View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Qatar University Institutional Repository

Multimedia Tools and Applications (2019) 78:22833–22859
https://doi.org/10.1007/s11042-019-7541-4

Text-to-picture tools, systems, and approaches: a survey

Jezia Zakraoui[1] • Moutaz Saleh[1] • Jihad Al Ja'am[1]

Received: 13 June 2018 / Revised: 6 December 2018 / Accepted: 20 March 2019 /
Published online: 25 April 2019
© The Author(s) 2019

Abstract

Text-to-picture systems attempt to facilitate high-level, user-friendly communication between humans and computers while promoting understanding of natural language. These systems interpret a natural language text and transform it into a visual format as pictures or images that are either static or dynamic. In this paper, we aim to identify current difficulties and the main problems faced by prior systems, and in particular, we seek to investigate the feasibility of automatic visualization of Arabic story text through multimedia. Hence, we analyzed a number of well-known text-to-picture systems, tools, and approaches. We showed their constituent steps, such as knowledge extraction, mapping, and image layout, as well as their performance and limitations. We also compared these systems based on a set of criteria, mainly natural language processing, natural language understanding, and input/output modalities. Our survey showed that currently emerging techniques in natural language processing tools and computer vision have made promising advances in analyzing general text and understanding images and videos. Furthermore, important remarks and findings have been deduced from these prior works, which would help in developing an effective text-to-picture system for learning and educational purposes.

Keywords Text-to-picture systems · Natural language processing · Natural language understanding · Text illustration · Text visualization · Multimedia

1 Introduction

A text-to-picture system is a system that automatically converts a natural language text into pictures representing the meaning of that text. The pictures can be static illustrations such as images or dynamic illustrations such as animations. Most of the very early work in text-to-

✉ Jihad Al Ja'am
jaam@qu.edu.qa

[1] Department of Computer Science and Engineering, College of Engineering, Qatar University, Doha, Qatar

picture systems concentrated on pictorially representing nouns and some spatial prepositions like maps and charts. For instance, the authors in [58] built the SPRINT system that generates 3D geometric models from natural language descriptions of a scene using spatial constraints extracted from the text. Throughout the last decade, many working text-to-picture systems have been developed. However, more efficient approaches and algorithms need to be developed. Joshi et al. [31] proposed a story picture engine that would depict the events and ideas conveyed by a piece of text in the form of a few representative pictures. Rada et al. [42] proposed a system for the automatic generation of pictorial representations of simple sentences that would use WordNet as a lexical resource for the automatic translation of an input text into pictures. Ustalov [12] developed a text-to-picture system called Utkus for the Russian language. Utkus has been enhanced to operate with ontology to allow loose coupling of the system's components, unifying the interacting objects' representation and behavior, and making possible verification of system information resources [13]. A system to create pictures to illustrate instructions for medical patients was developed by Duy et al. [15]. It has a pipeline of five processing phases: pre-processing, medication annotation, post-processing, image construction, and image rendering. More recently, a medical record summary system was developed by Ruan et al. [54]. It enables users to briefly acquire the patient's medical data, which is visualized spatially and temporarily based on the categorization of multiple classes consisting of event categories and six physiological systems.

A novel assisted instant messaging program to search for images in an offline database based on keywords has been proposed by Jiang et al. [29]. The final representation of the picture is constructed from a set of the most representative images. Jain et al. [28] proposed a Hindi natural language processor called Vishit, which aims to help with communication between cultures that use different languages at universities. The authors prepared an offline image repository module consisting of semantic feature tags that serve in the selection and representation of appropriate images, and it eventually displays illustrations linked with textual messages. Other important approaches [11, 23, 53] in the domain of news streaming have been proposed to usefully represent emotions and news stories The latter approach introduced new deep neural network architecture to combine text and image representations and address several tasks in the domain of news articles, including story illustration, source detection, geolocation and popularity prediction, and automatic captioning. All these technical contributions are evaluated on a newly prepared dataset.

According to [24], despite all these features of text-to-picture systems, they still have many limitations in terms of performing their assigned tasks. The authors pointed out that there is a possible way to improve the visualization to be more dynamic. They suggested directly creating the scene rather than showing representative pictures; this can be done via text-to-scene systems such as NALIG [1] and WordsEye [9], or text-to-animation systems such as animated pictures like text-to-picture Synthesis [21] and animations like Carsim [14], the latter of which converts narrative text about car accidents into 3D scenes using techniques for information extraction coupled with a planning and a visualization module. The CONFUCIUS system is also capable of converting single sentences into corresponding 3D animations [38]. Its successor, SceneMaker [22], expands CONFUCIUS by adding common-sense knowledge for genre specification, emotional expressions, and capturing emotions from the scripts.

Example of a common text-to-picture application is children's stories in which the pictures tell more of the story than the simple text [2, 5, 19, 57]. Huang et al. proposed VizStory in [25] as a way to visualize fairy tales by transforming the text to suitable pictures with consideration for the narrative structure and semantic contents of stories. Interactive storytelling systems have also been proposed, such as KidsRoom [4, 34, 56] and CONFUCIUS [18], the latter entails an interactive, multimodal storytelling system. For simple stories, in [46], the author proposed a system that can assist readers with intellectual disabilities in improving their understanding of short texts. A recent multimedia text-to-picture mobile system for Arabic stories based on conceptual graph matching has been proposed in [32]. For every input text, matching scores are calculated based on the intersection between the conceptual graph of the best selected keywords/sentences from the input text and the conceptual graphs of the pictures; in turn, matched pictures are assigned relative rankings. The best picture is selected based on the maximum intersection between the two graphs.

