Automated Emotion Recognition Through Graphical Cues on Comics at Character Scale

Théodose Ruddy, Burie Jean-Christophe, Revel Arnaud

Automated Emotion Recognition through Graphical Cues on Comics at Character Scale

Théodose Ruddy$^{1[0000-0001-9267-1351]}$, Burie Jean-Christophe$^{1[0000-0001-7323-2855]}$, and Revel Arnaud$^{1[0000-0002-9498-392X]}$

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

Abstract. Emotions are psychological reactions to external events. Characters represented in artistic works may manifest emotions in order to replicate credible and human-like behaviors in specific situations. They also provides important hints to better understand the stakes and the tone of story. In comics, markers of emotions can be found in the dialogues or through visual cues specifically drawn by the artists. While automated emotion extraction on textual information is an active research field, few works have addressed this topic through the graphical grammar of comics (and more generally on drawings). In this paper, we propose to review the different visual tools used by artists to convey expressiveness to their characters and how they can be exploited for automated processing. Some of those cues are strongly related to the human body, its representation and mechanisms. Consequently, we propose to study developed methods for those topics on photography or captured videos. Then, we suggest contributions that aimed at facilitating the transition between real and drawn domains.

Keywords: Emotion Recognition · Document Analysis · Machine Learning · Comic Analysis.

1 Introduction

Through digitization, comics books can now be processed and analysed by automatic algorithms like other types of documents. Goals are various : accessibility, archiving, facilitated translation... Multiple works have focus on extraction of comic core elements such as panels [19,14,27], speech bubbles [8,20,15], characters [6,24,30]... Most of them are restricted to the structural components analysis. Higher-level concepts such as layout understanding, story understanding and character analysis, because they rely on performances of low-level algorithms, remain less addressed. In every story, characters are the key elements. A scene evolves in a direction because of the events caused by the different protagonists. As they have to undergo specific situations along the story, the artists must illustrate their reactions to external factors as credible and life-like as possible.

Knowing how characters feel toward specific events can provide useful cues about the stakes of the story and establish a psychological profile of each character. For accessibility purposes, knowing which emotion is represented can improve speech synthesis systems by creating voices that fit better to the scene illustrated in the frame. Moreover, felt emotion at one moment can determine the linguistic choices of a speaker, and then can be a useful indication for text translation.

Emotion recognition is an active research field that spans across all human means of communication : voice, writing, photos... However, applications of this field on comics remain scarce. Comics are multimodal medias containing both textual and visual information. Dialogues written in speech bubbles relay important pieces of information that help to understand the speaker's state of mind. Techniques developed for emotion recognition on literature can be transferred to the dialogues. However, fewer works have attempted to benefit from the graphic specificities of comic stories.

In this paper, we propose to review the different visual tools used by artists for illustrating the emotional state of their characters and how they can be exploited by automated processes. We first describe, in section 2, how emotions can be represented in recognition problems and how the outputs can be expressed. Then, each category of graphic tools is introduced in section 3. For the tools that are closely related to real data, we present a glimpse of methods developed for captured photos and videos. Lastly, in section 4, we present methods that aims at making machine learning algorithms more robust to the transition between real and drawn images. Characters behaviours can be studied on multiple narrative levels from the story level to the character level. While the first levels induce higher levels of understanding on the story and interpersonal relations, they all rely on analysis of how each character reacts to situations. In this study we voluntarily omit social interactions and focus on indications linked to the reactions of each character independently of the others.

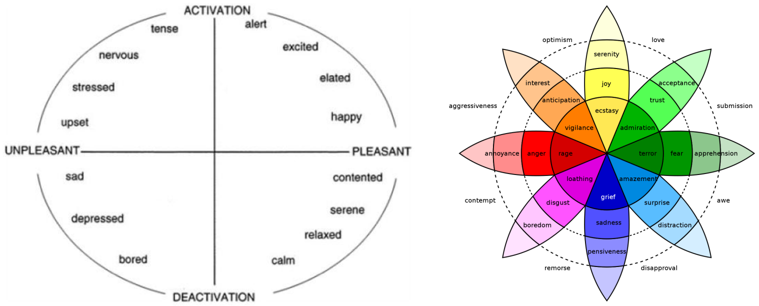
2 Models of emotion

One of the first essential decision in designing an emotion recognition system is choosing which emotions are considered and how they can be represented. Multiple computational models have been proposed to interpret the perceived emotion. These models can be grouped into multiple supersets : categorical models, dimensional models and hybrids models.

