VisEmoComic: Visual Emotion Recognition in Comics Image

Ruddy Théodose, Jean-Christophe Burie

VisEmoComic : Visual Emotion Recognition in Comics image

Théodose Ruddy[1 [0000−0001−9267−1351]] and Burie Jean-Christophe[1 [0000−0001−7323−2855]]

L3i Laboratory, SAIL joint Laboratory
La Rochelle Université
17042 La Rochelle CEDEX 1, France
{ruddy.theodose,jean-christophe.burie}@univ-lr.fr

Abstract. Emotion recognition in images have bean widely studied on captured data of real people but few works have been realized on drawn data. Among this category, comic books have become an important part of the of the popular culture. Whether realistic drawings or oversimplified designs, characters have to depict credible and understandable reactions to the events of the story they are included in. While human-like characters designs are often inspired by real face mechanisms, authors may include various graphic elements to emphasize those reactions to the events they undergo. In this paper, we propose VisEmoComic, an image-based dataset for emotion recognition on comics. Several annotators were invited to give their interpretation of the character emotions represented in given scenes. The image data comes from existing comic book datasets, dedicated to other tasks and from various origins, allowing to include cultural specificities. Additionally, for each sample, the face of the character of interest, its body and the frame where it was drawn are given to allow the use of the immediate spatial context for prediction. Collected samples were annotated by multiple annotators. Consequently, we proposed two schemes to generate labels that sum up the man-made labels and defined baselines using the built dataset.

Keywords: Emotion Recognition · Manga · Comics Analysis · Document Analysis.

1 Introduction

Non verbal communication is a key element to understand all the nuances a discourse can bring. When it comes to direct communication between persons, cues such as posture, gestures, voice intonation, gaze, or facial expressions allow to interpret details in the told speeches. Consequently the field of affective computing has explored the exploitation of these hints for various application topics such as behavioural analysis in crowds, human-machine interfaces or customers/testers satisfaction evaluation.

One subtopic of affective computing concerns the perception of the human agent's emotional state. While studies on captured data (acquired through sensors) have been widely investigated, with numerous methods and datasets [18,1],

pictorial images have drawn less attention. Among the various forms of visual arts, comics, also known as "sequential art" is a major part of the popular culture. They illustrate characters acting in a story told through successions of static panels. The design of drawn characters is often inspired by how real bodies look like and work, so we might suppose that techniques developed for captured data may also be applied to drawn characters. Moreover, besides the body related cues, comic artists have produced visual tools specific to the medium such as symbols or background effects in order to accentuate various aspects of the narration.

Comics analysis community have grown over the last decade. Detection of structural elements inside pages such as panels [16], characters [8,27], speech bubble and text lines [9] and even onomatopoeia [24] have been heavily investigated. Understanding what is described inside panels depends on the comprehension of social and cultural factors that may greatly vary between authors and between readers. In that direction, understanding the character's behaviour is a crucial topic to extract insights for higher goals such as the development of more efficient translation tools or for accessibility purposes.

In this paper, we introduce VisEmoComic, a new dataset for visual emotion recognition in comics.

Building on the foundation established by the Kangaiset dataset, this work continues the development of resources specifically for emotion analysis in comic media.

The interpretation of emotions remains a highly subjective topic, dependant on numerous components such as reader's cultural background. Consequently, the dataset was annotated by multiple annotators in order to gather possibly different opinions on the same data.

Moreover, we trained different networks on the built in order to create baselines. Multiple annotators were involved. One approach would be to train one model on each annotator's label. However, this approach is hardly scalable and can potentially overfit the biases of this annotator. Instead, we proposed to generate intermediate labels that attempt to summarize the "human" labels. The networks trained on generated labels are also evaluated on the original ones to analyze the proximity of the produced annotations with the man-made references. EmoRecCom challenge [29] have realized similar task by estimating the displayed emotions in the whole panel through the combination of text in bubbles and image. Most of the proposed methods combined image processing and language models. In this paper, we restrict to visual modality as we aim at studying the influence of graphical cues on methods initially developed for processing photos. Our contributions are:

- We have built an image-based dataset for emotion recognition on comics from various countries that provides boxes for face and body of the characters of interest;
- We trained and evaluated multiple models, initially designed for processing real captured data on this new dataset using two training strategies that consider the varying opinions of different annotators.

2 Related work

In this section, we first review existing datasets used in general comic analysis. Then we introduce several existing methods for emotion recognition based on real captured data, with a particular emphasis on methods that prioritize image-based analysis.

2.1 Emotion Recognition

Emotion Recognition has been heavily investigated on various types of data.

About image data, face images were considered to provide the most critical information for emotion recognition. When only face data is used, the acronym FER for Facial Expression/Emotion Recognition is often employed in the literature. Prior works used handcrafted features such as Local Binary Patterns [33] before exploring deep learning techniques [22]. While the task is related to object classification, specific techniques have been developed to better address its unique challenges. These techniques include new architectures such as ResMasking [30] that integrates U-Nets into ResNet blocks to generate attention maps, as well as various loss functions [6,12,11], and ensemble networks [5] etc.

