HUMAN-CENTERED EMOTION RECOGNITION IN ANIMATED GIFS

Zhengyuan Yang, Yixuan Zhang, Jiebo Luo

Department of Computer Science, University of Rochester, Rochester NY 14627, USA
{zyang39, jluo}@cs.rochester.edu, yzh215@ur.rochester.edu

ABSTRACT

As an intuitive way of expression emotion, the animated Graphical Interchange Format (GIF) images have been widely used on social media. Most previous studies on automated GIF emotion recognition fail to effectively utilize GIF's unique properties, and this potentially limits the recognition performance. In this study, we demonstrate the importance of human related information in GIFs and conduct human-centered GIF emotion recognition with a proposed Keypoint Attended Visual Attention Network (KAVAN). The framework consists of a facial attention module and a hierarchical segment temporal module. The facial attention module exploits the strong relationship between GIF contents and human characters, and extracts frame-level visual feature with a focus on human faces. The Hierarchical Segment LSTM (HS-LSTM) module is then proposed to better learn global GIF representations. Our proposed framework outperforms the state-of-the-art on the MIT GIFGIF dataset. Furthermore, the facial attention module provides reliable facial region mask predictions, which improves the model's interpretability.

Index Terms— Emotion Recognition, Affective Computing, Image Sequence Analysis, Visual Attention

1. INTRODUCTION

The animated Graphical Interchange Format (GIF) images have been widely used on social media for online chatting and emotion expression [1, 2]. The GIFs are short image sequences and are more light weighted compared to videos. Because of this, it can be used on social media with a lower time lag and required bandwidth. On the other hand, GIFs have a better ability to express emotions compared to still images because of the contained temporal information. By analyzing over 3.9 million posts on Tumblr, Bakhshi *et al.* [1] show that GIFs are significantly more engaging than other online media types. Because of GIF's popularity, many previous studies explire automated GIF emotion recognition. Most studies [3, 4] extract visual representations for emotion recognition with pre-defined features or convolutional neural networks. Although previous approaches provide feasible solutions for GIF emotion recognition, they process GIFs as general videos and fail to utilize GIF's unique properties. We show this potentially limit the recognition performance and propose the human-centered GIF emotion recognition.

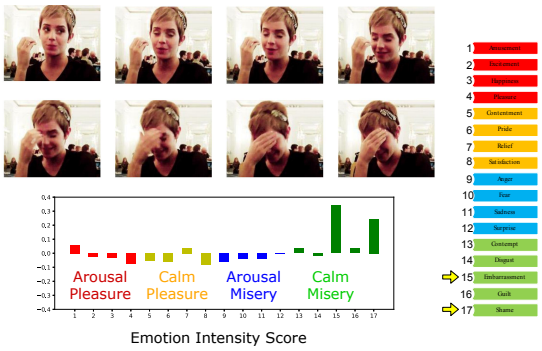

Fig. 1. We formulate GIF emotion recognition both as a classification task for coarse category prediction, and a regression task for emotion intensity score estimation. As shown in the bar chart, the four coarse GIF categories are represented by four different bar colors, and each bar shows the intensity score for one of the 17 annotated emotions.

Human and human-like characters play an importance role in GIFs. A sampling on a GIF search engine GIPHY[1] shows that a majority of GIFs contain clear human or cartoon faces. A previous study [5] also reveals the importance of human faces in expressing emotions. Motivated by this, we explore human-centered GIF emotion recognition and improve recognition performance by focusing on informative facial regions. To be specific, we design a side task of facial region prediction in the proposed facial attention module, where estimated facial keypoints are used to represent human information and are fused with frame-level visual features.