1.1 Domain application of text-to-picture systems

In text-to-picture systems, the visualized text can serve as a universal language for many applications such as education, language learning, literacy development, summarization of news articles, storytelling, data visualization, games, visual chat, rehabilitation of people with cerebral injuries, and children with delayed development. In the fields of education, language learning, and literacy development, an empirical study [6] strongly argues that text illustration with pictures generally enhances learners' performance and plays a significant role in a variety of text-based cognitive outcomes. For instance, an Arabic multimedia mobile educational system [32] has been proposed that allows users to access learning materials and mine illustrative pictures for sentences. Moreover, it has been shown in [43, 45] that representing and linking text to pictures can be very helpful for people to rapidly acquire knowledge and reduce the time needed to obtain such knowledge [10]. Language learning for children or for those who study a foreign language can also be improved through pictures [42].

Recently, studies on understanding learning behavior have suggested that the incorporation of digital visual material can greatly enhance learning [2] and promote imaginative capabilities, which is an important factor in inspiring creativity, as argued in [37]. In addition, the ability to encode information using pictures has benefits such as enabling communication to and from preliterate or non-literate people [29], improved language comprehension for people with language disorders, as argued in [42], and communication with children with autism [55]. Visualization and summarization of long text documents for rapid browsing, applications in literacy development [28], and electronic medical records [54] are also currently required. For instance, MindMapping, a well-known technique for taking notes and learning, has been introduced in work [16] as a multi-level visualization concept that takes a text input and generates its corresponding MindMap visualization. Yet, developing a text-to-picture system involves various requirements and challenges. The next section reviews some difficulties and challenges in developing text-to-picture systems.

1.2 Requirements and challenges

According to [24], in addition to the technical issues, difficulties and fundamental challenges in developing a text-to-picture system are rooted in natural language characteristics such as language being semi-structured, ambiguous, context-sensitive, and subjective. This work highlighted that designing a text-to-picture system capable of integrating a natural language interface and an interface for visualization purposes requires overcoming profound technical challenges in integrating artificial intelligence techniques, including natural language understanding (NLU), knowledge representation, planning, reasoning, and the integration of multiple knowledge sources and computer graphics techniques such as real-time rendering. With consideration for these requirements, the challenges of developing such systems are presented below.

1.2.1 Natural language understanding

Natural language understanding usually results in transforming natural languages from one representation into another [18]. Mapping must be developed in order to disambiguate a description, discover the hidden semantics within it, and convert it into a formal knowledge representation (i.e., semantic representation). This task presents a fundamental challenge; for more details, a brief overview of NLU issues can be found in [41]. Furthermore, language rarely mentions common-sense facts about a world that contains critically important spatial knowledge and negation [24]. Enabling a machine to understand natural language, which is variable, ambiguous, and imprecise, also involves feeding the machine grammatical structures (e.g., parts of speech), semantic relationships (e.g., emotional value and intensity), and visual descriptions (e.g., colors and motion direction) in order for it to match the language with suitable graphics [23]. The following figure (Fig. 1) shows the terminology of NLU and natural language processing (NLP) [41].

1.2.2 Knowledge representation, reasoning, and implicit knowledge

A text-to-picture system, which is designed to convert natural language descriptions to a visual representation, requires a knowledge representation component to represent the discovered semantics and use them to decide the actions to be taken. A reasoning mechanism embedded within the knowledge representation component can also help the system to derive implicit knowledge from available knowledge. Designing such a component is not a trivial task [24]. In addition, gathering the required tools and repositories, modeling, and obtaining the knowledge base (KB) is a challenge.

Despite all these difficulties, researchers are motivated to develop text-to-picture systems that can automatically convert the natural language descriptions to target visualizations either as a static picture or a dynamic animation. Furthermore, the current state of software and hardware for computer graphics technologies is highly advanced and can generate multimedia objects in real time, making the development of these systems an interesting challenge.

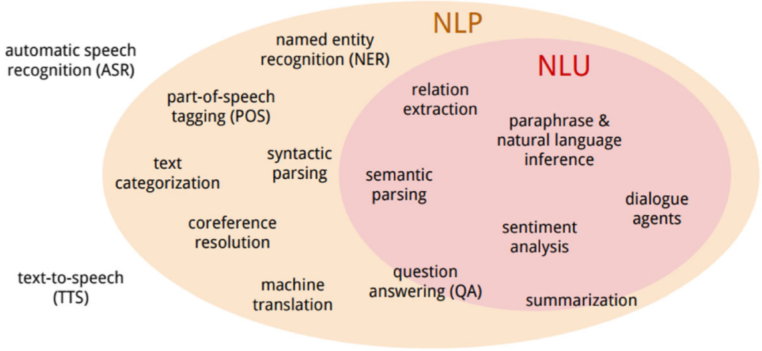


Fig. 1 Terminology of NLU versus NLP [41]

1.2.3 Loose image-text association

With the increased upload and usage of multimedia content, the pictures and texts involved can be loosely associated, and correspondence between pictures and texts cannot always be established, as highlighted in [59]. As a result, the association between pictures and texts in multimedia contexts can hardly be established using traditional methods, since the scale of the text alone can cover the entire natural language vocabulary. Therefore, there is a need for more powerful methods and techniques. The authors in [8] believe that there is now an increased difficulty in managing large multimedia sources to explore and retrieve relevant information. Unlike the English language, the Arabic language has complex morphological aspects and lacks both linguistic and semantic resources [27, 48, 49], yet another challenge to be addressed accordingly in image-text association.