Categorical Models The categorical models define a finite set of emotions. In that case, the goal of an emotion recognition system is then to assign to the studied signal the closest emotion of the set. The main advantage of these models is that they represent emotions with simple terms. However, such models do not allow to represent more complex states that are outside of the defined set. Among the categorical models, the one proposed by Ekman [9] is the most famous model. It defines 6 basic emotions besides the neutral emotion : Anger, Happiness, Disgust, Fear, Sadness and Surprise.

Dimensional Models In dimensional models, feelings are described in multidimensional continuous spaces, generally two for valence (positiveness or negativeness of a feeling) and arousal (the level of excitement from boredom or sleepiness to wild excitement). A feeling is then defined as a point in this multidimensional space. These models allow, in opposition with the categorical models, to represent a much broader and continuous spectrum of emotions. Such models can also allow the definition of metrics to better assess similarities between signals. However, defining emotion in a continuous space is much more complex. Each person has its own sensitivity regarding signals, hence emotion placement between participants may vary much more. The circumplex model [32] (Fig. 1a).the Self-Assessment Manikin (SAM) model [3] or the Positive Activation-Negative Activation (PANA) model [37] belong to this group of models and revolve around the concepts of valence and arousal.

Hybrid Models Between categorical and dimensional models, hybrid models organise emotions in a hierarchical structure and state that complex emotions are combinations of more simple/basic ones, allowing to extend the emotion spectrum. For example, the Plutchik model [28] defines *dyads* as mixtures of two of the eight primary emotions. For example, in Figure 1b, love, defined as a mixture of joy and trust is placed between them.



(a) Russells Dimensional Model [29]

(b) Plutchik model [28]

Fig. 1: Examples of emotion models

3 Main visual cues in comics

In this section, we review the tools employed by artists to convey emotion to their characters. They can be divided into two main groups : body related signals and artificial signals.

Body related signals include facial expression and body gesture. Emotion recognition through these two signals have already been studied for years on captured photos or videos. As drawings of human and human-like characters tend to be inspired by the mechanisms and the rules of real bodies, we introduce some trends found in the literature for both topics.

Contrary to body related signals that may be associated to the way real bodies work, artificial signals are invented tools, each one having a specific function. Hence, these signals have almost no equivalent in the real world. Artificial signals include the shape of speech bubbles, symbols or even stylized effects on the background. We also included the lighting effects because they belong to stylistic effects that have no equivalent in real contexts.

3.1 Facial Expression Recognition

Facial Expression Recognition or Facial Emotion Recognition (FER) aims at the identification of the emotion felt by a subject only through facial cues. On photos, these cues are visible through specific facial movements and muscular actions. In the psychology field, Ekman [10] defined the Facial Action Coding System (FACS). In this system, muscles contractions and releases that create specific movements are coded as Action Units (AU)(Fig. 2). Each emotion then correspond a specific combination of AUs. While useful for social sciences, these features are not easily obtainable for computer vision tasks. Advances in automated FER are essentially based on machine learning techniques, often applied on raw images or videos.

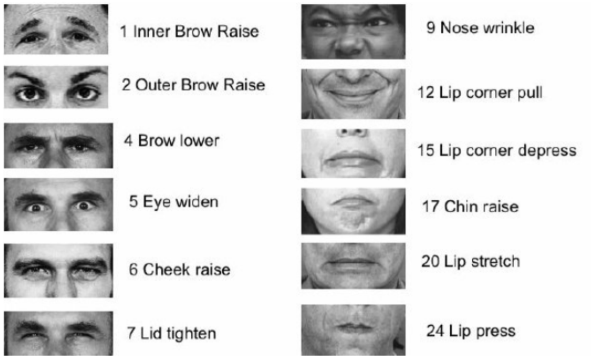


Fig. 2: Examples of Action Units (AU), extracted from [38]

Literature on FER FER on images can be generally split into multiple phases :

- pre-processing : if the image was captured in the wild, the first step is an alignment step in order to get the focus on the face. This step can be done with a traditional face detector [34] or a deep learning-oriented one such as Faster RCNN [31] or DETR [5]. In the case of photos, a pose or lighting normalization can also be applied ;

- feature extraction
- classification

With deep learning techniques, the two last stages are often merged. The basic approach is to train a standard classification network such as ResNet [13] on the target dataset. Some approaches aim at transferring knowledge from other datasets. Ng et al. [23] experimented multiple fine-tuning strategies based on FER2013 dataset on pretrained classification networks in order to improve the emotion classification task on the smaller dataset EmotiW. Knyazev et al. [17] claimed that networks pretrained on face recognition tasks, for example on Facenet dataset, tend to produce better results after fine-tuning on FER because the network has learned to extract identity specific features from the first training that are useful for the final task.