However, various works on psychology research [4] suggested the importance of contextual information for the realization of the task.

Following that assumption, research on (spatial) context aware emotion recognition (often abbreviated *CAER*) started to develop. Those methods tend to employ simultaneously multiple images. EMOTIC [19,20] is currently one of the largest public dataset on CAER. The authors of the dataset proposed a two-branch network, one branch focusing on extracting features related to the character of interest while the second branch deals with the whole image. A final fusion subnetwork is in charge of merging the extracted features to generate a prediction. CAER-Net [21] adopts a similar two-branch while estimating and applying weights to each branch before merging the weighted features. Mittal et al. [26] extends the multiple branch approach with EmotiCon by adding modalities other than image data such as character pose estimated by systems such as OpenPose [7] and depth image to better deal with spatial relationships in the scene. Context-Dependent Net (CD-Net) [35] computes global shared features from the entire image and use a transformer to aggregate face, body and scene information.

2.2 General comics databases

The most well-known public datasets for comics analysis and understanding mainly focus on detecting structural elements of pages, such as panels, characters, text, and speech bubbles. eBDThèque [13] was one of the earliest published dataset on the topic. It contains 100 pages from American, European and Japanese comics. By the same team, a dataset named DCM77 [28] focuses on American Golden Age comics available in the public domain in the Digital Comic Museum. Manga109 [2] is currently, to our knowledge, the largest dataset

on the task. It focuses on the japanese untranslated mangas. Multiple extensions were developed around this dataset in order to study other comic related tasks. For example, COO dataset [3] focused on the detection of onomatopoeia and Manga109Dialog [23] provides label data for comic speaker detection. IMCDB (Indian Mythological Comic Database) [14] is a collection of Indian comics translated in English that contains data for panel, speech bubbles and transcriptions of text lines. However, none of these datasets are specifically designed for the task of emotion recognition.

(a) Manga109 (b) EmoRecCom (c) IMCDB

Fig. 1: Unprocessed data from the studied datasets, Manga109 are double pages, EmoRecCom images are panels, IMCDB images are single pages.

3 Dataset

In this section, we describe the construction process of the VisEmoComic dataset and provide statistical information on the aggregated data.

3.1 Data construction

Data sources The dataset was created under the assumption that facial expressions of emotions are universal across cultures, even though psychologists are still debating about this topic [17]. Moreover, drawn representations may also greatly vary according the cultural background of the artists. For this reason, we extracted images from the Japanese mangas in Manga109 dataset [25,2], American Golden Age comics from the EmoRecCom Challenge dataset [29] and Indian comics from IMCDB dataset [14] into a single dataset for emotion recognition 1. The Manga109 and IMCDB datasets do not include emotion labels, as they were not created for the purpose of emotion recognition. The EmoRecCom challenge was created specifically for emotion recognition on comics. However, the provided labels consider the overall mood of the scene rather than emotions assigned to individual characters. Consequently, there are no specific emotion labels assigned to each character separately.

Dataset Structure In this paper, we call "character image" the image related to the character of interest, either the face image or the body image, in opposition to panel image. as illustrated in Figure 2. Each entry of the dataset represents a specific character of interest in a given situation represented by the panel. One sample then contains the panel, the location of the face and the body of the character of interest inside this panel. Manga109 dataset already provides the whole pages and for each panel, face and body boxes are given with the identifier of the represented character. For EmoRecCom, we used some of the panels extracted by the authors of the challenge. However, as aforementioned, neither face nor body boxes are available so we used a YOLOv5 network trained on Manga109 boxes to produce characters' boxes. The associations between face and body boxes were done manually. In IMCDB dataset, pages are available, but not all of them have panel boxes. Thus, the available panels were extracted and processed with the same approach as the one used for EmoRecCom panels. We built this dataset with the assumption that faces are required for the emotion estimation on a visual standpoint. Characters seen from the back were not integrated.

Fig. 2: Illustration of the different scales of study, rows : face, body and panel. Data extracted from Manga109 dataset. Columns in order : Appare Kappore ©Kanno Hiroshi, Hanzai Kousyounin Minegishi Eitarou ©Ki Takashi, Jiji Baba Fight ©Nishikawa Shinji, Karappo Highschool ©Takaguchi Satosumi

Emotion model Literature has modeled the concept of emotion in various ways. Methods such as Plutchik model [31] define a discrete and finite set while others such as the Russell's circumplex model [32] tend to represent emotions as points in multidimensional space. For the dataset, we use the Ekman model [10] that divide the emotion spectrum into 6 classes besides the neutral expression : anger, disgust, fear, joy, sadness and surprise. The inclusion of a "neutral" class is subject to debate since it could be perceived as the absence of a strong reaction, and therefore, not strictly classified as an emotion. While more complex models

such as the aforementioned ones exist, we wanted a set that are simple enough to be accessible to non expert annotators.