2 Motivation for the survey

In the previous section, we highlighted the importance of text-to-picture systems in different contexts. Nowadays, such systems are still needed, since they have demonstrated their effectiveness at helping users to communicate, learn, and interpret textual information efficiently [50]. In particular, in situations constrained by time, capability, or technology, these systems have demonstrated the ability to clarify meanings and improve learning outcomes in cases including students with cognitive deficiencies, learning disabilities, or learning difficulties [3]. Nevertheless, many everyday tasks require a combination of textual and visual information (e.g., understanding slides while listening to a lecture in a classroom or reading and understanding a story). Hence, this survey reviews well-known text-to-picture systems, tools, and approaches in order to investigate their performance and limitations. Our research study is also motivated by the emerging techniques in NLP tools, computer vision, and their combination, which have proven to have made great advances toward their respective goals of analyzing and generating text, and the understanding of images and videos.

In the past, text-to-picture systems used to be viewed as a translation approach from a text language to a visual language [51] with excessive manual efforts. Nowadays, text-to-picture systems are being seen as information retrieval systems [50], which intensively involve emerging deep learning techniques, specifically Web image classification [7, 47], generic data classification [26], image annotation [40], image feature extraction, and image captioning [17]. Therefore, automatic text illustration with multimedia systems has become more feasible even with minimal manual effort due to the massive availability of both Web multimedia content and open-resource tools. However, the feasibility of such systems requires a combination of different powerful techniques from different research areas to produce accurate results. In particular, with the new advances in deep convolutional neural networks and long short-term memory, neural networks are gradually enhancing these areas of research, and there are promising signs for developing successful text-to-picture systems.

Although there are many working systems and applications that automatically generate images from a given sentence, text-to-picture systems for Arabic text are limited. Hence, more studies, reviews, and tools for analyzing Arabic sentences are required to recognize the potential for automatic Arabic text illustration and to open new horizons for research into the Arabic language in general.

The main objectives of this survey are as follows:

- Identifying key problems in existing text-to-picture systems.
- Comparing existing text-to-picture systems and discussing their advantages, disadvantages, and performance.
- Presenting NLP capabilities using available text processing tools.
- Presenting natural language semantic analysis capabilities using available lexical resources.
- Detecting different types of datasets, lexical resources, databases, KBs, and corpora.
- Figuring out the connections between language's vocabulary and its visual representations.
- Identifying relevant topics from NLP, computer vision, and neural network literature for further research.
- Reflecting on the problem statement, particularly for Arabic text.
- Concluding findings and remarks.

Indeed, a key objective of this review is to investigate the feasibility of "automatic visualization of Arabic story text through multimedia" using available tools and resources, particularly the automatic mapping of Arabic text to multimedia using Arabic language processing capabilities, and developing a successful text-to-picture system for educational purposes.

It is also important to mention that there are many other systems, reviewed in [24], that can convert general texts into high-level graphical representations; e.g., text-to-scene and text-to-animation systems. In this work, we focus on text-to-picture systems, approaches, and tools, which is the simplest form of visualization, and we review those which have only been published in scientific journals and at conferences. Online tools and applications are totally out of the scope of this survey.

In the next section, we present a detailed overview of state-of-the-art text-to-picture systems. For each one, we will elaborate on the system inputs and outputs, design methodology, language processes, and knowledge resources, as well as discuss the advantages and disadvantages.

3 Specific text-to-picture systems

Many text-to-picture systems have been developed to date, most of which differ in their methodology, utilized language understanding approach, syntactic analysis, KB scheme, inputs, and outputs. Thus far, those systems' outputs have shown that verbs are difficult to visualize. One possibility for addressing this challenge is to use hand-drawn action icons and link the verb to its constituents as predicted by semantic role labeling [21]. There are also many other complications regarding the visualization of relationships and spatial and temporal information. In this section, we list specific text-to-picture systems and approaches whose features will be compared.

3.1 Story picturing engine

The story picturing engine refers to the process of illustrating a story with suitable pictures [31]. The system is a pipeline of three processes: story processing and image selection, estimation of similarity, and reinforcement-based ranking. During the first process, some descriptor keywords and proper nouns are extracted from the story to estimate a lexical similarity between keywords using WordNet. For this purpose, the stop words are eliminated using a manually crafted dictionary, and then a subset of the remaining words is selected based on a combination of a bag-of-words model and named-entity recognition. The utilized bag-of-words model uses WordNet[1] to determine the polysemy count of the words. Among them, nouns, adjectives, adverbs, and verbs with a low polysemy count (i.e., less ambiguity) are selected as descriptor keywords of a piece of text. Those with very high polysemy are eliminated because they offer little weight to the meaning conveyed by the story [30]. A simple named-entity recognizer is then used to extract the proper nouns. Those images that contain at least one keyword and one named entity are retrieved from a local, annotated image database that has been set as an initial image pool.

