IEEE Transactions on Multimedia

A Multi-task Neural Approach for Emotion Attribution, Classification and Summarization

Journal:	*IEEE Transactions on Multimedia*
Manuscript ID	Draft
Suggested Category:	Regular Paper
Date Submitted by the Author:	n/a
Complete List of Authors:	Tu, Guoyun; Fudan university, School of Physics Fu, Yanwei; Fudan University, School of Data Science Gao, Jiarui; Fudan University, School of Compute Science Li, Boyang; Walt Disney Co, Disney Research Jiang, Yu-Gang; Fudan University, School of Computer Science Xue, Xiangyang; School of Computer Science, Fudan University
EDICS:	3-MPIM Multimodal Perception, Integration, and Multisensory Fusion < 3 HUMAN CENTRIC MULTIMEDIA, 4-SEIM Social and Educational Issues In Multimedia < 4 MULTIMEDIA ENVIRONMENTS, 5-SOCL Social and Web Multimedia < 5 MULTIMEDIA DATABASES AND PROCESSING INFRASTRUCTURE, 9-DLMA Deep Learning for Multimedia Analysis < 9 EMERGING TOPICS IN MULTIMEDIA

Cover Letter

Guoyun Tu, Yanwei Fu, Jiarui Gao, Boyang Li, Yu-Gang Jiang and Xiangyang Xue

September 21, 2017

In this document we detail the differences between our ICMR'17 conference paper (*Frame-Transformer Emotion Classification Network*, Jiarui Gao, Yanwei Fu, Yu-Gang Jiang, and Xiangyang Xue) and the submitted IEEE TMM manuscript.

Our ICMR paper firstly studies the problem of *a new architecture–Frame-Transformer Emotion Classification Network (FT-EC-net)* to solve three highly correlated emotion analysis tasks: emotion recognition, emotion attribution and emotion-oriented summarization. The paper is attached as the supplementary of the documents.

This TMM manuscript introduces a new neural approach– Frame- Bi-stream Emotion Attribution-Classification Network (BEACNet) to solve these three problems. Particularly, we extend the MINMAX dynamic programming method to solve emotion-oriented summarization problem.

Methodological Improvements

1. Supervised Attribution Network

   ICMR Paper: The Attribution Network call Frame-Transformer Network in ICMR paper was unsupervised;

   TMM Manuscript: New sections (Section III.*B*Section III.*D*) introduce square loss function for regression and the joint loss function, enabling the Attribution Network under supervision.

   Outcome: Supervised Attribution Network performs better in tasks of emotion recognition(Tab 1) and emotion attribution(Fig 2).

2. Two-Stream Neural Architecture in Classification Network.

   ICMR Paper: Emotion Classification Network employed a simple CNN network;

   TMM Manuscript: Two-stream neural architecture composed of emotion stream and content stream is described in Section III.*C*

   Outcome: This improves the accuracy in the result of emotion recognition as shown in Tab 1.

3. MINMAX-DP For Emotion-Oriented Summarization.

   ICMR Paper: The result of attribution was taken as the result of summarization.

   TMM Manuscript: Section III.*E* demonstrates the approach which aims to maintain continuity between selected video frames while select as few frames as possible and focus on the emotional content.

   Outcome: User study(Fig 3) supports our method outperforms all baseline methods.

Experimental Additions

1. New condition of dataset.

TMM Manuscript: The experiment is conducted on the full Ekman-6 dataset while ICMR paper just compared the methods on two classes of Ekman-6 dataset.(Tab 1)

Outcome: The introduction of a more difficult dataset further proves the superior performance of BEACNet on the task of emotion recognition.

2. Ablation Studies:

TMM Manuscript:

ICMR Paper: Not present.

TMM Manuscript: We add three ablated networks: C-stream, E-stream and Unsup. E-stream for comparison.(Tab 1)

Outcome: The result shows the contribution of all our sub-networks.

Analysis of multi-modal attributes

Presentation Improvements

(a) **TMM Manuscript:** Gives formal mathematics expression and detailed function introduction on *The Attribution Network* (Section III.*B*);
ICMR Paper: Brief explanations about Frame-Transformer Network.

(b) **TMM Manuscript:** Explains the advantage of our emotion-oriented summarization compared with other methods in Section IV.*D*;
ICMR Paper: Brief explanations about the result.

A Multi-task Neural Approach for Emotion Attribution, Classification and Summarization

Guoyun Tu, Yanwei Fu, Jiarui Gao, Boyang Li, Yu-Gang Jiang and Xiangyang Xue

***Abstract*—Emotional content is a crucial ingredient in user-generated videos. However, the sparsely expressed emotions in the user-generated video cause difficulties to emotions analysis in videos. In this paper, we propose a new neural approach–Frame-Bi-stream Emotion Attribution-Classification Network(BEAC-Net) to solve three highly correlated emotion analysis tasks in an integrated framework: emotion recognition, emotion attribution and emotion-oriented summarization. We also contribute a new dataset for emotion attribution task by annotating the ground-truth labels of attribution segments. We apply the proposed networks on two video datasets and demonstrate the superior effectiveness of our framework in comparison to baseline methods.**

I. Introduction

The demand for computational understanding of visual media data has been increasing due to the explosive growth of user-generated videos. Great efforts and achievement have been made on the video content understanding, such as video activities and video actions. However, computational recognition and understanding of emotions from user-generated videos [1], [2] still remains a large problem to solve given the catholicity of emotional expressions in video content.

There are various real-world applications on computational understanding of the emotions expressed by video content.For instance, video recommendation services may be helped by matching users' attention on the emotions of video content; advertisers would be inclined to put ads alongside videos with homogeneous emotions to enhance the effect.

We consider three main challenges occupying the emotion understanding. First, the emotions are usually expressed by a small subset of frames in the total video, while the other parts supply context and background; thus the computational approach must be sensitive to the sparse emotional segment. Second, one video may contain several emotions with one dominant emotion; the discernment in video segment that has the major contribution to the overall emotion plays a significant role, which is identified as the video emotion attribution task [2]. Third, user-generated videos are varied in production quality, and contain diverse objects, scenes and events. This brings more and larger challenges compared with commercial videos with consistent production quality (e.g., illumination and camera positions)and restricted topic, such as news reports.

Guoyun Tu, Yanwei Fu, Jiarui Gao, Yu-gang Jiang and Xiangyang Xue are with Fudan University, Shanghai, China.

Boyang Li is with Disney Research, Pittsburgh, Pennsylvania, USA.

Yanwei Fu (corresponding author) is with the School of Data Science, Fudan University, Shanghai, China. Email:yanweifu@fudan.edu.cn

Considering these challenged, we contend that it is crucial to extract feature representations sensitive to emotions and unvarying under irrelevant conditions . In previous works, this is achieved by combining low-level and middle-level features [1], or by encoding video using a set of emotion-rich auxiliary images [2], [3]. These former work demonstrated that the effectiveness of these features could be obtained on emotion recognition, emotion attribution and emotion-oriented video summarization tasks. However, a major shortcoming is that the three tasks were handled separately and do not inform each other. We suppose that sharing information could improve the overall performance.

In this paper, we propose a multi-task neural architecture, the Bi-stream Emotion Attribution-Classification Network (BEAC-Net), which tackles both emotion attribution and classification at the same time, thereby allowing related tasks to reinforce each other. BEAC-Net is composed of an attribution network (A-Net) and a classification network (C-Net). The attribution network learns to detect the emotional video segments. The classification network processes the segments selected by the A-Net as well as the original frames of the video in a bi-stream architecture in order to recognize the emotion. In this setup, both the content information and the emotional information are retained to achieve high accuracy with a small number of CNN layers. Empirical evaluation on the Ekman-6 and the Emotion6 Video datasets demonstrate clear benefits of the joint approach and the complementary nature of the two streams.

The contributions of this work can be summarized as follows: (1) We propose BEAC-Net, an end-to-end trainable neural architecture that tackles emotion attribution and classification simultaneously with significant performance improvements. (2) We propose an efficient dynamic programming method for video summarization based on the output of A-Net. (3) To establish a good benchmark for emotion attribution, we re-annotate the Ekman-6 dataset with the most emotion-oriented segments which can be used as the ground-truth for the emotion attribution task.

II. Related Work

A. Image and Video Emotion Recognition

Extensive research has been performed on the problem of recognizing emotions from visual information. Most work follow psychological theories that lay out a fixed number of emotion categories, such as Ekman's six pan-cultural basic emotions [4] and Plutchik's wheel of emotion [5]. More recent work like DeepSentiBank [6] and zero-shot emotion recognition [2]

