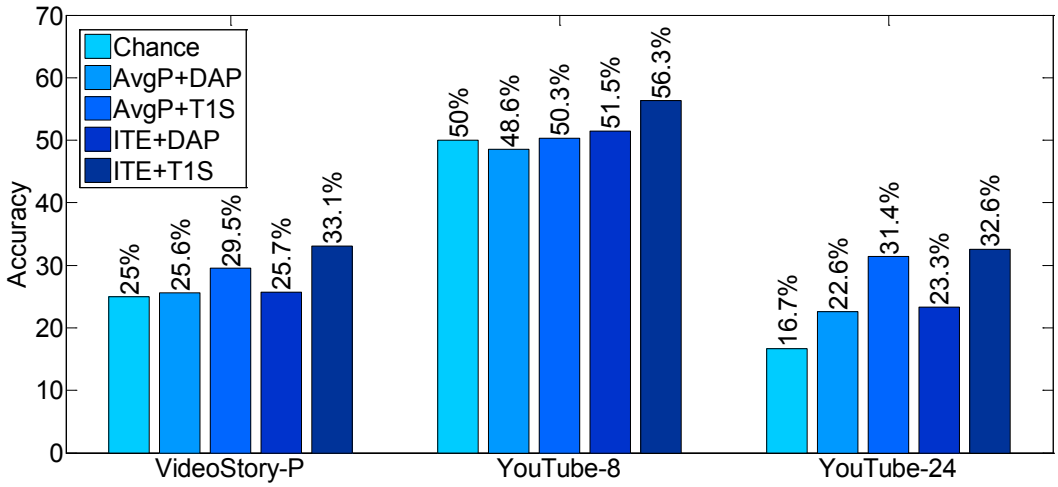

Fig. 4: Average accuracy of zero-shot emotion recognition on YouTube and VideoStory-P14 datasets.

Fig. 5: Some successful and failure examples of supervised emotion recognition on the YouTube-8 dataset. The ground truth categories are given at the top of each column; red labels indicate classification mistakes.

	VGG-16	VGG-19	GoogLeNet-22	AlexNet
YouTube-8	44.7	44.0	35.6	41.1
YF-E6	49.3	48.8	38.3	48.4

TABLE 4: VGG and GoogLeNet results. The AvgP is used here.

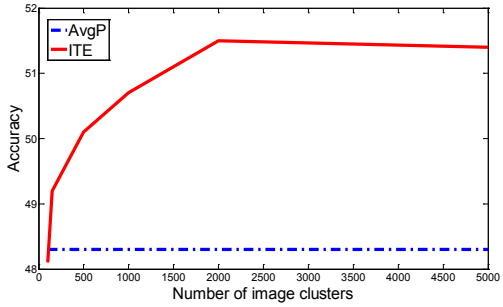

Fig. 6: Influence of varying the number of clusters of auxiliary images for ITE. The X-axis is varying the number of image clusters of auxiliary data; and y-axis is the accuracy of recognition tasks.

ImageNet challenge. Thus we conducted video emotion recognition using high layer features extracted from these architectures as descriptors. Table 4 presented the experimental results. We use $fc7$ of 16 and 19 layers VGG and $inception-5b$ of GoogLeNet. AvgP is used for all the deep architectures. The results of VGG-16 and VGG-19 are comparable to AlexNet, and outperform that of GoogLeNet-22. Although GoogLeNet gets promising results on image classification task, the lower results in Table 4 indicates that it may not extract the most useful features for video emotion recognition.

The number of auxiliary clusters. The number of clusters for the auxiliary images (i.e., D in Eq. 1) is a key parameter in our knowledge transfer framework. With more clusters, the framework is more capable of capturing the rich spectrum of emotional information, but is also prone to overfitting. Here, we empirically test how this parameter affects the ITE performance. The number of clusters is plotted against supervised performance in Figure 6. ITE results gradually improve when the number of clusters is increased from 100 to 2000. After 2000, the performance saturates with a slight drop as D keeps increasing. ITE outperforms AvgP over most settings for D, which indicates that our finding is robust to hyper-parameter settings. AvgP stays constant since varying D does not affect its performance.

