StorVi (Story Visualization): A Text-to-Image Conversion

Kim D. Alcantara, Jomar P. Calandria, Junika S. Calupas, Jean Paula R. Echas, and Ria A. Sagum

***Abstract*—Natural language is an easy and effective medium for describing visual ideas and mental images. thus, we forecast the appearance of language-based 2D scene generation systems to let ordinary users fast create 2D scenes without having to learn special software, obtain imaginative skills, or even touch a desktop window-oriented interface. This research presented by the researchers entitled “StorVi (story visualization): a text-to-image conversion”, is a system that can visualize stories of multiple framing in pictures. the system focus on fable stories for children ages 4-7 yrs. old. Recognizing the characters and partitioning of frames are the general problems of the study. In solving the two general problems, the researchers used classification algorithm/simple co-reference resolution algorithm an algorithm for recognizing the characters and used the rule to partition the frames by sentence with character/s for the partitioning of frames.**

***Index Terms*—2D Scene, database, StorVi, text-to-image, wordseye.**

I. Introduction

Story telling in children, particularly the use of visualizing or picturing stories, has become an essential part of telling the stories lives. It has widely developed in the school that is why the researchers developed a tool to visualize stories from text. [1]

Visualization is the process of representing abstract business or scientific data as images that can aid in understanding the *meaning* of the data. It has an indisputable capacity to represent and to communicate knowledge. As it has been frequently noted, it is often easier to explain physical phenomena, mathematic theorems, story, or structures of any kind using a drawing than words. Images can help understand ideas or situations and realize their complexity. They are

An effective means to describe and explain things especially for children. [2] Text- to-image is text associated with an image that serves the same purpose and conveys the same essential information as the image. The conversion consists in synthesizing a description from a text and in displaying it. [3] Ideally, a text-to-image converter would recreate mental images we form when we read a text. This represents a demanding task involving semantic and cognitive capabilities and too many people seems both a far off and surreal fantasy. [4]

The study is very helpful for people to easily picture a story. The study intends to develop software that can visualize a short fable story based on the application of WordsEye.

II. Background

In Development of visualize scene and elements involved in composing a virtual story scene, the construction of the environment or set, scene composition and the effect of genre styles are addressed in complete text-to-visual systems and scene directing systems. Scene visualization requires consideration of the interaction of actors and objects. SONAS constructs a three-dimensional virtual town according to the verbal descriptions of a human user. WordsEye depicts non-animated scenes with characters, objects, actions and environments. A database of graphical objects holds models, their attributes, poses and spatial relations. In CONFUCIUS, multimodal animations of single sentences are produced. [5]

To have an effective automatic and intelligent production of text story to visualize it are compose of two development stages; Detecting actors and action in the stories, Combining and positioning of actor to action.

In Detecting actors and action in the stories all actors express action states namely word choice, gestures, and body posture. In order to recognize scene in text and to create life-like scene images WordsEye is a system for converting from English text into three-dimensional graphical scenes that represent that text. WordsEye works by performing syntactic and semantic analysis on the input text, producing a description of the arrangement of objects in a scene. An image is then generated from this scene description. At the core of WordsEye is the notion of a “pose”, which can be loosely defined as a figure in a configuration suggestive of a particular action. [6]

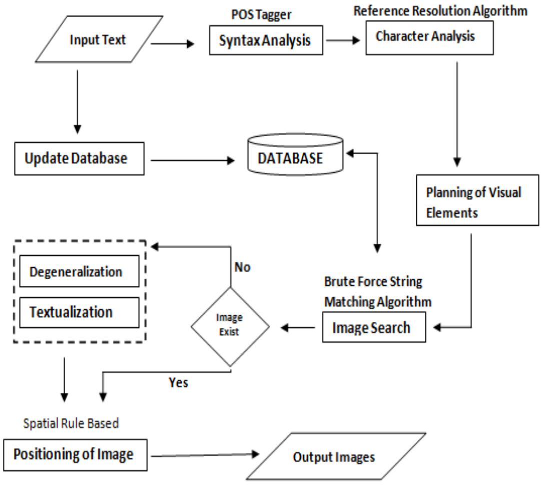

Fig. 1. System architecture.