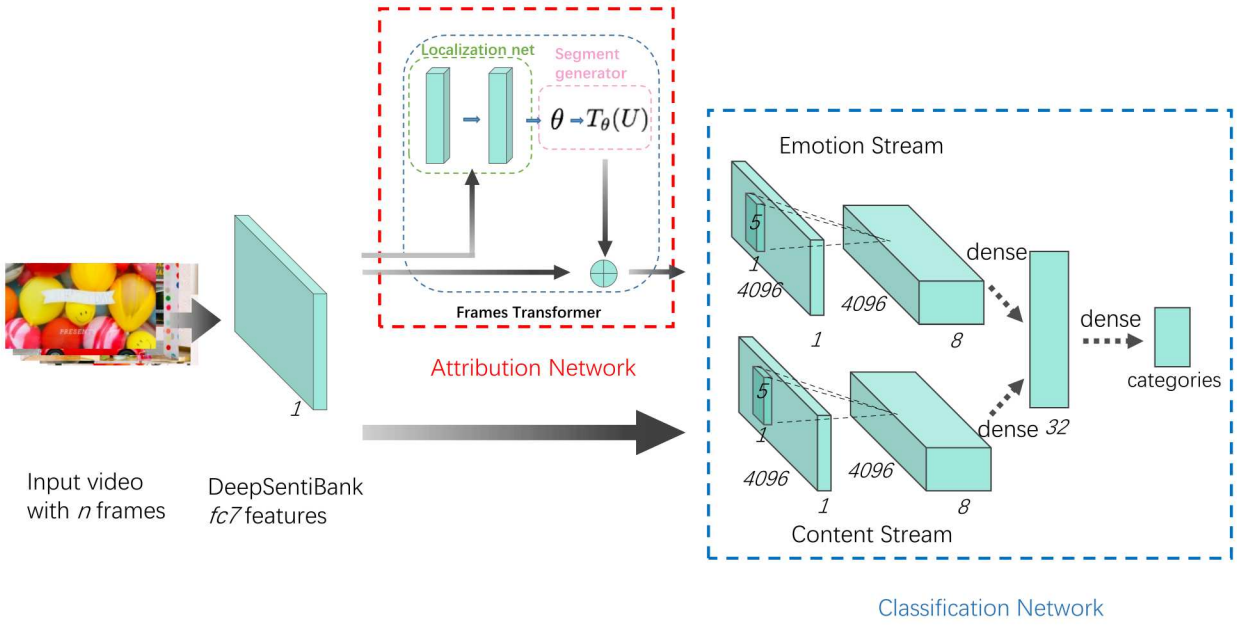


Fig. 1. An overview of the BEAC-Net neural architecture. We extract features from every video frame from the "fc7" layer of the DeepSentiBank convolutional neural network model. The attribution network extracts one video segment that expresses the dominant emotion, which is fed to the emotion stream of the classification network. The whole video is downsampled and fed to the content stream of the classification network.

deviate from this convention, as recent psychological theories [7]–[9] suggest the range of emotions are more varied than prescribed by previous theories due to the complex interaction between emotions and other cognitive processes.

Various researchers explored features for visual emotion recognition, such as features enlightened by psychology and art theory [10] and shape features [11]. A classifier such as a support vector machine (SVM) or K-nearest neighbors (KNN) is trained to distinguish video's emotions. Wang et al [12] adapted a variant of SVM with various audio-visual features to divide 2040 frames of 36 Hollywood movies into 7 emotions. Jou et al. [13] focused on animated GIF files. Yazdani et al. [14] used KNN to classify music video clips.

Since facial expressions are important expressions of emotion, many researchers focused on recognizing emotions from facial expressions. Joho et al. [15] paid close attention to viewers' facial signals for detection. Zhao et al. [16] extracted viewers' facial activities frame by frame and drew an emotional curve to classify each video into different sections. Preeminently, Liu et al. [17] construct expressionlet, a mid-level representation for dynamic facial expression recognition.

Deep neural networks have also been used for visual sentiment analysis [18], [19]. A massive scale of visual sentiment dataset was proposed in Sentibank [19] and DeepSentiBank [6]. Sentibank is composed of 1,533 adjective-noun pairs, such as "happy dog" and "beautiful sky". Subsequently, the authors used deep convolutional neural networks (CNN) dealing with images of strong sentiment and achieved better performance than the former models.

Progress in emotion recognition has been made on extracting different types of features, such as the low-level visual and audio features, attribute features in [1] as well as the mid-level audio-visual features in [20]. Emotions have also been researched in GIF files [21] which can be regarded as one type of short videos. For a more recent survey, we refer interested readers to [22].

Most existing work focus on emotion understanding from video focus on classification. In previous work [2], [3], we proposed two new tasks — video emotion attribution and emotion-oriented video summarization — that requires effective emotion understanding. In particular, the attribution task asks if we can identify video frames that express emotional content, which directly addresses the challenge that emotional expressions are sparse in a video. Noting the synergy between emotion attribution and recognition, we propose a multi-task neural network that tackles both tasks in this paper.

B. Emotion attribution and video summarization

Video summarization has been studied for more than two decades [23] and a detailed review is beyond the scope of this paper. In broad strokes, we can categorize summarization approaches into two major approaches: keyframes extraction and video skims. A large varienty of video features have been exploited, including visual saliency [24], motion cues [25], mid-level features [26], and semantic recognition [27].

Recently, we [2] introduced the task of emotion-oriented summarization which points at finishing video summarization task according to video emotion content. Inspired by the task of semantic attribution in text analysis, [2]the task of emotion attribution are defined as attributing the video's

overall emotion to its individual segments. However, [2] still processed the emotion recognition, summarization and attribution tasks separately. Intrinsically, the emotion recognition can remarkably benefit from emotion attribution and emotion-oriented summarization; and the results of emotion attribution can provide more information for emotion-oriented summary. Thus, our framework is designed to solve these three tasks simultaneously and mutually.

In the conference version [28], we only focused on the segment of high emotional value while neglected other frames which may contain content information. However, in this paper, the emotion segment and the entire content information will be combined with different emphasis.

C. Spatial transform networks

Spatial transform network (ST-net) [29] is firstly proposed for image (or feature map) classification. ST-Net provides the capability for spatial transformation, which helps various tasks such as co-localization [30] and spatial attention [31]. It is fully-differentiable and it can transform an image or a feature map with little time loss as an insert framework.

ST-net could be split into three parts: 1) Localization Network: employing an arbitrary form of function $f_{loc}()$ to generate the the transformation parameters θ which could adopt any dimensions. 2) Parameterised Sampling Grid: applying the generated θ to build a regular grid $G = G_i$ and avert the source feature map to the target coordination. 3) Differentiable Image Sampling: providing a (sub-)differentiable sampling mechanism and allowing the loss gradients to flow back to the whole network.

So far there are various variants and improvement of ST-net. Singh et al. [32]. adapted it for end-to-end facial learning framework and proposed a loss function for solving the problem that ST-net might lead to its output patch beyond the input boundaries. Lin et al. [33] improved upon ST-net by theoretically connecting it to the inverse compositional LK algorithm which exhibit better performance than original ST-net in various tasks.

The attribution network in BEAC-Net can be seen as performing spatial transformation on the temporal dimension. It enables the network to identify video segments that carry emotional content, which alleviates the sparsity of emotion content in video.

III. The Emotion Attribution-Classification Network

The BEAC-Net architecture is an end-to-end multi-task network that naturally handles the emotion attribution task and the emotion classification task. In this section, we describe its two constituents: the emotion attribution network (A-Net) and the emotion classification network (C-Net). The former extracts a segment from the video that contains emotional content, whereas the latter classifies the video into an emotion by using the extracted segment together with its context. Each input video is represented by features extracted using the DeepSentiBank network [6]. Fig. 1 provides an illustration of the network architecture.

A. Feature Extraction

We extract video features using the deep convolutional network provided by [6], which classifies images into adjective noun pairs (ANPs). Each ANP is a concept consists of an adjective followed by a noun, such as ''creepy house'' and ''dark places''. The network was trained on 867,919 images for classification into 2,089 ANPs. The network in [6] contains five convolutional layers and three fully connected layers. We take the 4096-dimensional activation from the second fully connected layer labeled as "fc7".

The classification into ANPs can be considered as the joint recognition of objects and the affects associated with the object. Indeed, [6] reports that initializing weights trained from solely object recognition on ImageNet substantially boosts performance. As a result, we believe the features extracted by this network retain both object and affective information from the images.

Formally, let us denote the whole dataset as $\mathcal{D} = \{(X_i, y_i, \boldsymbol{\alpha}_i)\}_{i=1,\ldots,N}$ where X_i denotes the i^{th} video , y_i denotes its emotion label, and $\boldsymbol{\alpha}_i$ denotes the supervision on emotion attribution, which is explained in the next section. The M frames of X_i are denoted as $\{\boldsymbol{x}_{i,j}\}_{j=1,\ldots,M}$. Let $\phi(\cdot)$ be the feature extraction function. The i^{th} video features are represented as $\{\phi(\boldsymbol{x}_{i,j})\}_{j=1,\ldots,M}$.

B. The Attribution Network

The emotion attribution task is to identify video frames responsible for a particular emotion in the video. The attribution network learns to select one continuous segment of L frames that accounts for the main emotion in the original video of M frames. The network performs regression to predict parameters $\boldsymbol{\alpha}$ for the affine transformation between the temporal coordinates of video segments.

The transformation defined by $\boldsymbol{\alpha}$ maps the indices, or temporal coordinates, of the selected video segment back to the full range of temporal coordinates. Formally, let the index set of frames in the video be $I = [1, M]$ and the index set of the selected frames be $I_s = [1, L]$. We let the indices be continuous due to the possibility of interpolation. After we select a video segment by finding its start time $s \in I$ and end time $e \in I$, we need an operation that maps s to 1 and e to L. In fact, the attribution task is finished as long as we find the mapping operation because frames mapped to indices outside $[1, L]$ will be discarded.

For this purpose, we define two transformational parameters $\boldsymbol{\alpha} = [\alpha_1\ \alpha_2]$.

$$\begin{aligned} \alpha_1 &= \frac{1}{M}(e - s) \\ \alpha_2 &= \frac{1}{M}(e + s - M) \end{aligned} \tag{1}$$

The relationship between the two index sets can be defined using an affine transformation. Letting $z \in I, s \le i \le e$ and $z_s \in I_s$, we have

$$z = \begin{bmatrix} \alpha_1 & \alpha_2 \end{bmatrix} \begin{pmatrix} z_s \\ 1 \end{pmatrix} \tag{2}$$

α_1 can be understood as the scale parameter which controls the length ratio between the input video and the selected segment;

α_2 is the translation parameter which indicates the offset of the segment over the video. When z is not an integer, it is rounded to the nearest integer $\lfloor z + 0.5 \rfloor$.

