A COMPUTATIONAL MODEL OF VISUAL NARRATIVE COMPREHENSION

by
Yi-Chun Chen

A Oral Prelim Proposal

Submitted to the
Department of Computer Science
North Carolina State University
In partial fulfillment of the requirement
For the degree of
Doctor of Philosophy in Computer Sciences
at
North Carolina State University
Jun 24, 2021

Chair: Dr. Arnav Jhala

Committee Members:
Dr. Ben Watson
Dr. Noboru Matsuda
Dr. Tianfu Wu

© 2021 Yi-Chun Chen

Abstract

Yi-Chun Chen
A COMPUTATIONAL MODEL OF VISUAL NARRATIVE COMPREHENSION
2020-2021
Dr. Arnav Jhala
Doctor of Philosophy in Computer Sciences

Comics present a rich media form for the study of human multi-modal communication that incorporates cultural narratives in a predominantly visual but discrete form compared to videos. In this work, we propose a computational model for the analysis of multi-modal narratives based on two datasets of western comics (Comics from Iyyer et al. 2017) and Japanese manga(Manga109 from 2017). We add annotations to the datasets by automatically generated visual narrative labels such as reading order analysis and panel transition sequences based on visual narrative comprehension theories proposed in the Cognitive Science literature(SPECT and PINS). We construct a hierarchical LSTM-based model for combining multiple modalities. Model parameters are evaluated using closure tests, commonly used to test human narrative comprehension, in which the model predicts the last frame in a sequence of story frames. Our contributions are threefold. First, we provide a systematic benchmark dataset of visually dominant multi-modal narratives. Second, we provide a baseline model of narrative comprehension based on cognitive science theories. Finally, we provide a rigorous evaluation metric for multi-modal narrative comprehension.

Table of Contents

Table of Contents (Continued)

List of Figures

List of Tables

Chapter 1

Introduction

Visual narrative (also visual storytelling) tells a story primarily through visual media; the forms of visual media have large diversity, such as illustrations, video, photography, and so on. Visual storytelling is about more than just using pictures to illustrate a story. It can be assisted with text, sounds, and graphical symbols hence having various branches. Since ancient times, transporting ideas through visual forms has been used; the cave paintings that vehicles meanings were even earlier than the invention of text. After which came animal hunt stories – all the way to today's visual arts such as picture books, comics, movies, video games, and so forth share some common advantages.

The first benefit of visual narrative is the effectiveness–graphical content's direct affinity with the visual representation of ideas that authors want to transport; they, therefore, can give audiences concrete images about described things. The second pros of using visual forms are efficiency. Not only have cognitive scientists shown that our brains respond first to visual input far faster than absorbing text or audio, but visual forms also compress complex details in one image and conveys them at a glance. The content encompasses complicated relations between scenes and the entities and preserved the very details of objects, spatial and temporal. Finally, visual narratives are excellent on the influences. Neurologists have documented that images are a direct pathway to our emotions that influence decision-making. The combination of color, light, and shapes create strong visual impressions that humans can feel the precise atmosphere and sentiment changes in the depicted story.

This helps that authors precisely transport the content they want to communicate. Therefore, especially nowadays, information grows rapidly; more and more stories told through visual media have been published through internet platforms. People love to use vi-

sual media to efficiently transport a large amount of information and their own idea clearly and then published to the platforms. Social media such as Instagram, Deviant art, Twitter, Pixiv, and other websites are full of either self-created visual media or visual stories published commercially.

According to this tide, automatically understanding visual content becomes a crucial issue. Topics like categorizing visual narratives to assist searching, exploring, and recommendations according to content that fits the audience's interests are all subsets of the area. Automatically understanding visual content faster has become more important than ever hence popped up as a popular topic in Artificial Intelligence. To assist people in finding suitable content, computer models or systems needs to have the ability that interprets the content detail first. Moreover, because visual forms used as vehicles of ideas do not always show up solely. When the idea they want to tell is more complicated such as stories, procedures, etc., they are more likely to be a sequence of images used to describe a complete event with a beginning, changes, and an ending in its context. Thus, the understanding that combines both visual and narrative attracted people's attention.

Speaking of visual media that transport complex ideas such as stories, in other words, the combination of visual media and narrative, the comic is the most famous form(section 1.1). It has been widely created and published through social media. This work aims to build a computational model to comprehend visual narratives; we then set our target specifically to comics.

1.1 Comic as a Visual Narrative Media

Comic (also manga) is a popular visual media that refers to sequential art and graphic novels, widely used to tell stories. It is even older than films. It is a particular format that integrates text expression, visual elements, and symbolic abstractions. The two most famous branches are western comics or Japanese Manga. They share many characteristics. They both have intertwined text and image to compose the story content, abstract

Table 1

Correspondingly, the examples from a western comic (left, book name, and author) and Japanese Manga (right, book name, and author).

Western Comic	Japanese Manga
	

symbols or lines to emphasize atmosphere or sentimental expression, variations on dialogue bubbles, and panel shapes to guide readers' attention. Also, the features are nested. Besides the detail in a single panel, comics also valued the layouts to construct panels into a fluent order hence put effects to the balance on a whole page. They use the well-designed composition of an image and the structuralized framing to guide readers' attention, hence achieving diverse discourses. While sharing common features, they also have many significant differences. In western comics, colorful image expressions are common; they often have relatively rigid or grid-like layouts. Textboxes are in a rather non-significant position of panels, such as the top side of a panel. On the contrary, Japanese manga usually has grayscale image expression, incorporate the blank-leaving technique from certain art styles, thus controlling the tempo, relatively flexible textboxes' positions to lead eye path. There are also some variants on the abstract expressions of sounds and effects such as atmosphere stipples, focus lines, motion lines, etc.

These common or various illustrations make comics interesting but complex visual narrative forms. The images in Table 1 give examples from a western comic and a Japanese manga correspondingly.

1.2 Comic Comprehension

The understanding of comics is a type of multimodality task, which is different from realizing other visual forms because it has multiple literacies within one medium. It involves the understanding of visual impressions and the interpreting of text representations, symbolic graphics, etc. The process, moreover, includes the reasoning of underlying relations such as casual, temporal, or spatial because of comics' sequential nature. Therefore, to study this complicated visual media, researchers from different fields approach the topic from many aspects. Understanding comics lies at the intersection of Cognitive science, Computer vision, Language processing, and Artificial intelligence.

To cracking the comic understanding tasks, cognitive scientists proposed models that correspond to the human's comprehension process that combine the steps from visual stimulation to attention shifting, then to the interactions between working memory and long-term memories [1, 2, 3] . Parallelly, theories related to how the visual representation of comics is composed sequentially also has been advanced. Other cognitive scientists concluded the narrative structures that decompose comic sequences and then proposed the grammar bridges the gutters between consecutive panels from existing comic sequences[4, 5, 6]. Meanwhile, the methods that comics use to organize visual content guide readers' attention and imply the spatial and temporal shifting through still images were summarized as theories [7, 8, 9]. Furthermore, the rules that represent the relations inside comic panels as hierarchical structures and graphs to construct events that happen in the story were also proposed by other works [10]. In addition to the discussions around comic elements, some computer scientists developed computational models based on that memory model in a neural network to capture sequential context representations of comic panels [11]. We will explain the details of these previous works in the literature review section (chapter 3).

Although previous research considered different aspects separately, we believe that methods from various points of view may need some integrations to tackle this interdis-

ciplinary topic better. Therefore, in this project, we propose and develop a computational model based on previous cognitive theories and integrate comics analyses to combine their advantages to understand comic content better.

1.3 Decomposing Comics

According to the previous research around comic understanding and observations, we can then define the elements needed for comic comprehension tasks and the composition of comics. They are listed below, and the sample images of composition elements are in Table 2 :

Book: it refers to the combination of comic pages, which usually consists of one or more titles, in other words, one or more stories; it is the common basic unit that dataset used to divide comic sets. The first row in the Table 2 shows the cover pages of comic books.

Page: The combination of comic panels use a different layout to frame the positions of panels.

Sequence: This is also a combination of comic panels; A sequence of panels refers to consecutive panels sort by reading orders with an arbitrary length and is different from a comic page, mainly by no page layout. However, in our discussions and experiments, the number of panels on the same page usually equals the length of the corresponding comic sequence. The sequence usually consists of panels from a single page since we take comic pages to divide the comic book's story. the third row of Table 2 shows the sequence that is extracted from a comic page, where the reading order of Western comics is from left to right; on the contrary, Japanese manga usually read from right to left.

Panel: The basic narrative unit that comic employed to convey messages. It possibly contains either visual representations, text representations, or both in the same frame. Each of them bears a short time slot of the story through a still image. The fourth row of the Table 2 gives examples of a comic panel from different types of comics.

Entity: The entities denote objects in a panel; it refers to characters, items, things in the

background, etc. In the fifth row of Table 2, the entities are labeled with red rectangles.

Textbox: The balloons that display the text messages in comics. They have diverse shapes that correspond to the atmosphere of content or what characters say or think. In the fifth row of Table 2, the textboxes are labeled with blue rectangles.

Scene: This denotes the location information that the story event happens. It is the background layer of a panel image. The sixth row of Table 2 gives examples of scenes in comic panels; the entities and textboxes are masked out.

Sound symbols: In comics, because there is no audio in this visual media, the sounds that happen when characters perform some actions such as laughing sound, crying sound, colliding sound, and so on will usually be presented as text images that mimic the sound. The seventh row of ref shows the sound symbols, which are marked with red rectangles.

Graphical symbols: In comics, some graphical symbols emphasize the effects of characters' actions, speed of motions, or an abstract metaphor that indicates changes of character emotion states and atmosphere such as nervousness, angriness, etc. In the seventh row of Table 2, the symbols are marked with blue rectangles. In the western comic panel, the focus line emphasizes the sound of opening the box, whereas the symbol in that manga panel emphasizes that the character closes the book.

Narrative Structure: The narrative structure here means how a story event is structured and told through a comic sequence.

Inter-panel relations: The inter-panel relations refer to the link that bridges the gap between two consecutive panels. Since stories in comics are told in discrete time frames rather than continuous frames, there are some things omitted and can only be inferred according to context. The link that describes the gutters between panels or says how the two-panel are connected, is the inter-panel relation.

Intra-panel relations: Things inside a panel can be described by intra-panel relations, such as the relations between an entity and another entity or the relations between an entity and the scene. For example, in the first image in the fourth row of ref, the character (entity)

sits (action) in front of (relation) his house (scene).

1.4 Dataset

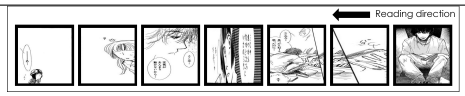
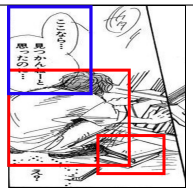
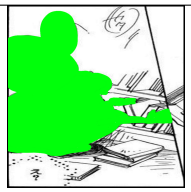
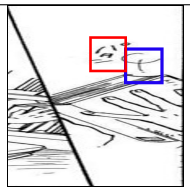
To support academic research about comics and fit the needs of data-hungry machine learning models, Iyyer et al. and Matsui et al. proposed two datasets for western comics and Japanese manga correspondingly in 2017–COMICS and Manga109. They both collected a massive amount of comic data from commercially published comic books.

1.4.0.1 COMICS This dataset includes 3,948 publicly available comic books published during the "Golden Age" of American comics (1938–1954) with diverse visual styles and genres, divided into 1.2 million panels in total. It was first introduced into the Computer Vision field by Iyyer et al. [11]. COMICS contains separated panel images plus around 500 raw pages that preserve the original layouts of comic style. This dataset provided annotations of text in comics generated through Optical Character Recognition (OCR) and the positions records of textboxes.

1.4.0.2 Manga109 The Manga109 dataset includes 109 titles of manga, which is a specialized art style of Japanese comics. It was first introduced into the Multimedia field by Matsui et al. in 2017 [12, 13]. The dataset consists of manga titles drawn and published in commercial manga magazines between the 1970s and 2010s by professional manga authors. Each title includes around 196 pages on average, with a total of 21,142 pages. The dataset includes entire pages for each manga title. It also includes annotations of textbox position, text content, and positions of character bodies and faces on the page. Reading order in manga is different from comics. Panels are to be read right to left and top to bottom. The mangas cover various kinds of genres: humor, battle, romantic comedy, animal, science fiction, sports, historical drama, fantasy, love romance, suspense, horror, and four-frame cartoons.

Table 2

Examples for each composition elements of comics.

	Western Comic	Japanese Manga
Book		
Page		
Sequence		
Panel		
Entities & textbox		
Scene		
Abstract symbols		

Chapter 2

Research Questions

To analyze the comic comprehension factors hence develop a comprehension model, the details of the topic need to be studied. Aiming to understand the content of comics well, we then ask four questions: What features in comics related to or influence understanding? What are the needed components for developing a comic comprehension model? How to integrate various aspects of comics to help comprehension? How to evaluate the comprehension model?

2.1 What features influence comic understanding?

As we described in the previous sections, comic comprehension is a complex problem. Besides the difficulties that originally exist in visual content understanding topics, the multi-modality of comics brings new aspects that need to be considered when exploring the topic. Furthermore, the composition elements of comics are complicated as well. Their sequence nature transforms the discussion from the semantic of a single still image to narrative reasoning, which involves causal, temporal, and spatial changes of objects. In addition, the content and the discourses of the stories and how stories are presented visually play an important role in comic comprehension. Therefore, we ask the question that "What features influence comic understanding?" to decompose features existing in this visual form and conduct experiments to observe whether or how these affect comprehension.

2.2 What are the components of a comic comprehension model?

After finding out the features that influence comprehension, the next question we are going to ask is that "What are the components of a comic comprehension model?" This

question is the basis of developing a comprehension model. Although the theories are many and having their own advantages in previous related work, they did not comprehensively consider distinct aspects of comic comprehension.

process that how human comprehends visual content, but how to transform it into a computational model remains an issue. Comic theories that decomposed how comics are formed and what technique authors use to guide readers' attention explained the underlying purposes that each comic panel is designed and why the content is shown through the representation we have seen. However, formulating the theories to make them fit into the computational model to achieve similar explanations automatically is also a problem. Narrative systems that described panel content in a structuralized way and the neural network that retrieved text and image mathematical representation also have insufficient parts. Although the former studied the narrative aspect of content, the scalability was bounded by that extracting details from panel content wasn't an automatical process. On the contrary, the latter handles a large amount of data at once but failed to ignore comics' narrative nature, which plays a significant role in understanding stories.

Therefore, when asking this research question, we aim to realize how the components in each theory fit into the cognitive process. Meanwhile, this question is to understand whether any elements should involve the comprehension process but be missed previously.

2.3 How to integrate various comics features that benefit comprehension?

After knowing the needed components, a question comes, "How to integrate various comics features that benefit comprehension?" how we embed the component from various comic theories into the comprehension model to simulate the cognitive process of comprehension? Follow the visual narrative cognitive models proposed by Magliano et al. and Cohn et al. [1], we plan to use it as the fundamental framework since it has already formalized the comprehension process. Then, we will fit the narrative event models and the neural network vision model into the framework and then adjunct the missing gap to

implement the cognitive model.

2.4 How to evaluate the comprehension model?

To estimate the results that our model can achieve, the final question will naturally be ”How to evaluate the comprehension model?” Metrics and tasks that assess the model’s ability are also important. They are the exams to check whether the goal is accomplished to not. Following the evaluation metrics from other content understanding researches, either answer questions about context or predict the followed content are common ways to test how well a comprehension model or system performs. Given that the sequential characteristic is one of the most important parts of comics, we will use cloze-tasks as the metrics to measure the model’s ability. The cloze-tasks ask models to predict possible next content based on the knowledge and the understanding it acquired from reading the previous context.

Chapter 3

Literature Reviews

This section will introduce the details of related works toward the same topic. It will cover four parts; the first part will describe what theories were proposed to frame how comics construct narratives. Next will be the conceptual models that formalized the cognitive process of narrative comprehension; The third part will describe the computational methods where the goal is visual narrative comprehension. Lastly, we will explain the differences between our proposed model and other methods and recount how our planned work related to comics theories and how to integrate the ideas proposed by other previous works.

Other related works which not directly link to comic comprehension but relate to our findings were scattered in chapter 4, chapter 5, chapter 6, and chapter 7.

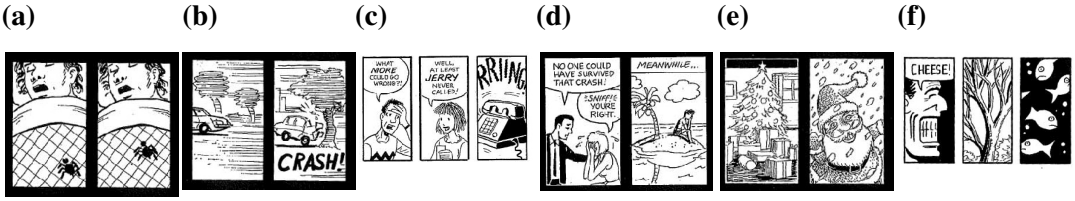
3.1 Theories of Narrative in Comic Sequences

Comic artist Will Eisner first defined ”sequential art” as art forms that deploy images in a specific order for graphic storytelling in his publication ”Comics and Sequential Art” for comic studies. Then, Scott McCloud’s book ”Understanding Comics: The Invisible Art” followed Eisner’s idea to define comics in terms of its essential defining attribute: ”sequence.” He extended the concept of sequential art by presenting his own sophisticated definition of ”comics” that pictures and other images are juxtaposed in deliberate sequence to convey information or produce an aesthetic response to readers [7, 9, 14].

In addition, he defined the inter-panel transitions (discussed in the followed subsection) to capture the operations of visual representation of comic panels toward understanding how they operate on viewers’ attentions. After that, Cohn et al. proposed the theory of Visual Narrative Grammar (VNG) in 2013 that packages meaning at a discourse

Figure 1

Annotation of inter-panel transitions based on Understanding Comics *[8].*



level. It assigns categorical roles to images based on prototypical correspondences with a conceptual structure of meaning [15].

3.1.1 Reading Between the Panels

Key to McCloud's theory is the space between the panels, the "gutter." He said that gutter bears much of the magic and mystery at the heart of comics because comic panels fracture time and space. They offer staccato rhythm of unconnected moments. Between the panels, the relations and changes between story events' states and content are committed. Readers can connect these fragment moments and mentally construct a continuous, unified reality–through closure

In order to develop an understanding of the narrative presented as the juxtaposition of comics panels, he then describes six possible patterns for the juxtaposition of comics panels. And, Iyyer et al. employed the patterns to annotate the feature of the western comic dataset (COMICS); we also did the same analysis for the manga dataset (Manga109) in our first findings (chapter 4). Moreover, the transitions that model the changes between panels was applied in our side project (chapter 7) as a technique to adjust the generated comic visual content. The Figure 1 gives examples of transitions from *Understanding Comics* [8] :

Moment-To-Moment: Moment-to-moment transition captures small changes with very little to no passage of time in the story world between the two panels. This requires little closure on the part of the reader because the state changes minimally, like successive

frames of a film. An example of this transition is shown in Figure 1(a).

