Look Over Here: Attention-Directing Composition of Manga Elements

Ying Cao Rynson W.H. Lau Antoni B. Chan

Department of Computer Science, City University of Kong Kong

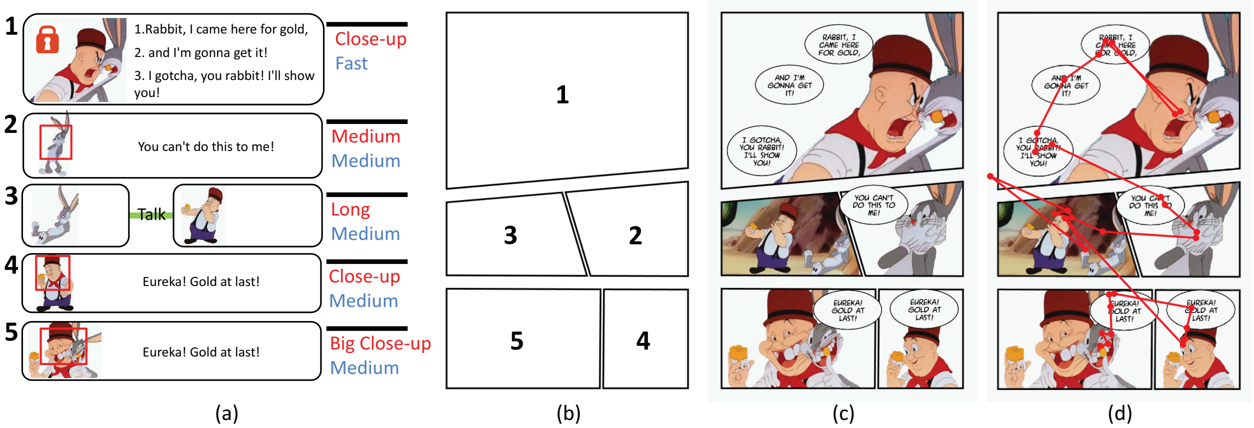


Figure 1: *Given a storyboard comprising input subjects and dialogue for each panel ("The Wacky Wabbit" in the public domain), interaction type between any two subjects (green line connecting two subjects in panel 3), and shot type (red text) and motion state (blue text) of each panel (a), our approach automatically generates a layout of panels (b), and then produces a storytelling composition of subjects and speech balloons on the layout (c), which effectively directs the viewer attention through the page. Red rectangles on the subjects represent regions of interest, and a red lock icon in panel 1 indicates that the subject is fixed in place by the user. The recorded viewer's eye movements is plotted as a red path (d). The background image is added to provide necessary context. The reading order of manga is right to left, and top to bottom.*

Abstract

Picture subjects and text balloons are basic elements in comics, working together to propel the story forward. Japanese comics artists often leverage a carefully designed composition of subjects and balloons (generally referred to as *panel elements*) to provide a continuous and fluid reading experience. However, such a composition is hard to produce for people without the required experience and knowledge. In this paper, we propose an approach for novices to synthesize a composition of panel elements that can effectively guide the reader's attention to convey the story. Our primary contribution is a probabilistic graphical model that describes the relationships among the artist's guiding path, the panel elements, and the viewer attention, which can be effectively learned from a small set of existing manga pages. We show that the proposed approach can measurably improve the readability, visual appeal, and communication of the story of the resulting pages, as compared to an existing method. We also demonstrate that the proposed approach enables novice users to create higher-quality compositions with less time, compared with commercially available programs.

CR Categories: I.3.8 [Computer Graphics]: Applications;

Keywords: manga, composition, probabilistic graphical model

Links: DL PDF

1 Introduction

Japanese comics has grown to become one of the most popular storytelling mediums across the world, with thousands of amateur artists creating their own strips and imagery. The success of manga can be attributed to the sophisticated utilization of unique storytelling techniques to amplify the sense of reader participation [McCloud 2006]. Among these techniques, the composition of foreground subjects and text balloons across a page is especially important for providing readers with a continuous and smooth reading experience. Unlike films, elements in comics are arranged in space rather than in time. Consequently, communication of the story heavily relies on the reader attending to the right place at the right time [Jain et al. 2012], e.g., the text balloons should be read in the correct order and associated with the correct subjects. Manga artists typically control the viewer attention via subject and balloon placement, so as to lead the reader continuously through the page [Folse 2010]. In this way, subjects and balloons, in addition to providing necessary information, can act as "road signs", guiding the readers through the artworks for better understanding of the story. We refer to this path through the page as the underlying *artist's guiding path (AGP)* and the viewer's eye-gaze path through the page as the actual *viewer attention*.

Such a skilled composition is a daunting task, requiring significant expertise and hands-on experience. It is deemed by some professional manga artists as one of the most difficult tasks in creating manga [SamuraiMangaWorkshop 2011]. Many state-of-the-art commercial comic maker programs [MangaStudio 2011; ComiPo 2012; Scott-Baron 2006] simplify the manga creation process using a *select-and-assemble* paradigm, especially for novices. Instead of drawing contents from scratch, users can directly *select* desired elements, such as characters, balloons and screen tones, from a library of pre-made objects, and *assemble* them on a page to yield the final artwork. Unfortunately, no existing programs support automated element composition, leaving the *assemble* operation to be done manually. Instead of using a fixed set of local rules as in western comics, manga artists compose the panel elements in a flexible and global way [Folse 2010]. As such, purely heuristic approaches for comic lettering [Kurlander et al. 1996; Chun et al. 2006] cannot be easily adapted to address this problem.

This paper proposes a probabilistic reasoning approach for automatic arrangement of panel elements, which permits novice users to create professional-looking manga composition with little effort. Given a simple storyboard comprising panel elements, i.e., pre-made subjects and their balloons, and semantic information (Fig. 1(a)), our approach first generates a layout of panels (Fig. 1(b)) and then places the elements on the layout so that the resulting composition functionally and stylistically looks like those produced by professional manga artists (Fig. 1(c)). At the core of our approach is a probabilistic graphical model for composition configuration, which captures relations among the panel elements, and relates them to AGP and viewer attention. Our key idea is to treat AGP, implicit to the artist, as a latent variable that is connected to all panel elements on the page. This permits interactions among the panel elements within the same panel as well as across multiple panels. In addition, by representing viewer attention as a result of the configuration of panel elements, our model can encode how viewer attention responds to variations in element configuration. We train our model using a set of manually annotated manga pages together with eye movement data of multiple readers captured from an eye-tracking system. With this trained model as a prior, we employ a *maximum a posteriori* (MAP) inference framework for panel element composition.

To evaluate the effectiveness of the proposed approach, we first perform a user study to compare our results with those of a traditional heuristic-based method. Experimental results indicate that our approach produces better composition in terms of readability and storytelling, and that our results are more successful in leading viewer gaze through the page. We then show with another user study that our approach allows novice users to rapidly create higher-quality compositions, as compared with a manual placement tool. Finally, we demonstrate how our model can be applied to automatically recover the AGP from a given manga page.

In summary, we introduce a novel probabilistic graphical model for subject-balloon composition. Based on this model, we propose an approach for placing a set of subjects and their balloons on a page, in response to high-level user specification, and evaluate its effectiveness through a series of visual perception studies.

2 Related Work

Graphical Element Composition. Composing discrete visual elements to produce a new form of visual medium has been an interest of the computer graphics community for a long time. Kim et al. [2002] created a jigsaw image mosaic by packing image tiles of arbitrary shapes into a given output shape domain. Huang et al. [2011] achieved a similar effect by matching Internet images with an input segmentation. AutoCollage [Rother et al. 2006] seamlessly composed representative images from a collection to build a compact photo collection representation. For the purpose of video navigation, video frames were composed in the form of multi-scale or dynamic images [Barnes et al. 2010; Correa and Ma 2010], allowing users to browse video content hierarchically and dynamically. While these techniques aim for a visually pleasing and compact composition of elements, our objective is to sequentially arrange elements, including subjects and balloons, on a page so that they effectively convey a story to the viewers.

Label Placement. Label placement aims at attaching text labels to point features on a map or diagram to maximize legibility, and is an important task for many applications such as automated cartography and geographical information systems [Christensen et al. 1995]. Since the label placement problem is known to be NP-hard, most algorithms follow a stochastic optimization framework [Christensen et al. 1994; Edmondson et al. 1996], where an objective function to measure labeling legibility is optimized. Treating subjects as feature points and their balloons as text labels, our composition problem can be cast as a type of label placement. However, as shown in [Chun et al. 2006], algorithms specifically designed for label placement fail to satisfy the reading order, and thus are not amenable to our problem. It is also not clear how to adapt such methods to improve their storytelling capabilities, as targeted by our approach.

Balloon Placement. There are only few works that address the problem of automatic positioning of speech balloons for comic strips. Kurlander et al. [1996] placed balloons at the top of the panels using a simple greedy strategy. This method preserves the reading order of balloons, but separates them far away from the speakers. Chun et al. [2006] considered relationships between balloons and corresponding speakers, as well as distances among the balloons, in order to make the balloons easier to read. However, all these techniques only address balloon placement in a single panel, using a set of local rules. They cannot reproduce manga-like subject and balloon compositions, which are designed with global consideration of a page. In contrast, our probabilistic graphical model captures long-range interactions among the elements across panels. This global strategy is more consistent with how manga artists position panel elements in practice, thus allowing us to yield reader-friendly compositions. In addition, the previous works do not perform subject placement. To the best of our knowledge, our work is the first to jointly place subjects and balloons in a principled framework, in order to form a attention-guiding coherent composition.