The attribution network employs two fully connected layers to project the video features $\{\phi(\boldsymbol{x}_{i,j})\}_{j=1,\ldots,M}$ to $\boldsymbol{\alpha}$. The network utilizes the following square loss function for regression, which computes the differences between the output of the network $\hat{\boldsymbol{\alpha}}_i$ and the externally supplied supervision $\boldsymbol{\alpha}_i$.

$$\mathcal{L}_i^A = \left((\alpha_{i,1} - \hat{\alpha}_{i,1})^2 + (\alpha_{i,2} - \hat{\alpha}_{i,2})^2\right) \tag{3}$$

To avoid numerical issues, the indices $z \in I$ and $z_s \in I_s$ are normalized to the range $(-1, 1)$ in practice.

In order to provide a solution to the emotion attribution task, we only need to perform the inverse operation of Eq. 1 and recover the start time s and end time e from the regression output $\hat{\boldsymbol{\alpha}}_i$:

$$\begin{aligned} \hat{s} &= \frac{M}{2}(\hat{\alpha}_2 - \hat{\alpha}_1 + 1) \\ \hat{e} &= \frac{M}{2}(\hat{\alpha}_1 + \hat{\alpha}_2 + 1) \end{aligned} \tag{4}$$

C. The Classification Network

The emotion classification task, as the name implies, categorizes the video as one of the emotions. We propose a novel two-stream neural architecture that employs the emotion segment selected by the attribution network in combination with the original video. This architecture allows us to focus on the dominant emotion, but also consider the context it appears in. It may also be seen as a variant of multi-resolution network, where we apply coarse resolution on the entire video and fine resolution on the emotional segment. We call the two streams the emotion stream and the content stream respectively.

The two streams have symmetrical architectures, containing one convolution layer before converging to two fully connected layer. In the content stream, the 100 frames from the original video are compressed to 20 frames, resulting in an input matrix of size 20×4096, which is identical with the input to the emotion stream. As we use the DeepSentiBank model to extract features for individual frames, the convolution layer can be seen as learning feature over the temporal dimension. The two streams share the same parameters for the convolutional layer, which accelerates training remarkably. Meanwhile, the interaction of both streams is the critical contributor to achieving high accuracy. The final output of the network comes from a softmax layer.

The classification network adopts the standard cross-entropy loss. For a K-class classification, the loss function is written as

$$\mathcal{L}_i^C = \sum_{k=1}^{K} -y_{ik} \cdot \log(\hat{y}_{ik}) \tag{5}$$

where y_i is a ground-truth one-hot vector and $\hat{y}_i$ is the output of the softmax classification.

D. Joint Training

In order to stabilize optimization, for every data point X_i, we introduce an indicator function $\mathbb{1}(o_i < \beta)$, where o_i indicates the temporal intersection over union (tIoU) between the A-Net's prediction $\hat{\boldsymbol{\alpha}}_i$ and ground truth $\boldsymbol{\alpha}_i$. β is a predefined threshold. That is, the indicator function returns 1 if and only if the attribution network is sufficiently accurate for X_i.

We combine the standard cross-entropy classification loss and the attribution regression loss to create the final loss function as

$$\mathcal{L} = \frac{1}{N} \sum_i \left[\mathbb{1}(o_i \geq \beta)\mathcal{L}_i^C + \mathbb{1}(o_i < \beta)\mathcal{L}_i^A\right] \tag{6}$$

In plain words, gradients from the classification loss are backpropagated only when the attribution network is accurate enough. Otherwise, gradients from the attribution loss are backpropagated. For each data point, the network focuses on training either the A-Net or the C-Net, but not both. We find this to stabilize training and improve performance.

E. Emotion-Oriented Summarization

In this section, based on the output of the emotion attribution network, we formulate the emotion-oriented summarization problem as a constrained optimization. The summarization aims to maintain continuity between selected video frames while select as few frames as possible and focus on the emotional content. This problem can be efficiently solved by the MINMAX dynamic programming method [34].

The emotion-oriented summarization problem can be formally stated as follows. From a video X_i containing M frames $\{\boldsymbol{x}_{i,j}\}_{j=1,\ldots,M}$, we want to select a subset of frames $h_1, \ldots, h_P \in \{1, \ldots, M\}$. Here P is not a predefined constant but determined by the optimization. We want to optimize the sum of individual frame's cost

$$\min \sum_p \text{cost}(h_p) \tag{7}$$

subject to the following constraints.

- $h_1 = 1, h_P = M$. Always select the first and the last frames.
- $h_{p+1} - h_p \leq K_{max}$. The frames are not too spaced out. The constant K_{max} is the maximum index difference for adjacent summary frames.
- $\forall h_p \leq i, j \leq h_{p+1}, d(\phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_j)) \leq D_{max}$, where $d(\cdot)$ is the Euclidean distance. In words, there is no large feature-space discontinuity ($\leq D_{max}$) between h_p and h_{p+1} in the video.

In other words, we minimize the total cost by selecting fewer frames, but we must also make sure removing a frame does not create a large gap in feature space. This avoids discontinuity that disrupts the viewing experience.

Based on the emotional segment identified by the A-Net, we encourage the inclusion of emotional frames in the summary by setting

$$\text{cost}(i) = \begin{cases} 1 & \text{if } \boldsymbol{x}_i \text{ in the emotional segment} \\ 2 & \text{if } \boldsymbol{x}_i \text{ not in the emotional segment} \end{cases}$$

We present a solution to the problem using the MINMAX dynamic programming technique [34]. We measure the discontinuity between adjacent frames in the video summary between the selected frames h_p and h_{p+1} as

$$D_p^{p+1} = \begin{cases} \max\limits_{i,j \in [h_p, h_{p+1}]} d\left(\phi(\boldsymbol{x}_{h_i}), \phi(\boldsymbol{x}_{h_j})\right) & if\ l_{t+1} - l_t \leq K_{max} \\ \infty & otherwise \end{cases}$$

where $d(\cdot)$ denotes the Euclidean distance between the features $\phi(\boldsymbol{x}_{h_i})$ and $\phi(\boldsymbol{x}_{h_j})$. The rate of this sequence segment is,

$$R_p^{p+1} = \begin{cases} \text{cost}(h_p) & \text{if } D_p^{p+1} \leq D_{max} \\ \infty & otherwise \end{cases}$$

This requires the discontinuity in every segment to be smaller than the maximum allowable amount D_{max}. If the sequence segment has admissible discontinuity, the cost of the segment is represented by the cost of the summary frame.

Using the dynamic programming technique, we define the quantity $N\left(t, h_{p+1}\right)$ as the minimum cost where t is the number of frame selected so far and h_{p+1} is the next frame to select. The recurrence equation is given by

$$N\left(t, h_{p+1}\right) = \min_{h_p \in [h_{p+1} - K_{max}, h_{p+1}]} \left\{ N\left(t-1, h_p\right) + R_p^{p+1} \right\} \quad (8)$$

for all $R_p^{p+1} < \infty$. To obtain the optimal solution, we find the minimum value $\min_t N(t, M)$ because we must include the last frame of the original value. The whole sequence is found by tracing through the intermediate minima back to the first frame. It is easy to see that the time complexity of the algorithm is $O(MK_{max}T_{max})$, where T_{max} is the maximum number of frames that the video summary can have.

IV. Experiments

A. Dataset and Preprocessing

We conduct experiments on two video emotion datasets.

The Emotion6 Video dataset. The Emotion6 dataset [35] contains 1980 images that are labeled with a distribution over 6 basic emotions (anger, surprise, fear, joy, disgust and sadness) and a neutral category. The images do not contain facial expressions or text directly associated with emotions. We consider the emotion category with the highest probability as the dominant emotion.

For our purpose, we create Emotion6 Video, a set of emotional videos using images from Emotion6. We collected an auxiliary image dataset online that has no strong emotion content, which was used as the neutral set. In order to create a video with a particular dominant emotion, we select images from Emotion6 that have the dominant emotion or from the neutral set. This allows us to create ground-truth emotion labels and duration annotations for the emotional segment. We created 3,600 videos in total.

The Ekman-6 dataset.. The Ekman-6 dataset is firstly collected by [2]which contains 6 basic types of emotions: anger, surprise, fear, joy, disgust and sadness; totally 1637 videos and roughly 220 video per class. In this paper, 1496 videos eliminating videos whose sizes are more than 45MiB. are employed for the evaluation. To further assess the tasks of attribution and video-oriented summarization, the dataset is annotated with the most emotion-based segment. Specifically, for each video, three annotators are invited to label no more than 3 key segments which has the most contribution to the overall emotion label of this video. The most longest overlapped segment labels between two annotators are thus saved as ground truth . Finally, we annotated 1-3 emotion related segments of each video. We will release these labels upon the acceptance.

As a preprocessing step, we uniformly sample 30 frames for each video in the Emotion6 Video dataset. Due to the fact that videos in the Ekman-6 dataset are slightly longer, we uniformly sample 100 frames from each video. Black frames are added if the video contains less than 100 frames.

We create two variations for the Ekman-6 dataset. The two-class condition focuses on the two largest emotion categories, anger and surprise. The second condition employs all videos in the dataset.

B. Hyperparameter Settings

We set the initial value of α to $[0.5, 0]$. The models are trained for 200 epochs. Dropout is employed here on all fully-connected layers and the keep ratio is set as 0.75. A softmax layer is employed to calculate loss and Adam [36] is utilized to accelerate learning. Our codes are deployed on tensorflow. For each dataset, the model is repeatedly trained for 5 times in order to reduce the variance; and the averaged performance is reported. The codes are downloadable[1].

The convolutional layer in the classification network has 8 convolutional filters with the size of 5×1. Intuitively, the $fc7$ features extracted by DeepSentiBank can be identified as a generic extractor for emotional related features. Nevertheless, the C-net does not provide a principled way to model the contextual information among the serial frames, which however is extremely significant for the emotion analysis. Thus this layer is utilized to understand these contextual information among each feature channel.