Combining human keypoints with appearance features has shown its effectiveness in related video analysis tasks [6, 7]. A majority of methods merge keypoints as an extra input modality, and thus require keypoints to be complete and accurate. However, the quality of keypoints often can not be guaranteed, especially when keypoints are machine estimated instead of manually labeled. In the facial attention module, we propose to take estimated facial keypoints as the supervision for a facial region prediction side task, and use predicted regions as attention weights to further refine extracted frame-

[1]https://giphy.com/

level visual features. As discussed in Section 3.1, the soft attention fusion is naturally robust against keypoint incompleteness. We further include the keypoint estimation confidence scores in the heatmap generation stage, and make KAVAN robust with respect to inaccurate keypoints. In short, the facial regions predicted by the side task refine the visual features by assigning higher weights to informative facial regions. Furthermore, the predicted facial regions improve the method's interpretability by reliably localizing facial regions.

Another unique property for GIFs is its temporal conciseness. Unlike videos that contain a portion of 'background frames' to better depict a complete story, GIFs are more compact and contain few 'redundant frames'. For example, emotions 'embarrassment' and 'shame' can only be correctly interpreted when jointly looking at all frames presented in Fig. 1. To better capture the temporal information from different segments of a GIF, we propose a Hierarchical Segment LSTM (HS-LSTM) structure as KAVAN's temporal module. GIFs are first evenly split into several temporal segments. The coarse local segment representation is then captured by HS-LSTM nodes. Finally a global GIF representation is learned with segment features from coarse- to fine-grained.

In this study, we propose the Keypoint Attended Visual Attention Network (KAVAN), which improves GIF emotion recognition performance by effectively utilizing GIF's unique properties. In the facial attention module, we utilize human information by merging estimated facial keypoints. Furthermore, we show that replacing the traditional LSTM layers in KAVAN with the proposed HS-LSTM structure can help better modeling temporal evolution in GIFs and further improve the recognition accuracy. Extensive experiments on the GIFGIF dataset prove the effectiveness of our methods.

2. RELATED WORK

GIF Analysis. Bakhshi *et al.* [1] show that animated GIFs are more engaging than other social media types by studying over 3.9 million posts on Tumblr. Gygli *et al.* [8] propose to automatically generate animated GIFs from videos with 100K user-generated GIFs and the corresponding video sources. The MIT's GIFGIF platform is frequently used for GIF emotion recognition studies. Jou *et al.* [3] recognize GIF emotions using color histograms, facial expressions, image based aesthetics and visual sentiment. Chen *et al.* [4] adopt 3D ConvNets to further improve the performance. The GIFGIF+ dataset [2] is a larger GIF emotion recognition dataset. At the time of this study, GIFGIF+ is not released.

Emotion Recognition. Emotion recognition [9, 10] has been an interesting topic for decades. On a large scale dataset [11], Rao *et al.* [12] propose a multi-level deep representations for emotion recognition. Multi-modal feature fusion [13] is also proved to be effective. Instead of modeling emotion recognition as a classification task [12, 11], Zhao *et al.* [13] propose to learn emotion distributions instead, which alleviates

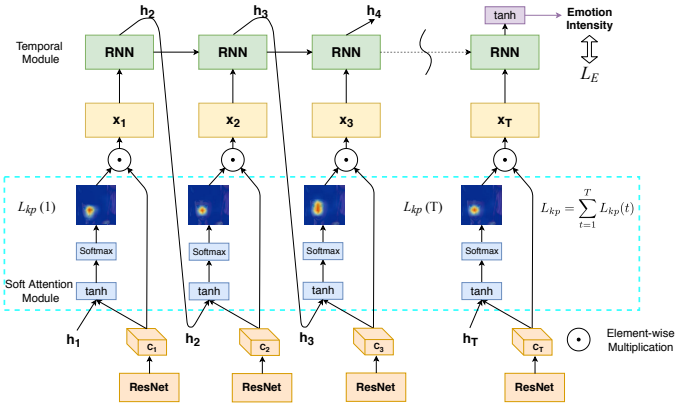

Fig. 2. The structure of the Keypoint Attended Visual Attention Network (KAVAN). Human centered visual feature X_i is first obtained with the soft attention module draw in blue. The RNN temporal module consists of either a single LSTM layer or the proposed HS-LSTM then learns a global GIF representation for emotion recognition.

the perception uncertainty problem that different people under different context may perceive different emotions from the same content. Regressing emotion intensity scores [3] is another effective approach. Han *et al.* [14] propose a soft prediction framework for the perception uncertainty problem.