Eye-tracking based Research. Eye-tracking has been widely exploited by different research communities as a means of capturing human attention during interaction between the users and various media such as images or videos. DeCarlo et al. [2002] determined visually meaningful regions in an image using eye fixations captured by an eye tracker, and produced an abstract image that retains visual details within the regions. Judd et al. [2009] collected a large database of eye fixations on natural images, on which a linear classifier was trained to predict saliency in an image. Recently, several researchers also analyzed eye movement recordings for scene understanding [Ramanathan et al. 2011].

Omori et al. [2004] investigated factors that may guide eye movements during manga reading, and found that there is a strong link between eye movements and balloon positions. Jain et al. [2012] experimentally confirmed that comic artists direct viewer attention through element composition, by measuring the consistency of eye movements across different viewers. However, none of these works present a model or algorithm for composing subjects and balloons. Recent works [Toyoura et al. 2012] used eye-tracking on film frames to place balloons to avoid salient regions. However, they only use eye fixations as a proxy for saliency, and do not investigate how viewer attention is affected by variations in composition. In contrast, our graphical model captures composition variations and its subtle interaction with viewer attention. In addition, unlike these methods, which use eye-tracking data as user input, our approach uses eye tracking data for offline training only.

Computational Manga. Due to its popularity, manga has attracted interest in the computer graphics community, with efforts to make manga production accessible to the general population. Some works have been devoted to developing techniques for manga coloring [Qu et al. 2006], screening [Qu et al. 2008] and layout [Cao et al. 2012]. Our proposed work is a continuation of this line of research, and addresses the composition of subjects and balloons.

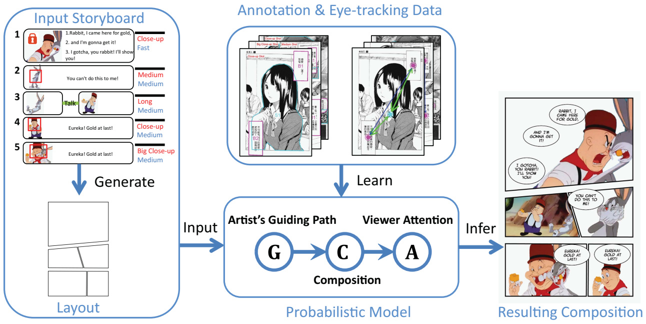


Figure 2: *Overview of our approach.*

3 Overview

Our approach is composed of two stages, as illustrated in Fig. 2. In the first stage, we learn a probabilistic model from a set of training examples. The training data set consists of manga pages from different manga series with representative and distinctive composition styles. For each page in the training set, we annotate all the main subjects and their balloons, and obtain eye movement data from multiple viewers reading the page using an eye-tracking system. In the second stage, we synthesize a composition in an interactive session, where the learned probabilistic model is used to generate a gallery of composition suggestions, in response to user-provided high-level specification. We have implemented an interactive prototype tool for composition synthesis. To synthesize a composition, the user only needs to make a storyboard. In particular, the user begins by specifying the number of panels. Then, for each panel, the user may choose the shot type and motion state (the amount of action, e.g., slow, medium and fast) of the panel, and add subjects along with their scripts. Last, the user may specify how two subjects inside a panel interact with each other. Given the input storyboard, our interface retrieves a layout of panels that best fits the input elements and semantics, from a database of labeled layouts, and then generates composition suggestions on the layout with a MAP inference framework. We incorporate well-known guidelines as a likelihood term and use the probabilistic model as a conditional prior. For computational efficiency, a sampling-based approximate inference method is employed to infer the most likely compositions, which are then presented as the suggestions.

4 Data Acquisition and Preprocessing

To train our probabilistic model, we have collected a data set comprising 80 manga pages from three chapters of three different series, "*Bakuman*", "*Vampire Knight*" and "*Fruit Basket*". These manga series have distinctive composition complexity and patterns, so that our data set is able to capture a wide range of composition styles used by manga artists. We manually annotate all the pages in our dataset. Each page is segmented into a set of panels. For each panel, we label its shot type (long, medium, close-up, or big close-up) [Kesting 2004] and motion state (slow, medium, and fast), and segment the foreground subjects (see Fig. 3 for an example). We further partition all segmented panels into three groups with similar geometric features, including aspect ratio (width/height) and size, using a Gaussian mixture model [Bishop 2006]. Grouping geometrically similar panels allows our model to learn composition patterns that vary with the panel shape.

To understand how manga artists control viewer attention via the composition of subjects and balloons, we use an eye-tracker to record the eye movements of 30 viewers as they read the manga pages in our dataset. The saccades (i.e., the rapid eye movements between eye fixations) in the eye movement data indicate how the viewer transition their attention among the segmented elements of interest (subjects and balloons). To compactly and visually represent such information, we preprocess the raw eye movement data to build an *element graph*, as shown in Fig. 3. In the graph, nodes represent panel elements, and each edge represents a transition of viewer attention between two elements. We define an attention transition from one element to another when more than 50% of the viewers transition through that route.

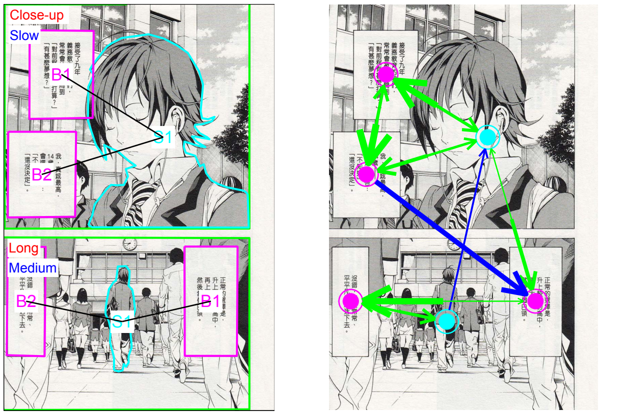


Figure 3: *Annotation and element graph. Left: a segmented and annotated manga page of "Bakuman" (© Tsugumi Ohba, Takeshi Obata / Shueisha Inc.). Foreground subjects (S) and balloons (B) are outlined in cyan and magenta. The shot type and motion state of each panel are labeled as red and blue, respectively, in the upper-left corner. Right: an example element graph, where nodes represent elements and edges denote transitions of viewer attention. A thicker arrow indicates that more viewers pass through that direction. Note that a transition can be bi-directional as the viewers might read back and forth to explore contents of interest.*

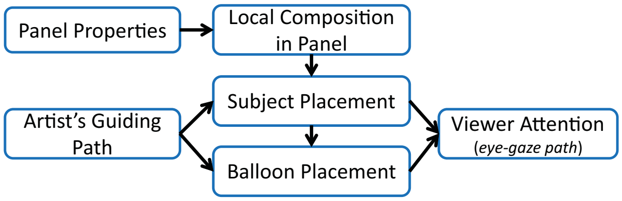
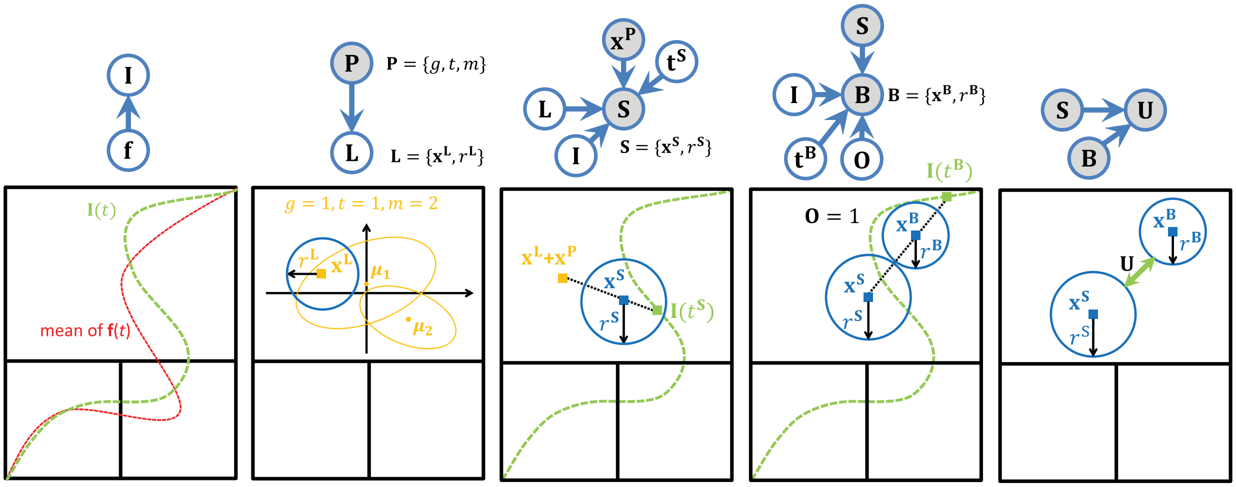


Figure 4: *Relationship between the 6 components in our probabilistic model.*

In summary, our training set $\mathcal{D}$ consists of manga pages with annotations for each panel, including panel properties (shot type, motion state, geometric style, and center location), subjects and balloons (center locations and radii), and viewer attention transitions between elements. More details on the data are in the supplemental.

5 Probabilistic Graphical Model

In our work, we aim at designing a model that can synthesize a composition that guides viewer attention over a set of panel elements. The desired model must characterize how subjects and balloons interact both locally within a panel and globally across the page, and also how the resulting composition relates to transitions in viewer attention. To this end, we propose a novel probabilistic graphical model to hierarchically connect subjects, balloons and viewer attention in a probabilistic network, and use it to generate subject and balloon compositions when trained on real-world manga pages. We abstract the artist's guiding path (AGP) as a continuous stochastic process, and represent it as a latent variable in our model. Modeling AGP explicitly enables our model to capture long-range interactions among the panel elements, making it possible to recover the AGP from a page.