The two fully connected layers has 32 units each. The threshold β in the loss function is set to 0.6.

C. Competing Baselines

Our model is compared against the following baseline models.

Image Transfer Encoding (ITE). Our model is compared against the state-of-the-art method – Image Transfer Encoding (ITE) [2], which using an emotion centric dictionary extracted from auxiliary images to encode videos. The encoding scheme has been shown to have a good performance in emotion recognition, attribution and emotion-oriented summarization.

We replicated the same setting of ITE as described in [2]: we cluster a set of auxiliary images into 2000 clusters using K-means on features extracted from AlexNet [37]. For each frame, we select K clusters whose center are the closest to the frame and the video feature vector is computed as the sum of the individual frames' similarity to the K clusters. Formally, let $\{\boldsymbol{c}_d\}_{d=1,...,2000}$ denote the cluster centers. The representation

[1] https://github.com/kittenish/Frame-Transformer-Network

TABLE I
EMOTION RECOGNITION RESULTS.

Dataset	Chance	SVM	ITE	C-Stream	Unsup. E-Stream	E-stream	BEAC-Net
Emotion6 Video	16.7%	80.0%	77.5%	81.3%	82.2%	99.5%	**99.7**%
Ekman-6 (two classes)	50.0%	62.8%	65.3%	59.5%	68.9%	70.4%	**71.6**%
Ekman-6 (full dataset)	16.7%	42.8%	43.2%	47.1%	41.7%	44.9%	**49.3**%

for the i^{th} video is a 2000-dimensional vector $\boldsymbol{s}_i$, which is computed as summation over all frames:

$$s_{i,d} = \sum_{j=1}^{M} \cos(\phi(x_{i,j}), c_d)\ \mathbb{1}\left(c_d \in \text{KNN}(x_{i,j})\right) \tag{9}$$

where the indicator function equals 1 if and only if the cluster center c_d is among the K nearest clusters of $x_{i,j}$ A linear SVM model is trained to solve the emotion recognition task. The attribution can be solved by selecting a sequence of frames whose similarities to video-level representation are greater than a set threshold while enduring no more than 10 frames out of threshold. And the frames with maximal similarities are adopted as the summarization result.

Support Vector Machine (SVM). The $fc7$ features are used to train a linear SVM classifier on each frame of the video. The final classification labels are extracted as the majority vote. The attribution results are obtained by selecting the longest segment classified as the same emotion and the summarization results are obtained by selecting the frames with the highest emotion scores.

The Content Stream Only (C-Stream). For the task of video emotion classification, we perform an ablation study by removing the attribution network and the associated emotion stream from the classification network. What remains is a single-stream, conventional convolutional neural network. We report the result for emotion classification only, as this network is not capable of emotion attribution.

Supervised Emotion Stream (E-Stream). As a second ablated network, we remove the content stream from the classification network. The attribution network and the associated emotion stream are kept intact. The attribution loss is also kept as part of the loss function.

Unsupervised Emotion Stream (Unsup. E-Stream). This is a third ablated network. Similar to the E-Stream version, we remove the content stream from the classification network. In addition, we also remove the attribution loss from the loss function. The A-Net and the emotion stream are kept intact. That is, we use the emotion stream for classification, but do not supply supervision to the attribution network.

D. Results and Discussion

Emotion recognition. We perform the emotion recognition on the Emotion6 Video dataset and the Ekman-6 dataset, where Ekman-6 has two experimental conditions. Table 1 reports the classification accuracy. The first column of Table 1 reports chance-level performance.

We observe that BEAC-Net achieves the best performance among all baseline models, including all ablated versions. Compared to the previous state-of-the-art method ITE, BEAC-Net improves classification performance by 22.2%, 6.3% and 6.1%, respectively.

The ablation study reveals the complementarity of all constituents of BEAC-Net. The C-Stream convolutional network underperforms BEAC-Net by 18.4%, 12.1%, and 2.2%. The E-Stream with attribution supervision underperforms by 0.2%, 1.2%, and 4.4%. Interestingly, the E-Stream beats the C-Stream on the Emotion6 and two-class Ekman-6, but underperforms on the full Ekman-6 dataset. This results indicate that the two streams indeed complement each other under different conditions and their co-existence is crucial for accurate emotion recognition. Adding supervision to the attribution network improves it by 17.3%, 1.5%, and 3.2%. While the improvements are admittedly significant, the unsupervised E-Stream still achieves the second best result on the first two conditions. This suggests the attribution network is capable of identifying emotional content even when supervision is absent.

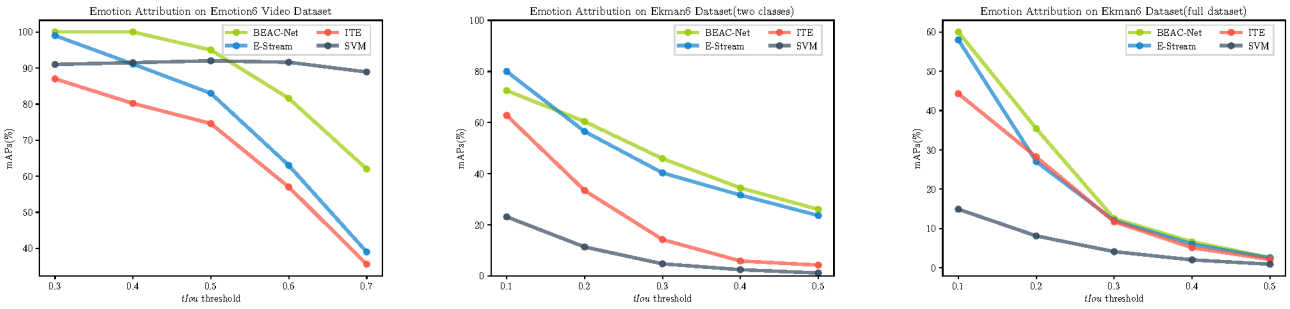
Finally, the three experimental conditions establish an easy-to-hard spectrum. The artificial Emotion6 Video dataset is the simplest, where a simple SVM technique can achieve 80% accuracy. The full Ekman-6 with all 6 emotions is the most difficult. It is worth noting that the effectiveness of the bi-stream architecture is the most obvious on the most difficult full Ekman-6 dataset, leaving a 2.2% gap between BEAC-Net and the second best technique. E-Stream is almost the same as BEAC-Net on the simplest conditions, but the gap widens as the task gets more difficult.

Emotion Attribution. We report the results on emotion attribution. Here the comparison baselines include ITE, SVM and Unsupervised E-Stream. For the SVM, the longest majority-voted video emotion segment is considered to be the extracted emotional segment.

We use mean average precision (mAP) to evaluate the performance of emotion attribution [38]. To calculate the mAP, we have to determine whether one video detection is true positive or not. Specifically,the overlap between the predicted video segment and the ground-truth segment are firstly calculated. This overlap is quantified by the temporal intersection over union ($tIoU$) score of predicted and ground-truth segment. The prediction will be marked as positive if the overlap is greater than a threshold. The three experimental conditions, Emotion6 Video, two-class Ekman-6, and full Ekman-6 are the same as previously. Fig. 2 shows the results, where the horizontal axis indicate different tIoU thresholds.

Once again, we observe strong performance from BEAC-Net, which achieves the best performance in almost all conditions. This validates that the A-Net can help identify the video segments that contribute the most to the overall emotion of one video. The unsupervised E-Stream performs worse than

Fig. 2. Emotion attribution results. We report the mAP scores for each dataset.



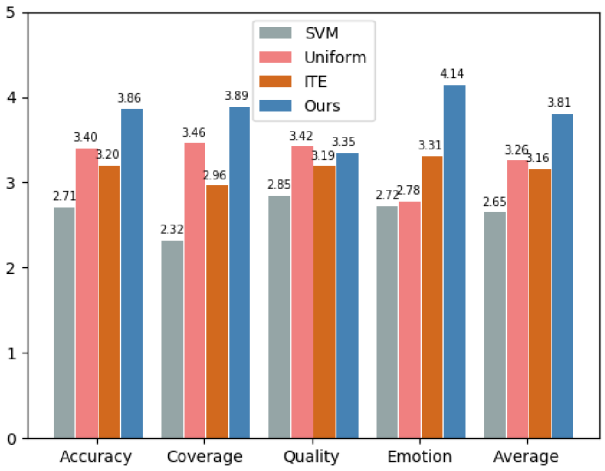


Fig. 3. The qualitative results of emotion-oriented video summarization. "Average" scores are the average of the other four scores.

BEAC-Stream, but remains a close second in the Ekman-6 experiments.

On the Emotion6 video dataset, BEAC-Net outperforms other methods except for the last two tIoU thresholds, where the SVM method has very good and stable performance. We hypothesize that this is because every frame in Emotion6 video dataset which is created from an image dataset has a definite label, while fundamentally the SVM method is based on classifying individual images. On the Ekman-6 dataset, the supervisions have been labeled for video segments instead of individual frames. Thus, not every frame in the emotional segment necessarily express the emotion. This is likely a reason why the frame-based SVM performs poorly in those conditions.

On the two-class Ekman-6 condition, BEAC-Net comfortably beats the rest, except for the very first tIoU setting. On the full Ekman-6 condition, BEAC-Net still outperforms the baselines, but the performance gap is smaller. This is consistent with our observation that the full Ekman-6 is the most difficult dataset.

Emotion-oriented Summarization. To quantitatively evaluate the video summaries, we carry out a user study where ten human participants viewed and rated summaries of twelve videos. To test the methods under different conditions, we compare video summaries containing 3 and 6 frames, respectively. The 12 videos were randomly assigned to the 3-frame and the 6-frame conditions. As the video summarization task is different from the two other tasks, we differed from the previous experiments and created 3 baseline techniques below.