The estimation of similarity between pairs of images based on their visual and lexical features is calculated based on a linear combination of integrated region matching distance [35] and WordNet hierarchy. Two forms of similarity measurement have been applied to consider visually similar images as well as images judged similar by annotations. Eventually, the images are ranked based on a mutual reinforcement method and the most highly ranked images are retrieved. This system is basically an image search engine that gets a given description as a query and retrieves and ranks the related images. Fig. 2 shows an example output for a story picturing engine.

Despite the good accuracy and performance of the story picturing engine, it only retrieves one picture for a given story and ignores many aspects such as temporal or spatial relationships. More advanced language processing techniques can be incorporated into the story picturing engine for richer performance; for instance, by integrating several image databases and building an online system that can accept stories provided by teachers and students [31].

3.2 Text-to-picture synthesis system

This system is a general-purpose text-to-picture system attempting to enhance communication. This system evolved and used semantic role labeling for its latest version [21] rather than keyword extraction with picturability, which measures the probability of finding a good image to represent the word [60]. Initially, the system starts with key phrase extraction to eliminate the stop words and then uses a Part of Speech POS tagger to extract the nouns, proper nouns, and adjectives. These words are then fed to a logistic regression model to decide the probability of their picturability based on the ratio of the frequencies under a regular Web search versus an image search. A TextRank summarization algorithm [44] is applied to the computed probabilities, and the top 20 keywords are selected and used to form the key phrases, each having an assigned importance score.

For image selection, the process is based on matching the extracted key phrases with the image annotations. First, the top 15 images for this key phrase are retrieved using Google Image Search. Next, each image is segmented into a set of disjointed regions using an image segmentation algorithm. Then, a vector of color features is calculated for all images and clustered in the feature space. Finally, the largest cluster is searched to find the region whose feature vector is closest to the center of this cluster. The image that contains this region is then selected as the best image for this key phrase.

[1] https://wordnet.princeton.edu/

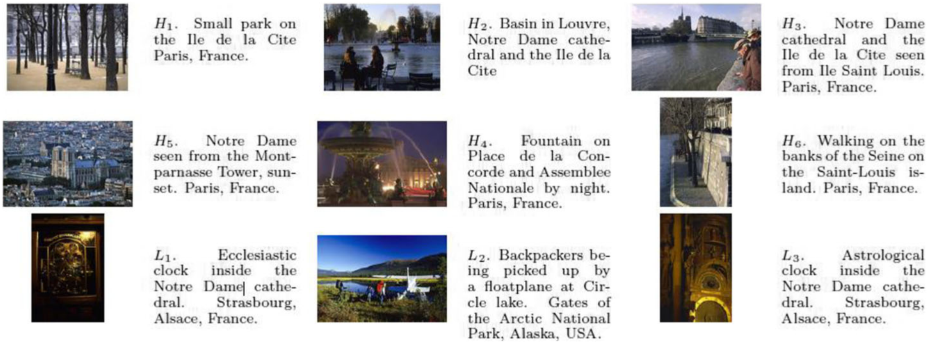

H_1. Small park on the Ile de la Cite Paris, France.
H_2. Basin in Louvre, Notre Dame cathedral and the Ile de la Cite
H_3. Notre Dame cathedral and the Ile de la Cite seen from Ile Saint Louis. Paris, France.
H_5. Notre Dame seen from the Montparnasse Tower, sunset. Paris, France.
H_4. Fountain on Place de la Concorde and Assemblee Nationale by night. Paris, France.
H_6. Walking on the banks of the Seine on the Saint-Louis island. Paris, France.
L_1. Ecclesiastic clock inside the Notre Dame cathedral. Strasbourg, Alsace, France.
L_2. Backpackers being picked up by a floatplane at Circle lake. Gates of the Arctic National Park, Alaska, USA.
L_3. Astrological clock inside the Notre Dame cathedral. Strasbourg, Alsace, France.

Fig. 2 Example output of a story picturing engine for the text “on walk through Paris”; H = highest ranked images, L = Lowest ranked images [31]

In the final stage, the system takes the text, the key phrases, and their associated images, and determines a 2D spatial layout that represents the meaning of the text by revealing the important objects and their relationships (see Fig. 3). The retrieved pictures are positioned based on three constraints: minimum overlap, centrality of important pictures, and closeness of the pictures in terms of the closeness of their associated key phrases. For that reason, the authors designed a so-called ABC layout, such that each word and its associated image is tagged as being in the A, B, or C region using a linear-chain conditional random field [60].

In contrast to the story picturing engine, this system associates a different picture with each extracted key phrase and presents the story as a sequence of related pictures. It treats the text-to-picture conversion problem as an optimization process, as mentioned in [24], and it still inherits the drawbacks of text-to-picture systems despite its performance compared to the story picturing engine. For complex sentences, the authors anticipate the use of text simplification to convert complex text into a set of appropriate inputs for their system. According to [24], the simplicity and the restriction to simple sentences may have prevented the system from reaching its goal because some elaborated steps possibly distort the meaning of the text. Moreover, the use of hand-drawn action icons only for visualization makes the system very restricted. This is where the true value of modern text-to-scene systems can be seen more efficiently, according to [24].