(a) Artist's guiding path (b) Local composition model (c) Subject placement (d) Balloon placement (e) Viewer attention

Figure 5: *Components of the probabilistic graphical model. (a)* ***Artist's Guiding Path (AGP)****: underlying and actual AGP ($\mathbf{f}(t)$ and $\mathbf{I}(t)$) are represented as smooth splines over the page. (b)* ***Panel Properties*** *and* ***Local Composition Model****: the local position $\mathbf{x}^{\mathbf{L}}$ of a subject w.r.t. its panel can be regarded as a sample from a mixture model, whose parameters depend on the panel's shot type $t \in \{long = 1, medium = 2, close\text{-}up = 3, big\ close\text{-}up = 4\}$, motion state $m \in \{slow = 1, medium = 2, fast = 3\}$, and shape $g \in \{geometric\ style\ 1 = 1, geometric\ style\ 2 = 2, geometric\ style\ 3 = 3\}$. In this example, $\mathbf{x}^{\mathbf{L}}$ is from the first component of the mixture distribution for a long-shot panel with geometric style 1. Its local size $r^{\mathbf{L}}$ w.r.t. its panel is only contingent on the panel's shot type. (c)* ***Subject Placement****: the actual placement $\mathbf{x}^{\mathbf{S}}$ of a subject is a mixture of its local position $\mathbf{x}^{\mathbf{L}}$ and an associated point $\mathbf{I}(t^{\mathbf{S}})$ on the global AGP. (d)* ***Balloon Placement****: the placement of a balloon depends on its subject's configuration $\{\mathbf{x}^{\mathbf{S}}, r^{\mathbf{S}}\}$, its size $r^{\mathbf{B}}$, and reader order O, as well as an associated point $\mathbf{I}(t^{\mathbf{B}})$ on the AGP. (e)* ***Viewer Attention Transitions****: the presence of transitions in viewer attention $\mathbf{U}$ between two elements relies on properties of the involved elements and their surrounding elements (e.g., $\mathbf{S}$ and $\mathbf{B}$ in this example).*

Our proposed model consists of 6 components, representing different factors that influence the placement of elements on the page. Globally, the AGP is a continuous curve through the page, which passes through each panel. Each panel is described by a set of properties, including the shot type (e.g., close-up or long shot) and the geometry. Locally within a panel, the candidate placements of a subject follow a local composition model, which depends on the type of panel (e.g., subjects in a close-up shot tend to be large and fill the panel). The actual placement of a given subject within a panel depends on both the AGP (i.e., the global path), and the local composition model of the panel. Balloons of a subject must be placed close to it, but also according to the AGP to preserve continuity of the global path and the reading order. Finally, the viewer attention is determined by the placement of subjects and balloons, and is observed through the eye movement data. Fig. 4 shows how the 6 model components are related to each other conceptually. Sections 5.1 and 5.2 describe them in more details.

5.1 Model Components and Variables

In our model, the i^{th} page consists of a set of J_i panels. Each panel has M_{ij} subjects, each of which has N_{ijm} balloons. Fig. 5 illustrates the constituent parts of our model.

Artist's Guiding Path (Fig. 5(a)). Manga artists often intend to lead viewer attention to continuously travel through a page, starting at the upper-right corner and exiting at the lower-left corner. Consequently, we represent the AGP as a continuous random process $f(t) = (f^x(t), f^y(t))^T$, which is a distribution over parametric curves on a 2D page, where $t \in [0, 1]$ is normalized arc length parameter. However, it is impractical to operate directly on the random process, which is an infinite set of random variables. Inspired by [Friedman and Nachman 2000], we approximate AGP using a finite subset of random variables of the process. Specifically, we uniformly sample l control points along the curve length, i.e., $\mathbf{f} = (f(t_1), \cdots, f(t_l))$. We denote $\mathbf{f}$ as the locations of the *underlying* AGP, and $\mathbf{I}$ as the *actual* AGP that is a noisy version of the underlying AGP, i.e., $\mathbf{I} = (I(t_1), \cdots, I(t_l))$.

Panel Properties (Fig. 5(b)). We consider both semantic (i.e., shot type and motion state) and geometric (i.e., rough shape) properties of the panels. Specifically, the shot type of the panel is represented by a discrete random variable t, and the possible shot types include "long", "medium", "close-up", and "big close-up". Motion state (the amount of action) is represented by m, taking three possible values of "slow", "medium" and "fast". The geometric style of the panel is denoted by g, and indicates the geometric style cluster (obtained in Section 4) that the panel belongs to. Finally, $\mathbf{x}^{\mathbf{P}}$ are the coordinates of the center location of the panel.

Local Composition Model (Fig. 5(b)). The local composition (i.e., spatial distribution) of the subjects within a panel relies on the geometric (g) and semantic (t, m) properties of the panel. For example, given a horizontal rectangular panel of long shot, subjects are more likely to spread sparsely along horizontal direction, instead of squeezing around the panel center. In addition, placing subjects off-center or along the diagonal could enhance the dynamics, and thus is more frequently applied when depicting fast action. We define $\mathbf{L} = \{\mathbf{x}^{\mathbf{L}}, r^{\mathbf{L}}\}$ as the possible subject locations and sizes according to the local composition in the panel, where $\mathbf{x}^{\mathbf{L}}$ is the subject's location relative to the panel center, and $r^{\mathbf{L}}$ is the subject's size relative to its panel. We represent subject size by $\sqrt{w * h}$, where w and h are the width and height of the subject's bounding box.

Subject Placement (Fig. 5(c)). Our model assumes that subject placement $\mathbf{x}^{\mathbf{S}}$ is governed by both local composition $\mathbf{x}^{\mathbf{L}}$ inside its panel, and global AGP $\mathbf{I}$ over the page. This assumption is motivated by the observation that professional artists first compose subjects using visual design rules (e.g., center principle or rule-of-thirds) with respect to the panel, and then adjust the composition slightly to accommodate their intention for directing viewer globally. Each subject is associated with a control point on the AGP $\mathbf{I}$, given by $t^{\mathbf{S}}$. The subject size $r^{\mathbf{S}}$ is directly related to its local size

Variable	Domain	Explanation
$\mathbf{f}$	$\mathbb{R}^{2\times l}$	underlying AGP.
$\mathbf{I}$	$\mathbb{R}^{2\times n}$	actual AGP.
$\mathbf{P} = \{g, t, m\}$	$g \in \{1,2,3\}$	geometric style of panel.
	$t \in \{1,2,3,4\}$	shot type of panel.
	$m \in \{1,2,3\}$	motion state of panel.
$\mathbf{x}^\mathbf{P}$	$\mathbb{R}^{2\times 1}$	center location of panel.
$\mathbf{L} = \{\mathbf{x}^\mathbf{L}, r^\mathbf{L}\}$	$\mathbf{x}^\mathbf{L} \in \mathbb{R}^{2\times 1}$	local position of subject.
	$r^\mathbf{L} \in \mathbb{R}$	local size of subject.
$\mathbf{S} = \{\mathbf{x}^\mathbf{S}, r^\mathbf{S}\}$	$\mathbf{x}^\mathbf{S} \in \mathbb{R}^{2\times 1}$	center location of subject.
	$r^\mathbf{S} \in \mathbb{R}$	size of subject.
$t^\mathbf{S}$	$\{1, \cdots, l\}$	location on AGP.
$\mathbf{B} = \{\mathbf{x}^\mathbf{B}, r^\mathbf{B}\}$	$\mathbf{x}^\mathbf{B} \in \mathbb{R}^{2\times 1}$	center location of balloon.
	$r^\mathbf{B} \in \mathbb{R}$	size of balloon.
O	$\{1, 2, ..., n\}$	reading order of balloon.
$t^\mathbf{B}$	$\{1, \cdots, l\}$	location on AGP.
$\mathbf{U} = \{U_{ij}\}$	$U_{ij} \in \{0, 1\}$	viewer attention transition.

Table 1: *Summary of the random variables used in our model.*

$r^\mathbf{L}$. We denote the subject's location and size as $\mathbf{S} = \{\mathbf{x}^\mathbf{S}, r^\mathbf{S}\}$.

Balloon Placement (Fig. 5(d)). Balloons are often positioned around their subject. Consequently, a balloon's location $\mathbf{x}^\mathbf{B}$ is determined by its own size $r^\mathbf{B}$ as well as its corresponding subject $\mathbf{S}$. The reading order O can also affect the balloon's position because placing a balloon around its subject may increase the probability of viewing it earlier than other counterparts. Finally, AGP will also guide balloon placement as it does for subjects. Hence, each balloon is associated with a control point on the AGP using index $t^\mathbf{B}$. We denote the balloon's position and size as $\mathbf{B} = \{\mathbf{x}^\mathbf{B}, r^\mathbf{B}\}$. It is worth noting that the latent AGP variable is connected to all elements on the page. The introduction of such a latent variable provides a way to capture complex interaction among the elements.

Viewer Attention Transitions (Fig. 5(e)). The final model component is a set of variables that represent the movement of viewer attention within a panel. For each panel, we define a set of binary variables $\mathbf{U} = \{U_{ij}\}$, where U_{ij} indicates that there is a viewer transition between elements i and j. To reflect the strong relationship between element composition and viewer attention, each U_{ij} is connected to all the elements ($\mathbf{S}$ and $\mathbf{B}$) inside the panel (see Fig. 6). This is because the movement of attention between any pair of elements is not only determined by element properties (e.g., location and scale), but also influenced by surrounding elements. In our context, we define the neighborhood of an element as all other elements that belong to the same panel.