4.3 Zero-Shot Emotion Recognition

We conduct zero-shot emotion recognition on YouTube-8, YouTube-24, and the VideoStory-P14 datasets. The YF-E6 dataset contains only 6 emotion classes. Splitting the 6 further into disjoint training classes and test classes will lead to difficulties for properly relating unknown classes to known classes. In the VideoStory-P14 dataset, we use *anger*, *joy*, *surprise*, and *terror* as testing classes, with a total of 300

testing instances. For YouTube-8, we use *fear* and *sadness* as the testing classes. For YouTube-24, we randomly split the 24 classes into 18 training and 6 testing classes with 5-round repeated experiments. In the zero-shot setting, no instances in test classes are seen during training.

We compare our T1S algorithm with Direct Attribution Prediction (DAP) [42], [43]. For DAP, at test time each dimension of the word vectors of each test sample is predicted, from which the test class labels are inferred. DAP can be understood as directly using Eq (8) without the word vector smoothing by Eq (7). Four variants are compared: (a) using different video-level feature representation (AvgP or ITE); (b) using different zero-shot learning algorithm (T1S or DAP).

Figure 4 shows the results. Our ITE+T1S approach produces the best accuracy, outperforming the second best baseline by 3.6, 4.8, and 1.2 absolute percentage points respectively and the random baseline by 8.1, 6.3, and 15.9 absolute percentage points. We observe that AvgP+T1S is the second best technique on VideoStory-P14 and YouTube-24, but ITE+DAP is the second best technique on the YouTube-8 dataset. An important difference between the two scenarios is that YouTube-8 contain less emotions than VideoStory-P14 and YouTube-24, so the semantic distance between individual emotions is greater in YouTube-8. This suggests the T1S technique contributes the biggest performance gain when the training classes bear some similarity to the unseen test classes. However, when the training classes are very different from the testing classes, the ITE encoding scheme plays an important role. It is also worth mentioning that the results of YouTube-24 have a largest margin improvement over baselines than the two other datasets. This result indicates zero-shot learning performs better when a larger variant set of emotions exist in the training set. Overall, the experiments show the combination of ITE+T1S is effective under different zero-shot learning conditions. Given the inherent difficulties of the zero-shot learning task, we consider the results to be very promising.

Qualitative results. In Figure 7, we show some successful examples of zero-shot emotion prediction. We highlight that even without any training examples on these categories, our method can still classify these video successfully using the encoded feature. Thus considering the difficulty of zero-shot emotion prediction, our results are very promising.

Note that Ekman dataset is not used for this tasks due to the small number of emotion classes. Specifically, in our work, each class-level emotion textual name $w \in \mathcal{V}$ is projected into a $\mathcal{K}$-dimensional embedding vector $\boldsymbol{\psi}_w \in \mathbb{R}^{\mathcal{K}}$ in the semantic word vector space; a regressor function $g(\cdot)$ is trained from video-level features to the corresponding embedding vector $\boldsymbol{\psi}_w$. In zero-shot learning scenarios, we need to further split the 6 emotion classes of Ekman dataset into auxiliary and testing dataset. In other words, we only have at most 4 embedding vectors $\boldsymbol{\psi}_w$ to train the regressor $g(\cdot)$ (in the split of 4 auxiliary and 2 testing classes). It is however extremely hard to train a reasonable regressor (without overfitting) with only 4 embedding vectors.

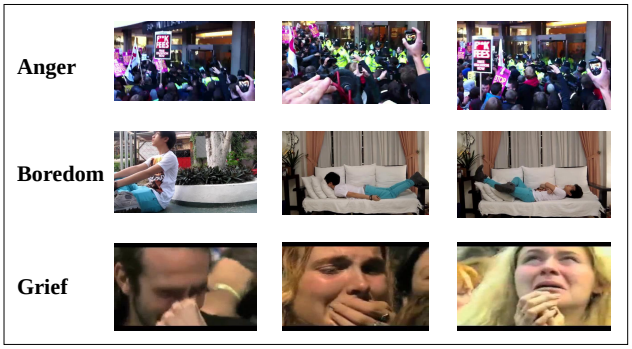

Fig. 7: Qualitative results of zero-shot emotion recognition. We show the keyframes of three successful cases: the frames of top row shows a video clip of an anger parade; the middle row is about a video of a boredom boy walking and lying on the couch; The bottom row is for the grief reaction of fans when their favorite football team lose the game.