Annotation For each dataset, 3 annotators were chosen. The three datasets were not necessarily processed by the same group because the annotators had to master the language of their assigned sub-dataset. Even if we didn't planned, at this stage, to use textual information, we considered that reading what was said by the characters could provide additional information on the events and help the annotators in their decision. Annotators received the panels showing the character of interest framed in a colored rectangle and were asked to choose between the given emotions. Multilabel classification was authorized meaning that annotators could select multiple choices. This decision was made in order to mitigate the smaller representation capability of a single label Ekman-based classification. Moreover, situations that the characters undergo induce complex reactions, so multilabel classification reduces the number of constraints imposed on the annotators who do not have to select a single most "predominant" emotion. Annotators were asked to select one or two classes, three in the most extreme cases.

3.2 Data statistics

Table 1 shows the composition of each sub-datasets per annotator. The first observation is the large class imbalance. Data is extracted directly from comic books meaning that characters are integrated inside a story and their actions and reactions have to be coherent with the events. For all annotators and all subsets, the "disgust" class is poorly represented while "neutral" and "joy" are the most dominant emotions.

Table 2 compares the Kangaiset dataset [34] with our VisEmoComic dataset, highlighting two main differences.

Firstly, Kangaiset confines its data to Manga109 pages, whereas we expanded our dataset to include images from the EmoRecCom and IMCDB datasets, resulting in greater diversity in graphic style representations.

Secondly, there is a distinction in the annotation process. The creators of Kangaiset used a single annotator for label creation, whereas our annotation process involved three specialized annotators.

3.3 Agreement between annotators

As the topic remains subjective, it is interesting to measure how similar are the labels produced by the different annotators. Several metrics exists to measure this similarity. Here, we use the Cohen's kappa and the Fleiss' kappa, two chance-corrected coefficients. Both produce a score between -1 and 1, -1 meaning a complete disagreement between annotators, 0 means that the labels were produced randomly and 1 a complete agreement. However, the interpretation of the intermediate values may vary between specialists. Cohen's kappa can be computed only between two annotators while Fleiss' kappa allows to compute

Table 1: Emotion count for the three datasets. Due to the multilabel setup, the sum of each row may not match the number of samples. Manga109 : 12616 samples; EmoRecCom : 3609 samples; IMCDB : 1036 samples

Dataset	AnnID	Anger	Disgust	Fear	Joy	Neutral	Sadness	Surprise
Manga109	1	2335	186	1483	3175	1625	1451	3113
	2	2616	60	1185	3117	1116	1924	3364
	3	2597	150	1397	3404	1758	1662	3258
EmoRecCom	1	817	174	787	849	311	190	783
	2	988	225	597	709	488	175	785
	3	815	176	639	970	628	272	425
IMCDB	1	164	11	61	324	231	45	267
	2	185	17	42	228	245	89	293
	3	155	7	67	303	325	91	123

Table 2: Comparison between Kangaiset and VisEmoComic. The three last columns represent the number of samples extracted from each datasets.

Dataset	Annotator Number	Manga109	EmoRecCom	IMCDB
Kangaiset	1	9387	/	/
VisEmoComic	3	12616	3609	1036

a score between two or more annotators. The computed scores are displayed in Table 3. For all classes, Manga109 images seem to be more consensual than EmoRecCom and IMCDB data. One assumption could be the the design and scenography trends. As illustrated in Figure 3 that display the histogram of the ratios face box size/panel size for the three subsets, Japanese mangas tend to prefer close shots with more refined faces meaning that characters are more easily recognizable. EmoRecCom and IMCDB images often include large shots meaning that actions and scenes tend to predominates over the characters themselves, at least on the visual aspect and for the selected data. Moreover, we can observe that the scores for the "disgust" class are consistently close to 0. This indicates a scarcity of annotations across all raters, making it challenging to discern a definitive trend. Consequently, the metrics suggest that these annotations are more likely to have been assigned randomly.

4 Experiments

4.1 Tested networks

In this paper, multiple methods were evaluated. First, we assessed FER methods that use only face images. We evaluated a Resnet34 [15], a Resnet34 with spatial and channel attention (CBAM) [36] and a ResnetMasking34 [30]. Then, we evaluated three CAER methods, which simultaneously use the image of a character (face or body) as well as data from its spatial context. This category includes the CAER-S network [21], the EMOTIC network [20], and the CD-Net [35].

Table 3: Agreement scores on the different datasets. "Cohen 1v2" is the Cohen's kappa computed with the labels of annotator n°1 and n°2

Dataset	Metric	Anger	Disgust	Fear	Joy	Neutral	Sadness	Surprise
Manga109	Cohen 1v2	0.7597	0.2876	0.4836	0.8916	0.517	0.6511	0.6049
	Cohen 1v3	0.7456	0.1073	0.4811	0.8892	0.5662	0.6971	0.5656
	Cohen 2v3	0.7043	0.08905	0.4232	0.8634	0.4397	0.6233	0.541
	Fleiss	0.736	0.1579	0.4631	0.8813	0.5091	0.6553	0.5703
EmoRecCom	Cohen 1v2	0.6575	0.2393	0.5354	0.773	0.4699	0.5528	0.5666
	Cohen 1v3	0.5091	0.1532	0.4823	0.6969	0.2742	0.3954	0.3591
	Cohen 2v3	0.5172	0.227	0.5686	0.7126	0.3912	0.3462	0.4459
	Fleiss	0.5619	0.2084	0.5264	0.7255	0.3712	0.4229	0.4601
IMCDB	Cohen 1v2	0.7555	0.204	0.3371	0.7118	0.6346	0.4931	0.6137
	Cohen 1v3	0.6999	0.1037	0.5005	0.7783	0.5475	0.4067	0.4672
	Cohen 2v3	0.6838	-0.009665	0.411	0.7158	0.5532	0.4524	0.4169
	Fleiss	0.7135	0.1042	0.4213	0.7354	0.5745	0.449	0.499