Manuscript received January 15, 2014; revised June 24, 2014. This work was supported in part by the Polytechnic University of the Philippines. StorVi(Story Visualization) : A Text-to-Image Conversion.

R. A. Sagum is with the College of Computer and Information Sciences, Polytechnic University of the Philippines (e-mail: riasagum31@yahoo.com).

K. D. Alcantara, J. S. Calupas, J. P. Calandria, J. R. Echas are with the undergraduates of the Polytechnic University of the Philippines (email: kdalcantara@yahoo.com, nikacalupas@yahoo.com, jomarcalandria@yahoo.com, heartjean2002@yahoo.com).

DOI: 10.7763/IJFCC.2014.V3.328

In Combining and positioning of actor to action research aiming to automatically model and animate scene with natural expressions character transformation. The exact manner of an active action depends on combining and positioning of the actors to action or vice versa. Spatial Relations in Text-to-Scene Conversion uses spatial tags and other spatial and functional properties on objects to resolve the meaning of spatial relations. We focus here on the interpretation of NPs containing spatial prepositions of the form \X-preposition-Y", where we will refer to X as the figure and Y as the ground. For example, in snow is on the roof, snow is the figure and roof is ground. The interpretation of the spatial relation often depends upon the types of the arguments to the preposition. There can be more than one interpretation of a spatial relation for a given preposition [7].

III. System Architecture

The key component is the Story Visualization module including syntax analysis, character analysis, planning of visual elements, image search, degeneralization, textualization and positioning of image.

A. Part-of-Speech Tagging (POST)

The researchers used part-of-speech tagging (POST) to process the marking up of a word in a text (corpus) as corresponding to a particular part of speech, based on both its definition, as well as its context, relationship with adjacent and related words in a phrase, sentence, or paragraph. A simplified form of this is commonly taught to school-age children, in the identification of words as nouns, verbs, adjectives, adverbs, etc. The researchers basically use the POST to identify only the part of speech because the result will be used later on by the other algorithm which is classification algorithm.

Example:

The boy is running. => The is article, boy is noun, is is verb be, and running is verb.

B. Classification Algorithm

The researchers used Classification Algorithm to identify the actors, actions, object, position and environment in each sentence. This algorithm is use because it is important to identify the characters and settings so that the system can generate a 2D representation orderly.

Example: *"A dog and a chair are in the forest."* In this example, the *dog* is the *actor,* the *chair* is the *object* and the *forest* is the *environment*.

The generated frame will look like this:

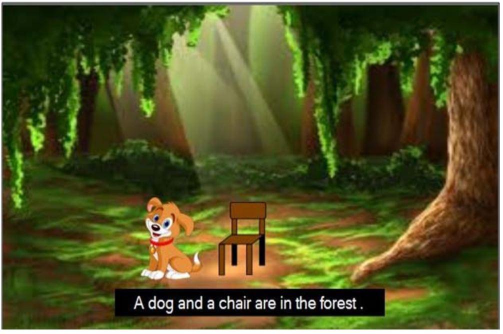

Fig. 2. Classification algorithm.

C. Simple Co-Reference Algorithm

The researchers use simple co-reference algorithm to solve co-reference or to derive the correct interpretation of text, or even to estimate the relative importance of various mentioned subjects, pronouns and other referring expressions that must be connected to the right individuals. In the system, **co-reference** occurs when multiple expressions in a sentence or document refer to the same thing; or in linguistic jargon, they have the same "referent." The algorithm intended to resolve co-references commonly look first for the nearest preceding individual that is compatible with the referring expression. For example, in the sentence "Mary said she would help me," *she* and *Mary* most likely refer to the same person or group, in which case they are *co referent*. Similarly, in "I saw Scott yesterday. He was fishing by the lake," *Scott* and *he* are most likely co referent.