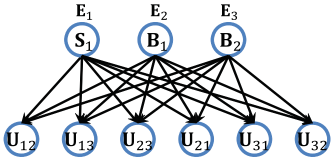


Figure 6: *A connectivity example between three element random variables* ($\mathbf{E}_1$, $\mathbf{E}_2$, $\mathbf{E}_3$) *and all random variables* $\{U_{ij}\}$ *of* $\mathbf{U}$.

Fig. 7 shows our complete probabilistic graphical model by putting the six model components together, and Table 1 summarizes the major variables.

5.2 Probability Distributions

Each random variable $\mathbf{X}_i$ in our model is associated with a conditional probability distribution (CPD), $p(\mathbf{X}_i|\mathbf{X}_{pa[i]})$, which represents the probability of observing $\mathbf{X}_i$ given its parents $\mathbf{X}_{pa[i]}$. We next describe the CPDs used for each variable in our model.

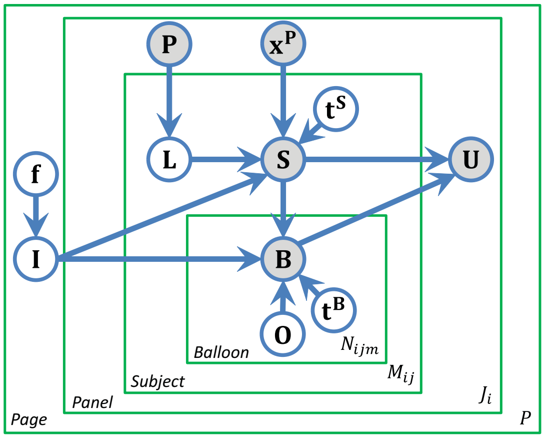


Figure 7: *The probabilistic graphical model for element composition. Each node is a random variable, with shaded-nodes corresponding to observed random variables and non-shaded nodes to latent (or hidden) random variable. Directed edges indicate conditional dependence between two random variables. A rectangular plate denotes that the enclosed random variables are duplicated the number times as indicated in the lower-right corner of the plate.*

Artist's Guiding Path ($\mathbf{f}$, $\mathbf{I}$). The two coordinate components of the curve are modeled as two independent Gaussian processes,

$$f^x(t) \sim \mathcal{GP}(m^x(t), k^x(t,t')),\ f^y(t) \sim \mathcal{GP}(m^y(t), k^y(t,t')),$$

where $k^x(t,t')$ and $k^y(t,t')$ are the squared exponential covariance functions, $k(t,t') = \alpha \exp[-\frac{1}{2}(\frac{t-t'}{\lambda})^2]$, where α and λ are hyperparameters. The mean functions $m^x(t)$ and $m^y(t)$ can be interpreted as the average AGP. The artist expects that the most basic behavior of the viewer is to sequentially visit all the panels. Therefore, we set the mean function by fitting a smooth parametric spline to the upper-right page corner, all the panel centers, and the lower-left page corner.

The actual AGP $\mathbf{I}$ is a noisy version of the underlying AGP $\mathbf{f}$, where

$$p(\mathbf{I}^x) = \mathcal{N}(\mathbf{I}^x; \mathbf{f}^x, \sigma_x^2 I),\ p(\mathbf{I}^y) = \mathcal{N}(\mathbf{I}^y; \mathbf{f}^y, \sigma_y^2 I), \qquad (1)$$

where I is the identity matrix, $\{\sigma_x^2, \sigma_y^2\}$ are the noise variances, and $\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes a multivariate Gaussian distribution of $\mathbf{x}$, with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$.

Panel Properties ($\mathbf{P}$). The shot type t, motion state m and geometric style g are all discrete random variables with categorical distributions,

$$p(t) = \mathrm{cat}(t; \lambda^t),\ p(m) = \mathrm{cat}(m; \lambda^m),\ p(g) = \mathrm{cat}(g; \lambda^G), \qquad (2)$$

where λ^t, λ^m and λ^G are their parameters.

Local Composition ($\mathbf{x}^L, r^\mathbf{L}$). The local composition is different for each shot type and panel geometric style. To describe the complexities of local foreground placement $\mathbf{x}^L$, we use a Gaussian mixture model (GMM),

$$p(\mathbf{x}^\mathbf{L}|g,t,m) = \textstyle\sum_{h=1}^{2} \mathcal{N}(\mathbf{x}^\mathbf{L}; \boldsymbol{\mu}^\mathbf{L}_{g,t,m,h}, \boldsymbol{\Sigma}^\mathbf{L}_{g,t,m,h}), \qquad (3)$$

with component means and covariances $\{\boldsymbol{\mu}^\mathbf{L}_{G,t,m,h}, \boldsymbol{\Sigma}^\mathbf{L}_{G,t,m,h}\}$. The local subject size $r^\mathbf{L}$ is Gaussian, $p(r^\mathbf{L}|t) = \mathcal{N}(r^\mathbf{L}; \boldsymbol{\mu}^\mathbf{L}_t, \boldsymbol{\Sigma}^\mathbf{L}_t)$.

Subjects and Balloons ($\mathbf{S}, \mathbf{B}$). The CPDs for the subjects and balloons are both conditional linear Gaussian distributions. Let $\mathbf{C}^\mathbf{S} = (\mathbf{x}^\mathbf{P}, \mathbf{x}^\mathbf{L}, \mathbf{I})$ be the continuous parent variables of $\mathbf{x}^\mathbf{S}$. For the subject $\mathbf{S}$, we have

$$p(\mathbf{x}^\mathbf{S}|\mathbf{C}^\mathbf{S}, t^\mathbf{S}) = \mathcal{N}(\mathbf{S}; \Phi_{t^\mathbf{S}} \cdot \mathbf{C}^\mathbf{S}, \boldsymbol{\Sigma}^\mathbf{S}_{t^\mathbf{S}}), \qquad (4)$$

where $\{\Phi_{t^\mathbf{S}}\}$ are regression parameters that determine the influence of various factors (i.e., local composition and AGP) in placing the subject. For subject size $r^{\mathbf{S}}$, we define $p(r^{\mathbf{S}}|r^{\mathbf{L}}) = \mathcal{N}(r^{\mathbf{S}}; \omega r^{\mathbf{L}}, \sigma^2)$, with ω and σ^2 being weight parameter and variance. Similarly, let $\mathbf{C}^{\mathbf{B}} = (\mathbf{x}^{\mathbf{S}}, \mathbf{I}, r^{\mathbf{B}}, r^{\mathbf{S}})$ be the continuous parent variables of $\mathbf{x}^{\mathbf{B}}$. For the balloon $\mathbf{B}$, we have

$$p(\mathbf{x}^{\mathbf{B}}|\mathbf{C}^{\mathbf{B}}, O, t^{\mathbf{B}}) = \mathcal{N}(\mathbf{B}; \Psi_{O,t^{\mathbf{B}}} \cdot \mathbf{C}^{\mathbf{B}}, \mathbf{\Sigma}^{\mathbf{B}}_{O,t^{\mathbf{B}}}), \qquad (5)$$

where $\mathbf{C}^{\mathbf{B}}$ are a similar set of parameters.

Viewer Attention Transitions ($\mathbf{U} = \{U_{ij}\}$). Let $\mathbf{O}_{ij}$ be a set of parent random variables of U_{ij}. We define the CPD of U_{ij} as

$$P(U_{ij} = 1|\mathbf{O}_{ij}) = \sigma(E(\mathbf{O}_{ij})), \qquad (6)$$

where $\sigma(\cdot)$ is a sigmoid function and $E(\mathbf{O}_{ij})$ is a potential function that measures how likely a viewer goes from elements i to j. We define $\mathbf{O}_{ij} = \{\mathbf{o}_i, \mathbf{o}_j, \mathbf{N}_{ij}\}$, where $\mathbf{o}_i$ and $\mathbf{o}_j$ are elements i and j, and $\mathbf{N}_{ij}$ is their neighbor set, i.e., all the other elements in the panel. The potential function is a linear combination of two terms,

$$E(\mathbf{o}_i, \mathbf{o}_j, \mathbf{N}_{ij}) = E_{pair}(\mathbf{o}_i, \mathbf{o}_j) + E_{context}(\mathbf{o}_i, \mathbf{N}_{ij}), \qquad (7)$$

where E_{pair} measures the likelihood of attention transition based on the relative information of i and j, while $E_{context}$ handles the influence of the surrounding context on the attention transition. E_{pair} is formulated as a weighted sum of four terms, which consider the identities, spatial distances, orientations and scales of the elements. Refer to the supplementary material for details on the terms.

6 Learning

The goal of the offline learning stage is to estimate the parameters $\boldsymbol{\theta}$ in the CPDs of all random variables in the probabilistic model, from the training set $\mathcal{D}$ obtained in Section 4. If all the random variables are observed in $\mathcal{D}$, we can estimate the parameters by maximizing the complete-data log likelihood $\ell(\boldsymbol{\theta}; \mathcal{D})$. Unfortunately, as some hidden random variables in our model are not observed, $\ell(\boldsymbol{\theta}; \mathcal{D})$ cannot be evaluated and thus maximized. Therefore, we estimate the parameters using the expectation-maximization (EM) algorithm [Bishop 2006].