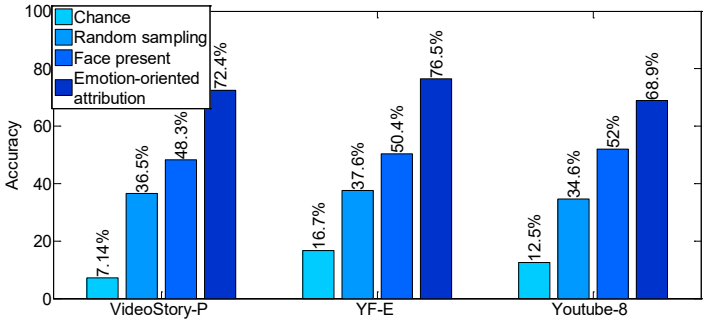

Fig. 8: Quantitative evaluation of video emotion attribution using the YouTube-8 dataset.

4.4 Video Emotion Attribution

As discussed earlier, another advantage of our encoding scheme is that we can identify the video clips that have high impact on the overall video emotion. A pilot study we performed indicated that emotions are sparsely expressed in videos. On average, around 10% of video frames are related to emotion in our three datasets.

As the first work on video emotion attribution, we define the evaluation protocol of user study to evaluate the performance of different algorithms for this task: Ten participants, unaware of project goals, were invited for the user study. Given all emotion keywords of the corresponding dataset and clip computed from the video, participants are asked to guess the name of the emotion expressed in the clip. These clips are generated by different baseline techniques, as discussed later. Since the ground-truth video emotion labels are known, we computed the fraction of participants who assigned the correct emotion label for each clip.

We randomly select 20 videos from each of the three datasets. For each video, we extract a 2-second video clip that contains the highest attribution towards video emotion, using Eq (10).

For comparison purposes, we created the following baselines: **Chance**, which is the probability of correctly guessing the emotion. **Random sampling**, where we first randomly

TABLE 5: Reliability evaluation of video emotion attribution experiments. We random split 10 annotators into 2 groups with 5 person each and compute the each group's score. The accuracy of each method is reported here.

	Group-1		
	Random	Face_present	Emotion
VideoStory-P14	35.1	41.7	**65.9**
YF-E6	34.4	54.1	**71.3**
YouTube	32.5	45.5	**62.8**
	Group-2		
	Random	Face_present	Emotion
VideoStory-P14	37.9	54.9	**78.9**
YF-E6	40.8	46.7	**81.7**
YouTube	36.7	58.5	**75.0**

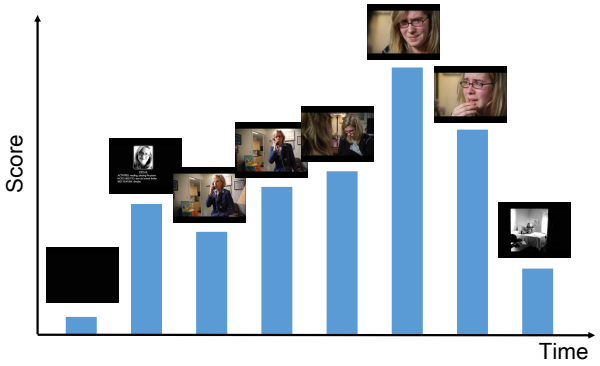

Fig. 9: An example of video emotion attribution.

sample 2 non-overlapping clips of 2 seconds each from the same video. We use both clips in the experiments and compute the average score as the results of this method. **Face presence**, where we use the "face_present" feature [82] to rank all the videos frames; frames with larger and more faces are ranked higher. One clip of 2-second length is generated for each video by using the top ranked frames.

The results are shown in Fig. 8. Our method achieve best accuracy, and outperforms the Face_present baseline by 16-26 absolute percentage points. Although the presence of a human face is often correlated with the expression of emotions, many user-generated videos in our datasets express emotions through other channels like body language or color. Thus, our technique compared more favorably to the face presence baseline. These results indicate that our method can consistently identify video clips that convey emotions recognizably similar to the emotion conveyed by the original video.

To further validate the reliability of the attribution experiments in this user study, we randomly split 10 participants into 2 groups and measure each group's score in Table 5. The measurement shows the consistency of our result; and in each group, our experimental conclusion still hold: the ITE result achieves best accuracy between two groups. .

A qualitative result of emotion attribution is shown in Figure 9, where the video is uniformly sampled every 10 frames. The bar chart shows scores of different frames, where the key frames are shown above the bars. The figure demonstrates that clips with stronger emotional contents are given higher scores of attribution, validating the effectiveness of our method.