Specifically designed blocks have also been developed in order to improve performance on FER task. For example, Zhao et al. [39] added to a feedforward network a new branch that takes the feature map of the backbone to generate a weighting map that is applied to this same feature map. Moreover, the network allows the use of a handcrafted mask in order to nullify the feature maps locations that fall into background. Supervised Scoring Ensemble (SSE) [16] defines three supervising blocks that are connected at different stages of a networks (shallow, intermediate and deep layers). The outputs of the supervision blocks are concatenated and processed by a fully connected layer to deliver the emotion output.

While facial muscles movements tend to be similar between individuals for each expression, each morphology has its own specific features that could affect the perception of some facial action. Hence, works like [21] developed multitask networks for emotion recognition and person re-identification in order to estimate features that fit better to the facial attributes instead of finding a common model for all morphologies.

Particularities of comics While this field is heavily investigated on photos and videos of real peoples, research on emotion recognition on drawings and comics is a lot scarcer. Most of the developed methods focus on multimodal approaches [25], with an important contribution of text. However, to the best of our knowledge, no published methods work exclusively exclusively with visual elements. First, there are few publicly available annotated datasets on comics that only target the detection of structural elements. As FER mostly relies on machine learning approaches, the availability of data is critical. Secondly, drawn characters can be illustrated in various ways, depending on the author style. Even if the most critical facial elements tend to be coherently placed on the face, visual features may vary on multiple aspects : shape, size, level of details... Nevertheless, facial actions tend to be similar between photos and drawn characters for the same expression as shown in Fig. 3.

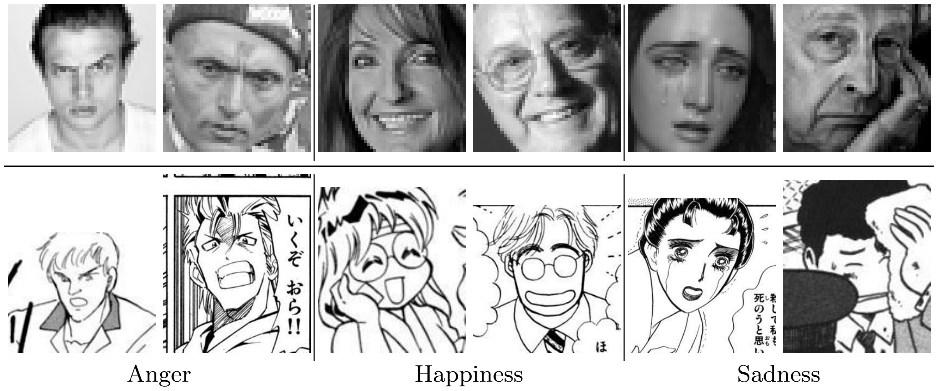


Fig. 3: Expressions on real people and drawn characters. Top images come from the FER2013 dataset, bottom images were extracted from Manga109 dataset (In order ©Nakamura Chisato, ©Shirai Sanjirou, ©Shirai Sanjirou, ©Konohana Akari, ©Ide Chikae, ©Ito Shinpei).

3.2 Pose Estimation and Action Recognition

While most of the visually perceptible information comes from facial expressions, body movements can also determine how the character is feeling. Most of the body gestures have functional purposes, for interacting with the environment for example or supporting the ideas of a speech, some of them tend to reflect directly the emotional state (Fig. 4). The topic of body gesture recognition and pose estimation has been less handled for the goal of emotion recognition than its facial counterpart. Hence, there are fewer data and methods developed on this topic, even on real acquisitions. Body gesture alone is a less stable modality for emotion recognition because, unlike facial actions that tend to have universal meanings [9], their meaning can greatly vary between cultures. For example, a thumb up can represent a validation gesture in some geographical areas and an offensive sign in other ones.

Literature on Emotion recognition from Body Gesture on real images/videos Most of the methods deal with temporal data as they do not only take the peak gesture into account but also analyze the motions (timings, intensity...) of the body parts. Hence they cannot be easily adapted to comic analysis as each frame is a snapshot of a scene.

Historically, like FER, body gesture analysis required preprocessing. The body is first detected and centered with standard detection methods. Then the different elements of the body are detected. On this step, two main paradigms exists : part based models and kinematic models. In the first one, body parts are detected independently without constraints. In the second one, the body is represented as a set of interconnected joints in order to reproduce the human body kinematics and constraints.