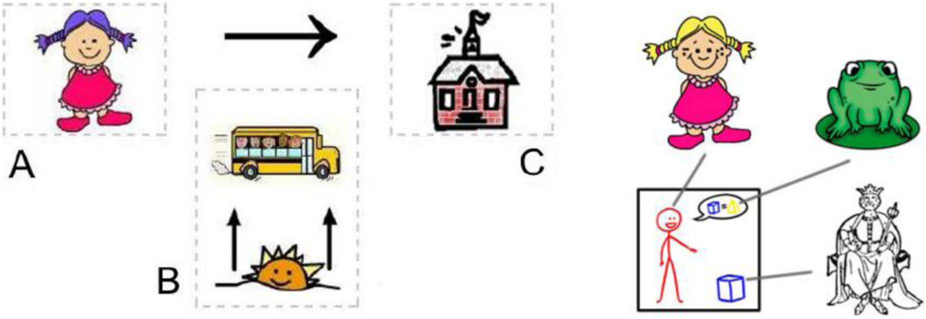


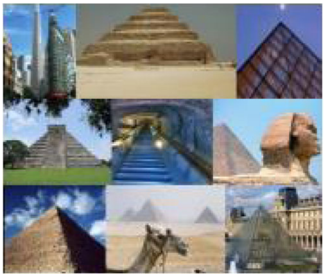
Fig. 3 Examples outputs from text-to-picture synthesis for the sentence, “The girl rides the bus to school in the morning” (left) [20], and for the sentence, “The girl called the king a frog” (right) [21]

3.3 Word2Image

This system proposed various techniques, including correlation analysis and semantic and visual clustering, in order to produce sets of images from Flickr [36]. The translation of a word into its visual counterpart with a set of high-quality, precise, diverse, and representative images undergoes following five steps.

1. Generating a diverse and precise image set in which a heuristic for diversity is attained through the collection of images from different groups, users, times, and locations, and which show enough variations on both semantic and visual levels.
2. Using a correlation analysis for precision to filter out the wrong and noisy images, accomplished through conducting a correlation analysis using Flickr's list of "related" tags, based on clustered usage analysis. An image is accepted as relevant only if the correlation score is greater than a certain threshold.
3. Using a semantic cluster where the system computes the saliency of each keyword in the set of tags and titles, keeping and utilizing only top-M keywords which are ranked at top M positions, to represent each image with an M-dimensional vector. An agglomerative algorithm is also used to hierarchically cluster the image set into different groups, and a cluster merging process is followed to combine the small clusters and form a larger cluster for later visual clustering.
4. Applying visual clustering to each semantically consistent cluster obtained from the previous step in order to divide them into visually coherent sub-clusters, and then selecting representative images for each cluster. All these clusters must follow certain criteria in order to compete for being selected to be shown to the user. Within each cluster, the images are also ranked according to their representativeness.
5. The system adopts the collage technique to construct a compact and visually appealing image collage from the top representative images of the top-K clusters which are ranked at top K positions. The images are resized according to the respective cluster's ranking score. To make the representative image more easily understandable, a large version of the original-size image will be shown when the user places the mouse over it, and the top-4 i.e. the top four keywords will be displayed to depict its content. Fig. 4 shows an example output for the concept of "pyramid."

Although Word2Image translates a concept into its visual counterpart with sets of high-quality, precise, diverse, and representative images, it inherits problems related to the input language, such as restricting the input to single words only.



Discovered topics:

- France- Paris- museum- Louvre
- Africa- Egypt- Cairo- desert
- Mexico- Yucatan- Maya- temple
- history- architecture- Giza- Sphinx

Fig. 4 Example output from Word2Image for the word "Pyramid" [36]

3.4 Enriching textbooks with images

This approach proposes techniques for finding images from the Web that are most relevant for augmenting a section of the textbook while also respecting the constraint that the same image is not repeated in different sections of the same chapter [2]. The techniques are comprised of optimizing the assignment of images to different sections within a chapter, mining images from the Web using multiple algorithms, and, finally, "ensembling" them. Upon image assignment, each section of the textbook is assigned the most relevant images such that the relevance score for the chapter is maximized while maintaining the constraints that no section has been assigned more than a certain maximum number of images (each section is augmented with at most k number of images) and no image is used more than once in the chapter (no image repeats across sections). A polynomial time algorithm is also used for implementing the optimizer.

For image mining, two algorithms are used for obtaining a ranked list of the top-k images, where k is the number of images, and their relevance scores for a given section where various possible variants of these algorithms are accepted, as well as additional image-mining algorithms that could be produced. The relevance score for an image is computed by analyzing the overlap between the concept phrases and the image metadata. The ranked lists of image assignments are then aggregated by image ensembling in order to produce the final result. Ensembling is done sequentially within a chapter, starting from the first section. Top images selected for a section are eliminated from the pool of available images for the remaining sections. The image assignment is then rerun, followed by ensembling for the next section. Fig. 5 summarizes the steps described above.