Action-To-Action: This transition shows a subject doing an action over a sequence of panels. In some cases, actions begin and end at different frames with a number of intermediate panels. In our first finding, when annotating this type of transition, we are only looking at pairs of panels in our annotations so we consider this annotation whenever any part (beginning, middle, end) of the same action is present in both frames. An example of this transition is shown in Figure 1(b).

Subject-To-Subject: This transition indicates change or introduction of subjects between panels. When annotating this type of transition for mangas, we found that authors often use a common method to guide the reader's attention: to change focus between a group of subjects. Because the focus changing matches the feature of this type of transition–juxtaposition while staying within a scene–where it is necessary to have readers' involvement in interpreting the transitions, we also consider these changes as members of this category even if the last focused subject may still exist in the next panel. In other words, as long as the focus is transferred from one subject to another, even if the previous subject remains in panels, it is a subject transition. An example of this transition is shown in Figure 1(c).

Scene-To-Scene: This transition transport the reader across a significant distance of time and space between two panels, which usually required the readers' deductive reasoning to link up the scenes. An example of this transition is shown in Figure 1(d).

Aspects-To-Aspects: This transition indicates a shift to an abstract scene similar to an interlude to indicate passing the time, such as a falling leaf. When annotating this transition for the manga dataset, manga uses this important category to slow down or em-

phasize characters' actions, feelings, and moments of an event to change the view angles of the same subject. An example of this transition is shown in Figure 1(e).

Non-sequitur: This transition indicates that there is no perceivable relationship between panels. An example of this transition is shown in Figure 1(f).

3.1.2 Visual Narrative Grammar

Neil Cohn separated the narrative and meaning for sequential images and formalized the narrative structures into theory Visual Narrative Grammar. He mapped comic sequences to event structures, and each panel is an aspect of the event. In other words, an event might extend across several panels, and the whole sequence formed the presentation of events. The sequence's narrative structure can be described by basic narrative categories and their canonical narrative arc [4]. In our planned method, the VNG plays the role that helps to analyze the structure of the target sequence (chapter 8).

The basic units of the VNG are five categories that composed a basic canonical schema. By extending the basic schema with recursively elaborated canonical arcs, more complex structures can be generated. The five categories are:

- **Establisher (E)** - Settings characters, place, etc. without action involved.
- **Initial (I)** - Beginning of story arc. Or the start of an action or an event.
- **Prolongation (L)** - Middle state of the story arc. Extend an action.
- **Peak (P)** - The highest story tension. the end of an action.
- **Release (R)** - releases the tension. The consequence of an action.

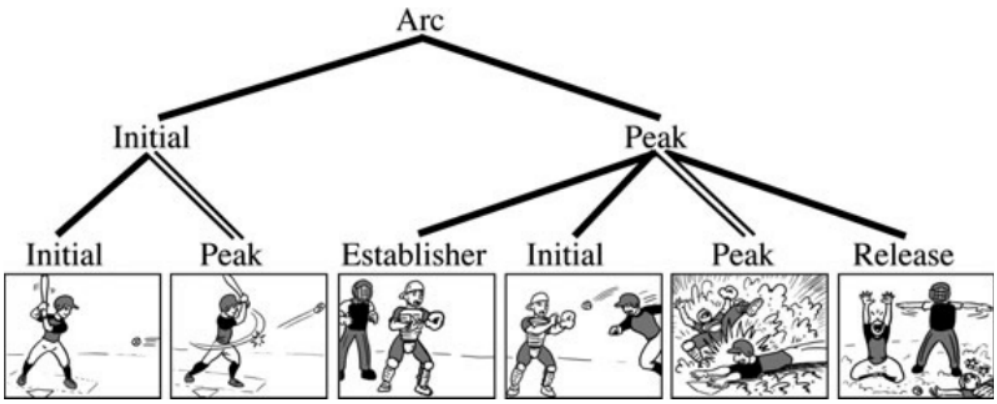
And the five categories form basic phases through linearly ordering the five categories:

Phase (Establisher) - (Initial(Prolongation)) - Peak - (Release)

The parentheses above imply these categories are optional. The importance of each category from high to low is Peak, Initial, Releases, Establishers, Prolongation. Furthermore, more complex combinatorial structures can be conjunct through several patterns of embeddings. The Figure 2 shows an example of how the narrative structure be obtained by expanding the canonical narrative arc.

Figure 2

Narrative structure example



In this example, the phase in the n basic canonical schema is ”Initial - Peak”. Then, the Initial was expanded through embedding the canonical narrative arc ”Initial - Peak” again, whereas the Peak in the first layer was embedded with the arc ”Establish - Initial - Peak - Release”.

3.2 Conceptual Models for Visual Narrative Comprehension

Toward comic comprehension, since visual narrative processing is a fundamental part of human cognition, cognitive scientists Cohn and Magliano developed conceptual models to frame the cognitive process. Magliano proposed the Scene Perception & Event Comprehension Theory (SPECT) and its theoretical framework. The model described the whole process from image stimulus to readers cognitive process. It distinguished the stim-

ulus features and front-end and back-end cognitive processes involved in the visual event and narrative cognition [1]. On the other hand, Cohn developed another conceptual model named Parallel Interfacing Narrative-Semantics (PINS) [3]. It modeled sequential image processing that frames the interplay between semantics and narrative structure.

Our target method will integrate the semantic processing of the visual narrative of PINS with the cognitive process that SPECT presented.

3.2.1 Scene Perception & Event Comprehension Theory (SPECT)

The Figure 3 shows the framework of SPECT theory. The SPECT starts from the stimulus that the eye can see from visual images of varying degrees of complexity composed in sequence. Due to the differences in complexity, the salience of primitive visual features guides the viewer's attention. It decides what potential information that viewer can get and is likely to influence how front-end and back-end processing the information.

The front-end processes of SPECT happens during single eye fixations, which extracting content and activate semantic representations. Depending on the detail degree of the stimulus, the front process performs attentional selection: dividing the information between the gist of the whole scene and the detail of entities inside the scene. On the other hand, the back-end processes the extracted content and then operates the information to construct an event model. Information from images is fed to the back-end, and the captured sequenced event helped construct and maintain the current event model, which represents what is happening now. If the event boundary is detected, in other words, the end of one event, the current event model will be stored into Episodic memory. When constructing the current event model, it cooperates with the Semantic memory (also prior knowledge) and the Episodic memory (stored event models) to complete the update.

3.2.2 Parallel Interfacing Narrative-Semantics (PINS)

The crucial of the PINS model is to combine two representational levels of semantics and a narrative structure in a parallel architecture to achieve sequential image comprehension. The Figure 4 provides the view of this model. It separated the semantic and narrative processing. By using Visual Narrative Grammar (VNG, subsection 3.1.2) to present the communicative (discursive) structure, it processes the syntax through access, prediction, and updating with forward-looking and backward-looking mechanisms. It, in parallel, interprets semantic visual forms such as spatial relations, connections between view layers of the depicted scene, and the underlying event models. Then the combination of the two processing layers achieves the understanding across different levels' interpretation.

3.3 Computational Model or systems about comic comprehension

Besides the researches that analyze the visual semantic and narrative of comics and the cognitive process that interprets sequential images, a few works proposed computational models to tackle the comprehension tasks: Iyyer et al. considered a large amount of comic data as the training basis of a hierarchical LSTM model to achieve the understanding of sequence content [11]. Martens et al. developed a Visual Narrative Engine that is based on Cohn's Visual Narrative Parallel Architecture (VNPA), which is the combination of VNG (subsection 3.1.2) and PINS (subsection 3.2.2) [10].

3.3.1 Visual Narrative Engine

The essence of the Visual Narrative Engine consists of two types of representations and the inference information, the event model and the spatial structure. The action status between entities and their relation with the scene are represented as logical literals. The predicates then construct the scene graph, which describes the composition of the graphic structure of a panel. Then, the event model (also situation model) combines the represen-

tations to form a hierarchical plan that can infer the semantics of the comic sequence. The Figure 5 show the scene graph representations of a comic sequence.

In the first panel, the ”at” relation describes the entity's status, and the action status ”falling” shows the entity's current state–it is an attribute. This composed the scene graph that represents the composition of the panel; similarly, the scene graphs of the second and third panels all contain objects, relations, and their attributes.

3.3.2 Hierarchical LSTM

While most theories and researches focused more on the understanding of comic's graphical side, Iyyer et al. took the multimodality of comic comprehension into considerations [11]. They proposed a hierarchical LSTM where separates the text information and graphical information into two layers of representations and then merges the feature vector representation through the inner product.

The Figure 6 gives the graphic of the hierarchical LSTM model. Since the LSTM is a recurrent neural network known for capturing sequence value, it accumulates the information gained by reading comic sequences. Because in a panel, the textual messages sometimes show up together, in other words, multiple textboxes at once in a panel, the first layer hance to take word-embedding vectors of the text as input. After that, the image feature description vector of the panel was obtained from the fully connected layer of VGG16–a pre-trained Convolutional Neural Network weight model–to be added into the second layer. They then use the inner product to combine both textual and graphical representations to achieve comprehension.

This task fits common computer vision question frames pretty well, which usually ask computational models or systems to train, predict, and learn from a massive amount of data. Their results were evaluated through closure tasks that tested on either text information, visual information, or the combination to achieve hire accuracy on predicting a correct closure end among three candidates.

In this work, to satisfy the data-hungry machine learning model, Iyyer et al. also introduced COMICS dataset, which was collected from western comics (details in subsubsection 1.4.0.1).

3.4 Gaps and Integration

Comic theories and the conceptual cognitive models interpret comic sequences from human perceptual aspects; they discussed visual stimulus, attention, or the narrative of story events. The semantics of comic sequences were acquired through consciously reasoning the spatial or other relations. They can comprehend complex content very well by analyzing and processing representations to retrieve syntax or semantic. However, scalability is often an issue because of the need to pre-processing target data to get meaningful representations about semantics. On the contrary, computational models that describe features as mathematical forms, although they may learn some characteristics through the training process, are usually unaware of the syntax or semantic meaning humans used to interpret the content. They, yet, excellent in processing a large amount of data automatically.

Aiming to understand the comics well and, at the same time, automatic the comprehension process for the rapid growth of visual narrative data. We plan to hybrid the advantages of the two types of methods. Therefore, in this work, we add annotations to the datasets by automatically generated visual narrative labels such as reading order analysis and panel transition sequences based on visual narrative comprehension theories proposed in the Cognitive Science literature(SPECT and PINS). We plan to construct a hierarchical LSTM-based model for combining multiple modalities. The detail of the proposed method will be in chapter 8.

Figure 3

Model of the Scene Perception & Event Comprehension Theory (SPECT) theoretical framework. The eye icon denotes the position of viewer gaze on the stimulus during a particular fixation.

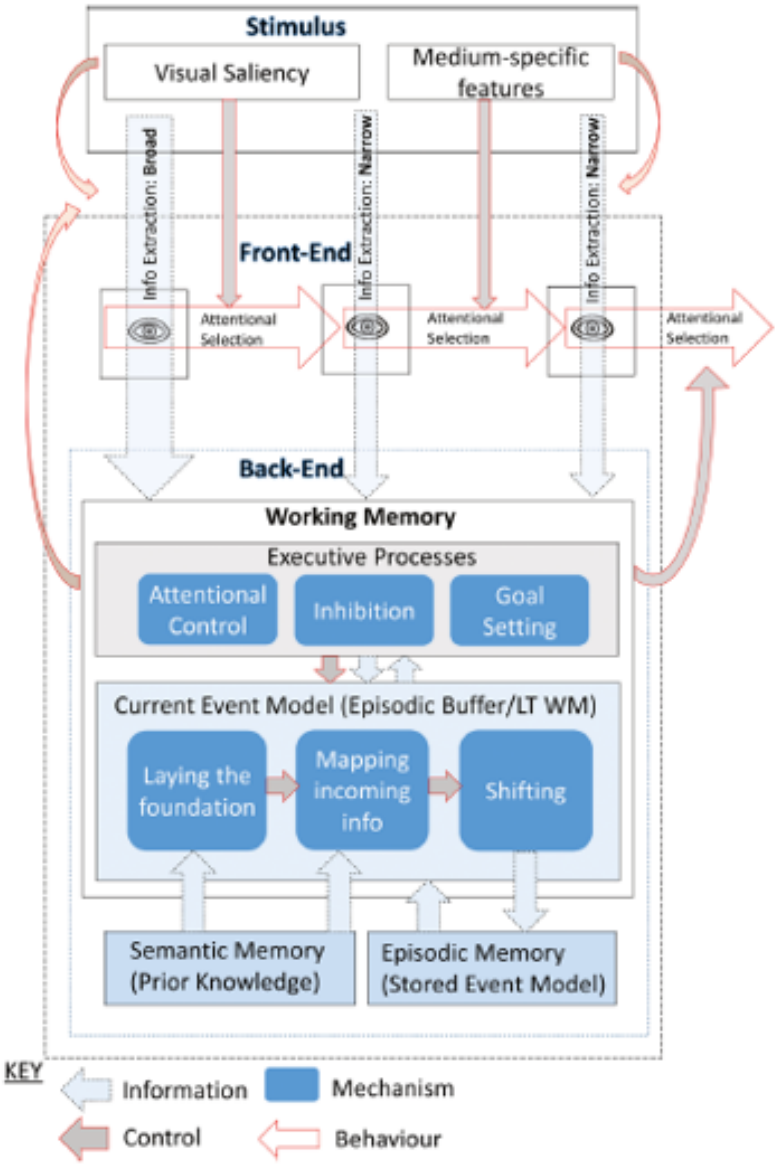


Figure 4

The structure of Parallel Interfacing Narrative-Semantics (PINS)

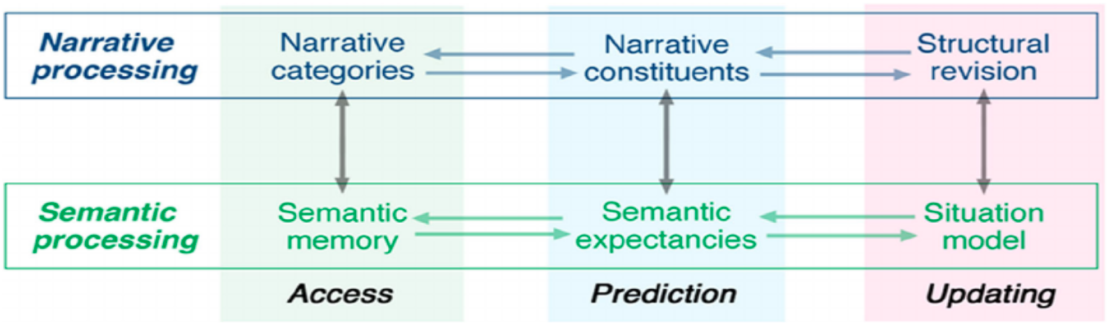


Figure 5

The scene representations (scene graphs) of a comic sequence.

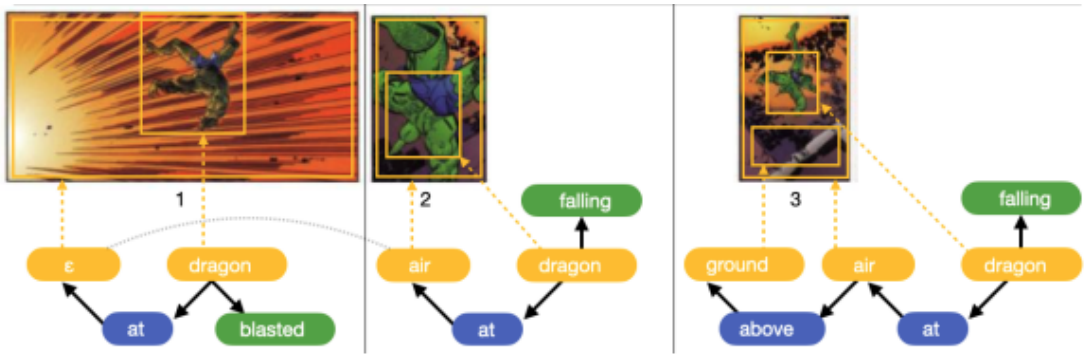
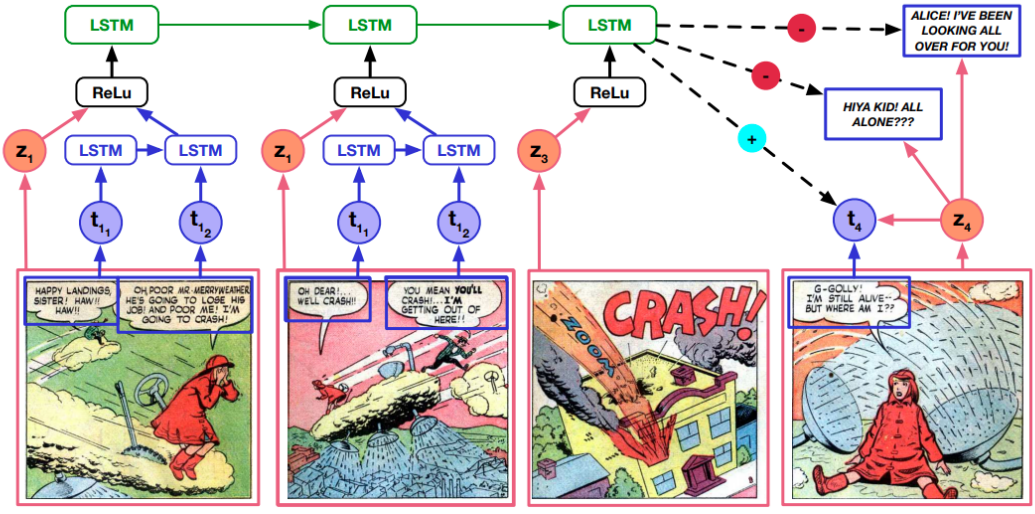


Figure 6

The view of Iyyer's Hierarchical LSTM



Chapter 4

Frist findings: Features influences understanding

Our first findings aim to investigate the possible features that influence comic comprehension. In this work, we focused on analyzing perceptual features that can be observed from comic visual representations, such as how complex the layout authors use to frame the story events, the proportion between textual expression and visual expression of panels, the ratio that authors convey ideas through images compares to transporting message by text. We annotated part of features manually and automatically analyzed others to learn the composition of the target dataset; meanwhile, we transform the features into vector representations to cooperate with the LSTM model. After that, the features were combined by a formula to describe the information amount that the visual aspect of a panel can vehicle. Then, the results were integrated with image features and were examined through cloze-tasks.