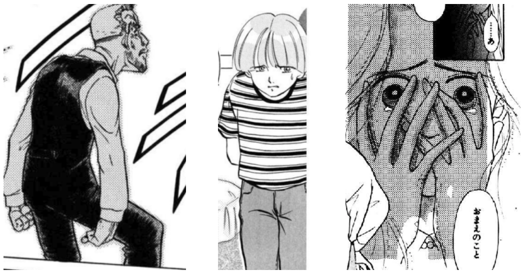


Fig. 4: Expressive body gestures. (Left) Head forward, clenched fists, ready for confrontation©Deguchi Ryusei ; (Middle) Hands behind, head dropped, feeling of fear/uncertainty ©Taira Masami ; (Right) Hands on face, shocking event ©Konohana Akari

Earlier works proposed multiple inputs types for their classification algorithms, most of them relying on geometric or dynamic values such as hands relative positions from a defined default pose were used for the upper body [12] or distances, accelerations, angles computed through upper body joints [33]. Neural networks fostered the use of learned features instead of handcrafted features. Barros et al. [1] exploited CNN in order to get more expressive features of the upper body and Botzheim et al. [2] used spiking networks to encode temporal events.

Regarding the output, a common trend was to simplify the emotion recognition problem. Glowinski et al. [11] chose to represent emotions according the Russells dimensional model. However, for the output of the recognition system, emotions that belong to the same quadrant were merged together.

Most of the time, body gesture signals are correlated with the ones delivered by the other communication media (speech, face). Consequently, a part of the literature aimed at data fusion problem in order to benefit from their complementarity, for example by associating audio and visual information through late fusion [26]. Caridakis et al. [4] first processed separately face and upper body images then compared the effects of middle and late fusions for the studied task.

Obstacles with comics The obstacles defined for FER also occur for pose estimation. Characters can have various shapes depending on the author's choices and the target audience. From realistic proportions to very stylized representations such as "Chibi" style or "Super Deformed" style, the main difficulty consists in identifying the different body parts across drawing styles and finding what is in common between characters drawn by different artists. Moreover, artists have to decide which type of shot to use for each frame. This choice can be determined by the mood they aim at or forced by limited space issues. Chosen shots must be as informative as possible. As faces convey most of the emotional information, the rest of the body tend to be less represented, leading to less data for learning algorithms.

3.3 Symbols

In order to illustrate or empathise feelings for the reader, symbols are often employed on the side of facial expression and text. Called "manpu" in japanese mangas, no consensual terminology was found to name these symbols that bring additional information about the characters. In Lexicon of Comicana [35], the author Mort Walker named symbols according to their relative position from the head. Symbols replacing the eyes are called *oculama* while icons which emerge from the head are called *emanata*.

There are numerous symbols to illustrate ideas or concepts as illustrated in figure 5. Some of them are integrated to the popular culture while other find their meaning in smaller communities. Consequently, symbols cannot be understood by everyone, and have to be initially explained to fully understand the reason why they were drawn for a specific situation.

Comics are generally defined by a genre and have a target audience. These two piece of information, that can be considered as metadata, may determine choices about graphic styles. Cohn and Ehly [7] show the influence of the target audience by studying the distribution of various visual morphemes on a corpus of 10 shonens and 10 shojos mangas

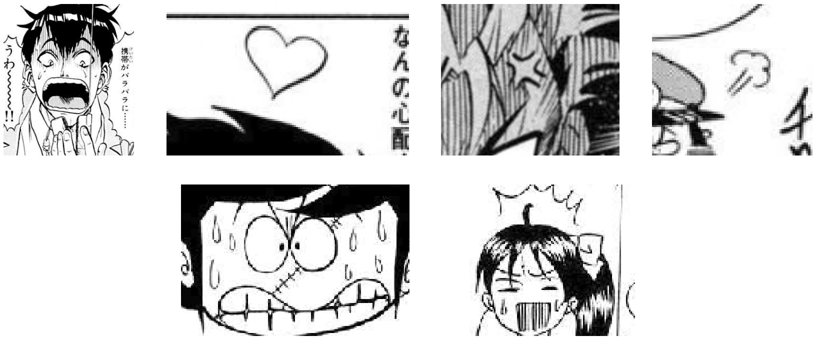


Fig. 5: Examples of "emotional" symbols. In order : Dropping lines for shock/depression (©Saijo Shinji), Heart for love or satisfaction (©Minamoto Tarou), Popping vein for anger (©Shirai Sanjirou), Smoke for annoyance (©Minamoto Tarou), Droplets for stressful situations (©Minamoto Tarou), Crown-like symbol for surprise (©Minamisawa Hishika)

Occasionally, symbols can be drawn in speech bubbles. This way, they do not link directly to the emotional state but act like intermediate representation of what is being told. For example, if the comic book targets younger audience, symbols can be replace offensive or insulting vocabulary without being explicit or semantically structured, as they carry the intention of the speaker.