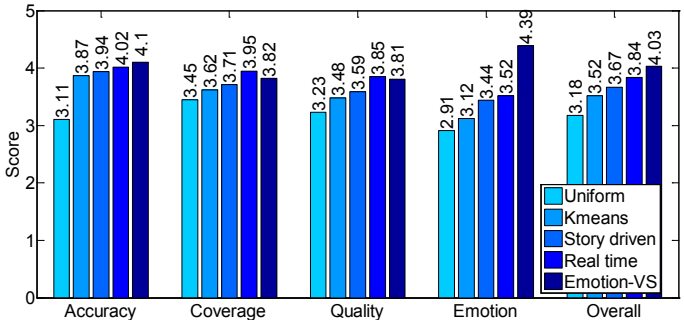

Fig. 10: User-study results on the video summarization.

4.5 Emotion-Oriented Video Summarization

Finally, we evaluate our framework on emotion-oriented video summarization. We compare with four baselines: (1) **Uniform sampling**, which uniformly samples several clips from video. (2) **K-means sampling**, which simply clusters the clips and selects a clip closest to each cluster centroid. (3) **Story-driven summarization [49]**. This approach was developed to summarize very long egocentric videos. We slightly modify the implementation and make the length of the summary controllable for our task. (4) **Real-time summarization [73]**, which is a technique aimed at efficient summarization of videos based on semantic content recognition results. For all the methods, the length of summary is fixed to 6 seconds if the original video is longer than 1 minute. For short videos, the length is fixed to 10% of the original video.

Following [68], we conduct a user study to evaluate different summarization methods. Ten subjects unfamiliar with the project participated in the study. We show the summary results of all the methods (without the audio information) to each participant. Participants are asked to rate each result on a five-point scale for each of the following evaluation metrics: (1) *Accuracy*: the summary accurately describes the "dominating high-level semantics of the original video"; (2) *Coverage*: the summary covers as much visual content using as few frames as possible. (3) *Quality*: the overall subjective quality of the summary; (4) *Emotion*: the summary conveys the same main emotion as the original video.

The results are shown in Figure 10. The average score is shown in the "Overall" column. Our method ("Emotion-VS") performs better than the other methods on the accuracy and the emotion metrics. On the emotion metric, we beat the best baseline by a margin of 0.87. Although we are doing slightly worse on the coverage metric (-0.13 compared to the best baseline), the drop in quality is minimal (-0.04 compared to the best baseline). The results suggest that the selection of emotional key frames and clips does not only capture the emotion of the original video, but also improves the overall accuracy of the summary, since emotional content plays an important role in an accurate summary. Our emotion-oriented summarization method significantly increases the amount of emotional contents captured by the summary without material loss on other quality measures.

We show a qualitative evaluation in Figure 11. At the top, the figure shows a video of an art therapist (the woman in green). Different from other methods, our summarization

Fig. 11: The qualitative results of emotion-oriented video summarization.

not only captured the therapy procedure, but also focused on the sadness of the therapist, which is the central emotion conveyed in this video. At the bottom, we illustrate a user-generated video where a father surprises his daughter during a baseball game by dressing as the catcher and revealing himself. All baseline methods are more focused on the baseball game itself, which is only marginally related to the emotion of this video. In contrast, our method clearly captures the reveal of the father, the surprised daughter, and the subsequent emotional hug.

Figure 12 demonstrates an example of using the video summary for retrieval on all available videos in terms of cosine similarity between videos and frames in Eq (10). The results shows the top retrieval results are all the same category and share some common visual characteristics. This also indicates the effectiveness of our method when finding the emotional clips.

5 Conclusions

Making a strong emotional appeal to viewers is the ultimate goal of many video producers. Therefore, being able to recognize a video's emotional impact is an important task for computer vision and affective computing. Given the diverse landscape of emotional expressions, we propose the first knowledge transfer framework for learning from heterogeneous sources for the task of understanding video emotion. Within the framework, we tackled interrelated video understanding problems including supervised and zero-shot emotion recognition, emotion attribution and emotion-oriented summarization.