If the distribution of a random variable is in the exponential family, computation of the two steps can be simplified [Yuille 2006]. In particular, in the E-step, the conditional expectation of the sufficient statistics of the parameters is computed. In the M-step, the expected sufficient statistics computed in the E-step are directly used in place of the sufficient statistics for the maximum likelihood solution of $\boldsymbol{\theta}$. This strategy can be employed for estimating parameters of CPDs of all the random variables in our model except $\mathbf{f}$. For $\mathbf{f}$, the Gaussian process is a non-parametric model without any sufficient statistics. We thus compute the parameter estimates for $\mathbf{f}$ using the following method: in the E-step, we approximate the expected likelihood of $\mathbf{f}$ using Monte Carlo integration; In the M-step, the hyperparameters are updated using a gradient-based optimization. Refer to the supplementary material for detailed derivations.

7 Interactive Composition Synthesis

The learned probabilistic model is used to synthesize a composition of input subjects and balloons, with respect to user-specified constraints. We have implemented an interactive tool for composition synthesis, which allows the user to intuitively and quickly specify a set of input elements and constraints, as demonstrated in the accompanying video. The inputs to our tool are the semantic properties of the panels, the input panel elements (i.e., subjects and balloons) and any user-specified constraints. We first generate a layout of panels that best suits input semantics and elements, and then compose the elements on the layout via MAP inference.

7.1 Layout Generation

When producing a manga page, artists normally begin by designing a panel layout, based on contents that they wish to present. As a result, given the number of panels N, input elements and semantics (i.e., shot type t and motion state m) for all the panels I, we need to determine a layout of panels whose configurations fit the input contents geometrically and semantically. We do this using a simple search algorithm to retrieve the best-fitting layout from our database of labeled pages. Our search algorithm first returns all the layouts with N panels as candidates $\{l\}$, and then ranks the candidate layouts using a compatibility score:

$$s(I, l) = \sum_{i}^{N} |t_i^I - t_i^l| + \sum_{i}^{N} |m_i^I - m_i^l| + \sum_{i}^{N} |n_i^I - n_i^l| \qquad (8)$$

where for i-th panel of the input and layout candidate, $t_i \in \{1, 2, 3, 4\}$ and $m_i \in \{1, 2, 3\}$, respectively, denote the shot type and motion state, and n_i denotes the number of elements. Finally, the top-ranked layout is selected as the best-fitting layout. Note that we do not use the layout method in [Cao et al. 2012]. This is because their method requires a sequence of images with existing compositions as input while our input elements are not composed in the panel at this stage.

7.2 Composition via MAP Inference

After generating the layout, our approach creates a composition of elements, which is compatible with the inputs and constraints, while also exhibiting properties exemplified by the training set. Note that we need to determine the positions of subjects and balloons, and the sizes of subjects, which is unknown in advance.

Formally, the inputs and constraints impose *evidence* on a subset of the random variables in the graphical model, denoted as $\mathbf{X}_E = \{\{t\}, \{m\}, \{G\}, \{r^{\mathbf{B}}\}, \{\mathbf{U}\}\}$. The most probable assignments of the unknown random variables for the panel elements, $\mathbf{X}_U = \{\{\mathbf{x}^{\mathbf{S}}\}, \{r^{\mathbf{S}}\}, \{\mathbf{x}^{\mathbf{B}}\}\}$, can then be inferred by maximizing their probability conditioned on the evidence, $p(\mathbf{X}_U|\mathbf{X}_E)$. Note that when the user fixes some elements, the corresponding $\mathbf{x}^{\mathbf{S}}$ or $\mathbf{x}^{\mathbf{B}}$ will become evidenced variables. However, not all samples from our graphical model are valid in terms of a set of constraints for a valid placement, including: 1) balloons should be placed in correct reading order; 2) elements should not significantly overlap [McCloud 1994; Chun et al. 2006]. These constraints are typically enforced to avoid ambiguities and obstructions in communication, although sometimes they are relaxed as a special effect. We do not incorporate these constraints in our model for two reasons: 1) imposing these constraints will induce direct connections between any two involved element variables, which makes the model impractical due to the exponential increase in the number of connections; 2) these constraints do not require a learning framework, as they are well-known rules that can be hard-coded directly.

Instead, we propose a MAP inference framework, where validity constraints and the graphical model are combined in a principled way. Let $\mathbf{Y}_C$ be validity constraints, and $\mathbf{X}_U$ and $\mathbf{X}_E$ be unknown and evidenced random variables in our model, respectively. The objective of MAP is to find a solution to $\mathbf{X}_U$ that maximizes the posterior probability,

$$\hat{\mathbf{X}}_U = \arg\max_{\mathbf{X}_U} \log p(\mathbf{X}_U|\mathbf{Y}_C, \mathbf{X}_E) \qquad (9)$$

$$= \arg\max_{\mathbf{X}_U} \underbrace{\log p(\mathbf{Y}_C|\mathbf{X}_U)}_{\text{constraint-based likelihood}} + \underbrace{\log p(\mathbf{X}_U|\mathbf{X}_E)}_{\text{conditional prior}}, \qquad (10)$$

where $\log p(\mathbf{Y}_C|\mathbf{X}_U)$ is a likelihood that measures how well the solution matches the validity constraints, and $p(\mathbf{X}_U|\mathbf{X}_E)$ is a conditional prior to determine how well the solution fits the learned probabilistic model. The likelihood is defined below. Since exact MAP inference is not tractable for our model, we perform approximate inference using a likelihood-weighted sampling method, which samples values of the unobserved random variables and weights each sample based on the likelihood of the observations [Murphy 1998].

Constraint-based Likelihood. The likelihood encodes four constraint terms,

$$\log p(\mathbf{Y}_C|\mathbf{X}_U) \propto \rho_1 C_{\text{overlap}} + \rho_2 C_{\text{order}} + \rho_3 C_{\text{bound}} + \rho_4 C_{\text{relation}},$$

where $\{\rho_i\}$ are weights controlling importance of different terms. Our implementation uses $\rho_1 = \rho_2 = 0.3, \rho_3 = \rho_4 = 0.2$. Individual terms are briefly described below, with the detailed formulation provided in the supplemental:

- The *overlap constraint term* (C_{overlap}) penalizes significant amounts of overlap between elements inside a panel.
- The *order constraint term* (C_{order}) penalizes configurations that violate the reading order of a sequence of balloons. Let b_i and b_j be two balloon in a panel, where b_j should be read *after* b_i. We define b_i and b_j as being in the correct order if b_i is located to the upper right side of b_j (for right-to-left reading).
- The *boundary constraint term* (C_{bound}) ensures that each element will be placed within the page and mostly within its panel. It takes a small value if any part of an element is outside of the page or more than 20% outside of the panel.
- The *subject relation constraint term* (C_{relative}) allows the user to specify how two subjects interact with each other. We support three types of common interactions: “talk”, “fight” and “none”. The constraint term is defined as $C_{\text{relation}} = C_{\text{size}} + C_{\text{interact}}$, where C_{size} encourages the interacting subjects to be of similar size, and C_{interact} measures how well the relative positions of subjects matches examples of the same interaction type in the training set.

Handling Background. Our tool allows user to insert a selected background image into a panel as the background scene. If there is any semantically important object in the background, the user can mark it out. We then treat the marked region as a foreground subject fixed in place, and run the same approach as above. This ensures that any important background object will not be largely occluded by foreground elements.

8 Evaluation and Results

In this section, we evaluate the effectiveness of our proposed model, by comparing compositions produced by our method with those produced by manga artists, an automatic heuristic method, and manual placement. We also show how our model can recover the underlying AGP from a manga page.

Implementation. To balance between the quality of results and efficiency, we choose 10 control points to approximate the underlying AGP. For learning, the parameters of $\mathbf{f}$ are estimated using the Gaussian Process for Machine Learning (GPML) toolbox [Rasmussen and Williams 2013], while parameter estimation for the remaining random variables is performed with the Bayesian Network Toolbox (BNT) [Murphy 2002]. For composition synthesis, each inference run generates 5,000 samples, which was found to perform well with reasonable running time. Our experiments were done on a PC with an Intel i7 3.1GHz CPU and 6GB RAM. Learning with 80 training examples takes $20h$. For composition synthesis, drawing 5,000 samples takes $30s$ on average. Evaluating the MAP values of all the instantiations takes about $15s$.