Question and exclamation marks are special cases of symbols. In fact, they are punctuation marks, meaning that they have to be analysed as text. However, they can be found alone, in speech bubbles or next to the head when the character is surprised (an exclamation mark) or confused (a question mark) as illustrated in Figure 6

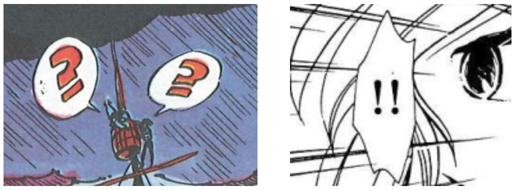


Fig. 6: Punctuation for emotion. (Left) Confusion (©Fred) (Right) Surprise (©Shimada Hirokazu)

3.4 Speech Bubbles

Speech bubbles are one of the main links between textual and visual information. While they do not directly illustrate emotion, the shape of speech bubbles provides indications on the way the speech is heard. Consequently, it would be preferable to combine them with other visual elements for emotion recognition. Figure 7 provides examples of bubbles shapes and their function. However, there is no normalization or convention about the topic, so each artist is free to choose how to draw and use speech balloon types.

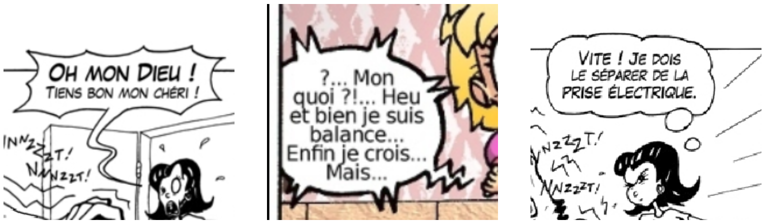
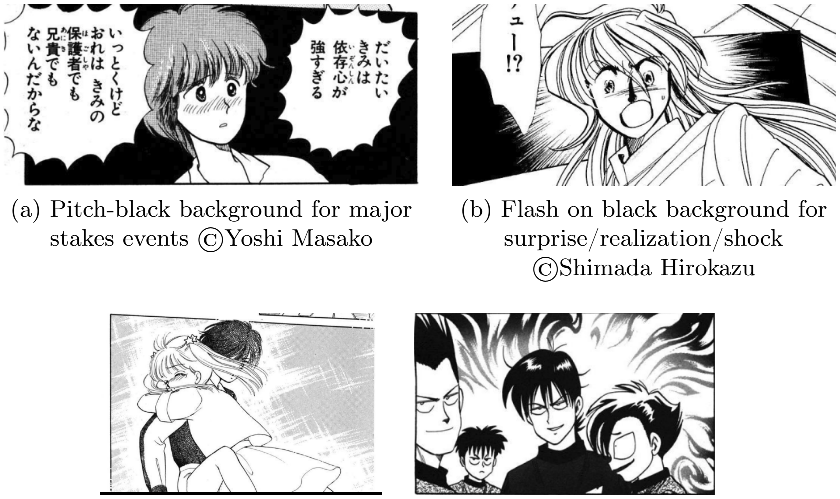

OH MON DIEU ! TIENS BON MON CHÉRI !
INNZZZT!
NNZZZT!
?... Mon quoi ?!... Heu et bien je suis balance... Enfin je crois... Mais...
VITE ! JE DOIS LE SÉPARER DE LA PRISE ÉLECTRIQUE.
NZZZT!
NZZZT!

Fig. 7: Common speech bubble shapes. (Left) Spiky bubble for loud or sudden speech ©Lamisseb (Middle) "Electric" bubble for speeches heard through electronic devices ©Studio Cyborga (Right) Cloud-shaped bubble for thoughts ©Lamisseb.

3.5 Background effects

While characters attributes remain the main tool for emotion analysis, the artists can employ background effects to highlight specific reactions. We show in Figure 8 some examples of effects used in different mangas.