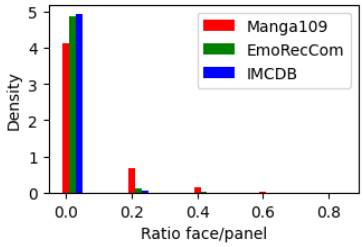


Fig. 3: Histograms of ratios face/panels for each dataset, x-axes represent the computed ratios, y-axes represent densities.

4.2 Training and testing setups

Manga109 and IMCDB provides information on the books where pages come from. EmoRecCom panels are not linked to any book information but the file naming convention provides hints on the original sources. The three subsets are split into a train and a test set. However, instead of splitting the whole batch of images, we split the books. As authors have their own specific art styles, splitting the books allows to reproduce the cases when new and unknown books are processed. In this experiment, we opted for a 7:3 ratio for train:test set.

Each sample have been annotated by multiple annotators. However, for the training, a label has to be defined. We planned two different training schemes. The first one, named *Perm* for "Permissive" consider that if a class is selected by at least one annotator, it is included in the training label. For the second one named *Maj* for "Majority", an emotion is included in the training label if at least two out of three annotators have selected the same emotion. Figure 4 provides examples of those generated labels according to the annotators' labels.

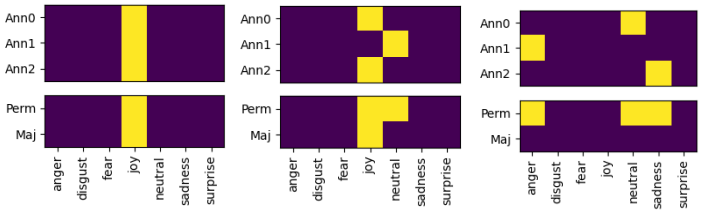


Fig. 4: (top) : Human annotations ; (bottom) : generated labels under the two schemes. Left :Unanimity case; Middle : Slight differences between annotators; Right : Complete disagreement

Figure 5 illustrates our training and testing processes. For the training phase (Fig. 5a), the annotation answers from each subset (Manga109, EmoRecCom, IMCDB) are merged into either majority or permissive labels then concatenated to create a complete training set. For the evaluation phase (Fig. 5b), the trained network is evaluated separately on each test subset using manual labels. We define $A = \{1, 2, 3\}$ the set of annotators, $D = \{\text{Manga109, EmoRecCom, IMCDB}\}$ the set of studied subsets and E the emotion classes (in our case, 6 emotions of the Ekman model + neutral). We note $S_{a,d,e}$ the test score for the annotator $a \in A$, , on the dataset $d \in D$, and the emotion $e \in E$.

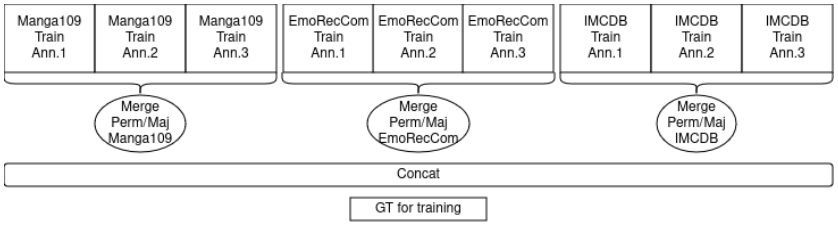
The metric used is the F1 score. Since the networks were trained for multilabel classification, we computed macro F1 scores (unweighted mean between F1 scores for positive and negative samples) for each emotion, consequently each category is processed independently as a binary classification.

Images related to characters are resized to 256 $\times$ 256 pixels. However, for panel images, since a scene can include multiple characters, maintaining the same size could compromise the visibility of smaller characters within the panel. Therefore, panel images are resized to 512 $\times$ 512 pixels. All networks were trained with the same hyperparameters : training lasted for 80 epochs using the focal loss function and the Adam optimizer, coupled with a one-cycle learning scheduler and a maximum learning rate set to 1e-4.