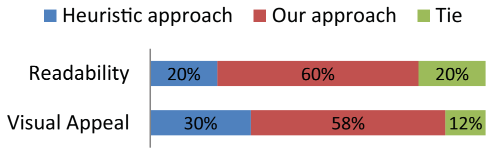


Figure 8: *Results from the visual perception study. Each number in the bar shows the percentage of votes from the participants. The participants show a strong preference for the compositions synthesized by our approach ($p < 0.05$, chi-squared test, 2 dof).*

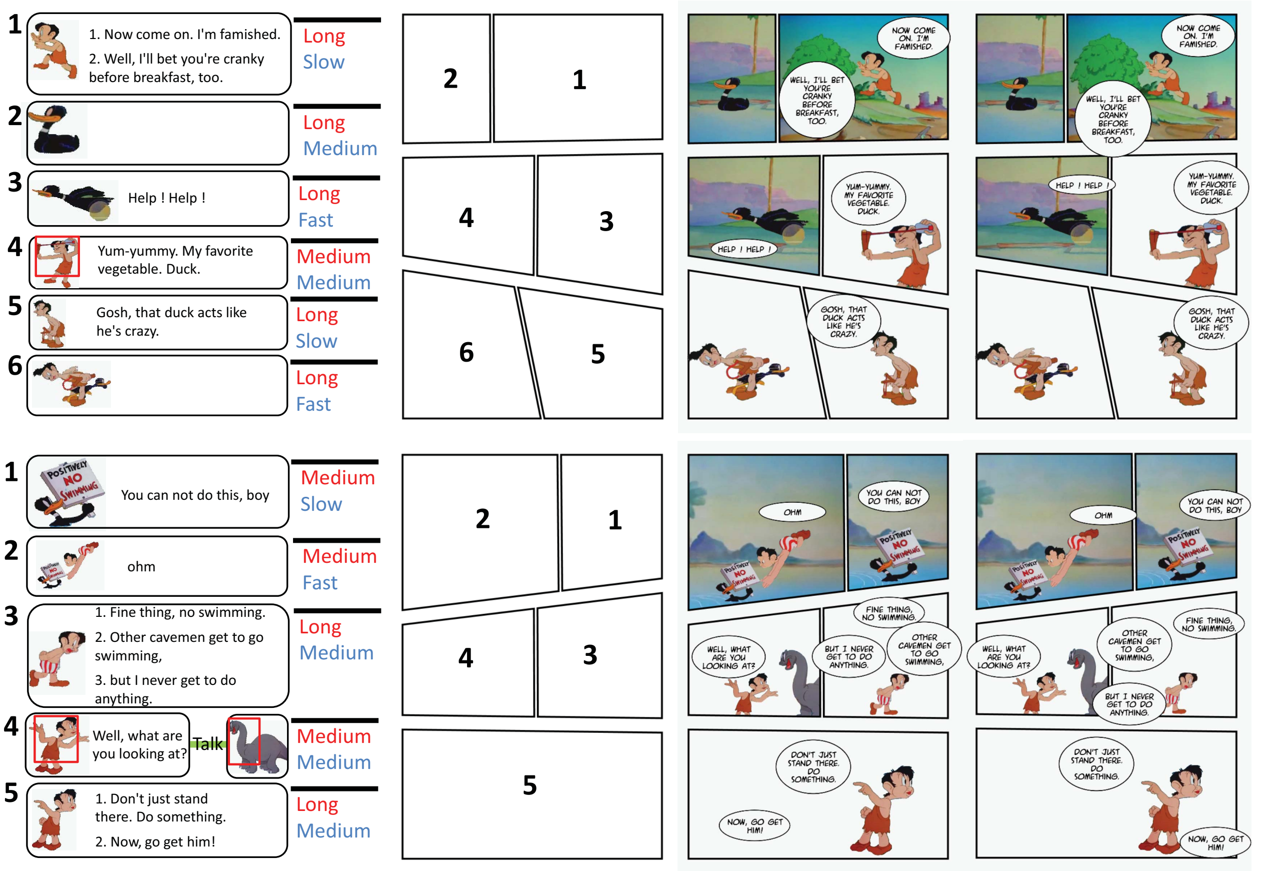
8.1 Comparison to Heuristic Method

We have compared our approach with a heuristic method for balloon placement [Chun et al. 2006], through a perceptual study and an eye-tracking experiment. The heuristic method places balloons around its subject according to a set of rules while keeping their reading order correct. Note that the heuristic method only handles balloon placement; it cannot address subject composition. As we are not aware of any previous approach for subject composition for the purpose of storytelling, for fair comparison, we run our approach first and use our subject composition as input to the heuristic method, which requires pre-defined subject positions.

Visual Perception Study. The goal of the visual perception study is to investigate if the participants have a strong preference for our results over those produced by the heuristic method. Ten participants were recruited and instructed to perform pair-wise comparisons in an online survey system. Half of the participants had manga reading experience, while the other half did not. Each participant was presented 11 pairs of compositions. Each pair included one composition by our approach and one by the heuristic method, and was shown side-by-side in randomized order. The participants were asked to choose the one that was better in terms of visual appeal and readability (“right better” or “left better”), or “tie” if they had no preference. In total, 110 comparisons were performed. The results are presented in Fig. 8, and suggest that the compositions produced by our approach are preferred by the participants.

Fig. 9 shows two examples used in the visual perception study. The heuristic method often uses the same patterns when placing balloons, which is due to the limited set of guidelines used. For example, the first balloon is always placed at the upper-right side of subject. In contrast, our approach is able to achieve a higher level of diversity, which demonstrates its flexibility. Our probabilistic model can capture the subtle dependencies between panel properties and balloon placement, as well as how balloons interact both locally and globally. This enables our approach to reproduce more naturally looking compositions that well adapt to various input configurations. In addition, the heuristic method places balloons on a per-panel basis, according to the local information inside each panel, which may result in a misleading composition. In the top example of Fig. 9(d), the left balloon (“Well, I’ll bet...”) in the first (top-right) panel and the balloon (“Help ! help !”) in the fourth panel are positioned very close to each other. With such a configuration, viewers are very likely to skip the second panel, and move directly to the fourth panel, which is undesirable. Our approach can eliminate this issue by using a more appropriate configuration to steer viewer attention along the desired route, as shown in Fig. 9(c).

Eye-tracking experiment and analysis. We next analyze eye movement data on the compositions to evaluate the effectiveness of our method in directing viewer attention, as compared with the heuristic method. If the viewers are successfully directed by a composition, we should observe higher consistency in eye movements across different viewers. We recorded the eye movements of 10 participants on the 11 pairs of compositions used in the visual perception study, with 5 on our results and another 5 on the results


1
1. Now come on. I'm famished.
2. Well, I'll bet you're cranky before breakfast, too.
Long
Slow
2
Long
Medium
3
Help ! Help !
Long
Fast
4
Yum-yummy. My favorite vegetable. Duck.
Medium
Medium
5
Gosh, that duck acts like he's crazy.
Long
Slow
6
Long
Fast
2
1
4
3
6
5
NOW COME ON. I'M FAMISHED.
WELL, I'LL BET YOU'RE CRANKY BEFORE BREAKFAST, TOO.
YUM-YUMMY. MY FAVORITE VEGETABLE. DUCK.
HELP ! HELP !
GOSH, THAT DUCK ACTS LIKE HE'S CRAZY.
1
POSITIVELY NO SWIMMING
You can not do this, boy
Medium
Slow
2
ohm
Medium
Fast
3
1. Fine thing, no swimming.
2. Other cavemen get to go swimming,
3. but I never get to do anything.
Long
Medium
4
Well, what are you looking at?
Talk
Medium
Medium
5
1. Don't just stand there. Do something.
2. Now, go get him!
Long
Medium
2
1
4
3
5
OHM
YOU CAN NOT DO THIS, BOY
FINE THING, NO SWIMMING.
WELL, WHAT ARE YOU LOOKING AT?
BUT I NEVER GET TO DO ANYTHING.
OTHER CAVEMEN GET TO GO SWIMMING,
DON'T JUST STAND THERE. DO SOMETHING.
NOW, GO GET HIM!

(a) Inputs (b) Retrieved layouts (c) Ours (d) Heuristic [Chun et al. 2006]

Figure 9: *Two example compositions by our approach and by the heuristic method. (a) Input subjects accompanied by their scripts ("Daffy Duck - Daffy Duck and the Dinosaur" in the public domain). (b) Automatically generated layout meshes. (c) Compositions by our approach. (d) Compositions by the heuristic method with locations of subjects determined by our approach.*

by the heuristic method. Following Jain et al. [2012], we measure the consistency in both unordered and ordered eye fixations across different viewers.

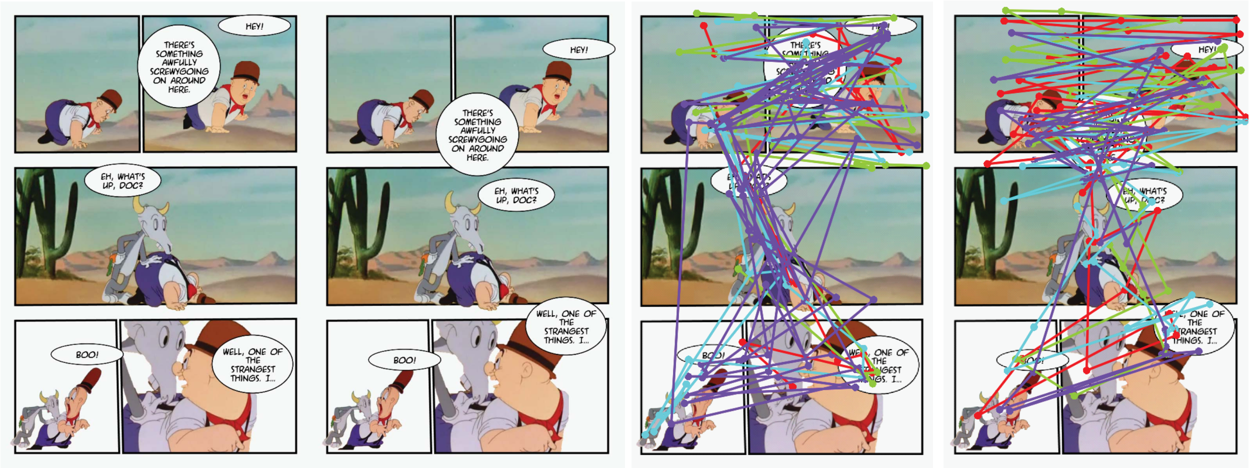
The similarity between unordered eye fixations of different viewers is based on the saliency map constructed from each viewer's eye fixation data. The thresholded saliency map from one viewer is used as a classifier, and the inlier percent [Judd et al. 2009], which measures the percentage of other viewers' fixations lying inside the saliency map, is computed. A higher inlier percent indicates that all the viewers fixate at similar locations. By using each viewer as the classifier and sweeping the threshold, we plot the average ROC (receiver operating characteristic) curves for the compositions by our approach and by the heuristic method in Fig. 11(a).