Like most of the previous graphical cues, the choice of using a background effect is up to the authors. There are no explicit rules on these effects, some techniques have been heavily used to the point that they have become tacit conventions across artists and readers. However, it is not possible to understand them without learning about them beforehand as they remain culture-specific symbols.


ュー!?
だいたいきみは依存心が強すぎる
いっとくけどおれはきみの保護者でも兄貴でもないんだからな

(a) Pitch-black background for major stakes events ©Yoshi Masako

(b) Flash on black background for surprise/realization/shock ©Shimada Hirokazu

(c) Glittering background for showing the shared happiness between the two characters ©Kuriki Shoko

(d) Twisty abstract background to illustrate the malicious intents of the characters ©Yabuno Tenya, Watanabe Tatsuya

Fig. 8: Examples of backgrounds effects used in mangas.

3.6 Lighting and colors

In other visual arts such as photography or cinema, lighting is a powerful tool to create moods. The way the lights and the shadows are placed can change the perception of the viewer on the same face. In comics (and painting), lighting is completely defined by the artist. Shading management can be used to empathise feelings or moods the same way background effect can do (Fig. 9).

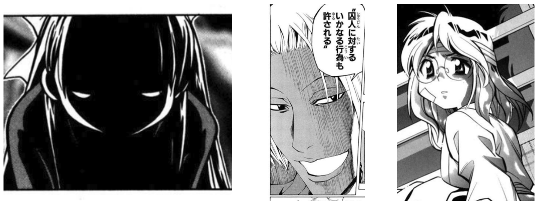


Fig. 9: Examples of specific shadings. (Left)Full shadow, no visible facial attribute for the harmful aspect ©Yagami Ken (Middle) Obscured face and smile for malicious intent ©Ishioka Shoei (Right) More detailed shading for impactful moment ©Shirai Sanjirou

In colored works, each character is illustrated with a default set of colors. However, to accentuate an effect, artists may change the default face color to another one that better matches the wanted emotion. In western comics, an angry character can be drawn with a red face, a sick or disgust one with green...

4 Robustness from real data to drawings

While speech bubbles and symbols have no real world equivalents, human like characters tend to be drawn according to real mechanisms that determine the behaviour of the human body. A drawing of an existing object is firstly a simplified representation. One may suggest that a convolutional network for object classification may return the same output with a photography and a drawing of a face as they illustrate the same concept. Hence, benefiting from larger datasets built on real images would be possible as there are few data for drawings and comics. However, convolutional networks runs in a hierarchical fashion, low level features are computed and used sequentially by higher layers to produce more complex and conceptual features. Low level features include contours, local details on textures... Those features differ a lot between drawings and photos. Consequently, these divergences are propagated through the network and generate different outputs.

However, a part of literature have begun to deal with this topic. Lagunas et al. [18] studied the effects of transfer learning on a dataset of cliparts with a VGG19 pretrained on ImageNet. Wang et al. [36] suggested that, during a standard training, local features acquire a great predictive power on the early layers, the subsequent layers then resting on their predicted hypotheses. Consequently, low level features tend to take much more importance than the illustrated concept itself. Their idea was then to penalize the predictive power in the early layers by maximizing the error on predictions computed with only low level features 10. In order to evaluate their methods, the researchers also built a dataset, ImageNet-Sketch, with only web-scraped drawings and the same classes as ImageNet dataset. With the idea of dampening the effects of details on final predictions, Mishra et al. [22] evaluated the effects of an anisotropic diffusion filtering on the input image for smoothing textures without altering sharp contours.

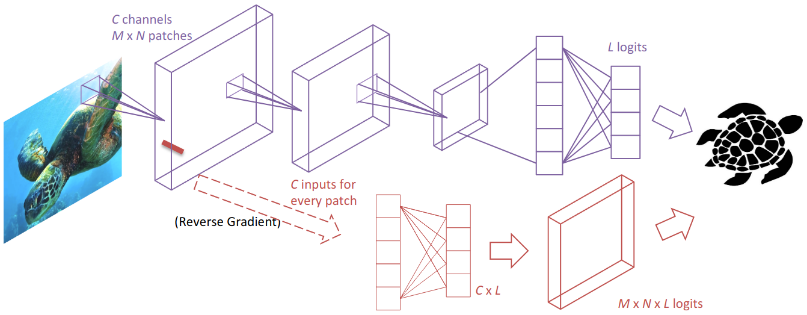


Fig. 10: Model presented in [36]. The model consists of two parts : the main global classifier (in purple), which takes into account all features locations, and a set of local classifiers (in red) for each location in early features maps. The goal is to fool the local classifiers while producing accurate predictions with the global classifier.