Example: "*The dog is in the forest. He decided to go to the beach."* In this example, the pronoun *He* is a co-reference with the actor which is the dog. Therefore, an image of a dog will also be used in the second frame.

The generated frames will look like this:

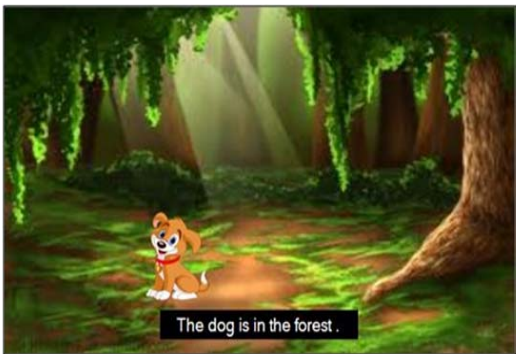

Fig. 3. (a). Simple co-reference algorithm.

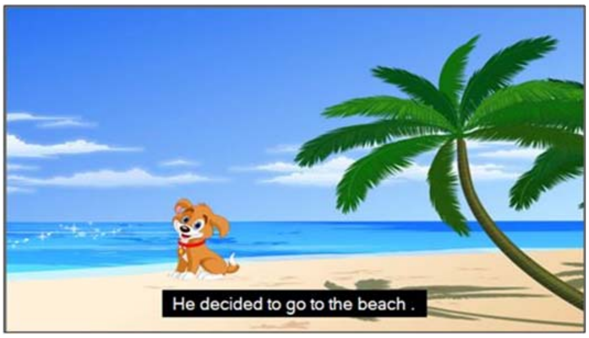

Fig. 3. (b) Simple co-reference algorithm.

D. Brute Force String Matching Algorithm

The researcher use Brute Force String Matching Algorithm. It is an algorithm that checks every single character from the text to match against the pattern. It is use in searching the equivalent images of the collected texts in the database. This algorithm is used by identifying the corresponding images of the possible characters or environment of the story in the database.

Example:

The cat is walking in the forest.

The system will undergo the process of Image Searching that uses Brute Force String Matching Algorithm to search the equivalent images of the actor and the environment in the database in this sentence. Cat is the actor and forest is the environment.

E. Degeneralization

The researchers used degeneralization. It is the act of specifying something that is general. . In this process, if the collected text is a general categorical terms, the system will search for the specific object instance of the same class.

Example: *"A cat is cleaning the furniture"* In this example, the collected word *furniture* is a general category. Since it doesn't have an equivalent image in the database, the image of a chair will be used because it is an instance of the collected word *furniture.*

The generated frame will look like this:

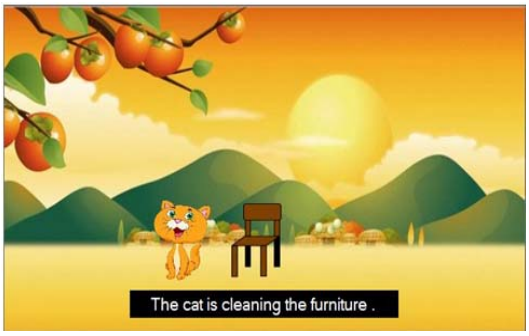

Fig. 4. Degeneralization.

F. Textualization

The researchers used textualization. It is the act or process of textualizing; rendering as text; the result of textualizing; a written version. It is used when the input word doesn't have an equal image on the database. The system will generate 2D extruded text of that particular word.

Example: "*A wall is in the forest.*" In this example, the system will generate a 2D text of the collected word *wall* since there is no other way to depict that entity.