For effective knowledge transfer, we learn an encoding scheme from a large-scale emotional image data set and a large, 7-billion-word text corpora. The encoding facilitates the creation of a representation conducive to the tasks of understanding video emotion. In zero-shot emotion recognition, an unknown emotional word is related to known emotion classes through the use of a distributed represen-

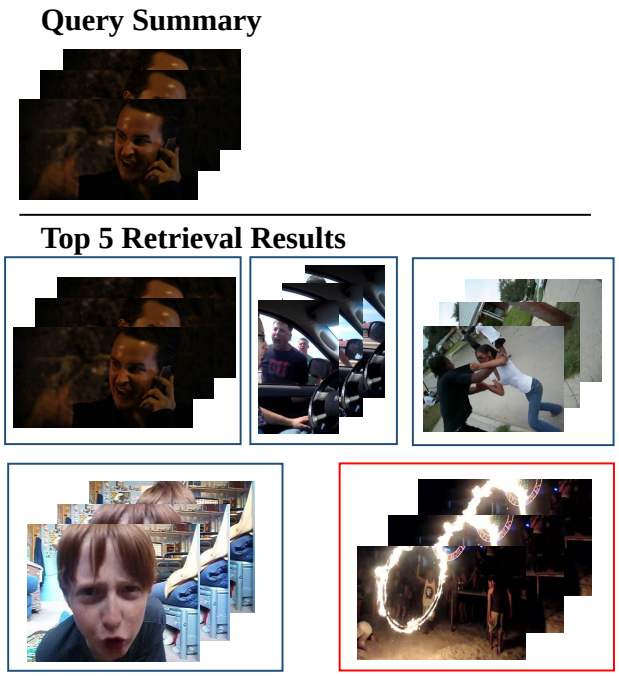

Fig. 12: A retrieval example using video summary. The query summary of an angry man is illustrated on the top. Top 5 retrieval results are showed on the bottom. The first 4 results (blue box) are all in the 'angry' category and the last result (red box) is a clip of magic show from 'joy' category.

tation in order to identify emotions unseen during training. Our experiments on three challenging datasets clearly demonstrate the benefit of utilizing external knowledge. The proposed framework also enables novel applications such as emotion attribution and emotion-oriented video summarization. A user study shows that our summaries accurately capture emotional content consistent with the overall emotion of the original video.

As future work, we will address the joint application of emotion-oriented summarization and story-driven summarization, which should allow us to create complete and emotionally compelling stories. The encoding of motion is another important topic that we leave for further investigation, partly due to the lack of a large-scale dataset that contain both emotional and motion information. The availability of such an auxiliary dataset would expand the applicability of the knowledge transfer framework as we defined in this paper. Computational understanding of human emotions exhibited in the video format holds the potential to the creation of intelligent systems that understand and interact naturally with humans.

6 Acknowledgement

We are grateful to the anonymous reviewers, whose suggestions considerably improved this paper. This work was supported in part by a China's National 863 Program ($\#2014AA015101$), and two grants from National Natural Science Foundation of China ($\#61572138$ and $\#U1509206$). This work is also funded by the PAPD of Jiangsu Higer Education Institutions and Jiangsu CICAEET.