4.1 Research Questions

The research question we asked here is, ”What features of comics influence understanding?” (section 2.1). To answer this question, we first needed to decide the features we would study. According to the cognitive model about visual narrative comprehensions such as SPECT and PINS, the semantic of images starts from the visual stimulus perceiving from visual forms. Therefore, excluding the authors' preferences on drawing style, which is different and too hard to study systematically, we focused on the perceptual features that commonly exist in every comic. Due to the way comics are to be published, book pages usually confine comic panels. The number of panels on the page has a limitation, and the panels should be arranged in an understandable order. Thus, the layout that authors employed to guide readers is a common feature. The use of layout follows some principles,

but they vary according to each page's content. This hence brings the reading order issue as well. After the use of textbox, in each panel, the ratio of the image compared to text expression is not alike, depending on which side transfer message is more efficient to current content. We then decided to study the page layout's structure complexity and the reading orders and estimate the information amount transported by panel image compared to text.

4.2 Method

4.2.1 Annotations

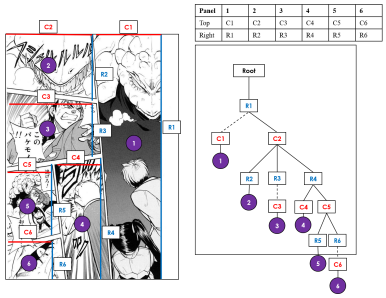
To interpret visual narrative content from a cognitively grounded perspective[3, 1], we augment the MANGA109 dataset with three types of annotations. The first type of annotation incorporates information about page layouts. The second type of annotation is the translated text. The third type of annotation is the transitions between panels, also grounded in the theory of comics[8, 11].

Layout and reading order annotations: The layout of the page influences the reading order of panels on a page. Western comics are generally laid out on a grid where the panels are read left to right and top to bottom, the subdivisions of the page as rectangles make it easier to determine reading order of panels. Manga, on the other hand, generally has more variety in shapes and placement of panels(see Figure 7). In order to capture the differences in reading order, we employ a tree representation of subdivision of the page to annotate individual panels and their relative positions based on reading order [16]. The algorithm for determining reading order annotations is described below and illustrated in Figure 7.

In Figure 7, the circled numbers denote panel number based on reading order of the panel on the page. The table beside the image in the figures shows the subdivision structure that is used to label the panel. The left side of the image denotes the horizontal row cut, so they are marked with R and the top borders of panels are the anchors of vertical column numbers marked with C. Determination of reading order happens by first moving along the

Figure 7

Tree representation of the layout from subdivisions on a manga page generated by the reading order algorithm. The top and right borders of panels are used as a reference for the structure annotations.



left border with annotation of a row cut whenever a horizontal panel edge is encountered on the edge. This is followed by movement along the top border with annotations for columns whenever a vertical panel edge is encountered. The process then recursively continues by moving the top border down to the first row edge and then annotating further rows followed by columns along that edge. The process terminates when all horizontal and vertical edges have been explored. For any node where no further subdivisions are found, a panel number is added as the leaf node in the structure. The resulting in-order traversal listing out leaf nodes gives the reading order of the page.

Automatic annotation of layouts: Due to the inherent complexity of parsing Manga pages, 72.5% of pages in our dataset were accurately annotated by the algorithm. The remaining 28.5% were manually annotated.Figure 8 shows two examples of unusual panel composition in terms of determination of reading order that present challenges in automatic annotation.

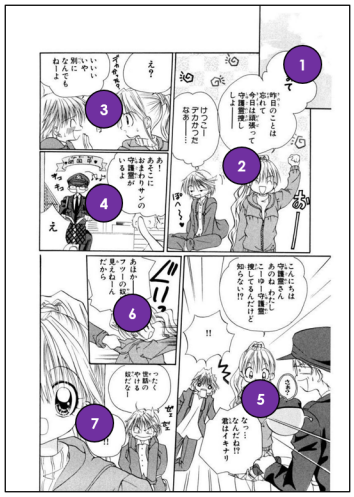
Text annotation: The final step of processing the manga data to directly compare with western comics is language translation. The text content transcribed in the Manga109 dataset is in Japanese. The text analysis models we use are trained on English so we employed the google-translate API to pre-process textual content.

Figure 8

Unusual layouts of panels on manga pages

(a) *Panels that appear to be arbitrarily arranged and spaced on the page.(AosugiruHaru ©Woquda, P49)*

(b) *Page that includes nested and overlapping panels.(BakuretsuKungFuGirl ©Ueda,)*



4.2.2 Reading Between the Panels

In order to develop an understanding of narrative that is presented as a sequence of static images and text panels, it is important to model the inferences that a reader makes in going from one panel to the next. Iyyer et al. [11] annotated the COMICS dataset with McCloud's six transitions to categorize the transitions between panels. These six categories were discussed in subsection 3.1.1.

Inter-rater reliability in the annotation process: We have 4 annotators who are non-expert readers of comics/manga involving in the annotation process. Selected panel pairs are divided to include duplicate pairs across annotators. We provide a graphical interface that displays the manga pages side by side and allow annotators to record transitions and layout structures by clicking options. In addition to 2228 pairs of panels in the Manga109 dataset, we also annotated 200 selected pairs from the COMICS dataset to test for agreement between the two sets in terms of these labels. The percentage of each type of

Table 3

Presence of different inter-panel annotation categories in the annotated pairs from COMICS and Manga109 datasets.

Transitions	COMICS (Iyyer's 250)	COMICS (New 200),1st	COMICS (New 200),2nd	Manga109 (2228)
Moment-to-moment	0.39%	4%	4%	12.6%
Action-to-Action	34.6%	25%	22%	33.2%
Subject-to-subject	32.7%	17%	26%	20.4%
Scene-to-scene	13.8%	26%	15%	10.1%
Aspect-to-aspect	–	4%	12%	8.3%
Non-sequitur	–	14%	15%	11.6%
Other	17.7%	10%	6%	3.5%

transitions are presented in the Table 3. Overall agreement among our annotators is 82%.

4.2.3 Layout and Text

Characterizing the complexity of page layout:

As was illustrated earlier in Figure 8, manga pages have sometimes unusual layouts. Based on the subdivision algorithm to determine reading order we can determine the complexity of panel composition in a given manga book. We do this with the following process. First, we take the standard regular grid layout of panel arrangement. This arrangement when presented to the subdivision algorithm yields a tree structure of depth 3 with terminal nodes at level 3. We label these as *average* complexity with a score of 0. If the layout of the page contains > 3 layers but can be unambiguously created then we label it as *complex* with a score of 1. If the structure cannot be represented with a unique parse tree then we give it the complexity score of 2 and label it as a *non-regular* page. If the page has

Table 4

The three examples of complexity scores of books from the manga dataset.

Manga examples	Non-regular layout	Simple layout	Average layout	Complex layout	Complex score
AosugiruHaru (Figure 8(a))	0.26	0.06	0.33	0.33	0.8
BakuretsuKungFuGirl (Figure 8(b))	0.42	0.11	0.21	0.26	1.0
BEMADER_P	0.04	0.13	0.41	0.45	0.4

a single row or column, or fewer than 3 panels then we label it with the simple category and a score of -1. Table 4 lists out the complexity scores calculated from some of the manga books in the dataset.

Text density: Next we characterize the usage of text within panels in manga. We calculate this as the proportion of the number of textboxes on a page. We then use the idea to separate text rich manga with image heavy pages. The average text proportion over the entire Manga109 dataset is 1.47, and the average over COMICS is 2.63. This indicates that every image will interact with 1.47 and 2.63 textboxes on average to convey the story content together correspondingly. This shows the importance of both modalities in being successful at the overall comprehension task.

Information amount in panels: If we interpret a panel as a unit that authors used to transfer stories, this information amount inside a panel may decide how easy it is to be understood. The formula (1) estimates the information score to separate between complex panels with simple ones. The higher score implies each panel in a chosen book may possess more information. The IA indicates how much area is covered by the image; it is the panel's size minus the textbox area. The LS stands for layout score, PN denotes the average number of panels on a page, and TN presents the textbox number inside the panel. The

formula is designed by assuming that as the layout complexity may influence the difficulty of interpreting reading orders, if more panels are on the same page, each panel area will be shrunk, so the panel's image may have lower complexity. Meanwhile, the text density implies the extent to which the author conveys details through a textbox rather than an image. The information score will be normalized between $[0, 1]$ when adding as features.

$$i\_score = w_1 * IA + w_2 * LS + (-1) * w_3 * PN + w_4 * TN \tag{1}$$

4.3 Closure Tasks and Comprehension Model

Closure is a process that involves understanding of individual panels and making connective and often complex inferences between consecutive panels. To characterize the differences in the performance of the computational model developed for COMICS on the Manga109 dataset, we employed two cloze tasks: text cloze and visual cloze [11]. A model's ability to understand narratives and characters given a few panels of context is tested through these tasks. In the *Reading between the panels* section, we observed differences in the proportion of different types of transitions between the two datasets.

Cloze Tasks: Here is how the two tasks are defined: Panels p_{i-1}, p_{i-2}, ..., p_{i-n} are given as context to the model. The model prediction for the content of p_i in terms of the respective aspect (text or visual) is recorded. Consistent with the original design of Cloze tasks [17], there is a single correct option among c candidates.

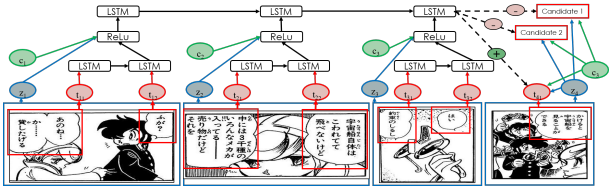
In the *text cloze* task, the model is asked to rank the a set of candidates on their likelihood to be in a particular textbox given preceding context panels (text and image) and the visual layout of the current panel image with text removed. The *visual cloze* task is set up in a similar way in terms of inputs but is provided with visual panels as options for ranking.

Model Description:

To conduct experiments on the analysis of how the stylistic features cause differences in the two sets of visual narratives, we incorporate the computational model that encodes panels with a hierarchical LSTM architecture [18]. The model is asked to predict the next panel from candidate panels after it gets the preceding n panels as context. The context information from a sequence of panels includes panel features based on annotations described earlier in the paper. It also includes features of elements within the panel including image and text features. The hierarchical LSTM uses a combination of these two sets of features for learning transitions.

Figure 9

Architecture of the model. p_i is the panel at i^{th} index. z_i is the corresponding visual image. t_{ix} is the x^{th} textbox on panel p_i. c_i is the style features. In our baseline, learned text features combine with image features in the hierarchical LSTM to get context representation. Then the style features are added as new information of context (BEMADER_P ©Hasegawa)



The hierarchical LSTM architecture is shown in Figure 9. It collects the element-level representations within the same panel in sequence. And then, the element representations are combined as panel-level representations, that are then passed into the second level as the context representation. The model consists of intra-panel LSTM and inter-panel LSTM. The text-only baseline has access to text within the panel. Each x^{th} text element of panel i, t_{ix}, is represented with a word embedding sum which is then combined with multiple textboxes through the intra-panel LSTM. For visual cloze, the pretrained $fc7$ layer of VGG-16 is projected down to the word embedding dimensionality with a fully connected layer. This overall context representation is then passed into the inter-panel LSTM.

Table 5

Experiment 1(b). Compare the performance of adding different style features

Task	W/ panel	W/ text	W/ layout	W/ info
text,text,easy,M	35.4	32.2	36.6	36.0
text,text,hard,M	33.5	37.8	37.2	36.6
image,vis,easy,M	70.1	70.3	70.1	**70.3**
image,vis,hard,M	67.8	64.1	67.6	**67.8**

4.3.1 Experiments

Our experiments have two main groups with four treatments per group. One is the comparison between the accuracies that the comprehension model can achieve on the text, visual, and multi-modal cloze tasks. The other one is the comparison between subsets that vary on stylistic features. Varying encodings of features that are different across the two styles allows us to characterize the contribution of these features on the performance of the model.

Feature Selection: The image features we used in context representation was the pre-trained VGG-16 fc7 features trained on ImageNet. As a next step, we compare the model that combines different style features to choose the most influential factor in later experiments. Four style features can be retrieved from our annotations: the average number of panels, text density in a panel, layout complexity, and the information score combine all the features with the weighted formula (1). The results are in the Table 5.

For each task, the model is asked to score the answer candidates given the preceding n panels as context. From the table, we can see that the accuracy is not significantly better than chance in the text-cloze tasks. Because the texts in the Manga109 dataset are written in Japanese, we suspect that language translation performance might be responsible for this.

Table 6

Experiment 2. the performance comparison between the baseline (without style features) and adding style features

Tasks	Base easy	Base hard	W/ style easy	W/ style hard
Text-only,text	42.4	51.5	36.0	36.6
Image-text,text	42.4	45.4	34.7	37.2
Image-only,visual	51.4	47.9	**70.3**	**67.8**
Image-text,visual	47.6	44.7	**60.9**	**55.8**

This would merit further investigation to see if expert translated manga would improve results over the baseline we present in this paper.

Manga109 comparison with style features: In the second experiment, we compare the performance between baseline context representation with the version involving style features.

The results in Table 6 indicate that trained models' accuracy is better than chance for all models. While the models on both context representation share this phenomenon, there are also interesting differences. Using style features performed around 10% 20% percent better than the baseline in visual-cloze tasks, but the performance in text-cloze led to lower accuracy. We suspect the main reason for it is the style feature we choose to discuss.

Reading order comparison: Our third experiment belongs to the second group that is trying to compare subsets with various stylistic features. In this experiment, we consider the reading order of panels. The comparison is done between the model that is provided reading order features against one that does not have this information.

The results of these experiments are presented in Table 7. When solving the text-cloze task, the reading order affects the text comprehension task. Reading order does not

Table 7

Experiment 3. the performance comparison between sorting panels with original recorded order and with analyzed orders. Similarly, the table show the comparison between baseline and with style features.

Tasks	Base random	Base ordered	W/ random	W/ ordered
Text-only, text	39.6	**51.5**	34.2	36.6
Image-text, text	35.9	**45.4**	34.3	37.2
Image-only, visual	54.6	47.9	63.7	**67.8**
Image-text, visual	48.8	44.7	53.6	**55.8**

Table 8

Experiment 3. the performance comparison between the subsets with lower page layout complexity and with higher layout complexity.

Model	Low complexity	High complexity	W/ low	W/ high
Image-only,vis	50.4	50.9	**58.3**	47.3
Image-text,vis	48.2	47.3	**51.7**	48.6

have a significant effect on visual comprehension tasks.

Layout complexity: Our fourth set of experiments captures differences in model predictions due to the complexity of panel layouts. As mentioned earlier, page layouts are more varied in the manga, often even within the same book. We incorporate the panel complexity we defined in the earlier section here. By scoring the complexity level in Table 8, we split the whole dataset with a threshold (complexity score ≤ 0.39 – the average score of the annotated dataset).

In Table 8, the style features on visual-cloze tasks aren't as significant as other divisions. only 8% higher on accuracy for low layout complexity set. Pages with lower

layout complexity tend to have larger panels or simpler panel arrangements. We considered these may carry more information through images than texts on the same page.

4.3.2 Discussions

In recent years, the preponderance of data and data-driven models have renewed interest in complex human capabilities like storytelling. Within Cognitive Science there has been interest in developing theoretical and computational models of how humans perceive and make sense of narrative through multiple modalities. Several datasets that focus on these questions have been proposed on related narrative operations (such as image description, summarization, etc.). Multi-modal narrative comprehension is a challenging task for computational models. We have provided an in depth account of investigation into the these challenges based on a systematically annotated dataset of two prominent styles of multi-modal narratives. This method is founded in well-established cognitive and phenomenological theories and evaluated based on rigorous comprehension metrics. We have further presented benchmarks with a baseline computational model and provided observations that could potentially lead to interesting advances both on the computational modeling side and on using computational models within cognitive systems to better understand human cognitive processes for interpretation and generation of multi-modal narratives.

The dataset, the feature set, and these benchmarks could be valuable in evaluating more sophisticated neural models as well as computational cognitive models of visual narrative comprehension.

Chapter 5

Second findings: Inter-panel relations affect content

In the second findings, we investigated the relevance between inter-panel transitions and the story content. The transitions capture the focus shifting on visual representations subsection 3.1.1). We aimed to know whether the selections on discourses–how stories are told–related to story content. Therefore, the first step we designed for this project was to employ a two-step Convolution Neural Network to learn transition labels to get the narrative structure for comic sequences. Moreover, since genres are often the compressed terms to describe story content, we chose the genres of comics as categories to observe the different patterns of transition uses.

5.1 Research Questions

The research questions we asked for this work are "Whether the discourse related to story content?" and "Is the narrative in comics follow some patterns? What are the patterns? " The question targeted to answer our second research questions for comic comprehension model, that is to study the aspects of narratives we should take into consideration to build the model. Hence we chose the inter-panel transition as the representative of comic discourse because they model the visual changes. In other words, transitions are the shifting authors selected deliberately to control the storytelling, to guide readers' attention. We studied the links between story genre, the short words that summarized story content, and transition sequences–the narrative sequence used to tell stories.

5.2 Related Work

Genre study based on content is a well established research area within multimedia analysis, especially in terms of recommender systems and for building knowledge graphs for information retrieval. It has been studied separately across multiple media forms including audio[19], textual narratives[20], web documents [21], films [22, 23, 24, 25], and comics[26, 27].

Among these film and comics are forms that combine both visual and textual/audio media.

In the movie area, genre identification has seen significant research by using different components for indexing, including generation of trailers [24, 25], posters [28], topological metadata [23], frame content [29], and the temporal structure[30, 22]. Analyses substantiated the theory that genres highly influence film structures; moreover, previous research showed that the structures helped retrieve genres. Although Comics, as a medium, is older than film and share similar sequential narrative characteristics[31], the genre detection in this area fails to have a similarly broad discussion across different aspects other than the image content.

In previous research, the comic genres were used to summarize story content that potentially serves the query based on the reader's interest [32]. Furthermore, to better represent the story detail, the image contents of pages in a comic book were described by sub-sequence that consist of genres [23]. While the genres encompass the image contents, the discussions did not consider the contiguous panels and the relationships between panels. Those narrative aspect features are usually rather neglected in comic researches, even if they are the nature of comics.

To bridge the gap between the image content, the usual emphasis of computer vision, and comics' narrative nature. Iyyer et al. used hierarchical LSTM to capture the image content sequentially and analyzed the "gutters" between panels based on McCloud's

definition of panel transitions [11]. Yet, the link between narrative features, derived in sequential content, was not discussed further. The theory of the narrative aspects features benefit sequential content understanding were demonstrate in previous researches. McCloud's books about comic understanding proposed transitions that cover the time and space shifts between panels [8, 7, 9]. Martens et al. then discussed the narrative and event structure helped on comic sensemaking [10]. These researches inspired our questions.