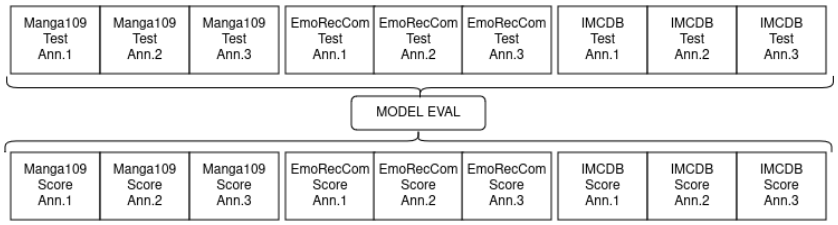
4.3 Global results

In this section, we first evaluate global mean on the tested networks. The average score for one emotion is computed with $S_e = \frac{1}{\#D\#A} \sum_{d \in D} \sum_{a \in A} S_{a,d,e}$ and the average on all emotions $S = \frac{1}{\#E} \sum_{e \in E} S_e$. Table 4 summarizes the initial input configuration of all the tested networks. For ResMasking and the three CAER methods, the inputs setups match the one introduced in the original papers. All the macro F1 scores are listed in Table 5.

As expected, minor classes such as "disgust" and "fear" exhibit lower detection rates while major emotions such as "joy" and "anger" illustrate good performances for the all the tested networks. However, the "neutral" class yields comparatively lower results. Given that emotions are typically associated with



(a) Training steps



(b) Testing steps, bottom blocks represent $S_{a,d,e}$

Fig. 5: Our proposed training and testing processes.

Table 4: Base Input Configuration for each experiment

Model	Face	Body	Panel
Resnet34	✓		
Resnet34CBAM	✓		
ResMasking34	✓		
CAER-S Net	✓		✓
EMOTIC Net		✓	✓
CD-Net	✓	✓	✓

facial expressions, the "neutral" class represents a challenge as it corresponds to the default facial expression without any emotional marker. Depending on one's perspective on the "neutral" class, finding positive examples for "neutral" emotions and negative examples for the other six basic emotions can be considered analogous tasks. Consequently, trying to separate the "neutral" class could potentially complicate the training process. Interestingly, similar dynamics can be observed between agreement scores and F1 scores for each emotion. Kappa scores indicate the level of consensus among classes, with low agreement scores suggesting greater classification difficulty. For the "disgust" class, the scores also depends on the quantity of positive samples, but for the "neutral" class which is one of the major class, the F1 scores show that it was not the easiest "emotion" to classify.

Although the ResNet34 network is considerably simpler compared to the others, it demonstrates comparable performance on the task. This suggests that the image of the character alone conveys the most crucial emotional informa-

Table 5: Global Macro F1 scores for each emotion S_e and average S.

Method	Anger	Disgust	Fear	Joy	Neutral	Sadness	Surprise	Mean
Resnet Perm	68.89	52.47	**61.94**	**79.25**	63.97	**62.55**	**64.67**	**64.82**
Resnet Maj	69.32	49.25	59.23	77.01	61.32	59.71	64.47	62.90
CBAM Perm	69.14	52.20	60.48	78.41	64.14	60.12	63.58	64.01
CBAM Maj	66.87	49.32	60.15	76.81	62.48	59.10	62.64	62.48
ResMasking Perm	69.05	49.32	60.40	77.19	62.45	60.65	64.44	63.36
ResMasking Maj	66.02	49.32	56.86	75.38	60.74	58.91	61.89	61.30
CAER-S Perm	**70.34**	50.31	61.30	78.52	**64.40**	61.49	64.30	64.38
CAER-S Maj	65.54	49.32	58.92	74.96	60.76	59.40	63.79	61.81
EMOTIC Perm	64.44	50.95	59.36	70.10	61.24	60.01	61.47	61.08
EMOTIC Maj	62.29	49.32	56.20	68.71	57.35	54.22	60.05	58.31
CDNet Perm	64.69	**53.81**	58.77	72.94	61.90	57.63	62.87	61.80
CDNet Maj	60.31	49.47	54.85	69.46	59.23	52.51	60.90	58.11

Table 6: F1 scores for **positive** and **negative** samples for the ResNet34 trained on permissive labels.

Metric	Anger	Disgust	Fear	Joy	Neutral	Sadness	Surprise	Mean
Neg F1	85.60	97.73	88.80	89.02	84.67	93.54	82.60	88.85
Pos F1	52.17	7.21	35.08	69.48	43.27	31.55	46.75	40.79
Macro F1	68.89	52.47	61.94	79.25	63.97	62.55	64.67	64.82

tion. The addition of contextual information appears to have a minor impact on performance. However, these results should be interpreted with caution due to the chosen training conditions. The same hyperparameters were applied to all networks, without taking into account the conditions presented in the original papers or the inherent complexity of each network. While all experiments converged, the simplicity of the ResNet34 compared to the others may have contributed to its generally better performance.

While we displayed the macro F1, it is interesting to also analyze the gap between positive and negative F1 scores, displayed in Table 6. Negative samples are much more abundant, making it easier to sets predictions to zero. Figure 6 illustrates some predictions on the test sets. Even if they were trained under the same conditions, network predictions can vary significantly when major cues are not visible.

4.4 Results on individual annotators' labels

In this section, we conduct a detailed analysis to examine the impact of the generated labels in relation to the individual annotators' labels across the different datasets. Given that ResNet34 produced the best overall performance in the previous section, our focus remains on this network. Table 7 presents the F1 scores for each dataset, annotator and emotion. ($S_{a,d,e}$ defined earlier).