The inline percent only measures the spatial similarity of eye fixations, regardless of their order. To evaluate both the spatial and temporal similarity in the eye movements of two viewers, we sequentially concatenate the fixation locations of each viewer into a vector, and compute the root mean squared distance (RMSD) between the two vectors. Since the number of fixations may be different for the two viewers, we use dynamic time warping [Sakoe and Chiba 1978] to warp all the vectors to the same length. Fig. 11(b) shows the mean RMSD for our approach and the heuristic method. The results in Fig. 11 indicate that our approach produces compositions with higher consistency in both unordered and ordered eye fixations, suggesting that our approach is more effective in directing viewer attention properly, as compared with the heuristic method. Fig. 10 shows example compositions with eye-tracking data. Refer to supplemental materials for more results.

8.2 Comparison to Manual Method

We next evaluate how well our approach facilitates manga compositions, as compared to existing comic maker programs that use manual placement. We conducted another user study with 10 participants who have no prior experience in manga production. Each participant was asked to perform 11 compositions tasks using the same set of inputs as in Section 8.1. For each task, starting with a storyboard as shown in Fig. 1 and an unorganized initial configuration of input elements (i.e., a given layout with randomly placed elements), the participants were asked to arrange the elements to produce a manga page according to their understanding of the storyboard. Each composition task was performed using either our tool or a manual tool as provided in commercial programs such as MangaStudio. Using the manual tool, the participant has to arrange all the elements manually. The interface was otherwise identical. All participants were given a short tutorial of the tool interface and a five-minute warm-up exercise before they start the tasks. In total, 110 compositions were produced, 55 by our tool and 55 by the manual one. As shown in Fig. 12, the participants spend an average of 140 seconds (std = 47 seconds) on producing a single composition using the manual tool, whereas each composition session with our tool only takes 45 seconds (std = 5 seconds) on average. This shows that our tool is almost three times faster than the manual one in assisting composition generation.

The compositions produced using our tool and the manual tool were evaluated by another 10 participants (evaluators), using the same method of pair-wise comparisons as in Section 8.1. All evalua-


HEY!
THERE'S SOMETHING AWFULLY SCREWYGOING ON AROUND HERE.
EH, WHAT'S UP, DOC?
BOO!
WELL, ONE OF THE STRANGEST THINGS. I...

(a) Ours (b) Heuristic [Chun et al. 2006] (c) Ours (d) Heuristic [Chun et al. 2006]

Figure 10: *Example compositions used in the eye-tracking experiment. (a) shows a composition produced by our approach, and (b) shows the one from the heuristic method with the positions of subjects determined by our approach. (c) and (d) show the captured eye movements of multiple viewers on (a) and (b), respectively. Note that the consistency of eye movements in (c) is higher than that in (d). Input elements and scripts are from "The Wacky Wabbit" in the public domain.*

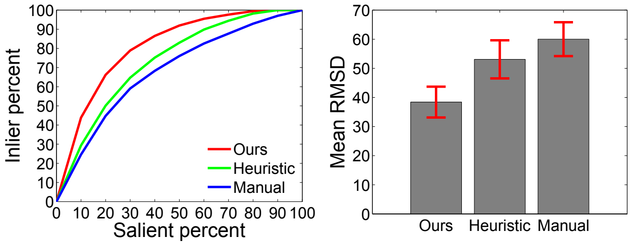


Figure 11: *Analysis of fixation consistency. (a) Mean ROC curves for our approach, the heuristic method and manual composition. Salient percent is the percentage of salient pixels after applying a threshold on the saliency map. The curve closer to the upper-left corner means higher consistency. (b) Mean RMSD values of the compositions by the three methods. A lower RMSD value implies a higher inter-viewer consistency in eye movements. Both results show that the compositions by our approach have higher consistency in eye movements across multiple viewers. From a paired t-test, the difference between our method and the two other methods is statistically significant ($p < 0.05$).*

tors were manga readers. A total of 275 pair-wise comparisons were performed, which were evenly distributed among the evaluators. Evaluation results are visualized in Fig. 12, which suggest that our results are significantly preferred by the evaluators. We have also selected a "best" group of the 11 compositions by the manual tool, each receiving the most votes from the evaluators. We then recorded eye movements of multiple viewers on these compositions, and analyzed the consistency of eye fixations, as in Section 8.1. The results are shown in Fig. 11 (blue curve). The results again suggest that our approach is better at guiding viewer attention than the manual method.

8.3 Comparison to Existing Manga Pages

We also evaluate how well our graphical model reproduces stylistic properties of the training examples, by comparing our results with those by the professional artists. In particular, we manually annotate elements and necessary semantics on several professionally-drawn manga pages, which are not in our training set. Taking such annotations as input, our approach automatically re-produces the compositions, which are then compared against the original ones. Fig. 13 shows two example comparisons. As can be seen, the compositions re-produced by our approach are functionally and stylistically close to the ones by the professional artists.

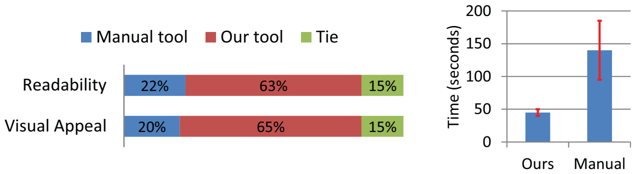


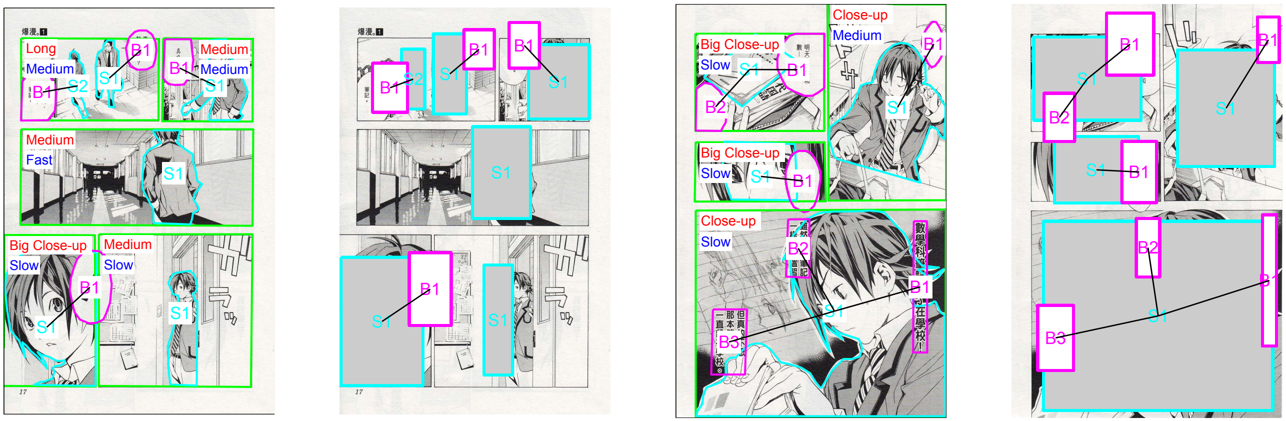
Figure 12: *Results from pairwise comparisons of compositions by our tool and the manual tool. Left:each number in the bar shows the percentage of votes from the evaluators. The evaluators preference for the compositions by our approach is statistically significant ($p < 0.05$, chi-squared test, 2 dof). Right: average time of generating one composition by our tool and manual tool.*

8.4 Recovering Artist's Guiding Path

Explicitly modeling the artist's guiding path in our graphical model allows for its automatic recovery from a labeled manga page. Given an existing manga page with panels, subjects and balloons segmented as in Section 4, the AGP can be recovered by performing inference on the probabilistic model for $\mathbf{I}$. Two examples are shown in Fig. 14. The recovered AGP can be used as a guiding cue for animating static comics. Recent work [Jain 2012] attempts to create a *Ken Burns* effect from comic panels using tracked eye movements. Our recovered AGP can serve as a reliable reference, in place of pre-captured eye-tracking data, for path planning of the viewport in such animations. This removes the requirement for real-time eye tracking, which is costly to perform.

8.5 Limitations

Our work has two limitations. First, our work assumes that the variations in spatial location and scale of elements are the only factors driving viewer attention. In practice, manga artists also manipulate visual patterns of a group of elements to steer viewer attention. For example, while a sharp contrast in intensity between two objects often induces a visual break, a gradual transition between the objects might be indicative of temporal continuity. Therefore, an interesting direction is to investigate how the appearance of foreground characters, i.e., intensity, color or local contrast, may change the way a composition is made, and integrate it into our



(a) original (b) synthesized (c) original (d) synthesized

Figure 13: *Comparison to artworks by professional artists. Using labeled elements and semantics on two existing manga pages ((a) and (c)) from "Bakuman" (© Tsugumi Ohba, Takeshi Obata / Shueisha Inc.), our approach is able to reproduce the compositions ((b) and (d)) that closely resemble the original artworks. For visual clarity, subjects and balloons in (b) and (d) are represented by gray and white bounding boxes, respectively. Balloons are connected to its subject via black solid lines.*

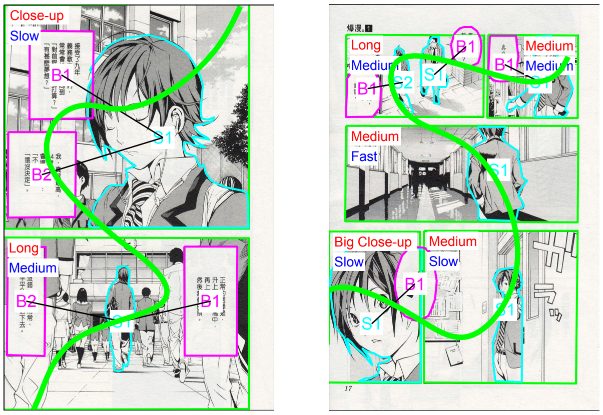


Figure 14: *Recovering artist's guiding path (AGP). The green solid curves represent the AGP estimated by our model, which was learned from annotated manga pages of "Bakuman" (© Tsugumi Ohba, Takeshi Obata / Shueisha Inc.). The paths in these two examples reflect the artists' general motivation to continuously direct viewers toward important subjects and balloons through the pages.*

graphical model. Second, for the panel with more than four subjects, our approach can fail to produce satisfying results automatically, as shown in Fig. 15, mainly because our training dataset lacks examples of such tightly packed panels. By analyzing our dataset, we note that there are less than 5% of panels with more than four subjects. A natural explanation is that placing too many subjects in a single panel might reduce visual clarity and, therefore, manga artists rarely do this. In these cases, manually freezing some of the subjects in place could alleviate the problem.