In film, work by Choroś *et al* [22] shows that structural analysis of videos can be useful in identification of genre. We take inspiration from this work and demonstrate that structural analysis of manga style can help with genre detection. In Cognitive Science, the Parallel Inference of Narrative Semantics (PINS) theory [2] proposes a structure of visual discourse presented in comics based on the role of panels in comprehension of the overall narrative. The Visual Narrative Grammar (VNG) represents the narrative structure level in the PINS model, it argues that the combinatorial structure functions to organize the meaningful information into comprehensible sequences. The VNG operates on sequential images as the syntactic structure in sentences. It gives image units a categorical roles like syntax decomposes sentences into nouns, verbs and so on. And then it organizes the categorical roles through a constituent structure to form the sequential images. This theory brings a more formal linguistic vocabulary to constituent elements of visual stories.

5.3 Method

We are trying to test the hypothesis that if panel transitions are intentionally chosen to indicate particular types of narrative events then a computational model will be able to learn these patterns of transitions. Further, we demonstrate the applicability of this result to show that genre, defined as a collection of artifacts with similar narrative arcs, can be better detected with the knowledge of transitions labels given to the system. To answer this question, we employed a corpus of visual narratives in the Manga109 dataset and then enhanced the dataset by adding panel transition narrative labels. After that, the labeled

features were used to retrieve clusters. Finally, the results were compared with the baseline provided in the original dataset.

5.3.1 Labeling for Narrative Transitions

To capture the medium's narrative aspects, we then analyzed different factors according to comic theories concluded by Scott McCloud [7, 8], which are usually underlaid inside comics. We then first took the sequence properties shared by the films and comics into consideration. Movies using camera shots to link frames and guiding readers by switching focus between content; an analog idea in comics is the transitions between comic panels. The transitions lead readers' minds by shifting between the content too. Comic panels fracture both time and space, offering a jagged staccato rhythm of unconnected moments like single frames. It is the audience's imagination to make the "gutters" closure; therefore, authors deliberately cooperate various transitions to convey their story in a rhythm that matches their mind and emphasis the content they want to raise readers' attention. The examples of the transitions and complex ones are in Figure 11.

5.3.1.1 Comic transitions There are six types of common transitions in total (details are in subsection 3.1.1); sometimes, the "gutter" between the panels combines multiple transitions simultaneously to transport a rather complex idea. We manually annotated the randomly selected 2228 pairs of consecutive panels from manga pages to analyze inter-panel transitions. Each pair of consecutive panels was labeled with the transition category that best describes the change between them. The annotated results are listed in Table 9.

The annotators involved in this process were 3 students familiar with comic reading but are not professional comic producers. Different chucks and overlap groups of selected manga pairs were assigned to different annotators. The agreement among the annotators on the overlapped pairs was in Table 10, and evaluated by Cohen's kappa score.

Table 9

Presence of different inter-panel annotation categories in the Manga109 datasets.

Transitions	Manga109 (2228)
Moment-to-moment	12.6%
Action-to-Action	33.2%
Subject-to-subject	20.4%
Scene-to-scene	10.1%
Aspect-to-aspect	8.3%
Non-sequitur	15.1%

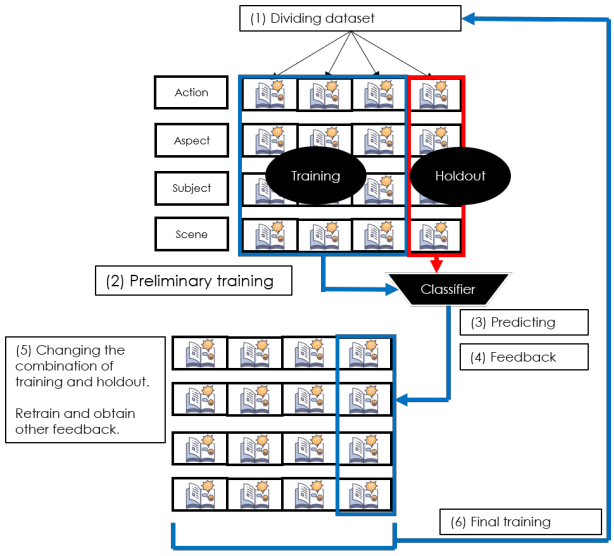
5.3.1.2 Story analysis method The framework was presented in Figure 10. The whole dataset was divided into two sets, one is holding out and the other is training samples. The labeled ground truth was the first set of training data, then some holding out samples were feed into model to obtain the feedback. After that, a group of unlabeled images was mixed into the framework, meanwhile, the combination of labeled data pool and the hold out set changed to obtain new feedback. Then the training process was launched again with the enlarged labeled pool. The model we used here was transfer learning based on CNN. On top of the model, two regular deeply connected neural network layers were added to transfer the layer output from feature descriptors to 256 units and then defined the output in the next layer to match the multiple classes we wanted. The images were described by features obtained from the fc7 layer of pre-trained ImageNet weights. The implementation details including loss function, activation functions in layers, and the optimizer will be discussed in experiment.

To provide the model sufficient information to identify the transitions, the images were considered in pairs. The input data combined two consecutive panels and the inter-

panel transition labels.

Figure 10

The labeling framework that shows how the dataset is organized across the steps of the iterative refinement process of labeling.



5.3.1.3 Narrative structures After obtained transition labels, our next step was to retrieve the narrative sequence from the results. Our transition analysis has two scopes: one was the summary of transitions that consist of a book. According to comic theories, the inter-panel transitions compressed space and time shifts, which connect the fractured parts into a complete storyline. Therefore, it is an analog concept that captures the temporal shifting relation. Our intuition here was that: like in cinema have different narrative pace in different movie genre, comics were likely to tend to use various transitions to convey the content to fit story pace. We showed that clusters of books could be coordinated by the distribution of transitions in the experiment section.

The other scope of narrative analysis was page-wise sequences. The same comic pages' transitions were extracted and formed a single sequence that represents the page's narrative structure. We then interpret the narrative structures from two reverse sides. We first do the clustering again to see whether the sequences could show some tendency and

form different groups—the next analysis laid on the genres from Manga109. We picked representative comics in different genres to see what sequences were used to transport the stories.

5.3.2 Clustering

The clustering method we used in this paper was K-means. The number of centroids was decided by comparing the distortion and inertia through the elbow method. After the clustering process finished, we then colored the books depends on their cluster labels and then compare the intersections between the cluster and the real genre groups to see whether the clustering results overlap with genres. A detailed description of experiment settings and results was in the experiment section. We choose an unsupervised clustering method here because we suspect that the content of a book might be a combination of genres and could not be described by only one word. Although the chosen label may close the main theme of a story, the content would likely be much complicated than the extent that a single word can describe. The previous comic genre research [26] also pointed out this similar thought. That is why we would like to use unsupervised clustering to approach the closest genres rather than direct label books with genres. This can describe the relatedness of books and can also see whether book genres have a correlation with the transition analysis through their overlapping.

5.4 Experiments and Results

This section will present our results with the parameter choices and then show whether the use of transitions related to comic genres.

We present our results through the following structure. First, we will analyze the ground truth labels used in our training. Then, the next sub-section evaluates the performance of the labeling framework and how well the automatically labeled results match with annotators' feedback. After that, the learned data are discussed by two scopes, through

Table 10

The agreement among annotators on evaluation set data

–	Annotator 1&2	Annotator 2&3	Annotator 1&3
reliability (Kappa score)	0.524	0.631	0.774

clustering and sequence analysis. Finally, we compare our results and the genres of Manga 109.

5.4.1 Ground truth analysis

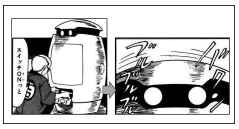
We have annotated 2228 pairs of panels as the ground truth about inter-panel transitions. In the annotating process, randomly selected pairs were dividing into an evaluation set (129 pairs) and a test set. Each annotator got both the whole evaluation set and part of the test set. Table 10 shows the statistic and agreement on the overlapped evaluation set. The agreements were evaluated through cohen's kappa [33]. It is a statistic that is used to measure inter-rater reliability categorical data. Cohen suggested the Kappa result be interpreted as follows: values ≤ 0 indicating no agreement and 0.01–0.20 as none to slight, 0.21–0.40 as fair, 0.41– 0.60 as moderate, 0.61–0.80 as substantial, and 0.81–1.00 as almost perfect agreement [34].

In some cases, the transition between a panel pair is a combination of multiple types of transitions rather than one that can be easily fit into a category. The multiple possibilities of labels were the reason why annotators might disagree. The Figure 11 gives a few examples. In Figure 11a, the subjects of the first panel are two characters (from left to right); in the next panel, the subject remained the same, the view angle changed to capture the characters' emotional reaction. Meanwhile, their action also different from the previous panel. In this situation, the annotators predicted the author's intention and chose a label they thought to capture the situation well. Another example in Figure 11b showed

Figure 11

Panel pairs with multiple transitions.

(a) *View angle changes happened together with action changes.*



(b) *Focused subject shifting happened together with camera shot changes.*

that when the focused subjects in the panel changed from two characters to one character, at the same time, the view shot become a close shot to emphasize what is happening to the character. Therefore, the transition between the panel can be interpreted as both subject-to-subject or aspect-to-aspect. The kappa scores in the Table 10 suggested the reliability between our annotator above or close to a substantial level.

5.4.2 Performance discussion on labeling model

In order to test whether the feedback-training framework benefits the learning process, in this experiment we compare both the model learning accuracy and labeling reliability between two sets.

In the first set, we remove the feedback-training process to train on simple transfer learning based on pre-trained imagenet weights. Through the experiment, we extended the training epochs to see whether the result get better through more learning iterations. After that, the trained model was used to predict the transition of randomly selected 100 comic panel pairs. After the prediction, our annotators were asked to label the predicted pairs as well. Therefore we can evaluate the reliability of automatically labeled transitions. The results were in Figure 12 and Figure 12. The reliability we used in here was also Cohen's kappa score.

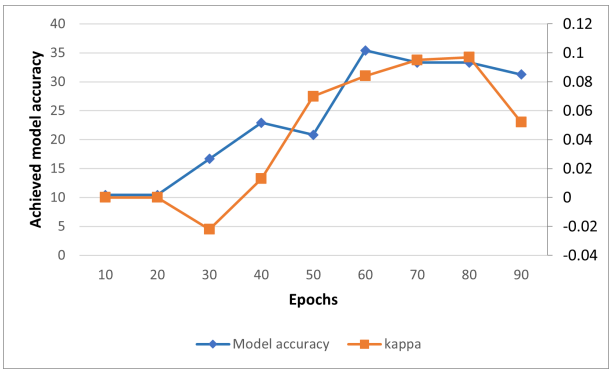
For the second set, the experiment was conducted on the labeling framework described above. We had started from our labeled ground truth, then after the training itera-

tions ended. The model tested on a holdout set, and it was asked to predict 100 randomly selected pairs from unlabeled panels. We then collected feedback as in the evaluation step of set 1. In the next round, according to the feedback, the predicted pairs become two groups. The correct group enlarged the labeled pool, and the combination of holdout set and training set changes based on 0.9, 0.1 proportion. Then the next round of training started. To observe the performance changing after similar iteration numbers, comparing with set 1, we set 10 epochs for each training round. The results were in Table 11 an Figure 13 provided a clearer view to interpret the result.

The settings shared in both sets were below: First, because it is a classification problem that uses the 6 types of transitions as categories, We evaluate the loss through categorical cross-entropy in the Keras framework, which computes the loss between the true labels and predictions. The optimizer for the model is the RMSprop algorithm, increasing our learning rate to converging faster. We had two regular deeply connected neural network layers after describing the input with the pre-trained VGG16 image-net weights; the top one transferred input into 256 units defined by the relu activation function. We then make the output fall into units with the same number of transition categories with the sigmoid activation function in the second layer.

Figure 12

The achieved accuracy after N training epochs and the kappa score between automatically labeled results and feedback.



Comparing the Table 11 and Table 12, it is clear that without the feedback-training

Table 11

The accuracy after different training rounds and the Cohen's kappa statistic between prediction and feedback.

labeling framework	10 epochs	20 epochs	30 epochs	40 epochs	50 epochs	60 epochs	70 epochs	80 epochs	90 epochs
Learning Accu	10.42%	10.42%	16.67%	22.92%	20.83%	35.42%	33.33%	33.33%	31.25%
Cohen's kappa	0.0	0.0	-0.022	0.013	0.07	0.084	0.095	0.097	0.052

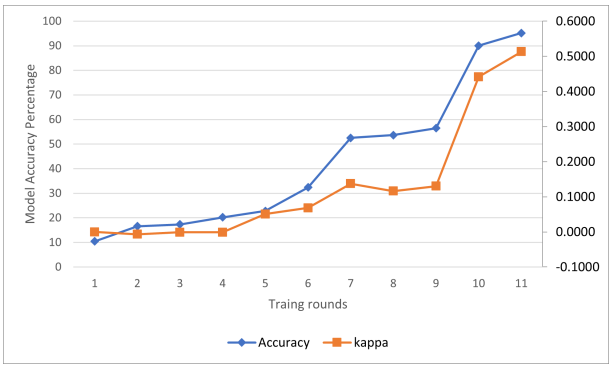
Table 12

The achieved accuracy after N training rounds, and each round has 10 epochs. And the kappa scores between prediction and feedback.

labeling frame-work	1 round	2 round	3 round	4 round	5 round	6 round	7 round	8 round	9 round	10 round	11 round
Learning Accu	10.41 %	16.56 %	17.32 %	20.23 %	22.81 %	32.36 %	52.53 %	53.67 %	56.49 %	90.01 %	95.19 %
Cohen's kappa	0.000	-0.007	-0.001	-0.001	0.051	0.069	0.137	0.116	0.130	0.442	0.513

Figure 13

The achieved accuracy after N training rounds, and each round has 10 epochs. And the kappa scores between prediction and feedback.



process, the achieved accuracy is much lower even after many iterations. Also, the reliability between prediction and feedback is always very low, indicating no agreement. However, the reliability of predictions for the model trained on the feedback-training process reached a moderate level in the end.

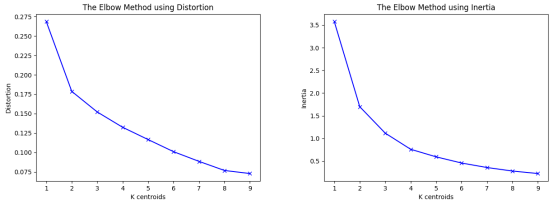
5.4.3 Transition analysis and clustering

After the model labeled transitions for the dataset, we then analyzed the transitions in two scopes. The first one was to summarize the transition uses in books. This helped us have a rough view about whether the way authors organize the narrative is different in various stories since the transition implied how separate fragments connected in the story and how readers were guided when reading. Moreover, we retrieve the narrative sequence formed by transitions and calculate the most frequent ones. This showed a characteristic of how authors tend to use transitions. The results were then compared with the real genre labels from the manga109 dataset.

The experiment of transition summarizing had two parts. The first part was to present the transition distribution as vectors and then perform a clustering algorithm to see whether there were any similarities between transition uses. If so, we then curious about if the clustering overlaps with the book genres? Therefore, the second part of the experiment

Figure 14

Elbow method using distortion and inertia to decide the feasible number is 4 centers, because the changes become smoother after 4.



(a) *Distortion* **(b)** *Inertia*

is to group books that share similar genres and compare the intersection between clustering and genre groups. For this experiment, 30 books in manga109 dataset, 22197 panel pairs, were fully labeled and analyzed.

The clustering method here was K-means, and the number of centroids was decided by be elbow method. The Figure 14 showed the distortion and inertia of n centers. We then decide 4 was the feasible number for our data.

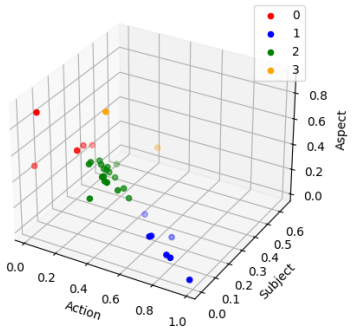
For the labeled book, we found that three types of transition were most significant. The Figure 15 represents the clustering result with 4 centroids using the three types of transitions as the axis. The feature vectors formed by transition use, we used to represent books and did the clustering, were normalized.

While Manga109 provided 12 genres, in our labeled books, some genres only have few books; if the amount of books is not sufficient, it is hard to observe the genre overlapping. We then combine the genres into five reasonable groups based on the common characteristic the genres have. The groups are ”Romantic,” which consists of Love-romance and romantic-comedy.” Fiction,” composed of Science-fiction and Fantasy. ”Action,” composed of Battle and Sports. ”Plot,” by Historical-drama, Suspense, Animals. And finally, remain the Four-panel cartoons unchanged because its’ style makes the category non-combinable with other types of comics.

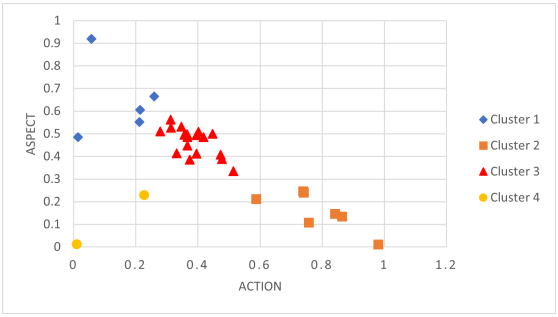
The Figure 16 displayed the comparison between transition statistics of various

Figure 15

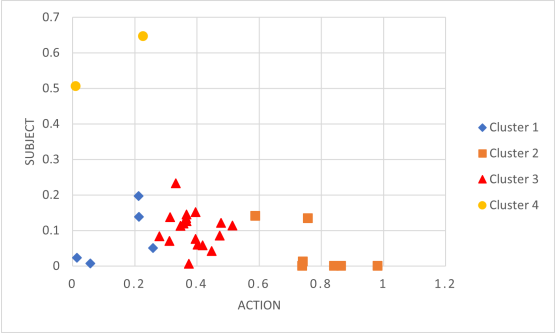
The clusering results with 4 centroids.