Uniform: uniformly sample the frames/clips from the videos;

SVM: the video is summarized by the scores of SVM prediction. The frames with top-6 or top-3 scores of each label are selected in practice.

ITE: we use the summarization method based on ITE, as described in [2].

We recruited ten volunteers for the user study and they were kept blind to the summarization techniques. We showed summary videos from all methods to the participants and asked them to rated the quality of summary on five-point Likert scale on the following four criteria [2]:

Accuracy: whether the summary accurately describes the main content of original video?

Coverage: how much visual content has been covered in the summary?

Quality: how is the overall subjective quality of summary?

Emotion: how many percentage of the same emotion has been conveyed from original video?

Fig. 3 shows the results from the user study, where the average column reports the average rating across four questions. On four out of the five measures (including the average), our method outperforms all baseline methods. The largest performance gap of 0.83 appears on the emotion criterion, suggesting our summaries covers emotional content substantially better than other methods. On the quality criterion, we perform slightly worse than the uniform method, but the gap is a mere 0.07.

The qualitative results, shown in Figure 4, exemplifies advantages of our comprehensive summarization method: (1) Coverage of sparsely positioned emotional expressions. Figure 4 b) contains an illuminating example. The original video contains an interview of two young lovers recalling their love stories. The vast majority of the video contains an interview with the couple sitting on a couch, as shown in the uniform row; emotional expressions are sparse and widely dispersed. Our model accurately chose multiple memory flashbacks as the summary while other methods give priority to the interview shots. (2) Capturing the main emotion segment. Benefiting

(a)

(b)

(c)

Fig. 4. The qualitative results of emotion-oriented video summarization. We compare several baselines.

from the results of the attribution framework, our summarization method focuses on clips that contain the main emotion of the video. Figure 4 a) shows a video with mainly angry content, and the video summary created by our method shows the fighting scenes. Figure 4 c) shows a video with sadness; our summary not only captures the crying but also the cause of sadness, a photo of a murdered child.

V. Conclusions

Computational understanding of emotions in user-generated video content is a challenging task because of the sparsity of emotional content, the presence of multiple emotions, and the variable quality of user-generated video. We suggest that the ability to locate emotional content in the video plays an important role in accurate understanding.

Toward this end, we present a multi-task neural network with a novel bi-stream architecture, which we call Bi-stream Emotion Attribution-Classification Network (BEAC-Net), which is end-to-end trainable and can solve the emotion attribution and recognition simultaneously. The attribution network can locate the emotional content, which is processed in parallel with the original video in the two streams. Empirical evidence shows that the bi-stream architecture to provide significant benefits for emotion recognition. In addition, we propose a video summarization technique based on the attribution provided by BEAC-Net. The technique outperformed existing baselines in a user study.

Emotions play an important role in the human cognitive system and day-to-day activities. An accurate understanding of human emotions could enable many interesting applications such as story generation based on visual information [39]. We believe this work represents a significant step in improving understanding emotional content in video.

References

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

[2] B. Xu, Y. Fu, Y. gang Jiang, B. Li, and L. Sigal, "Heterogeneous knowledge transfer in video emotion recognition, attribution and summarization," *IEEE TAC*, 2017.

[3] B. Xu, Y. Fu, Y. gang Jiang, B. Li, and L. Sigal, "Video emotion recognition with transferred deep feature encodings," in *ICMR*, 2016.

[4] P. Ekman, "Universals and cultural differences in facial expressions of emotion," *Nebrasak Symposium on Motivation*, vol. 19, pp. 207–284, 1972.

[5] R. Plutchik and H. Kellerman, *Emotion: Theory, research and experience. Vol. 1, Theories of emotion*. Academic Press, 1980.

[6] T. Chen, D. Borth, Darrell, and S.-F. Chang, "DeepSentiBank: Visual sentiment concept classification with deep convolutional neural networks," *CoRR*, 2014.

[7] L. F. Barrett, "Are emotions natural kinds?," *Perspectives on Psychological Science*, vol. 1, no. 1, pp. 28–58, 2006.

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

[9] J. J. Gross, "Emotion regulation: Affective, cognitive, and social consequences," *Psychophysiology*, vol. 39, no. 3, p. 281291, 2002.

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

[11] 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.

[12] H.-L. Wang and L.-F. Cheong, "Affective understanding in film," *IEEE TCSVT*, 2006.

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

[14] A. Yazdani, K. Kappeler, and T. Ebrahimi, "Affective content analysis of music video clips," in *Proc. 1st ACM workshop Music information retrieval with user-centered and multimodal strategies*, 2011.

[15] H. Joho, J. M. Jose, R. Valenti, and N. Sebe, "Exploiting facial expressions for affective video summarisation," in *Proc. ACM conference on Image and Video Retrieval*, 2009.

[16] S. Zhao, H. Yao, X. Sun, P. Xu, X. Liu, and R. Ji, "Video indexing and recommendation based on affective analysis of viewers," in *ACM MM*, 201.

[17] M. Liu, S. Shan, R. Wang, and X. Chen, "Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition," in *CVPR*, 2014.

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

[19] 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.

[20] E. Acar, F. Hopfgartner, and S. Albayrak, "A comprehensive study on mid-level representation and ensemble learning for emotional analysis of video material," *Multimedia Tools and Applications*, pp. 1–29, 2016.

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

[22] S. Wang and Q. Ji, "Video affective content analysis: a survey of state of the art methods," *IEEE TAC*, vol. PP, no. 99, pp. 1–1, 2015.

[23] B. T. Truong and S. Venkatesh, "Video abstraction: A systematic review and classification," *ACM TOMM*, vol. 3, no. 1, pp. 79–82, 2007.

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

[25] J.-L. Lai and Y. Yi, "Key frame extraction based on visual attention model," *Journal of Visual Communication and Image Representation*, vol. 23, no. 1, pp. 114–125, 2012.

[26] 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*, vol. 14, no. 4, pp. 975–985, 2012.

[27] 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.

[28] J. Gao, Y. Fu, Y. gang Jiang, and X. Xue, "Frame-transformer emotion classification network," in *ACM ICMR*, 2017.

[29] M. Jaderberg, K. Simonyan, A. Zisserman, *et al.*, "Spatial transformer networks," in *Advances in Neural Information Processing Systems*, pp. 2017–2025, 2015.

[30] K. K. Singh and Y. J. Lee, "End-to-end localization and ranking for relative attributes," in *European Conference on Computer Vision*, pp. 753–769, Springer, 2016.

[31] A. Show, "Tell: Neural image caption generation with visual attention," *Kelvin Xu et. al.. arXiv Pre-Print*, 2015.

[32] K. K. Singh and Y. J. Lee, "End-to-end localization and ranking for relative attributes," in *ECCV*, 2016.

[33] C.-H. Lin and S. Lucey, "Inverse compositional spatial transformer networks," in *CVPR*, 2017.

[34] Z. Li, G. M. Schuster, and A. K. Katsaggelos, "Minmax optimal video summarization," *IEEE TCSVT*, 2005.

[35] K.-C. Peng, T. Chen, A. Sadovnik, and A. Gallagher, "A mixed bag of emotions: Model, predict, and transfer emotion distributions," in *CVPR*, 2015.

[36] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

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

[38] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles, "Activitynet: A large-scale video benchmark for human activity understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 961–970, 2015.

[39] R. Cardona-Rivera and B. Li, "Plotshot: Generating discourse-constrained stories around photos," in *Proceedings of the 12th AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 2016.

Frame-Transformer Emotion Classification Network

Jiarui Gao[1], Yanwei Fu[2], Yu-Gang Jiang[1], and Xiangyang Xue[12*]
[1]School of Computer Science, Shanghai Key Laboratory of Intelligent Information Processing, Fudan University, China
[2]School of Data Science, Fudan University, China
{jrgao14,yanweifu,ygj,xyxue}@fudan.edu.cn

ABSTRACT

Emotional content is a key ingredient in user-generated videos. However, due to the emotions sparsely expressed in the user-generated video, it is very difficult to analayze emotions in videos. In this paper, we propose a new architecture–Frame-Transformer Emotion Classification Network (FT-EC-net) to solve three highly correlated emotion analysis tasks: emotion recognition, emotion attribution and emotion-oriented summarization. We also contribute a new dataset for emotion attribution task by annotating the ground-truth labels of attribution segments. A comprehensive set of experiments on two datasets demonstrate the effectiveness of our framework.

KEYWORDS

Video emotion recognition; video emotion attribution; video emotion-oriented summarization and spatial-transformer network

ACM Reference format:
Jiarui Gao[1], Yanwei Fu[2], Yu-Gang Jiang[1], and Xiangyang Xue[12]. 2017. Frame-Transformer Emotion Classification Network. In *Proceedings of ICMR '17, June 6–9, 2017, Bucharest, Romania,* , 7 pages.
DOI: http://dx.doi.org/10.1145/3078971.3079030

1 INTRODUCTION

The explosive growth of user-generated videos creates a great demand for computational understanding of visual media data. Great efforts and success have been made on the video content understanding, such as video actions and activities. Sentimental analysis on text data [10] and image data [4] have been studied recently. However, the ability to understand emotions from videos, to a large extent, remains an unaddressed problem, despite the fact that video content can convey strong emotional information to their viewers. Computational understanding of the emotions aroused by video content nevertheless has many real-world applications. For example, video recommendation services can benefit from matching users' interests with the emotions of video content.