9 Discussion

In this paper, we have proposed a probabilistic graphical model to represent the dependency among the artist's guiding path, panel elements, and viewer attention. The model enables interactive joint placement of subjects and balloons, producing effective storytelling compositions with respect to the user's given constraints. The results from a visual perception study and an eye-tracking experiment show that compositions from our approach are more visually appealing and provide a smoother reading experience, as compared to those by a heuristic method. We have also demonstrated that, in comparison to manual placement, our approach allows novices to create higher-quality compositions in less time. Our workflow does not necessarily match the workflow of current professional manga artists, where some artists directly draw subjects and balloons on canvas. However, our approach can still be used by professional artists for quick composition, once subjects and balloons are drawn on separate layers and input to our approach as pre-made objects.

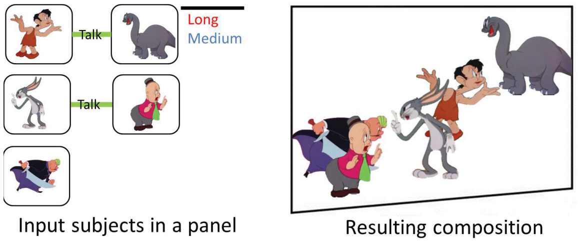


Figure 15: *When there are too many subjects within a single panel, our approach may produce an unreasonable composition with subjects that are cluttered together, leading to semantic ambiguity.*

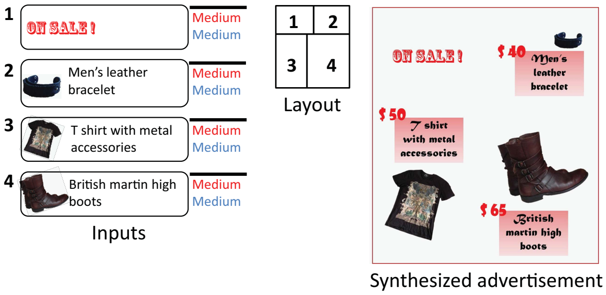


Figure 16: *Advertisement example synthesized by our approach.*

Extension to other graphic design. Our approach can also be extended to other graphic design domains. Similar to comics, various kinds of graphic design, e.g., print advertisement and magazine, also combine pictures and text to enhance the delivery of information to the audience. Successful design of such mediums should efficiently direct the audience's attention through important contents, so that the audience can quickly capture their ideas and effects. While a rudimentary result of directly applying our approach to advertisement design is shown in Fig. 16, facilitating more sophisticated designs in this domain is an area of future research.

Understanding readers' behaviors. Great research effort has been made to investigate how humans perceive visual information in various domains, such as print advertisement and film production. It is also important for manga artists to understand how readers move their attention over the composed artworks, and respond to visual effects designed by the artists. Unfortunately, works on systematically understanding reading behaviors of manga readers are limited. In this work, we have taken a step toward this objective, by understanding how the composition of manga elements interacts with reader attention using computational machinery. We hope this work can motivate further investigation regarding this direction, which would be of great practical value to the manga industry. For example, it would be interesting to develop an interactive system where manga artists are able to intuitively explore the joint space of composition and viewer attention. As artworks are being composed, the system would predict the reader's path of attention, which can be employed by manga artists to progressively improve their artworks.

Acknowledgements

We thank Janet Hsiao, Tianyin Liu, Tina Liu, and Cherry Cheng for their help with eye tracking data collection. This work was supported by the Research Grants Council of Hong Kong (CityU 123212 and CityU 115112).

References

Barnes, C., Goldman, D., Shechtman, E., and Finkelstein, A. 2010. Video tapestries with continuous temporal zoom. In *ACM SIGGRAPH'10*.

Bishop, C. 2006. *Pattern Recognition and Machine Learning*. Springer.

Cao, Y., Chan, A., and Lau, R. 2012. Automatic stylistic manga layout. In *ACM SIGGRAPH Asia'12*.

Christensen, J., Marks, J., and Shieber, S. 1994. *Placing text labels on maps and diagrams*. Academic Press.

Christensen, J., Marks, J., and Shieber, S. 1995. An empirical study of algorithms for point-feature label placement. *ACM Trans. on Graphics 14*, 203–232.

Chun, B., Ryu, D., Hwang, W., and Cho, H. 2006. An automated procedure for word balloon placement in cinema comics. *LNCS 4292*, 576–585.

ComiPo, 2012. http://www.comipo.com/en.

Correa, C., and Ma, K.-L. 2010. Dynamic video narratives. In *ACM SIGGRAPH'10*.

DeCarlo, D., and Santella, A. 2002. Stylization and abstraction of photograph. In *ACM SIGGRAPH'02*.

Edmondson, S., Christensen, J., Marks, J., and Shieber, S. 1996. A general cartographic labelling algorithm. *Computers & Geosciences 33*, 13–24.

Folse, S., 2010. http://hoodedutilitarian.com/visual-languages-of-manga-and-comics/.

Friedman, N., and Nachman, D. 2000. Gaussian process networks. In *Conf. on Uncertainty in Artificial Intelligence'00*.

Huang, H., Zhang, L., and Zhang, H. 2011. Arcimboldo-like collage using internet images. In *ACM SIGGRAPH Asia'11*.

Jain, E., Sheikh, Y., and Hodgins, J. 2012. Inferring artistic intention in comic art through viewer gaze. In *ACM SAP'12*.

Jain, E. 2012. *Attention-guided Algorithms to Retarget and Augment Animations, Stills and Videos*. PhD thesis, Carnegie Mellon University.

Judd, T., Ehinger, K., Durand, F., and Torralba, A. 2009. Learning to predict where humans look. In *ICCV'09*.

Kesting, M., 2004. Basic thoughts about visual composition. http://www.hippasus.com/resources/viscomp/index.html.

Kim, J., and Pellacini, F. 2002. Jigsaw image mosaics. In *ACM SIGGRAPH'02*.

Kurlander, D., Skelly, T., and Salesin, D. 1996. Comic chat. In *ACM SIGGRAPH'96*.

MangaStudio, 2011. http://manga.smithmicro.com/.

McCloud, S. 1994. *Understanding Comics: The Invisible Art*. William Morrow.

McCloud, S. 2006. *Making Comics: Storytelling Secrets of Comics, Manga and Graphic Novels*. William Morrow.

Murphy, K., 1998. A brief introduction to graphical models and bayesian networks. http://www.cs.ubc.ca/~murphyk/Bayes/bayes.html#reading.

Murphy, K., 2002. Bayes net toolbox for matlab. https://code.google.com/p/bnt/.

Omori, T., Igaki, T., Ishii, T., Kurata, K., and Masuda, N. 2004. Eye catchers in comics: Controlling eye movements in reading pictorial and textual media. In *Int'l Congress of Psychology'04*.

Qu, Y., Pang, W., Wong, T., and Heng, P. 2006. Richness-preserving manga screening. In *ACM SIGGRAPH'06*.

Qu, Y., Wong, T., and Heng, P. 2008. Manga colorization. In *ACM SIGGRAPH Asia'08*.

Ramanathan, S., Yanulevskaya, V., and Sebe, N. 2011. Can computers learn from humans to see better?: inferring scene semantics from viewers' eye movements. In *ACM MM'11*.

Rasmussen, C., and Williams, C., 2013. Gaussian processes for machine learning matlab code. http://www.gaussianprocess.org/gpml/code/matlab/doc/.

Rother, C., Bordeaux, L., Hamadi, Y., and Blake, A. 2006. Autocollage. In *ACM SIGGRAPH'06*.

Sakoe, H., and Chiba, S. 1978. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Trans. ASSP 26*, 43–49.

SamuraiMangaWorkshop, 2011. Manga artist interview - shin shinmoto - 1/2. https://www.youtube.com/watch?v=fxwcHnYcgLo.

Scott-Baron, H. 2006. *Manga Clip Art: Everything You Need to Create Your Own Professional-Looking Manga Artwork*. Andrews McMeel Publishing.

Toyoura, M., Sawada, T., Kunihiro, M., and Mao, X. 2012. Using eye-tracking data for automatic film comic creation. In *Symp. on Eye Tracking Research and Applications'12*.

Yuille, A., 2006. The EM algorithm. http://www.stat.ucla.edu/~yuille/courses/stat153/emtutorial.pdf.