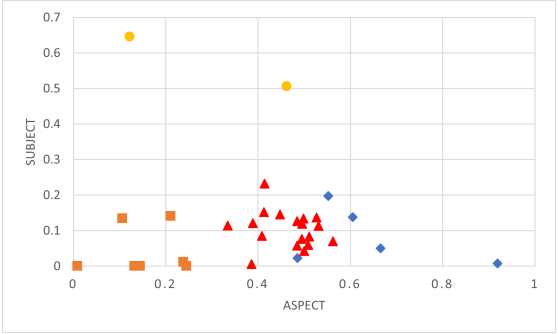
(a) *3D view of clustering results, using Action-to-Action, Aspect-to-Aspect, Subject-to-Subject as axis.*



(b) *2D view of clustering results, using Action-to-Action, Aspect-to-Aspect as axis.*



(c) *2D view of clustering results, using Action-to-Action, Subject-to-Subject as axis.*

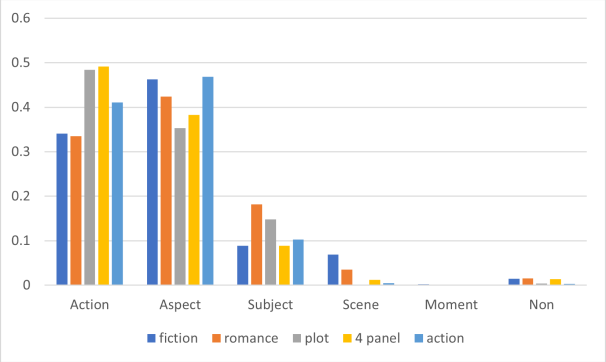


(d) *2D view of clustering results, using Aspect-to-Aspect, Subject-to-Subject as axis.*

groups of genres. Given that the number of transitions in each book is inconsistent, the matrix that employed six types of transitions as the axis was normalized before comparison. From the figure, we could observe that more Action-to-Action transitions were involved in storytelling in ”Action” ”Plot” types, while Aspect-to-aspect transitions seemed to show up more frequently when the story type is ”Fiction” or ”Romance.” Moreover, the ”Romance” stories employed more Subject-to-Subject transitions to connect panels.

Figure 16

The transition distributions on genres.The number represent the average of normalized distribution.



The intersection between genre groups and clusters was in the Table 13 to know whether the genre correlates with the use transitions. From the results, we could see that while the stories intersected with multiple clusters, each genre type of story seemed to focus more on some pattern of transition uses rather than evenly scatter in every cluster. For example, the ”Romance.” and ”Fiction.” groups overlap more with cluster 3, which has more Aspect-to-Aspect transition than other types. This also matches the bar chart figure which summarized the overall use of transitions.

5.4.3.1 Transition sequences analysis The last analysis we would like to provide from the generated transition labels was the transition sequences used most frequently in different stories. From various genre groups of books, we collected different lengths of transition sequences. The Table 14 contained the results.The statistic results said that the

Table 13

Intersections between clusters based on transition distribution and real genres

Genres	clus1	clus2	clus3	clus4
Action(Battle/Sports)	0.40	0.20	0.40	0.00
Romance(Love romance/Romantic comedy)	0.14	0.00	0.71	0.14
4 panels cartoon	0.00	0.25	0.75	0.00
fiction(Fantasy/Science fiction)	0.11	0.33	0.67	0.00
plot(Historical drama/Animal/Suspense)	0.00	0.50	0.50	0.00

sequences looks have many repetitions in them. For example, when an Aspect-to-Aspect transition showed up, it is highly possible, followed by another Aspect-to-Aspect transition. A similar phenomenon also applies to other types of transitions; we assess that maybe whenever an event is introduced in a story, it won't end immediately. The examples we had in mind are fighting scenes or conversation scenes, making the readers' focus switching between characters or view angles. Moreover, since the transitions bridge the inter-panel gaps, their repetition may imply the rhythm of storytelling. As shown in the Table 14, starting from sequences of length 3, the repeated phenomenon ends, and different types of transitions begin to mix in. Further analysis is a possibility that helps to studies the discourse of comic stories.

5.5 Discussion and Future Work

In this paper, we began by motivating the use of manga-style visual comics as an interesting domain for the study of narrative semantics in media. We started with a principled application of panel transition labels to automatically annotate the dataset from a model trained with input from human labels. We systematically analyzed the labeling re-

Table 14

The most frequent transition sequences used in different genres.The most frequent transition sequences used in different genres.Action-to-Action(ACT, AC), Aspect-to-Aspect(ASP, AS), Subject-to-Subject(SUB, SU), Scene-to-Scene(SCE,SC), Moment-to-Moment(MOM, MO), Non-sequitur(NON, NO).

Categories	Action	Romance	4 panels cartoon	Fiction	Plot
Top 1 sequence, with length 1	[ASP]	[ASP]	[ACT]	[ACT]	[ACT]
Top 2 sequence, with length 1	[ACT]	[ACT]	[ASP]	[ASP]	[ASP]
Top 3 sequence, with length 1	[SUB]	[SUB]	[SUB]	[SUB]	[SUB]
Top 1 sequence, with length 2	[ASP, ASP]	[ASP, ASP]	[ACT,ACT]	[ACT,ACT]	[ACT,ACT]
Top 2 sequence, with length 2	[ACT, ACT]	[ACT, ACT]	[ASP, ASP]	[ASP, ASP]	[ASP, ASP]
Top 3 sequence, with length 2	[SUB, SUB]	[SUB, SUB]	[SUB, SUB]	[SUB, SUB]	[SUB, SUB]
Top 1 sequence, with length 3	[ASP, ASP, ASP]	[ASP, ASP, ASP]	[ACT, ACT,ACT]	[ACT, ACT,ACT]	[ACT, ACT,ACT]
Top 2 sequence, with length 3	[ACT, ACT,ACT]	[SUB, SUB, SUB]	[ASP, ASP, ASP]	[ASP, ASP, ASP]	[ASP, ASP, ASP]
Top 3 sequence, with length 3	[SUB, SUB, SUB]	[ACT, ACT,ACT]	[SUB, SUB, SUB]	[SUB, SUB, SUB]	[SUB, SUB, SUB]
Top 1 sequence, with length 4	[AS, AS, AS, AS]	[AS, AS, AS, AS]	[AC, AC, AC, AC]	[AC, AC, AC, AC]	[AC, AC, AC, AC]
Top 2 sequence, with length 4	[AC, AC, AC, AC]	[AC, AC, AC, AC]	[AS, AS, AS, AS]	[AS, AS, AS, AS]	[AS, AS, AS, AS]
Top 3 sequence, with length 4	[SU, SU, SU, SU]	[SU, SU, SU, SU]	[SU, SU, SU, SU]	[SU, SU, SU, SU]	[SU, SU, SU, SU]

sults by clustering to observe the similarity that transition uses might have. And conducted a further comparison of results and real comic genres to show overlaps. This suggests the possible relationship between comic genres and panel transition features in addition to low-level compositional features extracted through computer vision. our results are promising and we hope by sharing the dataset, annotation methods, models, and the overall methodology with the community we will inspire productive work on media analysis. We will re-state the contributions of this paper in three aspects. First, a principled connection between practitioner specified best practices and semantic labelling. Second, a dataset and a workflow for annotations and analysis, Finally, a specific demonstration of detailed analysis of genre classification in manga as well as sequential patterns of panel transitions for analyzing the richness of expression in this medium.

Chapter 6

Third findings: Study the Visual representations and Intra-panel Relations

In this work, we investigated the intra-panel content. When the relations between entities and the scene are kept in a panel image, will semantics still be preserved if the drawing style transfers to another? This question is a response to the diversity of drawing styles of authors. At the same time, we studied the style features that comics shared; whether authors' art preferences influence the content understanding is unknown though that is also a characteristic of comics' visual representations. Therefore, we conducted experiments on either preserved intra-panel relations or unaware of the relations. To examine the effects of differences in styles and the extent to which semantics can be preserved, we applied for neural style transfer between western comics and Japanese manga. After that, we tested the understanding with closure tasks.

6.1 Research Questions

Our target questions for this task are two: first, whether the drawing style influences the preservation of semantics? Second, while neural style transfers are usually applied to the texture and stroke well, but not aware of the content. Does distinguishing the relations between entities and scenes help to preserve the semantic of the image? We, therefore, designed the experiments on applying different masks to images and observe whether the model can understand the style transferred content to demonstrate to what extent the sequence kept the semantics.

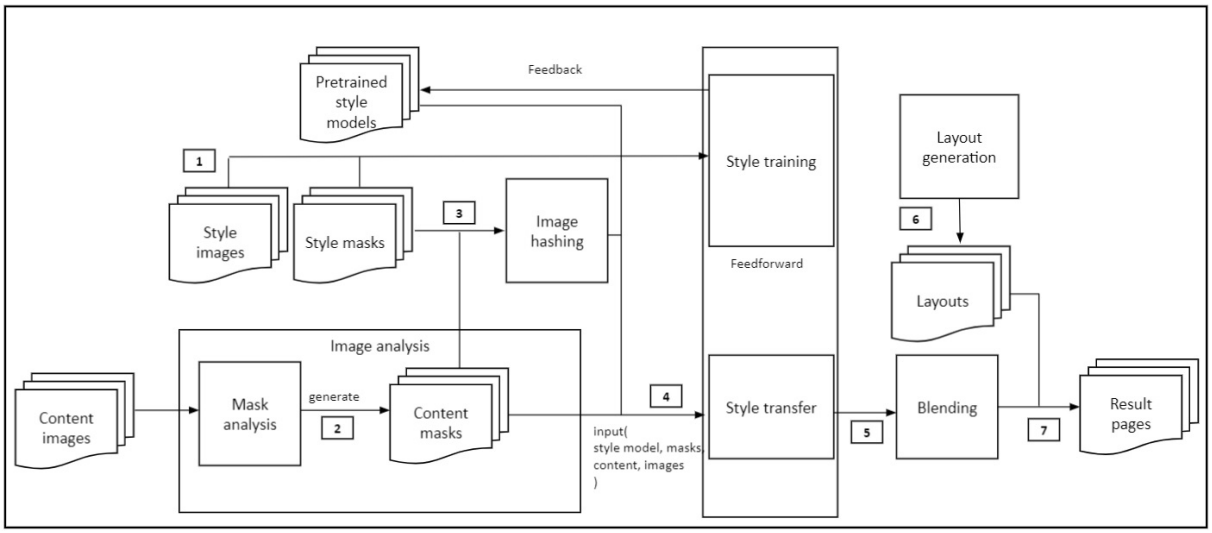
6.2 Related Work

Visual style transfer is studied as a generalized problem of texture synthesis, which is to extract and transfer the texture from a source to a target image [35, 36, 37, 38]. It has been used for non-photorealistic rendering [39, 40] to transfer scenes or objects into a particular artistic style. Gatys' et al. 2016 first introduced a method for using pre-trained convolutional neural network (CNN) to reproduce famous painting styles on natural images. Neural style transfer since that work has been a widely discussed topic [42]. This original approach has two limitations. First, the training process is not efficient. To address this, feedforward style transfer and other methods then were proposed to accelerate the process [43, 44]. These faster approaches allow multiple images to be processed faster making them feasible for processing multiple sequences of images such as in our dataset. Second, transferring the entire image ignores the relationships between content of the images. This poses a challenge for our application due to the presence of overlaid text within bubbles and the different sizes and layout of panels on a page. One way to approach the content problem is to use masking [45, 46] or to control the areas influenced during the transfer process [47]. Considering the nature and complexity of our target data we utilize masked variants of images according to content.

In terms of work related to comics, we incorporate two datasets in this project and augment them with annotations. The COMICS dataset which includes 1.2 million panels of western comics [11]. The Manga109 dataset includes 109 titles of Japanese manga [12, 13]. Our method is inspired by other content-aware style transfer, comic sense-making [48, 10], and content understanding with cloze-tasks and fine-tuned features [27] on narrative understanding.

Figure 17

The framework for the comic style transfer process. [1] destination style images and masks are used for training; [2] input images are masked; [3] inputs' masks and destination images' masks are hashed, and style are selected by similarity; [4] style transfer module transfers channels in parallel; [5] per-channel outputs are blended to form a single image;[6] generate comic layout; [7] output images are stored and combined with layout



6.3 Method

The process of style transfer requires processing of features of the target style trained on a dataset of target images and the processing of input image to be converted. The target style in our work is the western comic style and the input images are in manga style. We choose this setup because manga images do not have color making the comics to manga conversion less interesting for demonstration. Comparison of transfer in both directions is outside the scope of this paper and is planned as the next study in this project. Processing input features for target style is challenging because comic pages are complex. The scene, text, and layout layers have their own style features that are perceptually independent. To address this issue, we create masked versions of each panel image with layout feature annotations. These masked images are then passed through separate style transfer pipelines. For the input image to be converted, similar masking is done and the style is transferred to corresponding masked images. The style-converted images from channels

are finally blended together to construct the output image.

6.3.1 Computational Framework

Figure 17 shows the overall framework for style transfer. Our model follows the approach from Johnson et al.(2016) and Gordic et al.(2020), where is a feedforward style transfer (FFST). We try the general CNN style transfer model [50] before we choose this; however, the time cost is extremely high to get converted results for tens-thousands of manga panels. Thus, we use FFST, which reached similar qualitative results but is three orders of magnitude fast.

It has two components: image transformation network and loss network. The image transformation network is a deep residual convolutional neural network parameterized by weights. It maps input images into output images. The loss functions measures the difference between the output image and destination style or content images. They are defined through the loss network. The image transformation network is trained using the Adam algorithm, a replacement optimization algorithm for stochastic gradient descent to train deep learning models when handling sparse gradients on noisy problems. This is useful because masked images cause sparsity on feature representations. The optimizer minimizes a weighted combination of loss functions.

To better measure perceptual and semantic differences between images, the loss network is a 16-layer convolutional neural network [51] pre-trained on ImageNet (VGG-16) [52]. This is because it has already learned the perceptual and semantic information we want to measure in the loss functions. ImageNet is trained on real images and it has been documented that it is texture-biased[53] while our application is on stylized images. We acknowledge this as an issue. Future work could look at more specialized approaches and incorporate models that better fit this problem to improve benchmark performance.

The loss network defines two perceptual loss functions in the style transfer problem: feature reconstruction loss and style reconstruction loss. The former is the feature

difference between the output image and the content image, and the latter is the difference between the output image and the style image. They both use the feature representations in the different convolutional layers of the loss network to compute the loss. The squared and normalized Euclidean distance between the output image's feature representation and the content image's is the feature reconstruction loss.

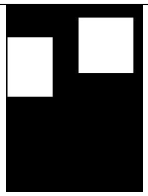
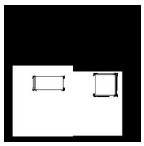
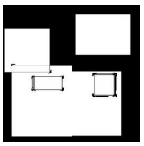
Given a feature map with $C_j \times H_j, \times W_j$ dimensions that represents the input images, where the H_j and W_j are the height and weight, and each point on its grid is described as C-dimensional features.Here is how the style loss is defined. First, a gram matrix is defined as a $C \times C$ matrix whose elements are the product of the feature matrix and then normalized by dividing with its dimension. This denotes the proportional to the uncentered covariance of the C dimensional features. Style reconstruction loss is defined as the squared Frobenius norm of the difference between the gram matrices of the output image and style image.

6.3.1.1 Image composition and masking Due to multi-modalities in manga, a panel can be split into two parts: scene content that carries visual information and textboxes that convey textual information. The scene content can be further divided into foreground and background. The major object that performs the story is counted as the foreground and the scene that carries additional information is the background. Based on this, we create three different masked images from the source image. Images with text bubbles masked, images with foreground characters masked, and images with background masked. These masked images are used to train independent style models. Examples are in Table 15.

After comparing the similarity between masks of both content and style images, it chooses the closest style image as the target for this specific panel. After that, the style transfer is applied to the corresponding masked images. Once all panels on the same page are transferred, the blending process reconstructs the style transferred images to get the final panel sequence. This reconstructed panel sequence is then mapped to a generated

Table 15

Examples of both rectangle and fit masks for textbox, foreground, and background masks are their combination.

Image	Textbox mask	Foreground mask	Background
			
			

layout to get the resulting comic page.

6.3.1.2 Selection of study images The compositions of a comic panel have a lot of diversity because of the changes in the number of objects, character actions, view angles, and the placement of text boxes. Therefore, we took the camera shots concept in films as the basis of how to divide compositions. Unlike other style transfer problems, we consider the style of comics on a book scale rather than take only an image. For each comic book, we considered all the panels of the same style and then chose panels with 6 different compositions as representative images of the style. We took the combination of close, medium shots with various numbers of objects in the panel image.

6.3.1.3 Content Similarity To address the bias of style transfer algorithms on texture features and focus on features of content, we use image hashing which to identify the similarities between image structures. We implement average hash [54] to down-sample the content- and style-masked images. This is then used to retrieve similar style images in terms of structure and image compositions to the content image.

6.3.1.4 Blending The resulting style transferred images from the masked channels are then combined to produce the final output. The blending is done in the order of the background image, followed by foreground image, and finally the text bubbles.

6.3.1.5 Comic layouts Comic layouts include multiple panels organized on a page with different layout parameters. For reconstructing the output panel in terms of a target layout, we first consider the number of panels on the page. This information is then used to map to a layout in the target style that corresponds to the number of panels. More sophisticated panel layouts can be mapped if panel sizes can be adjusted. This becomes problematic due to artifacts created by stretching the original images to fit the new panel shapes. For simplicity in this first iteration, we choose layout mapping according to number of panels.

6.4 Experiments and Evaluation

Our experimental evaluation is structured as follows. First, we take single panels and entire pages, and directly apply currently available image style transfer framework to set as a baseline. Next, we create style-transferred images with variants that include a final blended page with all components and with ablated versions that exclude mask channels. This allows us to visually see the contribution of different channels in the transfer process. Finally, we run a visual story cloze test to evaluate the differences between coherence scores between a sequence of panels before and after the style transfer process. Table 16 shows an example of input and output images and layouts for each treatment.

6.4.1 Single Panel Transfer

Our first run of the algorithm focuses on creating a baseline with art styles that are highlighted in currently popular style transfer work. We choose the art style used in the starry image frequently used in other style transfer research as the target to ensure

the quality of our implementation and to also highlight the challenges on our content. In the first three rows of Table 16, we show the content image, target art style, single panel transfer and page transfer. We notice in the starry night example (*AS,N-M* in Table 16) that the texture stroke is preserved in the resulting images. As expected, it treats the content image as a whole and does not distinguish between channels. In the next example (*AS,M* in Table 16), the model applies the dotted texture of the target style the color features from both foreground and background mix. The model treats text and text bubbles as a visual feature and also mixes its features in the visual scene.

Splitting the panels into foreground, background, and textbox. In the second experiment, we applied for the style transfer on two types of masks; one is the rectangle mask that we can parse from the dataset's annotations, the textboxs', objects' position; the other is the mask that fit the objects' shape. Table 16 showed the results of using various masks. Also, in the last column, we exhibit the results of using layout to combine several transferred panels into a page. The reference layout was a comic page that had the same number of panels as the content page. For each row of the selected layout, the panel widths were adjusted to fit content manga panels, and the content panels will be centered to the target positions.

The results show that the foreground or background color distribution transfers well, as expected (*CP-NM* in Table 16). Without masks there is noise due to mixing of features that affects the rendering of text bubbles. With the text mask, text is clear to read and arranged according to the style of comic text bubbles (*CP-RM and CP-FM* in Table 16). For full pages, blending separately style transferred panels and adding them in the new layout yields better results.

6.4.2 Panel Sequence and Layout

Our next experiment focuses on the coherence of sequences in terms of content. Our experiment is designed test whether coherence is influenced by style transfer given a

Table 16

The comparison between using art paintings as target style and using a content-rich image as target style. The settings are: with art style but no masks (AS, N_M), with art style and masks (AS, M), with comic panel but no masks (CP, N_M), with comic panels and rectangle masks (CP, R_M), with comic panels and fit masks (CP, F_M).