References

[1] J. Amores. Multiple instance classification: Review, taxonomy and comparative study. *Artif. Intell.*, 201(4):81–105, 2013.
[2] S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multiple-instance learning. In *NIPS*, pages 561–568, 2003.
[3] L. F. Barrett. Are emotions natural kinds? *Perspectives on Psychological Science*, 1(1):28–58, 2006.
[4] L. F. Barrett, K. A. Lindquist, and M. Gendron. Language as context for the perception of emotion. *Trends in cognitive sciences*, 11(8):327–332, 2007.
[5] D. Borth, R. Ji, T. Chen, T. M. Breuel, and S.-F. Chang. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *ACM MM*, 2013.
[6] J. M. Carroll and J. A. Russell. Do facial expressions signal specific emotions? Judging emotion from the face in context. *Journal of Personality and Social Psychology*, 70(2):205–218, 1996.
[7] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *BMVC*, 2014.
[8] T. Chen, D. Borth, Darrell, and S.-F. Chang. Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks. *arXiv preprint arXiv:1410.8586*, 2014.
[9] Y. Chen, J. Bi, and J. Z. Wang. Miles: Multiple-instance learning via embedded instance selection. *IEEE TPAMI*, 28(1):1931–1947, 2006.
[10] D. DeMenthon, V. Kobla, and D. Doermann. Video summarization by curve simplification. In *ACM MM*, 1998.
[11] A. Dhall and G. Roland. Group expression intensity estimation in videos via gaussian processes. In *ICPR*, 2012.
[12] R. J. Dolan. Emotion, cognition, and behavior. *Science*, 298(5596):1191–1194, 2002.
[13] P. Ekman. Universals and cultural differences in facial expressions of emotion. *Nebraska Symposium on Motivation*, 19:207–284, 1972.
[14] P. Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200, 1992.
[15] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov, et al. Devise: A deep visual-semantic embedding model. In *NIPS*, 2013.
[16] Y. Fu. Multi-view metric learning for multi-view video summarization. *arXiv preprint arXiv:1405.6434*, 2014.
[17] Y. Fu, Y. Guo, Y. Zhu, F. Liu, C. Song, and Z.-H. Zhou. Multi-view video summarization. *IEEE TMM*, 12(7):717–729, 2010.
[18] Y. Fu, T. M. Hospedales, T. Xiang, and S. Gong. Learning multi-modal latent attributes. *IEEE TPAMI*, 36(2):303–316, 2013.
[19] Y. Fu, T. M. Hospedales, T. Xiang, and S. Gong. Transductive multi-view zero-shot learning. *IEEE TPAMI*, to appear.
[20] Y. Fu, T. M. Hospedales, T. Xiang, S. Gong, and Y. Yao. Interestingness prediction by robust learning to rank. In *ECCV*, 2014.
[21] Y. Fu, T. M. Hospedales, T. Xiang, J. Xiong, S. Gong, Y. Wang, and Y. Yao. Robust subjective visual property prediction from crowdsourced pairwise labels. *IEEE TPAMI*, to appear.
[22] T. Gärtner, P. A. Flach, A. Kowalczyk, and A. J. Smola. Multi-instance kernels. In *ICML*, pages 179–186. Morgan Kaufmann, 2002.
[23] J. J. Gross. Emotion regulation: Affective, cognitive, and social consequences. *Psychophysiology*, 39(3):281–291, 2002.
[24] B. Gu and V. S. Sheng. A robust regularization path algorithm for support vector classification. *IEEE Transactions on Neural Networks & Learning Systems*, 2016.
[25] B. Gu, X. Sun, and V. S. Sheng. Structural minimax probability machine. *IEEE Transactions on Neural Networks & Learning Systems*, 2016.
[26] M. Gygli, H. Grabner, H. Riemenschneider, and L. V. Gool. Creating summaries from user videos. In *ECCV*, 2014.
[27] A. Habibian, T. Mensink, and C. G. M. Snoek. Videostory: A new multimedia embedding for few-example recognition and translation of events. In *ACM MM*, 2014.
[28] S. Hamann and T. Canli. Individual differences in emotion processing. *Current Opinion in Neurobiology*, 14(2):233–238, 2004.
[29] A. Hanjalic and H. Zhang. An integrated scheme for automated video abstraction based on unsupervised cluster-validity analysis. *IEEE TCSVT*, 9(8):1280–1289, 1999.
[30] Z. Harris. Distributional structure. *Word*, 10(23):146–162, 1954.
[31] J. A. Hartigan and M. A. Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979.

[32] G. Irie, T. Satou, A. Kojima, T. Yamasaki, and K. Aizawa. Affective audio-visual words and latent topic driving model for realizing movie affective scene classification. *IEEE TMM*, 12(6):523–535, Oct 2010.

[33] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM MM*, 2014.

[34] Y.-G. Jiang, B. Xu, and X. Xue. Predicting emotions in user-generated videos. In *AAAI*, 2014.

[35] Y.-G. Jiang, J. Yang, C.-W. Ngo, and A. G. Hauptmann. Representations of keypoint-based semantic concept detection: A comprehensive study. *IEEE TMM*, 12(1):42–53, 2010.

[36] Y.-G. Jiang, YanranWang, R. Feng, X. Xue, Y. Zheng, and H. Yang. Understanding and predicting interestingness of videos. In *AAAI*, 2013.

[37] B. Jou, S. Bhattacharya, and S.-F. Chang. Predicting viewer perceived emotions in animated gifs. In *ACM MM*, 2014.

[38] H.-B. Kang. Affective content detection using HMMs. In *ACM MM*, 2003.

[39] A. Kapoor, W. Burleson, and R. W. Picard. Automatic prediction of frustration. *International journal of human-computer studies*, 65(8):724–736, 2007.

[40] D. Kotzias, M. Denil, P. Blunsom, and N. de Freitas. Deep multi-instance transfer learning. *arXiv preprint arXiv:1411.3128*, 2014.

[41] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[42] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009.

[43] C. H. Lampert, H. Nickisch, and S. Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE TPAMI*, 36(3):453–465, 2013.