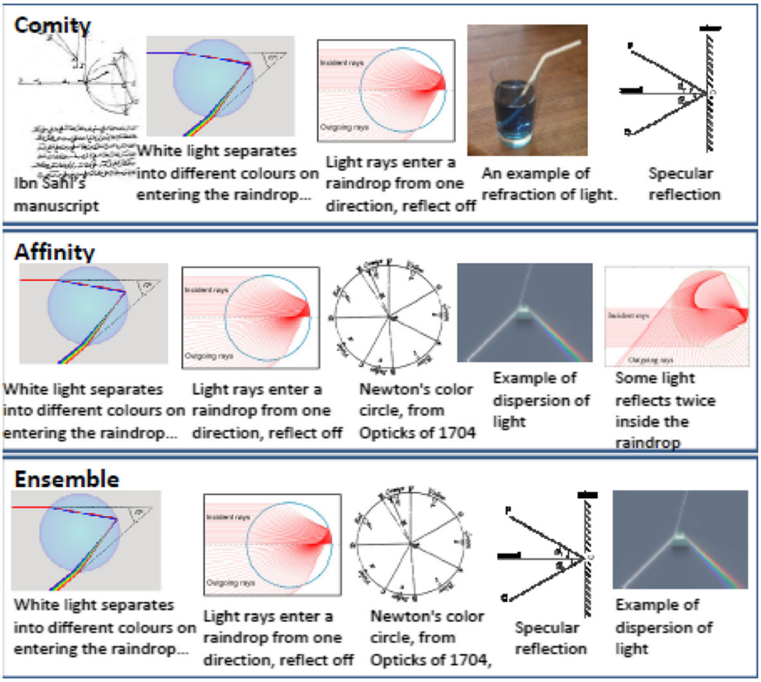

Comity
Ibn Sahl's manuscript
White light separates into different colours on entering the raindrop...
Light rays enter a raindrop from one direction, reflect off
An example of refraction of light.
Specular reflection
Affinity
White light separates into different colours on entering the raindrop...
Light rays enter a raindrop from one direction, reflect off
Newton's color circle, from Opticks of 1704
Example of dispersion of light
Some light reflects twice inside the raindrop
Ensemble
White light separates into different colours on entering the raindrop...
Light rays enter a raindrop from one direction, reflect off
Newton's color circle, from Opticks of 1704,
Specular reflection
Example of dispersion of light

Fig. 5 Example output for the section on "Dispersion of white light by a glass prism" [2]

The evaluation conducted through the Amazon Mechanical Turk platform showed promising results for the proposed system and indicated that the proposed techniques are able to obtain images that can help increase the understanding of the textbook material; however, deeper analysis to identify key concepts is needed. Despite its promises, the concepts in this work are divided into the two parts-of-speech categories: adjectives, nouns, and sometimes prepositions, thus ignoring all other categories of terms. Moreover, the proposed system does not support interactive modification of the images and does not embed any lexical or commonsense resources.

3.5 A text-to-picture system for the Russian language

Utkus is a text-to-picture synthesis system designed for Russian language processing and operates with the natural language analysis component, the stage processing component, and the rendering component [12]. The system evolved from its previous version to convey the gist of general, semantically unrestricted Russian language text. It consists of three processing stages: linguistic analysis, depictor generation, and picture synthesis. In first stage, the system uses shallow semantic analysis starting by tokenization using the Greeb[2] tokenizer, then moves to morphological parsing using the Myaso[3] analyzer, and, finally, performs syntactic parsing of the tokenized and morphologically annotated input text. A semantic representation of the input text is prepared. During the second stage, the obtained semantic representations are transformed into depictors, which are simple instructions to the Renderer block. Depiction rules are found using lemmatized verbs from the input expressions and mapping the expression arguments by defined rule operations. In the last stage, depictors are further executed by Renderer, which iterates across the depictor list, executes each depictor of the list, and stores it in the Renderer state, which is necessary for performing the picture layout. Renderer creates the output file and places the graphical primitives in the computed place and state according to the layout generation results; see Fig. 6 for an output scene.

According to [13], this approach provides loose coupling of ontology, thesaurus, gallery, and depiction rules. However, only verb phrases and related noun phrases are extracted from the dependency tree of each sentence of the text; other parts of speech, such as adjectives, pronouns, and numerals, are not considered. It should be noted that the Utkus system is currently unable to represent plural words or to solve the problem of semantic ambiguities. Many other deficits and several reasons for future works have been mentioned in [12].

3.6 Vishit: A visualizer for Hindi text

Vishit is an approach for processing Hindi texts for visualization purpose [28]. It consists of three major processing steps: language processing, KB creation, and scene generation. To perform language processing, the input text is analyzed sentence by sentence, which includes a morphological analyzer, POS tagger, and parser. The parsing of an input sentence forms a dependency tree with the identification of semantic arguments like subject and predicate along with their roles and action verbs. Here, the structure of the obtained dependency tree helps in resolving many co-references and in extracting subject–predicate information while the associated role with identified objects helps to identify semantic relationships between the objects.

[2] https://github.com/eveel/greeb

[3] https://github.com/eveel/myaso

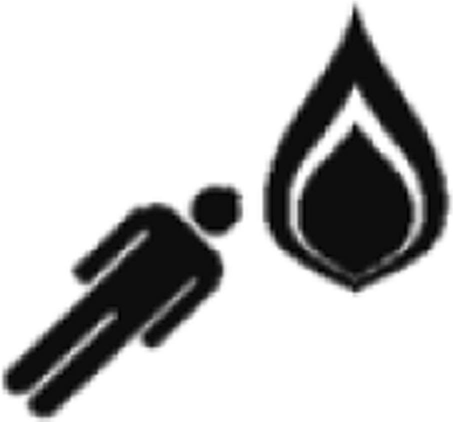


Fig. 6 Example output of for the sentence "A man has fallen into the fire" [12]

During KB creation, spatial relations are identified by recognizing the context of where objects are located with respect to the overall scene or relative to other objects and providing background and spatial information. A taxonomy repository with manual identification of semantic relations is created. In the scene generation step, a background is first selected for the scene, and objects, along with their relative positions, are determined. Then, an image is rendered by surface extraction and spatial partitioning along with the detection and removal of object collisions for positioning objects. Eventually, an offline image repository is used to stores various types of objects with semantic feature tags that later serve in the selection and representation of appropriate images.