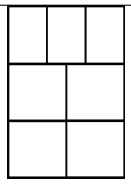
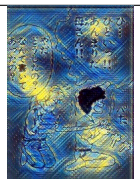
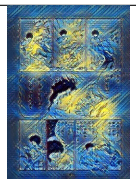
Content panel	Content page	Layout reference	Adjusted layout
			
Settings	Style	Panel	Page
AS, N_M			
AS, M			
CP, N_M			
CP, R_M			
CP, F_M			

Table 17

The cloze-test accuracy (the rate that our model got correct answers out of all questions) comparison. This showed two experiment groups, and each contained four sets. The groups used different fine-tuned features: one trained on COMICS, the other trained on Manga109. The 4 settings are no style transfer (N-S), style transfer with the whole image (T-W), style transfer with masks (T-M), and style transfer with composition features (T-C)

Settings	feature-C	feature-M
N-T	48.2	70.6
T-W	50.4	65.2
T-M	51.7	67.1
T-C	53.3	69.0

sequence of consecutive panels. The following subsections described how we design our comparison.

In this experiment, we transferred the style for 8 volumes of manga, 29,039 panels in total. These form the content of the visual story cloze task.

6.4.2.1 Cloze task Closure is a process that involves understanding of individual panels and making connective and often complex inferences between consecutive panels. To characterize the differences in the performance of the computational model developed for COMICS on the Manga109 dataset, we employed two cloze tasks: text cloze and visual cloze [11]. A model's ability to correctly predict the next panel in a given a sequence of panels of context is tested through these tasks. In our case, we use the cloze test on original sequence of images and compare the output with the style transferred image sequence. Given a sequence of panels as context, the model is asked to predict the most likely ending out of 3 candidates. While a comic panel is a combination of image and text content, we employed the visual tasks only. In this case, the model only observes visual features for prediction. Text features would be relevant but require proper translation or Japanese

language model. This avenue can be explored in future work.

6.4.2.2 LSTM model To answer the visual-cloze tasks, we employed an LSTM based comprehension model. It encodes the context images with the 4096-d fc7 layer of VGG-16. The fc7 features of each panel are feed to an LSTM. The model converts the sequence of context panels into fixed-length vectors and scores the answer candidates by taking the inner product of the candidates with the vectors then normalizing it with the soft-max function. The model projects both of the answer candidates and the context to 512-d representation. The feature descriptor we used in this experiment was a VGG16 pre-trained on image net data and then fine-tuned with either COMICS or Manga109. Therefore, our experiment has two sets: one used the features fine-tuned with COMICS to describe manga images, the other used the features with Manga109.

6.4.2.3 Results We divide the cloze-test into two groups, and each group had four sets. The groups used different feature descriptors to encode image context. The four sets are split according to whether they have style transfer, whether they used masks, and whether the target style chosen according to image composition?

Our target-style was comic panels from COMICS that represent Western comics, and the content images all came from Japanese manga books. The first row of Table 17 shows that the feature tuned on COMICS didn't capture the manga image as well as the features tuned on Manga109. The former's accuracy on the cloze-test is much lower than the latter's when the content images had no style transfer.

The results in both groups suggested that the transferred image still preserved the narrative characteristic to some extent. Because with the new data, the comprehension model can still predict the right answers for visual-cloze tasks. Besides, the two groups also show some interesting phenomena.

The results in the first column suggest a trend that when the manga images preserve more content after style transfer (to western comic style), the accuracy increases

accordingly. When the whole panel image transferred to comic style, the accuracy slightly increased. After the masks based on foreground, background, and so on were added to images, the accuracy increases slightly again. And then, after the model chose style according to compare the similarity between image composition, the accuracy increase once again.

In the study with manga features (second column), there is a drop in accuracy with whole image style transfer but improvement with the application of masks and composition features.

6.5 Discussions and Future Work

We presented the research problem of style transfer between comics and manga that represent multi-modal sequential narrative media. We described the feature set for this medium, highlighted challenges in terms of the interaction of layout, text, and scene features, and proposed a transfer framework. We introduced content masking with parallel style transfer for independent features and illustrated the differences between various single image compositions and combined the panels through an adjusted layout on both individual panels and full pages. To address the narrative communication aspect, we propose comprehension preserving transfer and evaluate style transfer modules based on visual cloze tests for narrative understanding. Results shown in this paper highlight the challenges for this medium and also set a baseline for future work on this topic.

As a new perspective on style transfer algorithms we hope to engage the narrative research community on interesting discussions around richer aspects of style transfer. This work also potentially has impact on future work related to interactivity in terms of visual interactive games and player styles and preferences as well as dynamic panel transitions. Our initial motivation for comic analysis was to study the use of composition for dynamic camera control in narrative based environments.

Chapter 7

Side Project: Applying Comic Theories to Generate Comics

In this study, our target is to learn how comic theories influence the comic's visual representations and to generate comics based on the theories. Therefore, we proposed a comic generating model with adjustable layers that aimed to expand the generator with new abilities, demonstrating the application of new comic theories.

7.1 Research Questions

This side project is related to our second research question for the comprehension model (section 2.2). To clarify the elements we need for building a comprehension model, we investigated the components in comics as the first step. When studying the narrative of comics, many pre-processing steps need human involvements to annotate labels and help express the semantic with knowledge representation forms; thus, computational models can process the content. Therefore, we proposed the generator to generate simpler comic data that matches comic theories for further research and the possibility and diversity for automatically comic generations.

7.2 Related Work

The discussion about comic composition can be divided into two aspects; the first part is the story that a comic wants to tell. The other is how to tell the story through the cooperation of graphical content and textual content. The relations between the two aspects had been curious by much previous research, hence some related analysis and theories. Some comic artists and cartoonists summarized some principles that comics used to tell the stories. Eisner's book discussed the diversities on comics' image representa-

tions [55]; whereas Carrier et al. studied the comic strips as a form of fine art [56]. In addition, McCloud's work encompassed many aspects of comics, such as inter-panel transitions that modeled the focus change between consecutive panels, the compressed time frame of comics, and the symbolic metaphors that made the unseen idea visible [7][8][9]. Besides the principles that applied to graphical aspects, the underlying structure of visual narrative also received focus. Cohn et al. modeled the story discourse in comic sequences as grammar categories and link the content with various conjunctions hence developed the theory Visual Narrative Grammar (VNG) [4][6]. The theory integrated the visual narrative structure with the concepts in narrative discourse[57][58][31].

Some research tried to develop models or systems to generate sequential art that incorporates different comics features based on the analyzed theories and ideas. Alves et al. discussed the graphical aspects of comic panels by describing the camera placement, background, independence, and so on [59]. In comparison, Nairat et al. focused more on the fictional part of comic [60]. The proposed inner character evolutions to form stories then rendered them as comics. Furthermore, Martens et al. focused on the discourse of comic sequences, utilized the conjunctions of VNG to model the changes between panels[61][62]. On the contrary, they simplified the graphical representation with abstract shapes.

Our work aims to provide a generator that integrates and tests multiple aspects of comics rather than focusing on only one element to tackle the challenge result from the complexity of this visual medium. Our generator divided the complex generating problem and tried to conquer various considerations by layered refinements on the result sequence. To achieve this, the generator supports the functions that can modify elements obtained from decomposing either comic's graphical composition and the structure as well as discourse elements, which made the generator extendable and applying different principles or theories into the generating process. We then showed how the generator integrated various theories and provided the sample symbolic material dataset set by retrieving common visual metaphors that make unseens visible in real comic datasets such as COMICS [11] and

Manga109[12][13].

7.3 Method

In this section, we will introduce the structure of our generator and the detail of the sample layers we applied to obtain a comic sequence.

We divided a comic strip into three aspects: image representation of the content, structure, and transitions that bridge the context and the narrative transported by the sequence. Therefore, we will discuss the abstraction and metaphor symbols of comics. Then applied comic theories to set the refinement layers. Finally, we show how the comic content is built up.

7.3.1 Generator Structure

Our generator modified the comic sequence based on iteratively applying refinement layers. The modification can be either adjust the whole sequence such as narrative structure or organize the content in a panel such as selecting characters' actions; Besides modify the setting of panels or sequence, some refinement layers enlarge the generating base by adding new elements to enrich the comic content or add more diversity on image representation. For example, a layer applies the mapping between textbox shapes with character actions' activation value. A layer provides different image composition templates; hence the image can be composed in more flexible ways. The generator provided a set of application program interfaces (API) to allow extendability. The structure of the generator and an example that combine refinement layers through API are in Table 18.

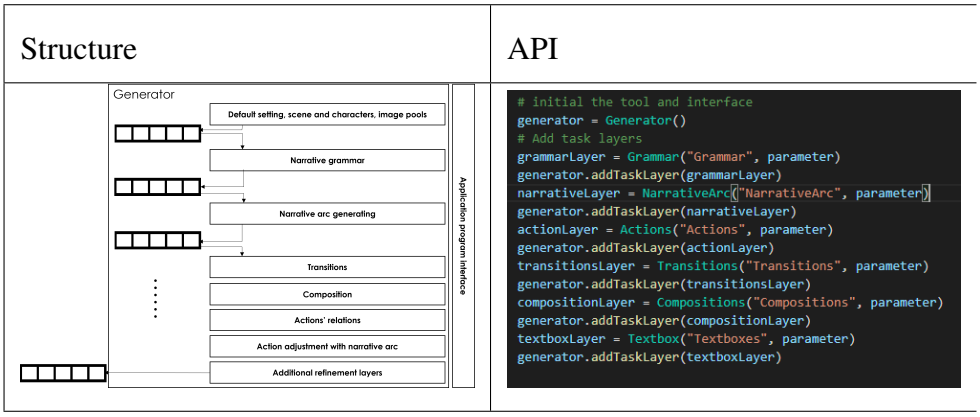
7.3.2 Graphical Content

.

To generate the image representation, we observed how existing comics represent specific actions and emotions with symbols. Then combine the abstracted symbols of com-

Table 18

The generator structure and how the implementation looks.



mon actions with character representation and scenes.

Factors in a comic panel are complicated, including the scene, characters, and symbols authors used to emphasize character states and emotions [59]. In this paper, however, we employed abstracted shapes to represent characters to simplify the complex detail of objects. On the contrary, keep the action and emotion symbol to show the sentiment and content changes influenced by narrative grammar and inter-panel transitions.

7.3.2.1 Scenes The first part of a comic panel is the scene that expresses a particular place that an event happens. In our model, five representative scenes are chosen as examples to support the generating process.

7.3.2.2 Abstractions Our first abstraction is the story characters. As Martens et al. used geometric shapes to represent characters in their comic generating work [62], we also employ simple shapes to represent the characters. Taking advantage of substituting characters with abstract shapes, some details of the characters, such as facial expressions and appearances, can be eliminated to reduce complexity.

7.3.2.3 Common Symbols In comics, authors usually employ several abstracted symbols to exaggerate characters' emotions and actions, such as speed lines to show the

Table 19

Examples of metaphor symbols of action and characters' emotions from real comics.AisazuNihaIrarenai© Yoshimasako, AkkeraKanjinchou© Kobayashiyuki,Akuhamu© Araisatoshi

Anger	Quick moving	Slow moving	Anxious	Collision	Relieved	Shock	Big shock
							

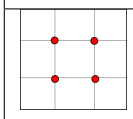
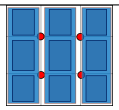
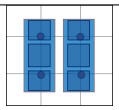
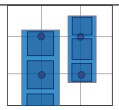
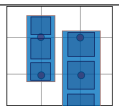
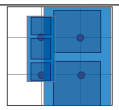
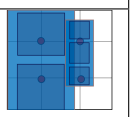
movement, explosion-shape to represent objects' collision, cross-shape symbol to emphasize anger, etc. Table 19 shows some examples from real comics. We analyzed the representations of actions and emotions in comic strips. The metaphor symbols and figure changes depict the status changes of a character throughout the story events. Those elements provide visible hints to readers about what is going on in a comic strip.

7.3.2.4 Image Compositions The second abstraction was the geometric positions of a character in a comic panel. Instead of using continuous coordinates to describe positions, we divided our panel into nine parts according to the basic composition method in photography, which is the rule of third. Based on the four intersection points, we then created two masks as the abstractions of character positions. Each mask was divided into 3 height levels; these are used to present the vertical physical position changes. The horizontal position changes were the shifts between masks set.

We used the arrangement of masks to achieve simple view angle changes and composition. Our model had two versions: the basic version provided the physical changes due to action changes; the other version then combined the view change to create more tension in the comic panel. Table 20 shows the masks for image compositions we used in our model.

Table 20

Examples of geometric abstractions and image compositions.

Rule of third	Basic	Parallel view	Left view with medium shot	Right view with medium shot	Left view with close shot	Right view with close shot
						

7.3.3 Global Modifications

To generate a comic sequence, what needs to be considered are not only the visual representations in panels but also the relation between consecutive content. Therefore, in this work, we combine the narrative grammar that describes the overall structure of a comic sequence with narrative arcs. And then integrate the sequence with transition in comic theories to decide how the narrative will be discoursed.

7.3.3.1 Narrative Grammar We formalize the generating process of new comics as two types of reasoning: global and local, mapping to factors that influence the whole sequence or only modify panel content.

To form a comic sequence, we first started by deciding the story discourse. Therefore, we generated the overall structure of the target comic sequence. Cooperating Cohn's narrative structures [4] that proposed an understandable comic follows a grammar that organizes its global structure with five categories (subsection 3.1.2)

Our global reasoning is the center-embedding; it expands a new structure by replacing a single category with a phase. We start with a basic phase (ex. E, I, P, R) and form a tree structure according to the chosen phase where the categories are leaf nodes. And then, referencing the center-embedded pattern of the theory, we expand each node based on choosing another phase with probability. Thus, we can get a structure tree whose

Table 21

Examples of a comic sequence that followed narrative grammar.

E	I	L	P	R
				

leaf nodes represent panels of a comic sequence. Table 21 gives an example that comic sequence follows the grammar structure.

7.3.4 Narrative Arc

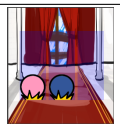
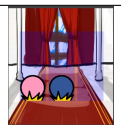
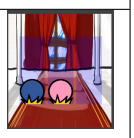
The five categories in narrative grammar reflect the concept of narrative arc, which describes a story's full progression. It implies that every story has a relatively calm beginning and then reaches the highest tension in the middle where character conflict and narrative momentum happens; In the end, conflict is resolved. This paper projected the five grammar categories to a value to get a curve that describes how the story tension changes.

7.3.4.1 Transitions In McCloud's comic theory [7], the transitions that bridge comic panels and describe the relation of gutters can be placed in several categories. These transitions are how the focus shifts in consecutive panels. We employed the following transitions to plan the shift between image content.

- **Action-to-Action** - This transition featuring a single subject in distinct action progression. We applied this to choose action for next panel.
- **Moment-to-Moment** - This transition captures slight changes in time and space. We use it to restrict action selections.
- **Aspect-to-Aspect** - This transition sets a wandering eye on different aspects of a place, or idea. When applying, we link it with image compositions.

Table 22

Examples of applying transitions between panels.

Action Pre	Action After	Aspect Pre	Aspect after	Scene Pre	Scene after	Moment Pre	Moment After
							

- **Scene-to-Scene** - This transition transport readers across a significant distance of space. Our model uses it to rearrange the scene.

Table 22 shows an example of how the image content changes after applying each transition correspondingly.

7.3.5 Local Modifications

Besides the global structure of the whole comic sequence and elements that consist of an image, the semantic in each panel are also important. This section will describe how we plan the local changes, in other words, content inside a panel with the help of the character's actions and several emotions.

7.3.5.1 Action Network By observing the common metaphor symbols in real comics, we recorded 19 actions and reactions. We defined an action relation net based on the actions, which is a graph that represents possible causal relations between actions. Table 23 lists part of the basic actions we have in our model and their reactions. The relation between actions and their reactions predicate the possible outcomes of the actions; hence can decide what will possibly happen in the next panel. For example, if the character is sitting, it will not start running immediately and will not start rolling or falling too. Similarly, if a character is falling, it will not fall asleep right after falling. Actions and their possible result form the action relation net in our model. It can be modified and extended

Table 23

Part of sample actions and their reactions.

Actions	Reactions
Stand	[Stand], [Sit], [Upset], [Laugh], [Mad], [Shock], [Walk], [Run], [Jump], [Dizzy], [Worry], [Think], [Relief]
Sit	[Stand], [Sit], [Upset], [Laugh], [Mad], [Shock], [Eat], [Drink], [Dizzy], [Worry], [Think], [Sleep], [Relief]
Fall	[Upset], [Mad], [Shock], [Collide], [Dizzy], [Worry]
Laugh	[Stand], [Sit], [Upset], [Laugh], [Mad], [Shock], [Jump] [Worry], [Think]
Collide	[Sit], [Upset], [Mad], [Shock]
Run	[Stand], [Roll], [Upset], [Laugh], [Mad], [Shock], [Collide], [Walk], [Run], [Jump]
Jump	[Stand], [Fall], [Collide]
...	...

by providing new actions and links.

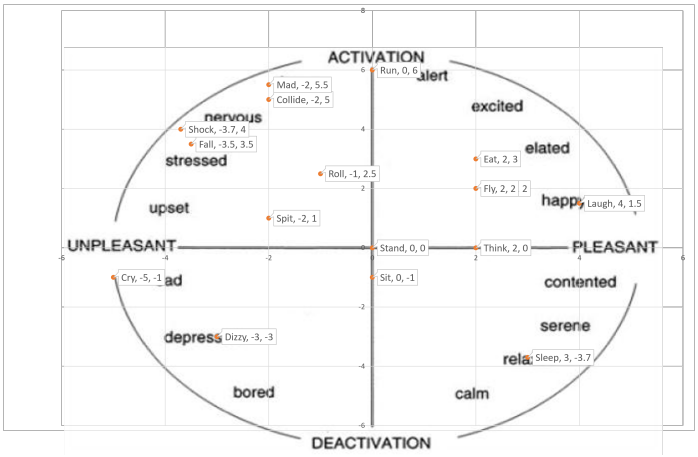
When constructing a new panel, besides characters and the scene, what characters are doing is also important. More accurately, what characters are doing, in other words, the actions, forms the whole story content and makes the sequence reasonable. Therefore, actions in each panel were planned according to the action relation net. The action in the previous panel decides the possible candidate of action in this panel.