[44] B. Li. A dynamic and dual-process theory of humor. In *The 3rd Annual Conference on Advances in Cognitive Systems*, 2015.

[45] K. A. Lindquist, T. D. Wager, H. Kober, E. Bliss-Moreau, and L. F. Barrett. The brain basis of emotion: a meta-analytic review. *Trends in cognitive sciences*, 35(3):121–143, 2012.

[46] D. Liu, G. Hua, and T. Chen. A hierarchical visual model for video object summarization. *IEEE TPAMI*, 32(12):2178–2190, 2009.

[47] G. Liu, J. Wu, and Z. Zhou. Key instance detection in multi-instance learning. In *ACML*, 2012.

[48] X. Lu, P. Suryanarayan, R. B. Adams, J. Li, M. G. Newman, and J. Z. Wang. On shape and the computability of emotions. In *ACM MM*, 2012.

[49] Z. Lu and K. Grauman. Story-driven summarization for egocentric video. In *CVPR*, 2013.

[50] T. Ma, J. Zhou, M. Tang, Y. Tian, A. Aldhelaan, M. Alrodhaan, and S. Lee. Social network and tag sources based augmenting collaborative recommender system. *IEICE Transactions on Information & Systems*, E98.D(4):902–910, 2015.

[51] Y.-F. Ma, L. Lu, H.-J. Zhang, and M. Li. A user attention model for video summarization. In *ACM MM*, 2002.

[52] J. Machajdik and A. Hanbury. Affective image classication using features inspired by psychology and art theory. In *ACM MM*, 2010.

[53] C. D. Manning, P. Raghavan, and H. Schutze. *Introduction to Information Retrieval*. Cambridge University Press, 2009.

[54] S. Marsella and J. Gratch. EMA: A process model of appraisal dynamics. *Journal of Cognitive Systems Research*, 10(1):70–90, 2009.

[55] D. McDuff, R. E. Kaliouby, J. F. Cohn, and R. Picard. Predicting ad liking and purchase intent: Large-scale analysis of facial responses to ads. *IEEE TAC*, 6(3):223–235, 2015.

[56] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, 2013.

[57] C.-W. Ngo, Y.-F. Ma, and H.-J. Zhang. Video summarization and scene detection by graph modeling. *IEEE TCSVT*, 15(2):296–305, 2005.

[58] A. Ortony, G. Clore, and A. Collins. *The Cognitive Structure of Emotions*. Cambridge University Press, 1988.

[59] Z. Pan, J. Lei, Y. Zhang, and X. Sun. Fast motion estimation based on content property for low-complexity h.265/hevc encoder. *IEEE Transactions on Broadcasting*, 2016.

[60] Z. Pan, Y. Zhang, and S. Kwong. Efficient motion and disparity estimation optimization for low complexity multiview video coding. *IEEE Transactions on Broadcasting*, 61(2):1–1, 2015.

[61] W.-T. Peng, W.-T. Chu, C.-H. Chang, C.-N. Chou, W.-J. Huang, W.-Y. Chang, and Y.-P. Hung. Editing by viewing: automatic home video summarization by viewing behavior analysis. *Multimedia, IEEE Transactions on*, 13(3):539–550, 2011.

[62] R. Plutchik. Emotion: Theory, research, and experience. In *Theories of Emotion*, volume 1. Academic Press, 1980.

[63] B. M. Sikka K, Dhall A. Weakly supervised pain localization using multiple instance learning. In *IEEE FG*, 2013.

[64] J. Sivic and A. Zisserman. Video google: a text retrieval approach to object matching in videos. In *ICCV*, pages 1470–1477, 2003.

[65] N. C. Smeeton. Early history of the kappa statistic. *Biometrics*, 41(3):795–795, 1984.

[66] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng. Zero-shot learning through cross-modal transfer. In *NIPS*, pages 935–943, 2013.

[67] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *CVPR*, 2015.

[68] B. T. Truong and S. Venkatesh. Video abstraction: A systematic review and classification. *ACM TOMM*, 3(1):79–82, 2007.

[69] H.-L. Wang and L.-F. Cheong. Affective understanding in film. *IEEE TCSVT*, 16(6):689–704, 2006.

[70] J. Wang and J.-D. Zucker. Solving the multiple-instance problem: A lazy learning approach. In *ICML*, pages 1119–1126, 2000.