3. METHODOLOGY

In this section, we introduce the proposed Keypoint Attended Visual Attention Network (KAVAN), which consists of a facial soft attention module and a temporal module. For clarity, the soft attention module is first introduced with a traditional LSTM temporal module in Section 3.1. We then introduce the novel temporal module in KAVAN, namely the Hierarchical Segment LSTM (HS-LSTM) in Section 3.2. Finally, we discuss the training objective and the refined problem setting for GIF emotion recognition.

3.1. Keypoint Attended Visual Attention Network

One unique property for GIFs is the frequent appearance of human and cartoon faces. More than 50% of the GIFs in the MIT GIFGIF dataset contain human faces. Moreover, many in the remaining portion contain cartoon or personated animal characters that also have abundant facial expressions. Previous studies [1, 15] also show a strong relationship between faces and the engagement level of social media contents. Motivated the importance of human faces in GIFs, we explore human-centered GIF emotion recognition.

We represent human information as estimated facial keypoints, and propose a facial soft attention module in the Keypoint Attended Visual Attention Network (KAVAN) to utilize the information by fusing keypoints with extracted frame-level visual features. A number of video action recognition

studies [6, 7] have explored the fusion of keypoints and appearance features. However, previous studies require manually labeled accurate keypoints and might collapse with noisy estimated keypoints. The major challenge is that estimated keypoints can be *inaccurate* and *incomplete*, i.e. certain estimates could be wrong or missing because of occlusions or algorithm failures. In order to solve this challenge, the soft attention module in KAVAN is proposed to fuse the two modalities with attention mechanism. We first introduce the side task of facial region prediction. The predicted facial masks are then processed as attention masks to refine visual features. The soft attention module helps focusing on informative facial regions and thus contributes to GIF emotion recognition.

The proposed KAVAN structure is shown in Fig. 2. Following Temporal Segments Network (TSN) [16], GIFs are first evenly split into T segments and one frame is randomly sampled from each segments as network inputs. At each timestamp t, a visual feature block C_t^{H*W*D} is extracted with the backbone network [17], and a facial region mask α_t is predicted. Extracted visual features are then refined by facial region mask α_t and fed into a temporal module for GIF emotion recognition. The temporal module can be as simple as a single LSTM layer, or other more effective structures as introduced in Section 3.2. For clarity, we first introduce the base KAVAN structure with a single LSTM layer:

$$\begin{pmatrix} i_t \\ f_t \\ o_t \\ g_t \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{pmatrix} T_{d+D,4d} \begin{pmatrix} h_{t-1} \\ x_t \end{pmatrix} \tag{1}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \tag{2}$$

$$h_t = o_t \odot \tanh(c_t) \tag{3}$$

$$x_t = \sum_{k=1}^{H*W} (\alpha_t(k) + w_{res})\ C_t(k) \tag{4}$$

where i_t, f_t, o_t, c_t, h_t are the input, forget, output, memory and hidden states. D is the channel length of visual feature blocks and d is the dimension of all LSTM states. x_t is the visual feature refined by estimated facial mask $\alpha_t(k)$. A residual link with adjustable weights w_{res} is included in the facial soft attention module.