The generated frame will look like this:

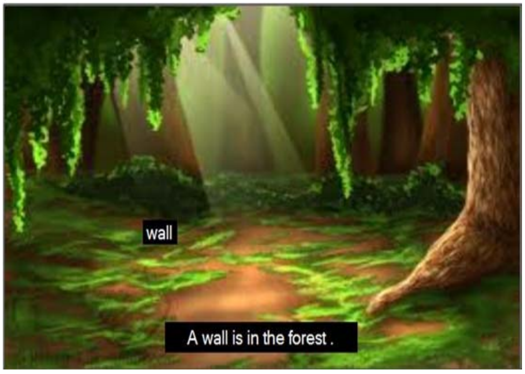

Fig. 5. Textualization.

G. Spatial Rule Based

The researcher use Spatial Rule Based. Spatial Relation specifies how some object is located in space in relation to some reference object. In this technique, the researcher come up with the rule that can identify the positions of the images and that is called Spatial Rule Based. It is use to know what are the default positions of the actors and objects in each sentence of the story. This rule by identifying the prepositions stated in the story like under, on, beyond, etc.

Example: "*The mouse is under the chair."* In this example, the actor which is the *mouse* has a spatial tag *under.* The system will now search for its target; in this case the target is the object *chair*.

The generated frame will look like this:

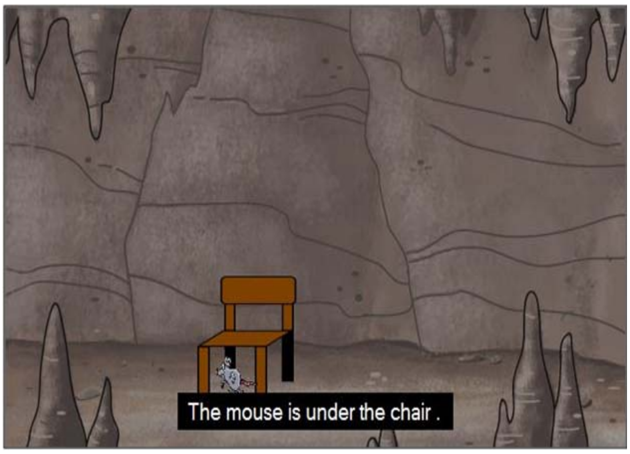

Fig. 6. Spatial rule based.

IV. Results

The study specifically seeks to solve the degree of acceptance in terms of the usability of the system, user-friendliness and content of the generated frames. The respondents of the study are teachers who teach preschool to 2nd grade pupils and parents who has a son or daughter that ages 4 to 7 years of age. The table below shows the results from the gathered data.

TABLE I: Overall Mean Results of the System

Degree of Acceptance in terms of:	Mean	Verbal Interpretation
Usability of the System	4.60	Agree
User-Friendliness	4.49	Agree
Content of the Generated Frames	4.33	Agree

Table I shows that the overall mean of usability of the system is 4.60 and the verbal interpretation of that is agree, while for the User-Friendliness, the mean is 4.49 and also the verbal interpretation of that is agree. Then lastly for the Content of the Generated Frames, the mean is 4.33 and it is also interprets that the respondents agreed.

V. Conclusion

The system StorVi(Story Visualization): A Text-to-Image Conversion was evaluated by the Teachers and Parents. Based on the Liker's Scale, both of them agreed on the Usability and User-Friendliness of the system, but on the other hand, Teachers agreed also on the Content of the generated frames but the Parents gave the researchers fair on the Content of the generated frames.

According to the respondents it is a helpful tool for their personal and daily use. No one among the respondents answered that the system is not efficient for them.

For the Usability of the System, the researchers concluded that the system need some improvements like animated images and sounds to capture more attention from the children. While for the User-Friendliness the researchers concluded that the objects inside the GUI are improperly organized to them. Then lastly for the Content of the

Generated Frames, the researchers concluded that the generated pictures are incomplete to them.

The researcher concluded it may be an affecting factor that the researcher presented the stories to the children and see the reaction of the children before the teachers answered the questionnaire. That's why the teachers gave a higher score rather than the parents because the teacher saw the reaction of the children but parents did not. The researcher also concluded that the system have a missing elements in presenting the output like sounds, moving images and in recognizing emotions of the characters in the story.