5 Conclusion

Emotion recognition analysis on comics is complex because multiple modalities can be intertwined in order to represent the state of mind of one character. When these ones are human-like, the face and body can by themselves provide enough information for automated processing, especially in case of realistic styles/proportions. Techniques specifically created for the comic book medium such as speech bubbles and symbols can form an important complement of information that can help to disambiguate situations. However, the tone of the work can forbid the use of "cartoony" assets. Among all the listed cues, facial information seems to be the most robust to culture changes.

We restricted the study on elements that could help to understand the state of a character during the moment of one frame. Hence, for one character, we studied the reactions to external events (other characters actions, speech...). However, the way this character affects the other was not addressed. On this topic, body gestures represent an important visual tool to analyze how characters interact between themselves. This analysis, if spread across frames, can also lead to the study of the social relationships.

Acknowledgements This work is supported by the Research National Agency (ANR) in the framework of the 2017 LabCom program (ANR 17-LCV2-0006-01) and the Region Nouvelle Aquitaine in the framework of the EmoRecCom project.

References

1. Barros, P., Jirak, D., Weber, C., Wermter, S.: Multimodal emotional state recognition using sequence-dependent deep hierarchical features. Neural Networks **72**, 140–151 (Dec 2015)
2. Botzheim, J., Woo, J., Tay Nuo Wi, N., Kubota, N., Yamaguchi, T.: Gestural and facial communication with smart phone based robot partner using emotional model. In: 2014 World Automation Congress (WAC). pp. 644–649 (Aug 2014)
3. Bradley, M.M., Lang, P.J.: Measuring emotion: the self-assessment manikin and the semantic differential. Journal of behavior therapy and experimental psychiatry **25**(1), 49–59 (1994)
4. Caridakis, G., Castellano, G., Kessous, L., Raouzaiou, A., Malatesta, L., Asteriadis, S., Karpouzis, K.: Multimodal emotion recognition from expressive faces, body gestures and speech. In: Boukis, C., Pnevmatikakis, A., Polymenakos, L. (eds.) Artificial Intelligence and Innovations 2007: from Theory to Applications. pp. 375–388. IFIP The International Federation for Information Processing, Springer US, Boston, MA (2007)
5. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020)
6. Chu, W.T., Li, W.W.: Manga FaceNet: Face Detection in Manga based on Deep Neural Network. In: Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval. pp. 412–415. ACM, Bucharest Romania (Jun 2017)
7. Cohn, N., Ehly, S.: The vocabulary of manga: Visual morphology in dialects of Japanese Visual Language. Journal of Pragmatics **92**, 17–29 (Jan 2016)
8. Dubray, D., Laubrock, J.: Deep cnn-based speech balloon detection and segmentation for comic books. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 1237–1243. IEEE (2019)
9. Ekman, P., Friesen, W.V.: Constants across cultures in the face and emotion. Journal of Personality and Social Psychology **17**(2), 124–129 (1971)
10. Ekman, P., Rosenberg, E.L. (eds.): What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS). Series in Affective Science, Oxford University Press, New York, 2 edn. (2005)
11. Glowinski, D., Mortillaro, M., Scherer, K., Dael, N., Volpe, G., Camurri, A.: Towards a minimal representation of affective gestures (Extended abstract). In: 2015 International Conference on Affective Computing and Intelligent Interaction (ACII). pp. 498–504 (Sep 2015)
12. Gunes, H., Piccardi, M., Jan, T.: Face and body gesture recognition for a vision-based multimodal analyzer. In: Proceedings of the Pan-Sydney area workshop on Visual information processing. pp. 19–28. VIP '05, Australian Computer Society, Inc., AUS (Jun 2004)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
14. He, Z., Zhou, Y., Wang, Y., Wang, S., Lu, X., Tang, Z., Cai, L.: An End-to-End Quadrilateral Regression Network for Comic Panel Extraction. In: Proceedings of the 26th ACM international conference on Multimedia. pp. 887–895. MM '18, Association for Computing Machinery, New York, NY, USA (Oct 2018)
15. Ho, A.K.N., Burie, J.C., Ogier, J.M.: Panel and Speech Balloon Extraction from Comic Books. In: 2012 10th IAPR International Workshop on Document Analysis Systems. pp. 424–428. IEEE, Gold Coast, Queenslands, TBD, Australia (Mar 2012)