Facial masks are learned with previous hidden state $\boldsymbol{h}_{t-1}$ and visual feature $\boldsymbol{C}_t(k)$. $\boldsymbol{v}$, $\boldsymbol{A}_h$ and $\boldsymbol{A}_c$ are learnable weights:

$$\tilde{\alpha}_t(k) = \boldsymbol{v} \tanh(\boldsymbol{A}_h \boldsymbol{h}_{t-1} + \boldsymbol{A}_c \boldsymbol{C}_t(k) + b) \tag{5}$$

$$\alpha_t(k) = \frac{\exp(\tilde{\alpha}_t(k))}{\sum_{j=1}^{H*W} \exp(\tilde{\alpha}_t(j))} \tag{6}$$

Different from previous self-attention studies [18], facial attention masks $\alpha_t(k)$ are learned with facial keypoint heatmap supervisions M_t and L2 losses as shown in Eq. 7,

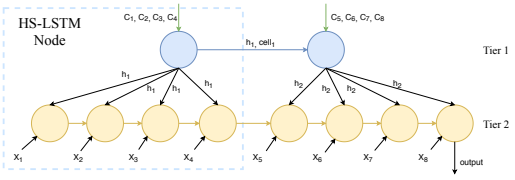

Fig. 3. A two-tier HS-LSTM network structure with two HS-LSTM nodes of size four, where X and C are attended and original visual features.

which provides the clear semantic meaning of facial regions to the learned attention masks.

$$L_{kp} = \sum_{t=1}^{T} \sum_{k=1}^{H*W} \left\| \left(M_t(k) - \alpha_t(k) \right) \right\|^2 \tag{7}$$

The heatmap M_t is converted from estimated keypoints:

$$M_t = \sum_{j=1}^{J} Conf_t^j * \mathcal{N}(Center_t^j, \sigma) \tag{8}$$

where each keypoint $Center_t^j$ is converted into a 2D Gaussian distribution centered at the keypoint. Keypoint estimation confidences $Conf_t^j$ provided by keypoint estimation algorithms are also included in heatmap generation to adjust the weights for Gaussian peaks. Finally, the overlay of keypoint heatmaps is normalized spatially with the softmax function.

The proposed soft attention fusion method is naturally robust against incomplete keypoints. Moreover, the inaccurate predictions with low confidence scores are depressed by the low keypoint estimation confidence included in the heatmap generation step. Therefore, the proposed approach is robust against both *incorrect* and *incomplete* estimated keypoints. Furthermore, it improves the method's interpretability by reflecting attended regions. Correct facial masks can even be generated on cartoon GIFs where no estimated facial keypoints are available. The entire framework is trained end-to-end with intermediate keypoint supervision loss L_{kp} added to main emotion recognition loss in Eq. 12:

$$\mathcal{L} = \mathcal{L}_E + w_{kp} * L_{kp} \tag{9}$$

3.2. Hierarchical Segment LSTM Network (HS-LSTM)

In Section 3.1, we introduce the base KAVAN with a single LSTM layer. Naive temporal networks tend to forget information in early stages [19]. This is not desired in GIF emotion recognition, because GIF contains less redundant frames and all frames are indispensable towards correct emotion recognition. Inspired by recent studies [20], we propose a Hierarchical Segment LSTM module (HS-LSTM) to better model long-term temporal dependencies.

Instead of learning global representations sequentially with LSTM layers, HS-LSTM first generates segment-level

Fig. 4. The four coarse emotion categories summarized from the 17 classes with the circumplex affect model [21].

representations for each segment in GIFs. The segment-level representations are then propagated through different tiers for global GIF-level representations. As shown in Fig. 3, HS-LSTM contains several tiers of LSTM layers that learns representations from coarse- to fine-grained. The first tier takes the stacked features in a segment to learn a coarse segment representation. Nodes in the next tier takes corresponding frame features and the coarse representations learned in the previous tier as input, and learns a refined representation. The representations learned at different temporal resolutions are then propagated through the HS-LSTM network for a final GIF representation. The number of tiers, HS-LSTM nodes and input frames can be adjusted flexibly based on data statistics.