The challenges of video emotion understanding come from three aspects. Firstly, the emotions are often sparsely expressed by a subset of the video, while the other parts of the video perform as the story context for video emotion. Secondly, there are several different types of emotions in one video, despite a single dominant emotion exists. It is essential to know, and yet difficult to answer which video segment contributes the most to the video's overall emotion, which is defined as video emotion attribution [24]. Thirdly, the user-generated videos are often captured in an uncontrolled environment and of high diverse content. The unconstrained space of objects, scenes, and events in user-generated videos, not only makes their content very complex to be analyzed; but also gets the user-generated videos more likely to suffer from the problems of occlusions of objects and illumination conditions than the commercial videos (e.g. movies, news and sports).

Previous efforts of understanding video emotion aim at solving the three tasks of emotion recognition, emotion attribution and emotion-oriented summarization, by either combining various of low-level and middle-level features [7]; or by using auxiliary image sentiment dataset to re-encode the features of video frames [23, 24]. These works still have to design a specific algorithm for each task, rather than a single framework solving all these three highly correlated tasks jointly.

In this paper, we propose our new architecture – Frame-Transformer Emotion Classification Network (FT-EC-net), which facilitates solving emotion recognition, emotion attribution and emotion-oriented summarization jointly. FT-EC-net is composed of FT-net and EC-net. Particularly, the FT-net is a variant of spatial transform networks (ST-net) [6]. It can learn to detect the emotional video segments; and thus facilitates both emotion attribution and emotion-oriented summarization. The EC-net is a classification network which further processes the results of FT-net for emotion recognition. Our architecture is thoroughly evaluated on Ekman6 and Emotion6 video dataset.

Contributions: (1) Our newly proposed FT-EC-net can solve the emotion recognition, emotion attribution and emotion-oriented summarization simultaneously. As a variant of ST-net, FT-net is firstly introduced in this paper to enable video emotion attribution and emotion-oriented summarization; (2) In establishing a good benchmark for emotion attribution task, we re-annotate the Ekman6 dataset with the most emotion-oriented segments which can be used as the ground-truth for emotion attribution task. (3) We also introduce a new evaluation metric to evaluate the video segments detected in attribution tasks.

2 RELATED WORK

Image and Video Emotion Recognition. Recently, inspired by the psychological theory, such as Ekman's six pan-cultural basic emotions and Plutchik's wheel of emotions, researchers have studied the problem of image emotion recognition extensively. Various features have been explored, such as the features inspired by psychology and art theory[15], and the shape features [13].

*Yanwei Fu is the corresponding authour.

ICMR '17, June 6–9, 2017, Bucharest, Romania
 ACM ISBN 978-1-4503-4701-3/17/06…$15.00
DOI: http://dx.doi.org/10.1145/3078971.3079030

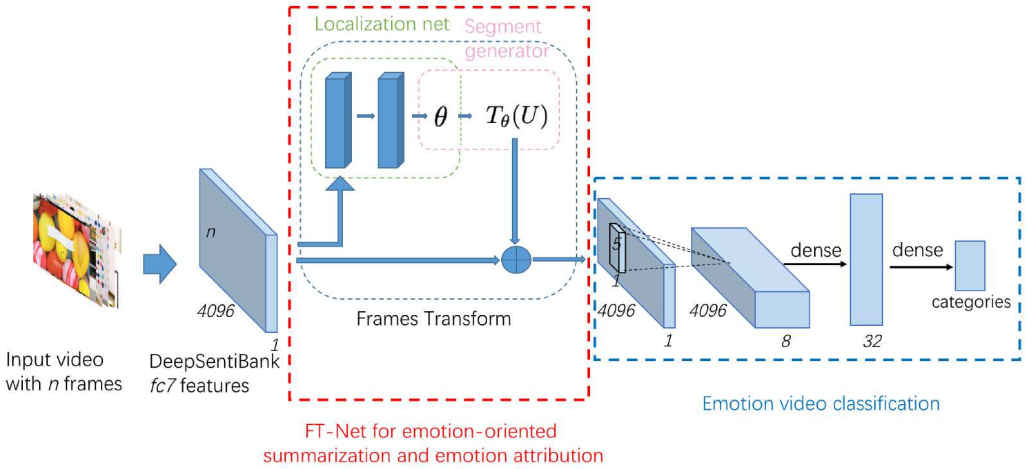


Figure 1: An overview of our framework. We use the DeepSentiBank to extract the features of each frame. With the extracted deep features, the localization network is thus regressed towards the parameters θ. The FT-EC-Net uses the extracted $fc7$ features as inputs, and is capable of emotion recognition (right part), emotion-oriented summarization and emotion attribution (middle part).

Recently with the renaissance of convolution neural networks, deep convolutional neural networks have been used for visual sentiment analysis [2, 25]. A large scale of visual sentiment dataset was proposed in Sentibank [2] and DeepSentiBank [4]. It contains a set of 1,533 adjective-noun pairs, such as "cute dog" and "happy wedding".

Video emotion recognition has been investigated recently and great progress has been made on extracting different types of features, such as the low-level visual and audio features, attribute features in [7] as well as the mid-level audio-visual features in [1]. Emotions have also been analyzed in GIF files [8] which can be taken as one type of short videos. The facial expressions have also been investigated on videos [5]. For a more recent survey, we refer to [21].

Most these existing works still formulate emotion understanding as a classification task. In our previous work [23, 24], we propose another two tasks – video emotion attribution and video emotion-oriented video summarization. Thus different from all previous works, we propose a new architecture that is able to solve these three tasks jointly.

Emotion attribution and emotion-oriented summarization. Video summarization has been explored for more than two decades [19] and the detailed review is beyond our scope. In general, the video summary has two forms, i.e., keyframes extraction and video skims; and to generate video summary, a set of features have been exploited, such as visual saliency[14], motion cues [12], mid-level features [20], and semantic recognition [22].

Recently, we [24] introduced the tasks of emotion-oriented summarization which aims at extracting video summarization according to more general video emotion content. Inspired by the task of semantic attribution in text analysis, [24] also defined the emotion attribution as attributing the video's overall emotion to its individual segments. However, [24] still treated the emotion recognition, summarization and attribution tasks as several separate tasks. Intrinsically, the emotion recognition can greatly help emotion attribution and emotion-oriented summarization; and the emotion-oriented summary can be selected from the results of emotion attribution. Thus, our framework is designed to solve these three tasks simultaneously.

Spatial transform networks. Spatial transform networks (ST-net) are firstly proposed in [6] for image classification. ST-Net provides the spatial transformation capabilities[6], which enables a wide variety of tasks such as co-localization [18], and spatial attention [17]. Our FT-net component is a variant of ST-net and it enables selectively learning to segment the emotional video segments.

3 FRAME-TRANSFORMER EMOTION CLASSIFICATION NETWORK

As illustrated in Fig. 1, this section introduces our framework which is composed of three parts: DeepSentibank, Frame-Transformer subnetwork (FT-net) and Emotion Classification subnetwork (EC-net).

Suppose we have n frames extracted from each video. The DeepSentiBank [4] is utilized to extract the features of each frame. It contains five convolutional layers and three fully-connected layers. Specifically, we use as features the $4096-dim$ output of $fc7$ layer of DeepSentiBank. The DeepSentiBank is trained on 2089 Adjective Noun Pairs(ANPs) (such as "sad eyes") with 867919 images. In practice, a particular number of frames are equally sampled from each video to formulate the input of the network.

The FT-net aims at learning to detect emotional video segments. It has the localization sub-network regressing the transformer parameter θ and normalized segment generator of partitioning the video segments from the data stream generated by DeepSentiBank.

The segments generated by FT-net can be used for emotion attribution and emotion-oriented summarization.The EC-net is constructed for video emotion recognition, with one convolutional layer and two fully connected layers. We will explain each part in the following subsections.

3.1 FT-net

The FT-Net can be further processed into localization sub-network and normalized segment generator.

Localization sub-Network. It has two fully connected layers to further project the extracted features of each frame into a non-linear representation. The localization network in Fig. 1 aims at regressing the sampling parameter $\theta = [\theta_1, \theta_2]$ for normalized segment generator. The parameter θ is calculated through backpropagation from the EC-net; thus no supervision is provided to the localization network.

Normalized Segment Generator. To represent the specific output frames, we set x_i^t as the target coordinate of each frame in the regular output segment and x_i^s as the source input coordinate of each frame along the input frames n. The output segment S is formed by L frames, *i.e.*, $S = \left[x_i^t\right]_{i=1}^{L}$. The segment generator enables projecting the emotional frames (with coordinate x_i^t, $i = 1, \cdots, L$) to the corresponding source frames x_i^s by the 2D affine transformation,

$$x_i^s = \begin{bmatrix}\theta_1 & \theta_2\end{bmatrix} \begin{pmatrix} x_i^t \\ 1 \end{pmatrix} \tag{1}$$

where θ_1 is the scale parameter which controls the length ratio in the input video projected from output segments; and θ_2 is the translation parameter which indicates the offset of the segment over the video. The reason why the computation above is backward (from x_i^t to x_i^s) is that it enables the length of the source segment from the original video to be unfixed. But this network can only detect one main segment due to its structure. In practice, to avoid the numerical issues, the coordinates x_i^t and x_i^s are normalized to $(-1, 1)$ $(-1 \leq x_i^t, x_i^s \leq 1)$.

This parameter θ is computed through the localization subnetwork according to the same input features as it is applied, so the output segment S can be taken as the attribution results of the original video. We can further generate the summary by the attribution results.

3.2 EC-net

The output of normalized segment generator is further processed by classification sub-network.

One Convolutional layer has 8 convolutional filters with the size of 5×1. Intuitively, the $fc7$ features extracted by DeepSentiBank can be taken as a generic extractor for sentimental related features. Nevertheless, the FT-net does not have a principle way to model the contextual information among the consecutive frames, which however is extremely important for our task of emotion analysis. Thus this layer is utilized to compute contextual information among each feature channel.