Initial results from Vishit seem encouraging, since the generation of a dependency tree with the identification of semantic arguments like subjects and predicates has the potential to be used for other input languages. However, this system is still in its prototype phase, and the image processing that is used attempts a manipulation of the image parameters that could result in changing the actual meaning of the text. Additionally, it requires annotated images to be prepared a priori, which is a labor-intensive and tedious task. Another problem is that this system is not interactive and does not exploit the user's feedback.

3.7 Chat with illustration

Chat with illustration (CWI) is a visually assisted instant messaging scheme that automatically presents users with visual messages associated with text messages [29]. The CWI scheme consists of the following five tasks.

1. Image database construction: An indexed image database is set and divided into two parts. One part corresponds to unpicturable keywords which have been labeled to images manually. The other part corresponds to the picturable keywords and was built automatically. For the latter part, a two-level image filtering strategy is performed. Then, semantic and visual similarities of images are exploited, and an advanced clustering algorithm is applied to hierarchically cluster the images to semantically and visually consistent groups.
2. Dialogue analysis: This module detects meaningful keywords and analyzes grammatical and logical relationships. Meaningful keywords reflect users' intent in chat and are used as

query words in image selection. The analyzed relationships are used as the foundation for the visual layout.

3. Image selection and ranking: A set of representative images for obtained meaningful keywords of dialogue are searched and selected from the image database. For unpicturable keywords, the representative images are searched manually, whereas for picturable keywords, a two-step approach has been developed. First, the most proper sub-cluster is selected from all sub-clusters that are clustered with specific semantic aspects. Then, with the help of visual and tag information, images in the selected sub-cluster are ranked, and the most representative image for the keyword in the specific dialogue is selected.
4. Visual layout: A template-based visual layout scheme is prepared in order to present the obtained representative images for meaningful keywords in the dialogue from the previous step. Some image layout templates have been designed based on grammatical relationships between words and logical relationships between clauses. Each template stands for a certain semantic unit and is simple and very easy to understand.
5. Visual dialogue summarization: When the chat is finished, a visual dialogue summarization is made by illustrating the main picturable concepts in the dialogue. The image size, centrality, and distance define how the images are integrated in one illustration. The locations of all images are formulated as an optimization problem and solved by the use of a Monte Carlo randomized algorithm. See the following figure (Fig. 7) for visualized dialogue.

Despite the accuracy and the intuitiveness of the visual layout, CWI relies on making excessive preparations of image resources. Furthermore, only a few unpicturable concepts such as verbs, adjectives, fixed phrases, and interrogatives are considered and labeled to images manually, which makes the chat directed more toward concrete concepts. Another disadvantage of this system is its limited capability in terms of language processing and grammatical and logical relationships.

3.8 Illustrate it! An Arabic multimedia text-to-picture m-learning system

Illustrate It! is an Arabic multimedia text-to-picture mobile learning system that is based on conceptual graph matching [32]. To build a multimedia repository, the system uses the Scribd[4] online book library in order to collect educational stories which are then stored locally in binary format and marked for text extraction. An educational ontology is built to provide educational resources covering different domains such as the domain of animals' stories, in particular it describes the story's structure, the question's semantic structure and the grammatical tree structure.

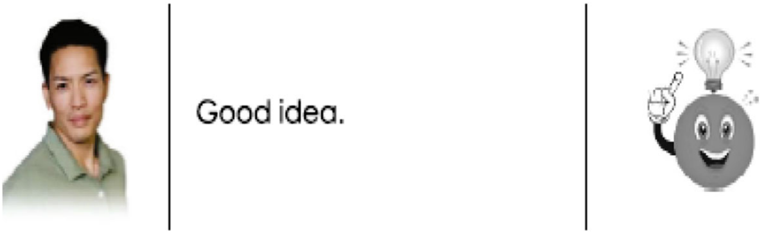


Fig. 7 Example output for the message "Good idea" [28]

For text processing, relationships between entities in the story text are extracted using a basic formal concept analysis approach based on the composition of entity-property matrices. A conceptual graph for the best selected sentence and for the best selected keywords is also built. The obtained graph is used to select the best picture based on the maximum intersection between this graph and the conceptual graphs of the pictures in the multimedia repository.

The proposed system solves the limitations of existing systems by providing pedagogic illustrations. However, the current implementation cannot automatically find illustrations that do not have annotations or textual content, and cannot locate annotated elements in the picture. Moreover, the system uses only cartoon images and disregards essential educational benefits from other multimedia types, only focuses on entities and relationships, and ignores spatial and temporal relationships and other story clues that can be used to infer the implicit knowledge.

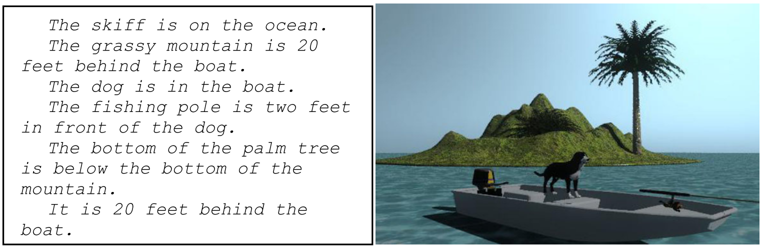
3.9 WordsEye

WordsEye is a text-to-scene system that can automatically convert input text into representative, static, 3D scenes [9]. The system consists of two main components: a linguistic analyzer and a scene depicter.