Finally, we show the complete KAVAN structure with HS-LSTM module integrated. The keypoint attended visual attention is only conducted in the last tier:

$$\tilde{\alpha}_t(k) = \boldsymbol{v} \tanh\left(\boldsymbol{A}_h \boldsymbol{h}_{t-1} + \boldsymbol{A}_H \sum_{l=1}^{Tier-1} \boldsymbol{H}_l^i + \boldsymbol{A}_c \boldsymbol{C}_t(k) + b\right) \tag{10}$$

where $\boldsymbol{H}_l^i$ is the output segment representation in a same segment i at all previous tiers l. The input visual feature to the last tier is weight-averaged by the generated attention mask. The inputs to all other tiers remain unchanged.

3.3. Problem Formulation

In this section, we introduce the problem formulation and training objective for GIF emotion recognition. The base task is modeled as the regression of emotion intensities on all labeled emotion classes. Normalized mean squared error L_{RG} is used for regression, which can avoid over or under prediction [3] compared to the MSE loss. The normalized mean squared error ($nMSE$) is defined as the mean squared error divided by the variance of the target vector.

Although intensity score regression is a good formulation for emotion recognition, it becomes increasingly challenging when the number of emotion classes increases. To alleviate this problem and meanwhile achieve a reliable understanding about coarse GIF emotion categories, we divide all labeled emotions into four coarse categories based on the circumflex affect model [21, 22]. The circumflex affect model proposes that emotions are distributed in a 2D circular space, where the vertical axis represents 'arousal' and the horizontal axis represents 'valence'. With the two axes, we divide the emotions into four categories as shown in Fig. 4. We conduct a four-class-classification with cross entropy loss L_C alongside the main regression task. Introducing the categorical emotion classification task has two advantages. First, predicted coarse emotion labels provide extra prior knowledge to the regression branch and make regression easier. Second, a reliable classification branch guarantees correct understanding for the coarse emotion type. For example, confusing 'Happiness' with 'Pleasure' is a smaller error compared to interpreting 'Happiness' as a negative emotion.

Finally, we include a ranking loss L_{RANK} to preserve the rank from the strongest emotion to the most unlikely one. We show that predicting the ranking order of emotion intensity scores could also help the regression task. The proposed ranking loss L_{RANK} is consist of the sum of pairwise ranking loss that is designed to penalize the incorrect orders:

$$L_{RANK}(i) = \left|\left\{(k,l) : \hat{f}_{ik} < \hat{f}_{il}, k < l, k \; in \; K\right\}\right| \tag{11}$$

where $\hat{f}$ is the emotion intensity and K is the total number of emotions. The final loss $\mathcal{L}_E$ for emotion recognition is:

$$\mathcal{L}_E = L_{RG} + w_C * L_C + w_{RANK} * L_{RANK} \tag{12}$$

4. EXPERIMENTS

We first introduce the GIFGIF dataset and facial keypoints pre-processing methods. The proposed framework is then evaluated with both classification and regression metrics.

4.1. Experiment Settings

The data used in this study is collected from a website built by MIT Media Lab named GIFGIF, and is referred to as 'the GIFGIF dataset'. Extending from previous definition of eight emotions, 17 emotions as shown in Figure 4 are labeled to study the more detailed emotions. The dataset is labeled by distributed online users. The annotator is presented with a pair of GIFs and is asked whether GIF A, B or neither expresses a specific emotion. At the time of our data collection, we collect 6,119 GIFs with more than 3.2 million user votes. The massive user votes are converted to a 17-dimensional soft emotion intensity score with the TrueSkill algorithm [23]. Each output emotion intensity score ranges in $[0, 50]$, which is then linearly normalized into $[-1, 1]$.

Besides the appearance feature, estimated facial keypoints are integrated for GIF emotion recognition. 70 facial keypoints are estimated with OpenPose [24]. We then convert the 70 keypoints in each frame into heatmaps following Eq. 8. Each keypoint corresponds to a Gaussian distribution with $\sigma = 5$. The weight of each Gaussian is adjusted by the prediction confidence of the keypoints that $Conf_t^j * \mathcal{N}(Center_t^j, \sigma)$.