[71] M. Wang, R. Hong, G. Li, Z.-J. Zha, S. Yan, and T.-S. Chua. Event driven web video summarization by tag localization and key-shot identification. *IEEE TMM*, 14(4):975–985, 2012.

[72] S. Wang and Q. Ji. Video affective content analysis: a survey of state of the art methods. *IEEE TAC*, 6(99):1–1, 2015.

[73] X. Wang, Y. Jiang, Z. Chai, Z. Gu, X. Du, and D. Wang. Real-time summarization of user-generated videos based on semantic recognition. In *ACM MM*, 2014.

[74] X.-S. Wei, J. Wu, and Z.-H. Zhou. Scalable multi-instance learning. In *ICDM*, 2014.

[75] X. Wen, L. Shao, Y. Xue, and W. Fang. A rapid learning algorithm for vehicle classification. *Information Sciences*, 295:395–406, 2015.

[76] C. Xu, S. Cetintas, K.-C. Lee, and L.-J. Li. Visual sentiment prediction with deep convolutional neural networks. *arXiv preprint arXiv:1411.5731*, 2014.

[77] X. Xu and E. Frank. Logistic regression and boosting for labeled bags of instances. In *8th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2004.

[78] Z. Xu, Y. Yang, and A. G. Hauptmann. A discriminative CNN video representation for event detection. In *CVPR*, 2015.

[79] Q. You, J. Luo, H. Jin, and J. Yang. Robust image sentiment analysis using progressively trained and domain transferred deep networks. In *AAAI*, 2015.

[80] Z. Zhou, Y. Wang, Q. M. J. Wu, C. N. Yang, and X. Sun. Effective and efficient global context verification for image copy detection. *IEEE Transactions on information forensics and security*, 2016.

[81] Z.-H. Zhou and M.-L. Zhang. Solving multi-instance problems with classifier ensemble based on constructive clustering. *Knowledge and Information Systems*, 11(2):155–170, 2007.

[82] R. D. Zhu X. Face detection, pose estimation, and landmark localization in the wild. In *CVPR*, 2012.

Baohan Xu received the BS degree from Fudan University, Shanghai, China, in 2014. She is now pursuing her MS degree of Computer Science at Fudan University. Her research interests include computer vision and video emotion analysis.

Yanwei Fu received the PhD degree from Queen Mary University of London in 2014, and the MEng degree in the Department of Computer Science & Technology at Nanjing University in 2011, China. He worked as a Post-doc in Disney Research at Pittsburgh from 2015-2016. He is currently an Assistant Professor at Fudan University. His research interest is image and video understanding, and life-long learning.

Yu-Gang Jiang is a Professor in School of Computer Science, Fudan University, China. His Lab for Big Video Data Analytics conducts research on all aspects of extracting high-level information from big video data, such as video event recognition, object/scene recognition and large-scale visual search. He is the lead architect of a few best-performing video analytic systems in worldwide competitions such as the annual U.S. NIST TRECVID evaluation. His visual concept detector library (VIREO-374) and video datasets (e.g., CCV and FCVID) are widely used resources in the research community. His work has led to many awards, including "emerging leader in multimedia" award from IBM T.J. Watson Research in 2009, early career faculty award from Intel and China Computer Federation in 2013, the 2014 ACM China Rising Star Award, and the 2015 ACM SIGMM Rising Star Award. He holds a PhD in Computer Science from City University of Hong Kong and spent three years working at Columbia University before joining Fudan in 2011.

Boyang Li is a Research Scientist at Disney Research, where he directs the Narrative Intelligence group. He obtained his Ph.D. in Computer Science from Georgia Institute of Technology in 2014, and his B. Eng. from Nanyang Technological University, Singapore in 2008. His research interests include computational narrative intelligence, or the creation of Artificial Intelligence that can understand, craft, tell, direct, and respond appropriately to narratives, and understanding how human cognition comprehends narratives and produces narrative-related affects.

Leonid Sigal is a Senior Research Scientist at Disney Research Pittsburgh and an adjunct faculty at Carnegie Mellon University. Prior to this he was a postdoctoral fellow in the Department of Computer Science at University of Toronto. He completed his Ph.D. at Brown University in 2008; he received his B.Sc. degrees in Computer Science and Mathematics from Boston University (1999), his M.A. from Boston University (1999), and his M.S. from Brown University (2003). From 1999 to 2001, he worked as a senior vision engineer at Cognex Corporation, where he developed industrial vision applications for pattern analysis and verification.