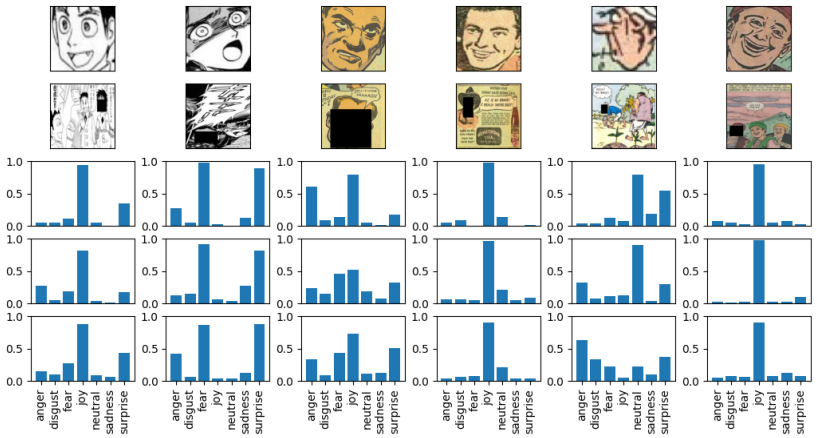


Fig. 6: Predictions from the three networks. Column 1-2 : Manga109 data; Column 3-4 : EmoRecCom data; Column 5-6 : IMCDB data. Row 1 : Face Image; Row 2 : Panel image, the face is masked; Row 3-5 : outputs from ResNet, CAER-S Net and EMOTIC Net in that order.

We first observe that, for the same experiment, results tend to be better on Manga109 compared to EmoRecCom and IMCDB. This is likely due to the larger size of the Manga109 dataset used for training, which allows the network to specialize more effectively in this type of data. Additionally, faces in the EmoRecCom and IMCDB datasets are often less prominent, making it more challenging to analyze facial expressions when characters are not depicted in close-up shots.

For the three datasets, networks trained on "Permissive" labels demonstrate better alignment with individual opinions. In fact, a label set to "positive" by the "Majority" scheme indicates the number of annotators who have chosen this emotion, suggesting a higher level of consensus and thus easier fitting. However, this approach can also ignore the opinions expressed by the minority. If we consider the annotated emotions as sets, the “permissive” set encompasses the “majority” set for each image. One can suppose that permissive labels, by adapting to marginal judgments, may increase the risk of classification error for certain annotators. However, under the given training and testing conditions, the inclusion of all opinions seems to overtake the effects of potential noise.

4.5 Difference between face and body images

Historically, facial features were considered as the main features for emotion recognition, leading to the development of the "FER" terminology for methods centered on this modality. However more recent context-aware and multimodal methods integrate bodily features such as pose or gait. In this section, we compare the impact of face and body images on prediction results. All the tested

Table 7: Macro F1 scores $S_{a,d,e}$ of ResNet34 on each subset and each annoator separately. "Dset" : Dataset; "AnnID": Annotator ID; "M109": Manga109; "ERC": EmoRecCom

Dset	AnnID	Label	Anger	Disgust	Fear	Joy	Neutral	Sadness	Surprise	Mean
M109	1	Perm	78.02	53.41	65.70	89.27	73.96	73.12	75.07	72.65
		Maj	78.29	49.58	60.98	89.38	68.13	74.82	73.88	70.72
	2	Perm	78.20	49.78	63.65	88.50	66.23	72.04	74.45	70.41
		Maj	77.00	49.88	59.00	88.32	65.29	70.20	73.23	68.99
	3	Perm	77.70	49.60	63.58	88.98	71.45	74.76	72.76	71.26
		Maj	76.99	49.71	60.20	88.08	65.47	72.30	70.76	69.07
ERC	1	Perm	60.81	54.03	62.63	71.24	54.69	54.89	58.67	59.57
		Maj	62.73	48.47	59.06	71.39	59.25	49.58	56.90	58.20
	2	Perm	63.21	51.90	60.56	71.43	61.17	52.41	60.30	60.14
		Maj	62.83	48.10	62.78	72.65	55.43	50.13	55.47	58.20
	3	Perm	63.20	53.53	63.59	75.15	65.32	54.68	54.21	61.38
		Maj	64.20	48.64	63.94	71.91	55.66	49.09	56.84	58.61
IMCDB	1	Perm	65.39	49.18	60.34	74.82	59.21	60.97	65.83	62.25
		Maj	67.10	49.76	57.46	69.04	59.06	59.06	67.56	61.29
	2	Perm	66.67	61.63	57.51	75.90	61.53	58.62	63.41	63.61
		Maj	65.94	49.43	55.68	73.40	63.39	56.00	63.75	61.08
	3	Perm	66.80	49.13	59.92	77.97	62.21	61.43	57.37	62.12
		Maj	68.80	49.71	53.99	68.89	60.25	56.18	61.85	59.95

networks, except CD-Net which already use both, were trained on either face or body images, with permissive labels. Table 8 shows the global F1-scores S_e for each experiment.