The keypoints around lips are denser then other facial regions according to the 70-point facial keypoint definition [24]. Therefore, the weights for the Gaussian distributions around lips is further reduced by 50%. The initial heatmap resolution is 64×64 and is later converted to 7×7 after overlaying all Gaussian peaks. We randomly split 80% of data for training and the rest for testing. The averaged performance on five random splits is reported. The processed data will be released [2].

4.2. Categorical Emotion Classification

We first evaluate the proposed modules with the coarse emotion category classification task. The emotion categories are generated based on the most significant emotion in a GIF. The number of GIFs in each category is $1507/1275/2004/1333$. As shown in Table 4.3, we start with a baseline that uses the *ResNet-50 + LSTM* structure and only the regression loss L_{RG}. A baseline accuracy of 61.47% is achieved. We then evaluate the effectiveness of the proposed soft attention module and HS-LSTM module separately, which are referred to as *Soft-Att+LSTM* and *ResNet-50 + HS-LSTM*. The soft attention module learns an keypoint guided attention mask defined in Eq. 5. The dimension of $\boldsymbol{v}$ is 1×32. $\boldsymbol{A}_h$ and $\boldsymbol{A}_c$ has a parameter size of $32 \times L_{hidden}$ and $32 \times D_{conv}$, which is both 32×2048 in this study. L_{hidden} is the hidden size of the LSTM and D_{conv} is the channel number of visual features. With purely the soft attention module, the recognition accuracy improves from 61.47% to 63.55%. In the HS-LSTM module experiment, we adopt a two-tier structure with two HS-LSTM nodes of size four as shown in Fig. 3. With purely the HS-LSTM module, the accuracy improves from 61.47% to 62.55%. When combining the soft attention module with HS-LSTM, the KAVAN framework achieves an accuracy of 65.95%, which is better than both separate modules. Furthermore, by incorporating multi-task learning with the loss proposed in Eq. 9, an extra 2.32% improvements is obtained and the accuracy reaches 68.27%. This proves the effectiveness of the proposed MTL setting on the classification task.

4.3. Multi-task Emotion Regression

We then show the effectiveness of the proposed modules and the multi-task learning setting with regression metrics. As shown in Table 4.3, the baseline model *ResNet-50 + LSTM* achieves an *nMSE* of 1.0675. With the same parameters in Section 4.2, soft attention module *Soft-Att + LSTM* achieves an *nMSE* of 1.0100, which is significantly better than the baseline. HS-LSTM module *ResNet-50 + HS-LSTM* along also outperforms the baseline LSTM by reaching an *nMSE* of 1.0277. Finally, we evaluate the full framework with both the soft attention module and HS-LSTM adopted. The method reaches an *nMSE* of 0.9987. Furthermore, we fuse the regression framework with the classification branch by conducting multi-task learning. A weighted sum of the *nMSE* loss, the *CrossEntropy* loss and the *ranking* loss is adopted to train the framework. An extra improvement is obtained and the *nMSE* reaches 0.9863.

[2] https://github.com/zyang-ur/human-centered-GIF

Table 1. The coarse emotion category classification accuracy.

Methods	Accuracy $\%$
ResNet-50 + LSTM	61.47%
Soft-Att + LSTM	62.55%
ResNet-50 + HS-LSTM	63.55%
Soft-Att + HS-LSTM	65.95%
MTL Soft-Att + HS-LSTM	$\mathbf{68.27}\%$

Table 2. The predicted Normalized Mean Squared Error (nMSE) on the GIFGIF dataset.