Recommendation

The researchers would like to recommend the following for further improvements on the system and to research. For the future study of the future researchers, the researchers would like to enhance the system to capture the children's attention by the following elements:

1) Adding sounds or sound effects.
2) Moving objects or images.
3) Recognizing emotion of the characters

References

[1] Storytelling. [Online]. Available: http://www.en.wikipedia.org/wiki/Storytelling,/

[2] Hefreedictionary. [Online]. Available: http://www.hefreedictionary.com/visualize, copyright

[3] J. Allbeck, N. Badler, R. Bindiganavale, A. Joshi, M. Palmer, and W. Schuler, "Dynamically altering agent behaviors using natural language instructions," *Autonomous Agents*, pp. 293-300, 2000.

[4] T. Tokunaga, K. Funakoshi, and H. Tanaka. "K2: animated agents that understand speech commands and perform actions," in *Proc. The 8th Pacific Rim International Conference on Artificial Intelligence 2004*, pp. 635-643, 2004.

[5] B. Coyne and R. Sproat. "WordsEye: An Automatic Text-to-Scene Conversion System," In AT&T Labs— Research, 2000.

[6] E. Hanser1, P. M. Kevitt1, T. Lunney1, J. Condell1, and M. Ma2. "SceneMaker: Multimodal Visualisation of Natural Language Film Scripts", University of Ulster, Magee Derry/Londonderry BT48 7JL, Northern Ireland, None.

[7] B. Coyne, R. Sproat, and J. Hirschberg, *Spatial Relations in Text-to-Scene Conversion, Columbia University*, New York NY, USA, None.

Kim D. Alcantara was born on October 3, 1993. He is 4th year student of Polytechnic University of the Philippines and taken up bachelor of science in computer science. Has a basic knowledge in computer programming in different languages such as C++, Java and C#. He is also proficient in using some Microsoft Office tools. He has a good communication skill and a good documentary.

Jomar P. Calandria was born on January 27, 1994. He is a 4th year student, currently studying at Polytechnic University of the Philippines and taking up bachelor of science in computer science. He is interested in computer programming using several programming languages such as C++, Java and C#. He is also interested in Web development using PHP and ASP.net. He has also basic knowledge in SQL and android development.

Junika S. Calupas was born on August 12, 1994. She is recently taking up her college degree in computer science from Polytechnic University of the Philippines, Manila. She knows how to code in Turbo C programming, C#, and Java. She also has knowledge in database MS Access as well as MS SQL and has a background in using Microsoft Office tools such as MS Word, Publisher, Excel, etc.

Jean Paula R. Echas was born in Munioz, Quezon City on January 12, 1995. She is a 4th year student of Polytechnic University of the Philippines, Manila and taking up bachelor of science in computer science. She is interested in web designing, photo editing and has knowledge in Database MS Access as well as MS SQL.

Ria A. Sagum was born in Laguna, Philippines on August 31, 1969. She took up bachelor of computer data processing management from the Polytechnic University of the Philippines and professional education at the Eulogio Amang Rodriguez Institute of Science and Technology. She received her master's degree in computer science from the De La Salle University in 2012.

She is currently teaching at the Department of Computer Science, College of Computer and Information Sciences, Polytechnic University of the Philippines in Sta. Mesa, Manila and a lecturer at the information and computer studies, Faculty of Engineering, University of Santo Tomas in Manila.

Ms. Sagum has been a presenter at different conferences, including the 2012 International Conference on e- Commerce, e-Administration, e-Society, e-Education, and e-Technology and National Natural Language Processing Research Symposium. She is a member of different professional associations including ACMCSTA and an active member of the computing society of the Philippines- Natural Language Processing Special Interest Group. She strives as a leader of a promising start-up software development group, Optika Studios.