16. Hu, P., Cai, D., Wang, S., Yao, A., Chen, Y.: Learning supervised scoring ensemble for emotion recognition in the wild. In: Proceedings of the 19th ACM International Conference on Multimodal Interaction. pp. 553–560. ACM, Glasgow UK (Nov 2017)
17. Knyazev, B., Shvetsov, R., Efremova, N., Kuharenko, A.: Convolutional neural networks pretrained on large face recognition datasets for emotion classification from video (Nov 2017), `http://arxiv.org/abs/1711.04598`
18. Lagunas, M., Garces, E.: Transfer Learning for Illustration Classification. Spanish Computer Graphics Conference (CEIG) p. 9 pages (2017)
19. Li, L., Wang, Y., Tang, Z., Gao, L.: Automatic comic page segmentation based on polygon detection. Multimedia Tools and Applications **69**(1), 171–197 (Mar 2014)
20. Liu, X., Li, C., Zhu, H., Wong, T.T., Xu, X.: Text-aware balloon extraction from manga. The Visual Computer: International Journal of Computer Graphics **32**(4), 501–511 (Apr 2016)
21. Meng, Z., Liu, P., Cai, J., Han, S., Tong, Y.: Identity-Aware Convolutional Neural Network for Facial Expression Recognition. In: 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017). pp. 558–565. IEEE, Washington, DC, DC, USA (May 2017)
22. Mishra, S., Shah, A., Bansal, A., Choi, J., Shrivastava, A., Sharma, A., Jacobs, D.: Learning Visual Representations for Transfer Learning by Suppressing Texture. arXiv:2011.01901 [cs] (Nov 2020)
23. Ng, H.W., Nguyen, V.D., Vonikakis, V., Winkler, S.: Deep Learning for Emotion Recognition on Small Datasets using Transfer Learning. In: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction. pp. 443–449. ACM, Seattle Washington USA (Nov 2015)
24. Nguyen, N.V., Rigaud, C., Burie, J.C.: Comic Characters Detection Using Deep Learning. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). pp. 41–46. IEEE, Kyoto (Nov 2017)
25. Nguyen, N.V., Vu, X.S., Rigaud, C., Jiang, L., Burie, J.C.: Icdar 2021 competition on multimodal emotion recognition on comics scenes. In: International Conference on Document Analysis and Recognition. pp. 767–782. Springer (2021)
26. Noroozi, F., Marjanovic, M., Njegus, A., Escalera, S., Anbarjafari, G.: Audio-Visual Emotion Recognition in Video Clips. IEEE Transactions on Affective Computing **10**(1), 60–75 (Jan 2019)
27. Pang, X., Cao, Y., Lau, R.W., Chan, A.B.: A Robust Panel Extraction Method for Manga. In: Proceedings of the 22nd ACM international conference on Multimedia. pp. 1125–1128. ACM, Orlando Florida USA (Nov 2014)
28. Plutchik, R.: 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)
29. Posner, J., Russell, J.A., Peterson, B.S.: The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. Development and psychopathology **17**(3), 715–734 (2005)
30. Qin, X., Zhou, Y., He, Z., Wang, Y., Tang, Z.: A Faster R-CNN Based Method for Comic Characters Face Detection. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). vol. 01, pp. 1074–1080 (Nov 2017)
31. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28** (2015)
32. Russell, J.: A Circumplex Model of Affect. Journal of Personality and Social Psychology **39**, 1161–1178 (Dec 1980)

33. Saha, S., Datta, S., Konar, A., Janarthanan, R.: A study on emotion recognition from body gestures using Kinect sensor. In: 2014 International Conference on Communication and Signal Processing. pp. 056–060 (Apr 2014)
34. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001. vol. 1, pp. I–511–I–518. IEEE Comput. Soc, Kauai, HI, USA (2001)
35. Walker, M.: The Lexicon of Comicana. iUniverse (2000)
36. Wang, H., Ge, S., Lipton, Z., Xing, E.P.: Learning robust global representations by penalizing local predictive power. Advances in Neural Information Processing Systems **32** (2019)
37. Watson, D., Tellegen, A.: Toward a consensual structure of mood. Psychological bulletin **98**(2), 219 (1985)
38. Wu, T., Butko, N.J., Ruvulo, P., Bartlett, M.S., Movellan, J.R.: Learning to make facial expressions. In: 2009 IEEE 8th International Conference on Development and Learning. pp. 1–6. IEEE (2009)
39. Zhao, S., Cai, H., Liu, H., Zhang, J., Chen, S.: Feature selection mechanism in cnns for facial expression recognition. In: BMVC. p. 317 (2018)