Methods	Loss	nMSE
Color Histogram [3]	nMSE	1.2623 ± 0.0615
Face Expression [3]	nMSE	1.0000 ± 0.0064
ResNet-50 + LSTM	nMSE	1.0675 ± 0.0305
Soft-Att + LSTM	nMSE	1.0100 ± 0.0085
ResNet-50 + HS-LSTM	nMSE	1.0277 ± 0.0147
Soft-Att + HS-LSTM	nMSE	0.9987 ± 0.0033
Soft-Att + HS-LSTM	MTL	$\mathbf{0.9863 \pm 0.0084}$

We then compare our results to other state-of-the-art on MIT GIFGIF. Because the GIFGIF dataset keeps growing, the version we collect with 6,119 GIFs is larger than the one in previous study [3] with 3,858 GIFs. Therefore, a direct comparison is unfair as the task becomes more challenging with more ambiguous GIFs included. Based on our re-implementation as shown in Table 4.3, the *Face Expression + Ordinary Least Squares Regression* approach [3] works the best and achieves a *nMSE* of 1.0000. OpenCV's haar feature-based cascade classifiers are used for face detection. CNN+SVM facial expression features [5, 3] pretrained on a facial emotion dataset [5] are extracted on the largest detected face. Our proposed KAVAN framework achieves a *nMSE* of 0.9863, which is better than the best re-implemented statr-of-the-art of 1.0000.

4.4. Qualitative Results

As shown in Fig. 5, good qualitative results are observed. For example, the upper-left GIF in Fig. 5 belongs to category 'Misery Arousal' that is represented in blue, and is predicted correctly. Fittingly, the predicted emotion intensity of 'anger', 'fear' and 'supervise' are the highest. Furthermore, ideal results on facial region estimation is also observed, as shown in Fig. 6. The larger unmasked image on the left of each sub-figure is the first sampled input frame, and the remaining eight smaller images are the overlay of input frames and facial keypoint heatmaps. The upper two sub-figures in Fig. 6 visualize the supervision heatmaps generated with estimated facial keypoints, which may be inaccurate or incomplete. As shown in the lower two sub-figures in Fig. 6, the attention masks predicted by KAVAN accurately focus on correct fa-

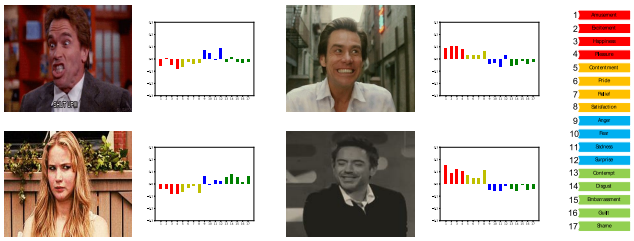

Fig. 5. The visualization of GIF emotion recognition results.

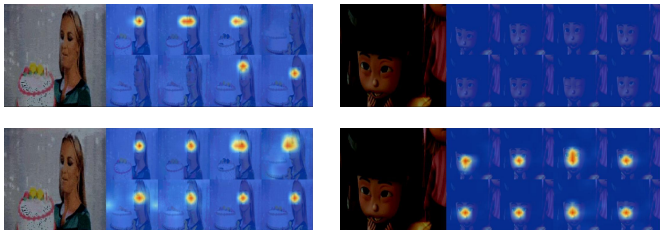

Fig. 6. The visualization of predicted facial attention masks. Upper figures show supervision heatmaps and lower figures visualize the estimated facial masks predicted by KAVAN.

cial regions even when no original keypoint annotations are available, such as in cartoon GIFs. Experiments show that the proposed approach well utilizes the keypoints information and is robust against missing or inaccurate annotations. Furthermore, the predicted facial region masks improve the framework's interpretability.

5. CONCLUSION

Motivated by GIF's unique properties, we focus on human-centered GIF emotion recognition and propose a Keypoint Attended Visual Attention Network (KAVAN). In the facial attention module, we learn facial region masks with estimated facial keypoints to guide the GIF frame representation extraction. In the temporal module, we propose a novel Hierarchical Segment LSTM (HS-LSTM) structure to better represent the temporal evolution and learn better global representations. Experiments on the GIFGIF dataset validate the effectiveness of the proposed framework.

Acknowledgement. This work is partially supported by NSF awards #1704309, #1722847, and #1813709.