Two fully connected layers have 32 neurons individually. The final softmax layer is followed to classify the emotions. Due to the difficulty of emotion analysis task, the fully connected layers are added to increase the non-linearity of the classification network.

Note that the convolutional layer follows the fully connected layers in DeepSentiBank and FT-net, which however do not have negative effects on learning the parameters of convolutional layers. (1) The DeepSentiBank is pre-trained beforehand and its parameters are fixed when we extract the $fc7$ features of frames. (2) The FT-net has two fully connected layers to regress the parameter θ in its localization subnetwork. Once the θ is learned to segment the input features, our convolutional layer still processes the data extracted by DeepSentiBank as shown in Fig. 1.

3.3 Our tasks

A very nice property of our network is that it can enable the emotion recognition, emotion attribution and emotion-oriented summarization simultaneously. Specifically,

Emotion recognition. The final softmax layer of EC-net is to identify the emotion. It thus outputs the likelihood of one video belonging to which type of emotion.

Emotion attribution. The video segments of high emotion scores, i.e. emotion attribution, can be directly computed by using the θ values computed by FT-net. With the well trained FT-EC-net and θ computed, the emotion attribution segment is computed as follows. Suppose the length of candidate video is l, the selected segment starts at t_s and ends at t_e;

$$t_s = \frac{l}{2} \cdot (\theta_2 - \theta_1 + 1) \tag{2}$$

$$t_e = \frac{l}{2} \cdot (\theta_1 + \theta_2 + 1) \tag{3}$$

The Eq (2) and Eq (3) are derived from Eq (1).

Emotion-oriented summarization. By utilizing the results of video attribution, the summary of video skims can be directly generated from the selected segments. We also uniformly sample the frames/video clips from selected segments as the keyframes summary/video skims.

4 EXPERIMENTS

4.1 Dataset

We conduct experiments on two video emotion datasets.

Emotion6 video dataset. Inspired by [6], we augment the Emotion6 dataset [16] and create Emotion6 video dataset as the testbed for our tasks. Emotion6 dataset consists of 6 basic emotions. Each image has been labelled with a distribution of all these emotions as well as one domaint emotion. To construct the dataset, we randomly crawled an auxiliary image dataset online which has no strong evoked emotions and uses as the noise set. The frames of each video are selected either from Emotion6 dataset with the dominant evoked emotion, or from the auxiliary image set. The corresponding emotion labels from Emotion6 dataset.

Ekman6 dataset. Ekman6 dataset is firstly collected by [24]. It contains 6 different types of emotions; totally 1637 videos with around 220 videos per class. To further evaluate the tasks of emotion-oriented video summarization and attribution, all the videos are annotated with the most emotion-oriented segments. Specifically,

for each video, three annotators are invited to label the key segments which contribute the most to the overall emotion label of this video. The overlapped segment labels beween two annotators are thus saved as ground-truth segments. Finally, we obtain $1-3$ emotion related segments of each video. In this paper, we employ as the evaluation the anger and surprise classes, which have most number of video instances. We will release these annotations upon the acceptance.

Features. DeepSentibank [4] is utilized to extract the features of the $fc7$ layer of each frame. DeepSentibank is initialized with the weights trained from ImageNet and fine-tuned on the Sentibank dataset.

4.2 Competitors

Our model is compared against the following models.

SVM by Majority Voting (SVM) . The $fc7$ features of Emotion6 training set are used to train a linear SVM classifier. The emotion scores can be thus predicted on each individual frame of testing videos. The final classification labels are voted majoritively; and the attribution/summarization results are obtained by selecting the segments with the highest emotion scores.

Image Transfer Encoding (ITE). Our model is compared with the state-of-the-art method – ITE [24]. Particularly, we use the same setting of ITE as [24]: we cluster 2000 centers from the emotion-rich auxiliary image dataset; and by using these 2000 centers as bins, we encode the frames of one video into video-level representation; a linear SVM model is trained to solve the emotion recognition task. The attribution/summarization can also be solved by selecting the frames that have the smallest distances to video-level representation.

Convolutional neural networks (CNN). In video emotion classification, our model is compared with convolutional neural networks. We remove the FT-net from our framework and get a pure CNN structure capable for emotion recognition task.

4.3 Settings and Evaluation

Emotion recognition. The predicted labels are compared against the ground truth labels to calculate the classification accuracy.

Emotion Attribution. We employ the evaluation metrics – mean Average Precision (mAP) for video detection [3].

Emotion-oriented summarization. User study is employed to quantitatively compare our results against the competitors.

Parameter Settings. We empirically set the initial value of ; We empirically set the bias of the second fully connected layer in localization network, which is also the value of $\theta = [0.2, 0]$, to be $[0.2, 0]$. The models are trained for 1000 epochs with the batch size of 40 on Emotion6 video dataset, and the batch size of 20 on Ekman6 dataset. Dropout is employed here on all fully connected layers and the ratio is set as 0.75. A softmax layer is utilized to calculate loss and Adam[9] is used to accelerate learning. Our codes are implemented on tensorflow. For each dataset, the model is repeatedly trained from 5 times to reduce the variance; and the averaged performance is reported. The codes are downloadable from https://github.com/kittenish/Frame-Transformer-Network.

Table 1: Emotion recognition results.

Dataset	Chance	SVM	ITE	CNN	Ours
Emotion6 video	16.7%	80.0%	77.5%	81.3%	**82.2%**
Ekman6	50.0%	60.1%	67.0%	74.0%	**74.4%**

Preprocessing of dataset. On Emotion6 video dataset, we uniformly sample 30 frames for each video. Due to the fact that videos in Ekman6 dataset are slightly longer, 100 frames are uniformly sampled from each video. Zero-value frames are added if the total frames of one video is less than 100 .

4.4 Results on Emotion Recognition

We perform the emotion recognition on two datasets. The results are reported in Tab. 1. We compare with SVM, ITE, and CNN methods. The experimental results show that our framework outperforms all the other baselines on both datasets. This validates the effectiveness of our methods.

We also observe that, (1) comparing with CNN methods, our model has 0.9 and 0.4 absolute percentage improvement thanks to the FT-net which can attribute the emotional segments. (2) The improvement of our model over ITE and SVM shows that by integrating the attribution results, our EC-net can get better emotion classification results. (3) The worst performance comes from SVM. This might be due to the fact that emotion is intrinsically sparsely expressed in the video; and thus the trained linear SVM classifier is not reliable to predict confident emotion scores of frames.

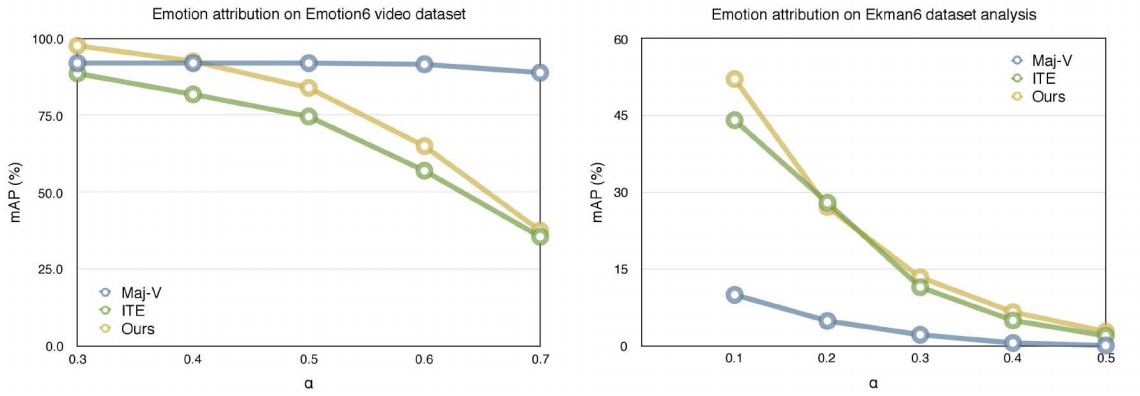
4.5 Results on Emotion Attribution

We conduct the experiments on emotion attribution. Specifically we compare with the methods of ITE and SVM. For SVM method, the emotion segments can be sampled from the longest majority-voted video emotion segments.

The mAP score is introduced here to evaluate the performance of emotion attribution[3]. To compare the mAP, we have to decide whether one video detection is true positive or not. In particular, we firstly calculate the overlap between the prediction of temporal video segment and the ground-truth segment. This overlap is measured by the temporal intersection over union ($tIou$) score of predicted and ground-truth segment. The prediction is marked as positive if the overlapping is greater than a threshold α. The mAP score can thus be computed.

We compare the emotion attribution results on the two classes of Ekman6 and Emotion6 video dataset in Fig. 2. We vary the threshold α on these two datasets. On Ekman6 dataset, our results are comparable or better than those of ITE method. This validates that our FT-net network can actually help identify the video segments that contribute the most to the overall emotion of one video. Furthermore, note that the good results of FT-net also benefit greatly from the recognition task of EC-net, since FT-net is basically learning to sample the video segments by the classification results. On the Emotion6 video dataset, our results still consistently outperform those of ITE method. On this dataset, we also notice that the SVM method has very good and stable performance; this is because our SVM classifier is exclusively trained on Emotion6 images. However,

Figure 2: Emotion attribution results. We report the mAP scores for each dataset.