In most cases, using body images instead of face images results in poorer performance.

While character poses may convey emotional cues, they are often more indicative of actions rather than emotions relevant to the task. When body images lack close-ups, crucial facial information becomes diluted in the "body" context, making the input data less relevant for the emotion recognition.

5 Conclusion

In this paper, we introduce VisEmoComic, a novel dataset designed for visual emotion recognition in comics. The data collection process was curated from various sources to examine how emotions are represented across different cultures. Each character is analyzed at three levels, face, body and panel, in order to evaluate its emotional state. Given the subjectivity inherent in emotion recognition, we requested the opinion of multiple annotators to build the dataset and proposed schemes to train a system based on the "median" annotator's perspective rather than fitting it to a specific annotator's style. We established initial baselines using various networks, some of which incorporate the spatial context

Table 8: F1 scores comparison between face and body inputs.

Method	Anger	Disgust	Fear	Joy	Neutral	Sadness	Surprise	Mean
Resnet Face	68.89	52.47	61.94	79.25	63.97	62.55	64.67	64.82
Resnet Body	65.57	50.65	60.91	71.85	61.16	59.63	62.92	61.81
CBAM Face	69.14	52.20	60.48	78.41	64.14	60.12	63.58	64.01
CBAM Body	60.99	50.19	56.89	71.25	59.73	59.01	62.30	60.05
ResMasking Face	69.05	49.32	60.40	77.19	62.45	60.65	64.44	63.36
ResMasking Body	56.98	49.32	55.36	69.75	57.92	52.74	59.05	57.30
CAER-S Face	70.34	50.31	61.30	78.52	64.40	61.49	64.30	64.38
CAER-S Body	64.88	49.78	59.21	71.52	60.93	60.15	61.46	61.13
EMOTIC Face	70.62	52.91	60.71	79.09	63.34	61.02	63.61	64.47
EMOTIC Body	64.44	50.95	59.36	70.10	61.24	60.01	61.47	61.08

of the character of interest. While annotators were strongly encouraged to consider all panel elements, including text in speech bubbles, for labeling decisions, our primary aim is to explore what can be inferred solely from the image for emotion recognition. Future research may explore approaches that combine text and image processing to harness insights from both modalities.

Acknowledgements This work benefited from access to the computing resources of the L3i laboratory, operated and hosted by the University of La Rochelle and the computing resources of the "CALI 3" cluster, operated and hosted by the University of Limoges, part of the HPC network in the Nouvelle-Aquitaine Region. Both are financed by the State and the Region.

References

1. Naveed Ahmed, Zaher Al Aghbari, and Shini Girija. A systematic survey on multimodal emotion recognition using learning algorithms. Intelligent Systems with Applications, 17:200171, 2023.
2. Kiyoharu Aizawa, Azuma Fujimoto, Atsushi Otsubo, Toru Ogawa, Yusuke Matsui, Koki Tsubota, and Hikaru Ikuta. Building a manga dataset "manga109" with annotations for multimedia applications. IEEE MultiMedia, 27(2):8–18, 2020.
3. Jeonghun Baek, Yusuke Matsui, and Kiyoharu Aizawa. Coo: Comic onomatopoeia dataset for recognizing arbitrary or truncated texts. In European Conference on Computer Vision, pages 267–283. Springer, 2022.
4. Lisa Feldman Barrett, Batja Mesquita, and Maria Gendron. Context in emotion perception. Current directions in psychological science, 20(5):286–290, 2011.
5. Christian Białek, Andrzej Matiolański, and Michał Grega. An efficient approach to face emotion recognition with convolutional neural networks. Electronics, 12(12):2707, 2023.
6. Jie Cai, Zibo Meng, Ahmed Shehab Khan, Zhiyuan Li, James O'Reilly, and Yan Tong. Island loss for learning discriminative features in facial expression recognition. In 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), pages 302–309. IEEE, 2018.

7. Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 7291–7299, 2017.
8. Wei-Ta Chu and Wei-Wei Li. Manga FaceNet: Face Detection in Manga based on Deep Neural Network. In Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval, pages 412–415, Bucharest Romania, June 2017. ACM.
9. David Dubray and Jochen Laubrock. Deep cnn-based speech balloon detection and segmentation for comic books. In 2019 International Conference on Document Analysis and Recognition (ICDAR), pages 1237–1243. IEEE, 2019.
10. Paul Ekman and Wallace V. Friesen. Constants across cultures in the face and emotion. Journal of Personality and Social Psychology, 17(2):124–129, 1971.
11. Ali Pourramezan Fard and Mohammad H Mahoor. Ad-corre: Adaptive correlation-based loss for facial expression recognition in the wild. IEEE Access, 10:26756–26768, 2022.
12. Amir Hossein Farzaneh and Xiaojun Qi. Facial expression recognition in the wild via deep attentive center loss. In Proceedings of the IEEE/CVF winter conference on applications of computer vision, pages 2402–2411, 2021.
13. Clément Guérin, Christophe Rigaud, Antoine Mercier, Farid Ammar-Boudjelal, Karell Bertet, Alain Bouju, Jean-Christophe Burie, Georges Louis, Jean-Marc Ogier, and Arnaud Revel. ebdtheque: a representative database of comics. In 2013 12th International Conference on Document Analysis and Recognition, pages 1145–1149. IEEE, 2013.
14. Vaibhavi Gupta, Vinay Detani, Vivek Khokar, and Chiranjoy Chattopadhyay. C2vnet: A deep learning framework towards comic strip to audio-visual scene synthesis. In Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part II 16, pages 160–175. Springer, 2021.
15. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.
16. Zheqi He, Yafeng Zhou, Yongtao Wang, Siwei Wang, Xiaoqing Lu, Zhi Tang, and Ling Cai. An End-to-End Quadrilateral Regression Network for Comic Panel Extraction. In Proceedings of the 26th ACM international conference on Multimedia, MM '18, pages 887–895, New York, NY, USA, October 2018. Association for Computing Machinery.
17. Rachael E Jack, Oliver GB Garrod, Hui Yu, Roberto Caldara, and Philippe G Schyns. Facial expressions of emotion are not culturally universal. Proceedings of the National Academy of Sciences, 109(19):7241–7244, 2012.
18. Smith K Khare, Victoria Blanes-Vidal, Esmaeil S Nadimi, and U Rajendra Acharya. Emotion recognition and artificial intelligence: A systematic review (2014–2023) and research recommendations. Information Fusion, page 102019, 2023.
19. Ronak Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza. Emotion recognition in context. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1667–1675, 2017.
20. Ronak Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza. Context based emotion recognition using emotic dataset. IEEE transactions on pattern analysis and machine intelligence, 42(11):2755–2766, 2019.

21. Jiyoung Lee, Seungryong Kim, Sunok Kim, Jungin Park, and Kwanghoon Sohn. Context-aware emotion recognition networks. In Proceedings of the IEEE/CVF international conference on computer vision, pages 10143–10152, 2019.
22. Shan Li and Weihong Deng. Deep facial expression recognition: A survey. IEEE transactions on affective computing, 13(3):1195–1215, 2020.
23. Yingxuan Li, Kiyoharu Aizawa, and Yusuke Matsui. Manga109dialog a large-scale dialogue dataset for comics speaker detection. preprint arXiv:2306.17469, 2023.
24. John Benson Louis and Jean-Christophe Burie. Detection of Buried Complex Text. Case of Onomatopoeia in Comics Books. In M. Coustaty and A. Fornés, editors, Document Analysis and Recognition – ICDAR 2023 Workshops, Lecture Notes in Computer Science, pages 177–191, Cham, 2023. Springer Nature Switzerland.
25. Yusuke Matsui, Kota Ito, Yuji Aramaki, Azuma Fujimoto, Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa. Sketch-based manga retrieval using manga109 dataset. Multimedia Tools and Applications, 76(20):21811–21838, 2017.
26. Trisha Mittal, Pooja Guhan, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. Emoticon: Context-aware multimodal emotion recognition using frege's principle. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 14234–14243, 2020.
27. Nhu-Van Nguyen, Christophe Rigaud, and Jean-Christophe Burie. Comic characters detection using deep learning. In 2017 14th IAPR international conference on document analysis and recognition (ICDAR), volume 3, pages 41–46. IEEE, 2017.
28. Nhu-Van Nguyen, Christophe Rigaud, and Jean-Christophe Burie. Digital comics image indexing based on deep learning. Journal of Imaging, 4(7), 2018.
29. Nhu-Van Nguyen, Xuan-Son Vu, Christophe Rigaud, Lili Jiang, and Jean-Christophe Burie. Icdar 2021 competition on multimodal emotion recognition on comics scenes. In International Conference on Document Analysis and Recognition, pages 767–782. Springer, 2021.
30. Luan Pham, The Huynh Vu, and Tuan Anh Tran. Facial expression recognition using residual masking network. In 2020 25Th international conference on pattern recognition (ICPR), pages 4513–4519. IEEE, 2021.
31. Robert Plutchik. The Nature of Emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. American Scientist, 89(4):344–350, 2001.
32. James Russell. A Circumplex Model of Affect. Journal of Personality and Social Psychology, 39:1161–1178, December 1980.
33. Caifeng Shan, Shaogang Gong, and Peter W McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. Image and vision Computing, 27(6):803–816, 2009.
34. Ruddy Théodose and Jean-Christophe Burie. KangaiSet: A Dataset for Visual Emotion Recognition on Manga. In M. Coustaty and A. Fornés, editors, Document Analysis and Recognition – ICDAR 2023 Workshops, Lecture Notes in Computer Science, pages 120–134, Cham, 2023. Springer Nature Switzerland.
35. Zili Wang, Lingjie Lao, Xiaoya Zhang, Yong Li, Tong Zhang, and Zhen Cui. Context-dependent emotion recognition. Journal of Visual Communication and Image Representation, 89:103679, 2022.
36. Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In Proceedings of the European conference on computer vision (ECCV), pages 3–19, 2018.