6. REFERENCES

[1] Saeideh Bakhshi, David A Shamma, Lyndon Kennedy, Yale Song, Paloma de Juan, and Joseph'Jofish' Kaye, "Fast, cheap, and good: Why animated gifs engage us," in *CHI*. ACM, 2016, pp. 575–586.

[2] Weixuan Chen, Ognjen Oggi Rudovic, and Rosalind W Picard, "Gifgif+: Collecting emotional animated gifs with clustered multi-task learning," in *ACII*. IEEE, 2017, pp. 410–417.

[3] Brendan Jou, Subhabrata Bhattacharya, and Shih-Fu Chang, "Predicting viewer perceived emotions in animated gifs," in *ACM MM*. ACM, 2014, pp. 213–216.

[4] Weixuan Chen and Rosalind W Picard, "Predicting perceived emotions in animated gifs with 3d convolutional neural networks," in *ISM*. IEEE, 2016, pp. 367–368.

[5] Yichuan Tang, "Deep learning using linear support vector machines," *arXiv preprint arXiv:1306.0239*, 2013.

[6] Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael J Black, "Towards understanding action recognition," in *ICCV*. IEEE, 2013, pp. 3192–3199.

[7] Guilhem Chéron, Ivan Laptev, and Cordelia Schmid, "P-cnn: Pose-based cnn features for action recognition," in *ICCV*, 2015, pp. 3218–3226.

[8] Michael Gygli, Yale Song, and Liangliang Cao, "Video2gif: Automatic generation of animated gifs from video," in *CVPR*, 2016.

[9] Sicheng Zhao, Yue Gao, Xiaolei Jiang, Hongxun Yao, Tat-Seng Chua, and Xiaoshuai Sun, "Exploring principles-of-art features for image emotion recognition," in *ACM MM*. ACM, 2014, pp. 47–56.

[10] Sicheng Zhao, Hongxun Yao, Yue Gao, Rongrong Ji, and Guiguang Ding, "Continuous probability distribution prediction of image emotions via multitask shared sparse regression," *IEEE Transactions on Multimedia*, vol. 19, no. 3, pp. 632–645, 2017.

[11] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang, "Building a large scale dataset for image emotion recognition: The fine print and the benchmark.," in *AAAI*, 2016, pp. 308–314.

[12] Tianrong Rao, Min Xu, and Dong Xu, "Learning multi-level deep representations for image emotion classification," *arXiv preprint arXiv:1611.07145*, 2016.

[13] Sicheng Zhao, Guiguang Ding, Yue Gao, and Jungong Han, "Learning visual emotion distributions via multi-modal features fusion," in *ACM MM*. ACM, 2017, pp. 369–377.

[14] Jing Han, Zixing Zhang, Maximilian Schmitt, Maja Pantic, and Björn Schuller, "From hard to soft: Towards more human-like emotion recognition by modelling the perception uncertainty," in *ACM MM*. ACM, 2017, pp. 890–897.

[15] Saeideh Bakhshi, David A Shamma, and Eric Gilbert, "Faces engage us: Photos with faces attract more likes and comments on instagram," in *CHI*. ACM, 2014, pp. 965–974.

[16] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *ECCV*. Springer, 2016, pp. 20–36.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[18] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *ICML*, 2015, pp. 2048–2057.

[19] Shaojie Bai, J Zico Kolter, and Vladlen Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271*, 2018.

[20] Junyoung Chung, Sungjin Ahn, and Yoshua Bengio, "Hierarchical multiscale recurrent neural networks," in *ICLR*, 2017.

[21] James A Russell, "A circumplex model of affect.," *Journal of personality and social psychology*, vol. 39, no. 6, pp. 1161, 1980.

[22] Albert Mehrabian, "Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament," *Current Psychology*, vol. 14, no. 4, pp. 261–292, 1996.

[23] Ralf Herbrich, Tom Minka, and Thore Graepel, "Trueskill: a bayesian skill rating system," in *NIPS*, 2007, pp. 569–576.

[24] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *CVPR*, 2017.