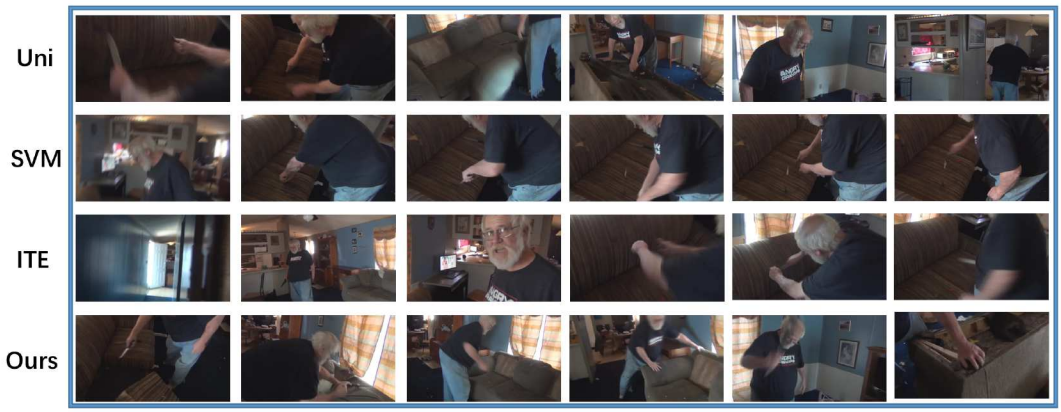


Figure 3: The qualitative results of emotion-oriented video summarization. We compare several baselines.

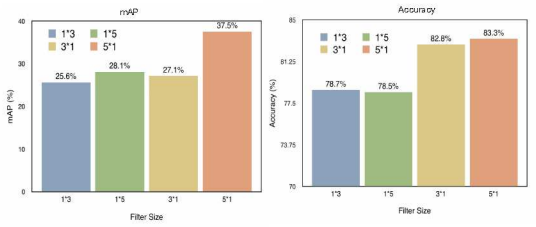


Figure 4: The classification accuracy and mAP scores of ablation study on the filter size of convolutional layer. We use $\alpha = 0.7$ to compute the mAP score.

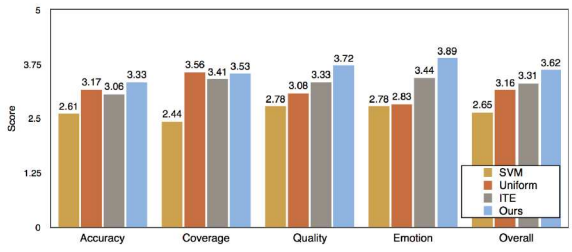


Figure 5: The qualitative results of emotion-oriented video summarization. "Overall" scores are the averaged sum of all the other four scores.

when we empirically validate training a SVM predictor which is weakly supervised on Ekman6 dataset, the performance is very poor.

4.6 Video Emotion-oriented Summarization

The emotion-oriented video summarization is evaluated finally. Several different baselines are compared against our results: (1) Uniform (Uni): uniformly sample the frames/clips from the videos; (2) SVM: the video is summarized by the scores of SVM prediction. The frames with top-6 scores of each label are selected in practice. (3) ITE: we use the method based on ITE [24] for video summarization. To make the results more comparable, the length of summary is fixed to the same length.

User study is utilized to evaluate the results of summarization. Five students kept unknown from our project participate in the

study. We show the results of all methods to each participant who rate the quality of summary on five-point scale by answering the following four questions [24]: (1) Accuracy: whether the summary accurately describes the main content of original video?(2) Coverage: how much visual content has been covered in the summary?(3) Quality: how is the overall subjective quality of summary? (4) Emotion: how many percentage of the same emotion has been conveyed from original video?

The user-study results are shown in Fig. 5. Our framework achieves good performance on all the scores. This shows the good summary results that our model can generate. Figure 3 gives the qualitative illustration of the summary generated by different methods. The example is an "angry" video, which capture the process of an angry man smashing everything in the room. The Uniform and SVM methods are not good enough to grasp the video summary of angry emotion; and it makes the viewers think that the man is finding something. In contrast, both ITE and ours can help identify emotional segments. Especially, our summary has the shot of the man smashing the sofa with a hammer.

4.7 Ablation Study of the Convolutional Layer

The convolutional layer in EC-net performs the role of connecting FT-net and EC-net in our architecture. This layer processes the output of FT-net for the classification in EC-net. Previous work [11] used symmetric convolutional filters. In contrast, with the $fc7$ input features, our convolutional layer aims at modeling the contextual information within consecutive frames of each feature dimension. Considering the importance of the convolutional layer, we conduct the ablation study to verify different filter sizes of this layer. We compare the tasks of emotion recognition and emotion attribution with the filter sizes of 1×3, 1×5, 3×1, and 5×1 respectively on Emotion6 video dataset.

The results are reported in Fig.4. The results with the filter size of 5×1 can beat all the other choices on both classification accuracy and mAP score. This suggests that our choice of 5×1 filter size can successfully capture the contextual emotion information which can finally help both emotion classification and attribution tasks. In general, the convolutional filter sizes of 3×1 and 5×1 capture the information on the same feature channel (totally 4096 channels) along consecutive frames. In contrast, the filter sizes of 1×3 and 1×5 extract information from the same frame of neighboring feature channels, which however have limited contextual information usable for EC-net. That explains the low results of emotion recognition accuracy of filter sizes 1×3 and 1×5.

5 CONCLUSION

In this paper, we present a new architecture– Frame-Transformer Emotion Classification Network (FT-EC-net), which can solve the emotion recognition, emotion attribution and emotion-oriented summarization tasks simultaneously. A new labelled emotion attribution dataset is contributed. We also introduce an objective evaluation metric of emotion attribution task. The experiments on Emotion6 video and Ekman6 dataset have validated the effectiveness of our framework.

6 ACKNOWLEDGEMENT

This work was supported in part by three NSFC projects (#U1611461, #U1509206, #61572134) and a grant from STCSM, Shanghai, China (#16JC1420401), and Shanghai Sailing Program (17YF1427500).

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] Damian Borth, Rongrong Ji, Tao Chen, Thomas M. Breuel, and Shih-Fu Chang. 2013. Large-scale Visual Sentiment Ontology and Detectors using Adjective Noun Pairs. In *ACM MM*.
[3] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 961–970.
[4] Tao Chen, Damian Borth, Darrell, and Shih-Fu Chang. 2014. DeepSentiBank: Visual Sentiment Concept Classification with Deep Convolutional Neural Networks. *CoRR* (2014).
[5] Abhinav Dhall, Roland Goecke, Jyoti Joshi, Jesse Hoey, and Tom Gedeon. 2016. Emotiw 2016: Video and group-level emotion recognition challenges. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 427–432.
[6] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and others. 2015. Spatial transformer networks. In *Advances in Neural Information Processing Systems*. 2017–2025.
[7] Yu-Gang Jiang, Baohan Xu, and Xiangyang Xue. 2014. Predicting Emotions in User-Generated Videos. In *AAAI*.
[8] Brendan Jou, Subhabrata Bhattacharya, and Shih-Fu Chang. 2014. Predicting Viewer Perceived Emotions in Animated GIFs. In *ACM MM*.
[9] Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
[10] Dimitrios Kotzias, Misha Denil, Phil Blunsom, and Nando de Freitas. 2014. Deep Multi-Instance Transfer Learning. *CoRR* (2014).
[11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *NIPS*.
[12] Jie-Ling Lai and Yang Yi. 2012. Key frame extraction based on visual attention model. *Journal of Visual Communication and Image Representation* 23, 1 (2012), 114–125.
[13] X. Lu, P. Suryanarayan, R. B. Adams, J. Li, M. G. Newman, and J. Z Wang. 2012. On shape and the computability of emotions. In *ACM MM*.
[14] Yu-Fei Ma, Lie Lu, Hong-Jiang Zhang, and Mingjing Li. 2002. A User Attention Model for Video Summarization. In *ACM MM*.
[15] J. Machajdik and A. Hanbury. 2010. Affective image classication using features inspired by psychology and art theory. In *ACM MM*.
[16] Kuan-Chuan Peng, Tsuhan Chen, Amir Sadovnik, and Andrew Gallagher. 2015. A Mixed Bag of Emotions: Model, Predict, and Transfer Emotion Distributions. In *CVPR*.
[17] Attend Show. 2015. Tell: Neural Image Caption Generation with Visual Attention. *Kelvin Xu et. al.. arXiv Pre-Print* (2015).
[18] Krishna Kumar Singh and Yong Jae Lee. 2016. End-to-end localization and ranking for relative attributes. In *European Conference on Computer Vision*. Springer, 753–769.
[19] Ba Tu Truong and Svetha Venkatesh. 2007. Video Abstraction: A Systematic Review and Classification. *ACM TOMM* 3, 1 (2007), 79–82.
[20] Meng Wang, R. Hong, Guangda Li, Zheng-Jun Zha, Shuicheng Yan, and Tat-Seng Chua. 2012. Event Driven Web Video Summarization by Tag Localization and Key-Shot Identification. *IEEE TMM* 14, 4 (2012), 975–985.
[21] S. Wang and Q. Ji. 2015. Video affective content analysis: a survey of state of the art methods. *IEEE TAC* PP, 99 (2015), 1–1. DOI:https://doi.org/10.1109/TAFFC.2015.2432791
[22] Xi Wang, Yu-Gang Jiang, Zhenhua Chai, Zichen Gu, Xinyu Du, and Dong Wang. 2014. Real-time summarization of user-generated videos based on semantic recognition. In *ACM MM*.
[23] Baohan Xu, Yanwei Fu, Yu gang Jiang, Boyang Li, and Leonid Sigal. 2016. Video Emotion Recognition with Transferred Deep Feature Encodings. In *ICMR*.
[24] Baohan Xu, Yanwei Fu, Yu gang Jiang, Boyang Li, and Leonid Sigal. 2017. Heterogeneous Knowledge Transfer in Video Emotion Recognition, Attribution and Summarization. *IEEE TAC* (2017).
[25] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. 2015. Robust Image Sentiment Analysis Using Progressively Trained and Domain Transferred Deep Networks. In *AAAI*.