7.3.5.2 Circumplex model and Modification To bridge the actions with narrative structure and narrative arc, map the actions with narrative arc score. We considered the possible emotion related to the actions and then borrowed the circumplex model of affect used in emotion classification to quantify the actions. In a circumplex model of affect [63],

the horizontal axis representing the valence dimension, and the vertical axis representing the arousal or activation dimension, we borrow the concept to imply how much tension, if it is denoted as a number, an action could change. Figure 18 shows how we linked the actions with emotion. For example, ”laughing” is an expression of ”happy,” so their coordinates are arranged in the same position. Similarly, ”relief” and ”sleep” are close to ”relax” because they are a common reaction when people feel relax. In contrast, action like ”stand” is rather neutral hence to be set to (0,0).

Figure 18

Circumplex model of affect, and the actions mapping according to their related emotion.

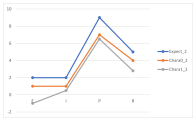
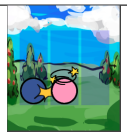
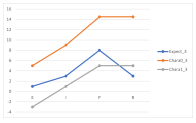


Then, we used the actions to adjust the current state to approach the narrative arc. Table 24 shows an example: a comic sequence with narrative structure [E, L, P, R]; if we map the structure into scores on a scale of 1 to 10 and use it to decide narrative arc, it will be [2, 5, 8, 3] $+(-1)^k \times l$, where l denotes a small modification range to create some oscillations on the curve that formed by tension score; and the k is a random number. The Establish(E) and the Release(R) mapped to the Exposition and Falling Action correspondingly in Freytag's Pyramid structure [57] of narrative arc, so their base score is relatively lower. On the contrary, the Prolongation(L) and Peak(P) were linked to the rising action and climate, hence higher scores.

In Table 24, starting from the first panel, which scored $2 + modification$, the char-

Table 24

Examples of applying action modification according to the narrative arc; and the action selection results follow tension score changes.

Set#	E	L	P	R	narra-arc
1					
2					

acter performs "stand" action. Among the possible action list, because the score in the next panel is $5 + modification$, which implies the tension should go up. The extent should be close to the score difference of the two panels. "run" and "jump" will have a higher chance of being chosen because they increase the tension score, which is closer to the wanted trend of the narrative arc. We were choosing actions based on possibility instead of selecting greedily because we want to keep the chance that the model can expand the different stories' content. The probability that an action will be chosen is $\frac{1}{distance(\frac{|s_{i+1}-s_i|^2}{s_{i+1}-s_i}, a_j)} / \sum \frac{1}{distance}$.The s_i represents the expected score of panel i, and a_j denotes the activation value of action j. And the distance between the wanted score and action value decides the probability. The s represent The figures in Table 24 also show the expected score trend and the result of selected actions. We can see that the action selection of each character followed the curve trend.

7.4 Experiments

We designed structure refinement layers through VNG and mapped the categories to curves that express narrative-arc to improve action selection for creating content. And another refinement layer that employed comic transitions to increase visual variety on panels. Moreover, we applied two extension layers to add richness to image content; Which

led to four experiments.

In the first experiment, we compared generated sequences based on a narrative arc decided by VGN or sequences or generated freely in the first experiment. This experiment aims to compare the content differences caused by following the narrative arc. The second experiment displayed the changes in content-led by the underlying rules applied when constructing content. In other words, this aims to show the possibility that alters the constructing layer to influence content details. Besides the overall structure and detail of panels, we also wanted to test the comic sequence's possibility by modifying the links between panels. The third experiment showed the flexibility in control the discourse that was affected because of inter-panel adjustment. The fourth experiment showed the possibility of extending the generator to enhance the richness of content expression.

7.4.1 Grammar Layer and Action Selection Based on Narrative Arcs

Our first two refinement layers implement the five categories of VGN and expand the center-embedded structure through probability for the whole comic panel. Then, according to the usages that the categories suggested, we mapped the categories to stages in narrative arcs. Then the action candidates that in our dataset utilized the activation and deactivation concept to get a corresponding coordinate in the Circumplex model, Figure 18. The coordinate implies whether an action raises tensions or drops tensions.

Therefore, the expanded narrative grammar tree basically decides how the story looks in the generated comic sequence. The Table 25 shows the results of whether the grammar and narrative arc layer is on or off.

The sequence with the narrative-arc layer suggests that the action selection follows a usual pattern. The action has higher tension such as collide and run showed up in peak and then tension relieved by more relaxed actions. And both the action of the characters follow the trend of the tension curve. On the contrary, the characters' action selection becomes rather arbitrary in the sequence without refining the narrative-arc layer. Even in the story

Table 25

Examples of with or without narrative grammar and narrative structures.

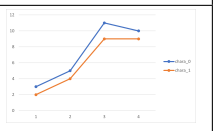
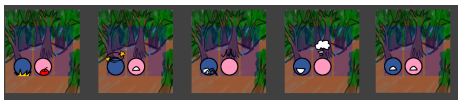
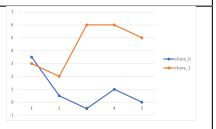
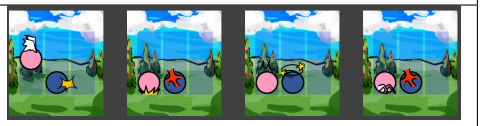
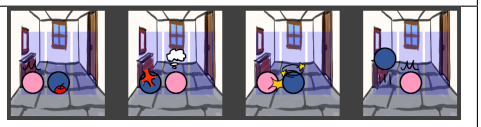
Settings	Layer: Narrative Grammar and Narrative Arc	Narrative Trend
with		
without		

Table 26

Examples of with or without action relation network.

Settings	Layer: Action Relations Network Layer
with	
without	

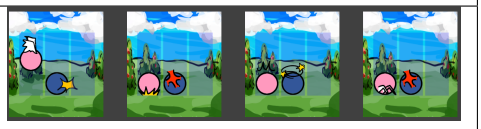
peak, the characters take relaxed actions.

7.4.2 Action relations network

The next refinement layer implements are action-reaction relation network of possible actions. For example, the ”fall” happens after ”jump” or ”fly,” ”dizzy” happens after ”collide,” and so on. The Table 26 suggests the results where the action relation network was on or off. The sequence with the action relation network can easily link the panel with a continuous story, whereas the other can not. In the former, the blue character ”collide” something, so it becomes ”angry” and then ”dizzy The pink character ”fly” and then ”falls.” After that, it is ”shocked” and then takes ”a rest.” On the contrary, characters in the later

Table 27

Examples of with or without transition network.

Settings	Layer: Transition layer
with	
transition sequence	[Aspect-to-Aspect, Action-to-Action, Scene-to-Scene]
without	

sequence select actions randomly.

7.4.3 Transition layer

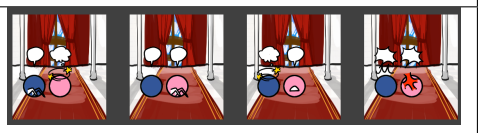
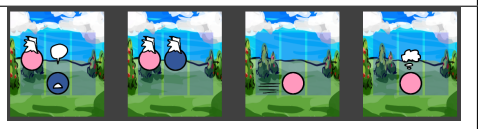
The transition refinement layer uses the transitions to modify the story presentation of the comic sequence. For example, in the Table 27, the sequence linked by [Aspect, Action, Scene] transitions, adjusted the image's composition, modifies the chosen actions, and transports a significant space. The view angle changes make the story a bit more interesting. The modification makes the story seem to be interpreted as two characters meeting in a garden, and one suddenly leaves the place. When the other follows the character, it saw the character sit in the forest.

7.4.4 Customize layers

The final two refinement layers are the cases that extend the generator with either image set or functions. The textbox layer provides three types of text balloon where the sharpness degree of ballon's edge help to emphasize the emotion of characters. The action with higher activation scores will be pairing the sharp ballon when the mild action will

Table 28

Examples that apply additional refinement layers

Settings	
Textbox layer	
Display layer	

be paired with the normal talking ballon. Furthermore, the shapes present whether the character is talking or thinking. The display layer implemented the function that flips the display flag of characters on or off by probability. In the last two panels, as the blue character's display flag was turned off, the whole sequence then shows a story that two characters flying together, but one disappeared eventually. Both of the extended layers enrich the image content and add more detail to the presented story. The results are in Table 28

7.5 Discussion

In the experiments, we demonstrated the effect of different comic theories. VNG helped the progression of comic content follow certain patterns. The use of action relations enhanced the reasonability of interpretations of the comic sequence; furthermore, the transitions controlled the pace of content changing and benefit shifting the space in generated comic—meanwhile, the composition templates and the collected symbolic metaphor increases the dynamic of panel images. With examples generated by these layers, the ability of our generator to project comic theories to content and the interestingness of randomized sequential content are clear.

The possibility of combing the refinement layers is many. We can estimate this by discussing the possible modifications of each layer. For the overall progression pattern, because our narrative arc depended on the center-embedded structures generated from VNG, the possibility of structure can be presented as $2^4 \times (2^5 \times 2^4)^{n-1}$. Given that except the Peak(P), all other categories can be either exist or not, this leads to the 4 power of 2. And, each category can potentially expand further with grammar sequence to get a new layer, thus causes each new layer to have 5 power of 2's possibility.

With our sample action relation network, suppose we have p actions and represent the network as a directed graph; the upper bound of the possible action sequence that constructs the content will be $\frac{(p-2)!}{(p-2-k)!}$ if the sequence length is k. The reasonable links between action and their possible reactions will graph a subset of the complete graph with p nodes. Therefore, the actions can only create at most the number of possible paths between two verticles with length k. In our sample network, the possibility will be $\frac{17!}{12!}$

Chapter 8

Project Design

Our target comprehension model will be developed based on Scene Perception & Event Comprehension theory (SPECT, subsection 3.2.1) since it formalized the cognitive process about visual narrative comprehension. We are going to adjust the SPECT by combining scene representation from VNE (subsection 3.3.1) and the modified hierarchical LSTM, which is inherited from Iyyer's version for multi-modal understanding.

8.1 Overall Structure of Modified SPECT

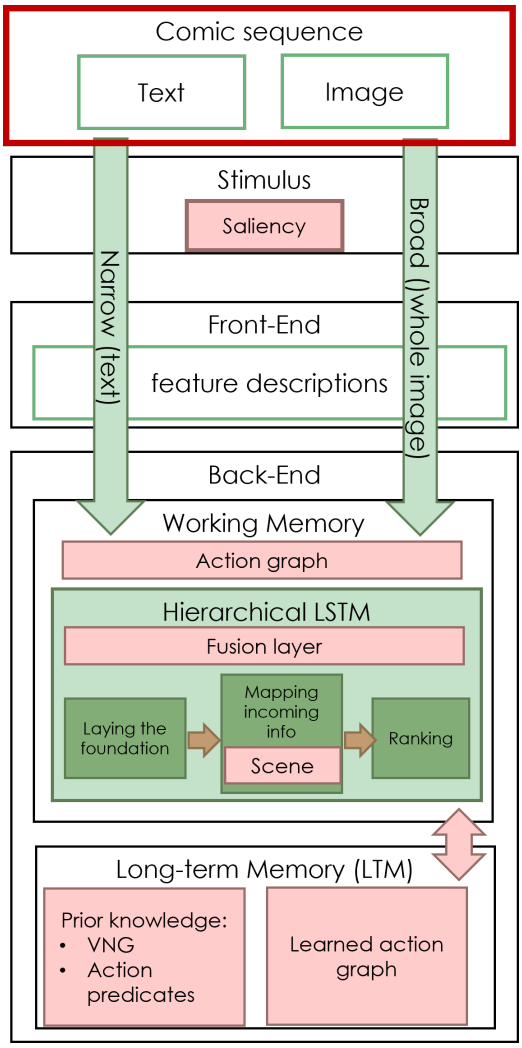
The Figure 19 gives the graphic of our expected model. For the three main components of the model: Stimulus, Front-End, and Back-End, the following sections will provide an overview of each part and planned details about how to achieve these.

8.1.1 Stimulus to Front-end: Attention Selection and Features Description

From the Stimulus part to the Front-end side, we map the attention selection process to the multi-modal information of comics, while the comic sequences are still the stimulus sources. The original SPECT model divided the information from narrow to broad according to the degrees of details because of the needed times of fixations. In our modified version, we map the text information and the frame, object annotations as the narrow part, whereas the features from the whole panel become the broad part. We reframe the attention selection process in the front-end to divide the information type and encrypt the information with feature vectors instead. The features descriptors in our first plan will be the word-embedding and pre-trained CNN weights for texts and images correspondingly.

Figure 19

The proposed model which modified the SPECT with scene representation from VNE and the hierarchical LSTM.



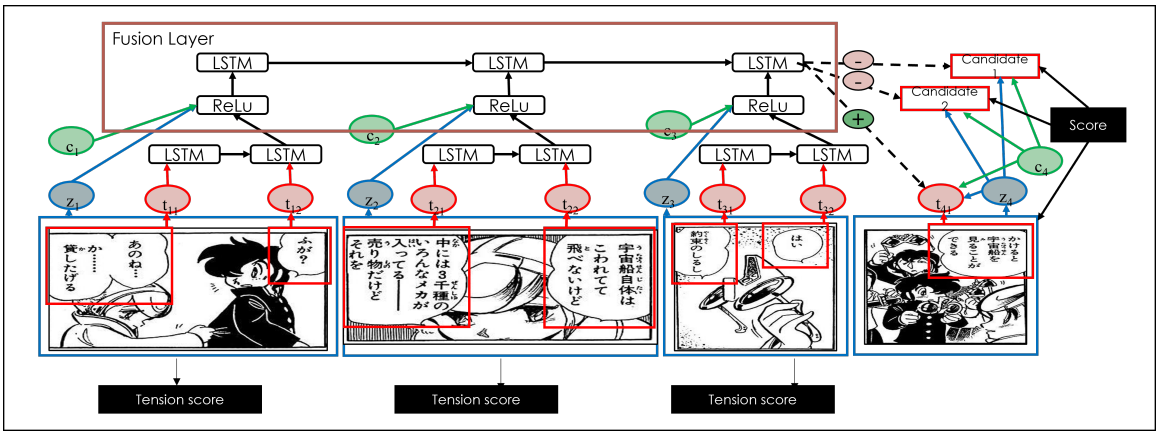
8.1.2 Back-End: Scene Representations, Action Graphs, and Hierarchical LSTM

In the back-end, we use hierarchical LSTM and the scene representation to implement the current event model.

8.1.2.1 Hierarchical LSTM The Figure 20 shows the view of our improved hierarchical LSTM model. We adjust Iyyer's version by adding the scene representations, narrative structure predictions and plugging a fusion layer to combine the features through

Figure 20

Our version of hierarchical LSTM.



a weighted function.

In the original hierarchical LSTM, the visual feature vectors were injected into the second layer as panel context. We preserve this part and mix the scene graphs–relation representation of panel content–into this layer. Therefore, we can mix narrative aspects' knowledge as features to describe the context. Meanwhile, when processing the panel context, we plan to ask the model to predict a tension score, which in the end, will form a tension curve for the sequence. In our side project (subsection 7.3.4), we discussed how the narrative structure project to a numerical value which links to narrative arcs of story discourse. The tension score curve of the read comic sequence can reflect the possible visual narrative grammar sequence. We can reversely deduct the possible canonical narrative arc. Taking the sequence in refNarrativeGrammar as an example, the result VNG sequence is "I-P-E-I-P-R," it can trace back to several possible ways that can lead to the sequence, including the "I-R" pattern that the comic sequence actually used.

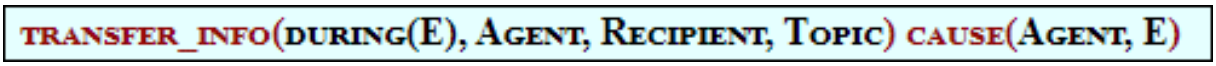
Then, outside the second layer, we plan to plug in a fusion layer that will use a weighted function to balance text features with visual and narrative features.

8.1.2.2 Scene Graphs and Action Graphs Our model will read each comic sequence twice to construct the scene representation and the comic panel's scene graph. The

first time retrieves the relations between actions to construct a graph where actions are the vertex, and the links are the edges. Then the model cooperates with the verb predicates provided by VerbNet [64], where the VerbNet defines the predicates through combining Thematic Roles with conditions and results for performing the actions. For example, the predicates of "read" is the predicate in Figure 21. This can be interpreted as that the agent causes the event E, and during the event E it is a recipient for information transferring about a topic.

Figure 21

The VerbNet's predicate for verb "read".

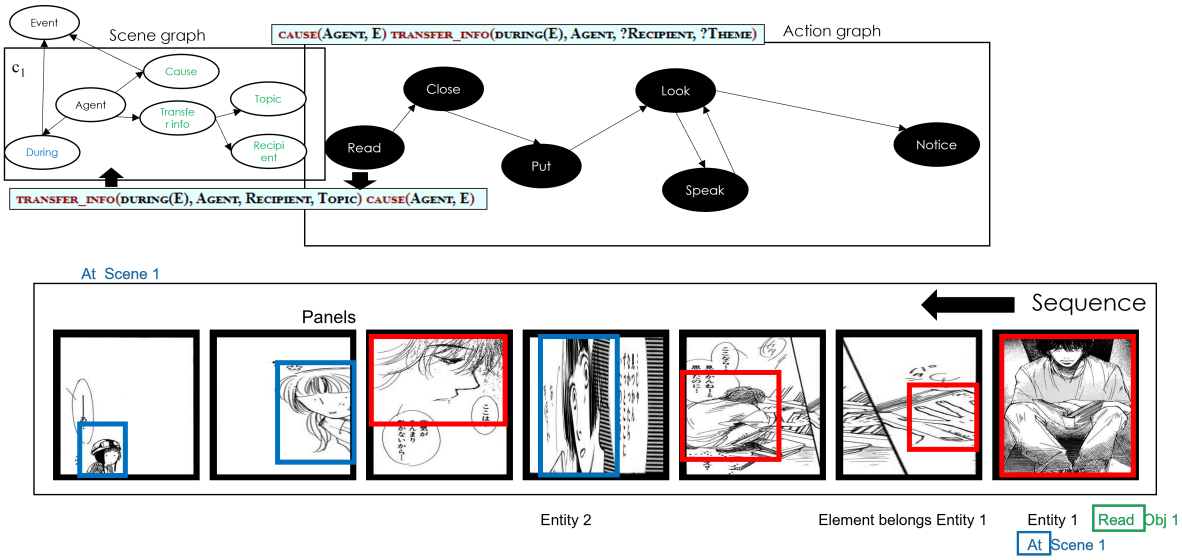


We then will use the action graph and the predicate to build the scene graphs of comic panels. The Figure 22 gives an example of the action graph and the scene graph obtained from the predicate. The action performer, in other words, who acts as the current panel, is explicit information in the dataset's annotation. Therefore, only the actions that show up in comic panels need to be detected. The Figure 23 gives the annotated information we can acquire from the dataset. The missing part is the labels about scenes and actions; as a result, we plan different developing stages to solve this issue. section 8.2 will discuss the details.

8.1.2.3 Prior Knowledge Mapping to the prior knowledge part in SPECT, which is stored in Long-term memory. We consider the verb predicate pool of VerbNet and the patterns of VNG as the knowledge that helps our model create scene graphs and predict possible narrative structures. Moreover, whenever the model reads a new sequence, it will merge the newly constructed action graph into previously stored ones. By doing so, we will get a possibly deliverable action network that records the causal relation or the order of various actions. This accumulates the knowledge acquired from reading comic sequences.

Figure 22

Action relation graph from comic sequence.



8.2 Stages

To smooth the difficulties we might encounter during implementation, we divide the development into a few stages.

8.2.1 Possible Tasks and Difficulties

The tasks toward this project are listed below:

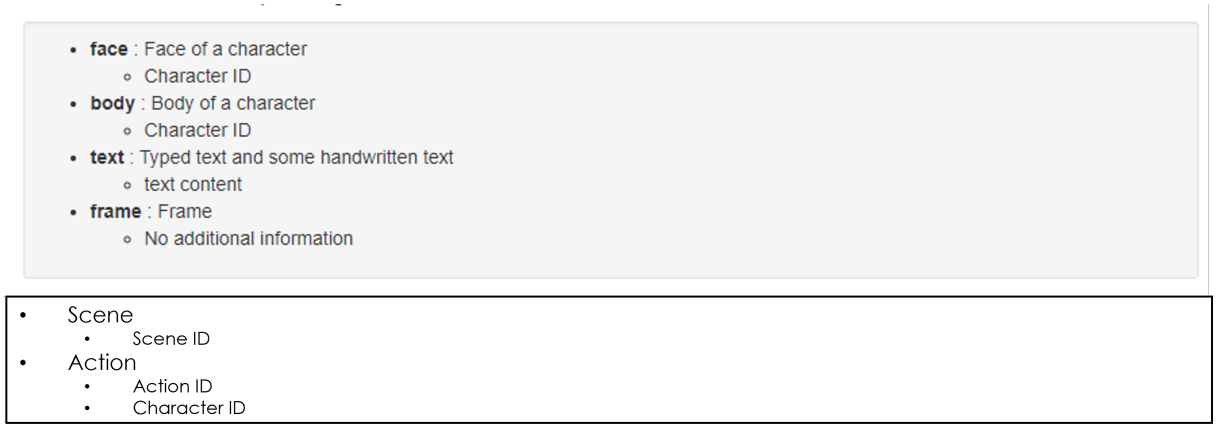
Detect actions from comic panel: The problem in this part involves research questions about how to detect actions in non-photorealistic images. Although the related topic has been studied in computer vision for photos, we are unsure whether a similar method will be applicable or how well the results will be. Therefore, this is the biggest challenge in our estimations.

Construct action graph: Once the actions can be detected, the most effort for this part will be the implementation of defining and designed graph operation on action graphs.

Build scene graph: the question in this part is to create scene graphs according to predicates of actions. Given that the VerbNet already provides Application Program

Figure 23

Annotations in Manga109 dataset.



Interface for retrieving predicates, the challenge here will be parsing and frame the known entities into the roles in predicates. Furthermore, in this task, the ambiguity of entities should be resolved because there is no guarantee that we can get a complete scene, entity information from every comic panel.

Rank the tension degree of comic panels: to tackle the task, we plan to use the low-level features in the pre-trained CNN model as the features descriptor of panel images because the textures, strokes such as the lines or visual effects that show atmosphere or emphasize actions is highly close to lower-level features. Furthermore, in our current plan, we will employ a ranking algorithm like ”trueskill” to digitize comic panels' tension.

Integrate scene representation with LSTM model: the challenge in this task is to transform the scene graph into a one-dimensional representation like a vector; hence the narrative knowledge can be used in the LSTM model. The remaining part is the implementation to put all the fragments together.

To divide the challenges, the planned development will be three stages.

8.2.2 Prototype

Our first stage of the model development is a prototype that tested on the results from our comic creation side project (chapter 7). The benefit of using generated comics

as test cases are several: first, the generated comics followed the comic theories that we employed in this work; this prevents the troubles about handling possible exceptions in real comics.

Second, the comic generation is actually the dual direction of comic comprehension. While comic generation employs narrative structure and action relations to create content, the comprehension side learns the unseen relations from perceived actions and figures the underlying narrative structure of the story event. Therefore, before we tackle the action detection problem, the clear action annotations from generated comics assist the construct of model functions rather than waiting for action labels on real comics. So, the generated comics can be used as the testing data for our very first stage.

8.2.3 Minimal Model

The second stage of the model development tests the functions on a limited set of real comic examples. Compared to using the generated comics as cases, real comics are likely to have a broader set of possible actions. The action graph constructing process we used for generated comics might encounter difficulties when applying to real comics.

In this stage, to solve the issue that lacks action labels, we plan to choose two or three comics from certain two genres separately. And then manually annotate all or some panels with action labels. This step, although exhausted, can demonstrate whether our theory works well on real comics.

8.2.4 Final Model

The final stage happened at the moment we successfully solve the action detection problem. We then plan to shift all the functions to a real comic dataset. This step can achieve the true scalability of the comic comprehension model.

Chapter 9

Timeline and Expected Outcomes and Impacts

Our estimation for finishing the project is around six to eight months.

9.1 Expected Timeline

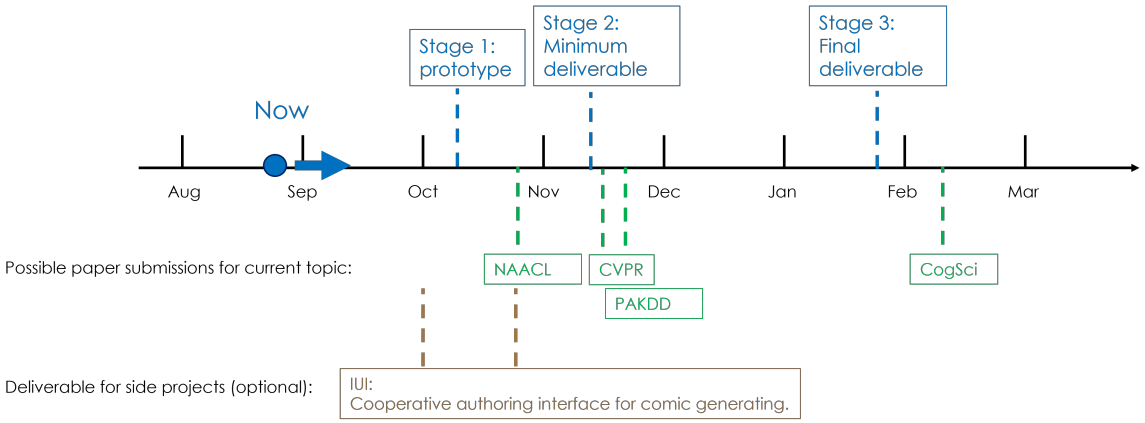
The first stage is planned to be finished around mid-October because the challenges we will face in this stage are transforming the action graph structures to a one-dimensional form and implementing the model and the syntax parsing for predicates to form scene graphs. Then the second stage is expected to be finished in one or two months after the first stages' implementation. The main task for this stage is to annotate actions labels for a confined set of real comics. In addition, the potential difficulties are caused by the difference between a larger action set and a limited prototyped set. The third stage is expected to take two to three months, which will end in February or March, due to the action detection problem and the scaling of model implementation.

If the progress goes well, we plan to submit the results of each development stage to conferences that are close to the estimated finish dates. Meanwhile, we plan to enhance the side project (chapter 7) with an intelligent interface for generating comics. But this extra implementation is only expected when the development stages catch the progress. If some unsure difficulties are encountered, the expected minimum deliverable should be the finished model of the second stage.

The Figure 24 shows the estimated timeline for finishing each stage and the potential conference targets.

Figure 24

The expected timeline.



9.2 Expected Results and Contributions

We expect the result will be a narrative comprehension model based on cognitive science theories. This will provide a baseline to compare the performance of comic comprehension tasks. Besides, the test dataset in each stage will potentially become the systematic benchmark of multi-modal narratives. Meanwhile, the learned annotations or labels that enhance existing comic datasets will benefit further research on comic semantic understanding. Finally, the techniques we use for tackling potential difficulties will also include the tool for solving similar problems, such as extracting semantic from comic panels or the action network learned through the training process.

References

[1] L. C. Loschky *et al.*, "The scene perception & event comprehension theory (spect) applied to visual narratives," *Topics in cognitive science*, vol. 12, no. 1, pp. 311–351, 2020.

[2] N. Cohn, "Your brain on comics: A cognitive model of visual narrative comprehension.," *Topics in cognitive science*, 2019.

[3] N. Cohn, "Visual narratives and the mind: Comprehension, cognition, and learning," in *Psychology of learning and motivation*, vol. 70, Elsevier, 2019, pp. 97–127.

[4] ——, "Visual narrative structure," *Cognitive science*, vol. 37, no. 3, pp. 413–452, 2013.

[5] ——, "The architecture of visual narrative comprehension: The interaction of narrative structure and page layout in understanding comics," *Frontiers in Psychology*, vol. 5, p. 680, 2014.

[6] N. Cohn, R. Jackendoff, P. J. Holcomb, and G. R. Kuperberg, "The grammar of visual narrative: Neural evidence for constituent structure in sequential image comprehension," *Neuropsychologia*, vol. 64, pp. 63–70, 2014.

[7] S. McCloud and M. Martin, *Understanding comics: The invisible art*. Kitchen sink press Northampton, MA, 1993, vol. 106.

[8] S. McCloud and A. Manning, "Understanding comics: The invisible art," *IEEE Transactions on Professional Communications*, vol. 41, no. 1, pp. 66–69, 1998.

[9] S. McCloud, *Making comics: Storytelling secrets of comics, manga and graphic novels*. Harper New York, 2006.

[10] C. Martens, N. EDU, R. E. Cardona-Rivera, and U. EDU, "The visual narrative engine: A computational model of the visual narrative parallel architecture," in *8th Annual Conference on Advances in Cognitive Systems*, 2020.

[11] M. Iyyer, V. Manjunatha, A. Guha, Y. Vyas, J. Boyd-Graber, H. Daume, and L. S. Davis, "The amazing mysteries of the gutter: Drawing inferences between panels in comic book narratives," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7186–7195.

[12] R. Narita, K. Tsubota, T. Yamasaki, and K. Aizawa, "Sketch-based manga retrieval using deep features," in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, IEEE, vol. 3, 2017, pp. 49–53.

[13] Y. Matsui, K. Ito, Y. Aramaki, A. Fujimoto, T. Ogawa, T. Yamasaki, and K. Aizawa, "Sketch-based manga retrieval using manga109 dataset," *Multimedia Tools and Applications*, vol. 76, no. 20, pp. 21 811–21 838, 2017.

[14] M. P. Graves, "The comics meet the quakers: Scott mccloud's theory of the comics employed to unravel the quaker tapestry," *Studies in Popular Culture*, vol. 23, no. 1, pp. 75–90, 2000.

[15] N. Cohn, "The architecture of visual narrative comprehension: The interaction of narrative structure and page layout in understanding comics," *Frontiers in Psychology*, vol. 5, p. 680, 2014, ISSN: 1664-1078. DOI: 10.3389/fpsyg.2014.00680. [Online]. Available: https://www.frontiersin.org/article/10.3389/fpsyg.2014.00680.

[16] T. Tanaka *et al.*, "Layout analysis of tree-structured scene frames in comic images.," in *IJCAI*, vol. 7, 2007, pp. 2885–2890.

[17] W. L. Taylor, "Cloze procedure: A new tool for measuring readability," *Journalism quarterly*, vol. 30, no. 4, pp. 415–433, 1953.

[18] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[19] P. Simeone, R. Santos-Rodrguez, M. McVicar, J. Lijffijt, and T. De Bie, "Hierarchical novelty detection," in *International Symposium on Intelligent Data Analysis*, Springer, 2017, pp. 310–321.

[20] S. Tafreshi and M. Diab, "Emotion detection and classification in a multigenre corpus with joint multi-task deep learning," in *Proceedings of the 27th international conference on computational linguistics*, 2018, pp. 2905–2913.

[21] D. Pritsos, A. Rocha, and E. Stamatatos, "Open-set web genre identification using distributional features and nearest neighbors distance ratio," in *European Conference on Information Retrieval*, Springer, 2019, pp. 3–11.

[22] K. Choroś, "Video genre classification based on length analysis of temporally aggregated video shots," in *International Conference on Computational Collective Intelligence*, Springer, 2018, pp. 509–518.

[23] P. Doshi and W. Zadrozny, "Movie genre detection using topological data analysis," in *International Conference on Statistical Language and Speech Processing*, Springer, 2018, pp. 117–128.

[24] P. G. Shambharkar, A. Anand, and A. Kumar, "A survey paper on movie trailer genre detection," in *2020 International Conference on Computing and Data Science (CDS)*, IEEE, 2020, pp. 238–244.

[25] A. Yadav and D. K. Vishwakarma, "A unified framework of deep networks for genre classification using movie trailer," *Applied Soft Computing*, vol. 96, p. 106 624, 2020.

[26] Y. Daiku, M. Iwata, O. Augereau, and K. Kise, "Comics story representation system based on genre," in *2018 13th IAPR International Workshop on Document Analysis Systems (DAS)*, IEEE, 2018, pp. 257–262.

[27] B. Park and M. Matsushita, "Estimating comic content from the book cover information using fine-tuned vgg model for comic search," in *International Conference on Multimedia Modeling*, Springer, 2019, pp. 650–661.

[28] K. Kundalia, Y. Patel, and M. Shah, "Multi-label movie genre detection from a movie poster using knowledge transfer learning," *Augmented Human Research*, vol. 5, no. 1, pp. 1–9, 2020.

[29] M. Sreeja and B. C. Kovoor, "Towards genre-specific frameworks for video summarisation: A survey," *Journal of Visual Communication and Image Representation*, vol. 62, pp. 340–358, 2019.

[30] Y. Yu, Z. Lu, Y. Li, and D. Liu, "Asts: Attention based spatio-temporal sequential framework for movie trailer genre classification," *Multimedia Tools and Applications*, pp. 1–16, 2020.

[31] H. J. Pratt, "Narrative in comics," *The Journal of Aesthetics and Art Criticism*, vol. 67, no. 1, pp. 107–117, 2009.

[32] Y. Daiku, O. Augereau, M. Iwata, and K. Kise, "Comic story analysis based on genre classification," in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, IEEE, vol. 3, 2017, pp. 60–65.

[33] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[34] M. L. McHugh, "Interrater reliability: The kappa statistic," *Biochemia medica*, vol. 22, no. 3, pp. 276–282, 2012.

[35] N. Ashikhmin, "Fast texture transfer," *IEEE Computer Graphics and Applications*, vol. 23, no. 4, pp. 38–43, 2003.

[36] Y. Shih, S. Paris, C. Barnes, W. T. Freeman, and F. Durand, "Style transfer for headshot portraits," 2014.

[37] W. Zhang, C. Cao, S. Chen, J. Liu, and X. Tang, "Style transfer via image component analysis," *IEEE Transactions on multimedia*, vol. 15, no. 7, pp. 1594–1601, 2013.

[38] P. Rosin and J. Collomosse, *Image and video-based artistic stylisation*. Springer Science & Business Media, 2012, vol. 42.

[39] S. Bruckner and M. E. Gröller, "Style transfer functions for illustrative volume rendering," in *Computer Graphics Forum*, Wiley Online Library, vol. 26, 2007, pp. 715–724.

[40] C. Ma, H. Huang, A. Sheffer, E. Kalogerakis, and R. Wang, "Analogy-driven 3d style transfer," in *Computer Graphics Forum*, Wiley Online Library, vol. 33, 2014, pp. 175–184.

[41] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2414–2423.

[42] Y. Jing, Y. Yang, Z. Feng, J. Ye, Y. Yu, and M. Song, "Neural style transfer: A review," *IEEE transactions on visualization and computer graphics*, vol. 26, no. 11, pp. 3365–3385, 2019.

[43] T. Q. Chen and M. Schmidt, "Fast patch-based style transfer of arbitrary style," *arXiv preprint arXiv:1612.04337*, 2016.

[44] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision*, Springer, 2016, pp. 694–711.

[45] X. Shen, A. Hertzmann, J. Jia, S. Paris, B. Price, E. Shechtman, and I. Sachs, "Automatic portrait segmentation for image stylization," in *Computer Graphics Forum*, Wiley Online Library, vol. 35, 2016, pp. 93–102.

[46] C. Castillo, S. De, X. Han, B. Singh, A. K. Yadav, and T. Goldstein, "Son of zorn's lemma: Targeted style transfer using instance-aware semantic segmentation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2017, pp. 1348–1352.

[47] L. A. Gatys, A. S. Ecker, M. Bethge, A. Hertzmann, and E. Shechtman, "Controlling perceptual factors in neural style transfer," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3985–3993.

[48] O. Augereau, M. Iwata, and K. Kise, "A survey of comics research in computer science," *Journal of Imaging*, vol. 4, no. 7, p. 87, 2018.

[49] A. Gordić, *Pytorch-nst-feedforward*, https://github.com/gordicaleksa/pytorch-nst-feedforward, 2020.

[50] L. A. Gatys, A. S. Ecker, and M. Bethge, "A neural algorithm of artistic style," *arXiv preprint arXiv:1508.06576*, 2015.

[51] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[52] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

[53] P. Chen, C. Agarwal, and A. Nguyen, "The shape and simplicity biases of adversarially robust imagenet-trained cnns," *arXiv preprint arXiv:2006.09373*, 2020.

[54] C. Zauner, "Implementation and benchmarking of perceptual image hash functions," 2010.

[55] W. Eisner, *Comics and sequential art: Principles and practices from the legendary cartoonist*. WW Norton & Company, 2008.

[56] D. Carrier and M. A. Oliker, *The aesthetics of comics*, 2001.

[57] G. Freytag, *Freytag's technique of the drama: an exposition of dramatic composition and art*. Scott, Foresman and Company, 1908.

[58] J. Stern, *Making shapely fiction*. WW Norton & Company, 2000.

[59] T. Alves, A. McMichael, A. Simões, M. Vala, A. Paiva, and R. Aylett, "Generating graphical content for comics,"

[60] M. Nairat, M. Nordahl, and P. Dahlstedt, "Generative comics: A character evolution approach for creating fictional comics," *Digital Creativity*, vol. 31, no. 4, pp. 284–301, 2020.

[61] C. Martens and R. E. Cardona-Rivera, "Discourse-driven comic generation," in *Proc. International Conference on Computational Creativity*, 2016.

[62] ——, "Generating abstract comics," in *International Conference on Interactive Digital Storytelling*, Springer, 2016, pp. 168–175.

[63] J. Posner, J. A. Russell, and B. S. Peterson, "The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology," *Development and psychopathology*, vol. 17, no. 3, pp. 715–734, 2005.

[64] K. K. Schuler, *VerbNet: A broad-coverage, comprehensive verb lexicon*. University of Pennsylvania, 2005.