First, the input text that can include information about actions, spatial relationships, and object attributes is parsed, and a dependency structure is constructed that represents the dependencies among the words to facilitate the semantic analysis. This structure is then utilized to construct a semantic representation in which objects, actions, and relationships are represented in terms of semantic frames.

Then, the depiction module converts the semantic frames into a set of low-level graphical specifications. For this purpose, a set of depiction rules is used to convert the objects, actions, relationships, and attributes from the extracted semantic representation to their realizable visual counterparts. The geometric information of the objects is manually tagged and attached to the 3D models. This component also employs a set of transduction rules to add implicit constraints and resolve conflicting constraints. Finally, once the layout is completed, the static scene is rendered using OpenGL similar to the example output shown in Fig. 8.

Although WordsEye has achieved a good degree of success, the allowed input language for describing the scenes is stilted, as mentioned in [24]. It is not interactive and does not exploit the user’s feedback. Moreover, WordsEye relies on its huge offline rule base and data repositories containing different geometric shapes, types, and similar attributes. These elements are manually annotated, meaning that WordsEye lacks an automatic image annotation task.



```
   The skiff is on the ocean.
   The grassy mountain is 20
feet behind the boat.
   The dog is in the boat.
   The fishing pole is two feet
in front of the dog.
   The bottom of the palm tree
is below the bottom of the
mountain.
   It is 20 feet behind the
boat.
```

Fig. 8 Example output for the text on the left side [9]

3.10 Confucius

CONFUCIUS is a multi-modal text-to-animation conversion system that can generate animations from a single input sentence containing an action verb and synchronize it with speech [18]. It is composed of several modules with different tasks to accomplish; we briefly mention the relevant modules:

1. Knowledge base: Encompasses a lexicon, a parser, and a visual database that contains a very limited set of 3D models and action animations.
2. Language processor: Uses a Connexor functional-dependency grammar parser, WordNet, and a lexical conceptual structure database to parse the input sentence, analyzes the semantics, and outputs lexical visual semantic representation.
3. Media allocator: Exploits the acquired semantics to generate an XML representation of three modalities: an animation engine, a speech engine, and narration.
4. Animation engine: Uses the generated XML representation and the visual database to generate 3D animations, including sound effects.
5. Text-to-speech engine: Uses the XML representation to generate speech.
6. Story narrator: Uses the XML representation to initialize the narrator agent.
7. Synchronizer: Integrates these modalities into a virtual reality modelling language file that is later used to render the animation.

CONFUCIUS can address the temporal relationships between actions. It integrates the notion of visual valency [39], a framework for deeper semantics of verbs, and uses it as a primary criterion to re-categorize eventive verbs for the animation. It utilizes the humanoid animation standard for modeling and animating the virtual humans. As seen in Fig. 9, CONFUCIUS supports lip synchronization, facial expressions, and parallel animation of the upper and lower body of human models.

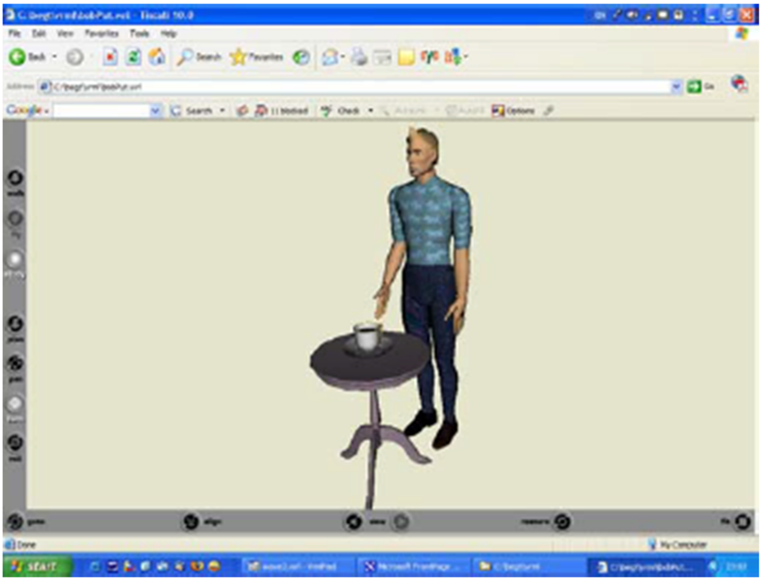


Fig. 9 The output animation of "John put a cup on the table" [18]

Generally, CONFUCIUS is not interactive in the sense that it does not let the user modify the generated animation [24]. In addition, only a single sentence is allowed in each input and only a restricted format of the input sentences (i.e., one action verb per sentence, and only simple verbs are considered) is permitted, hence the user is restricted in expressing the intended description.

4 Discussion

The reviewed text-to-picture systems treat the problem of mapping natural language descriptions to a visual representation as an image retrieval and ranking problem [24]. The authors in [51] see the problem from another perspective; namely, as a knowledge extraction and translation problem. Together, these systems extract concepts from the input text, then match them against the image annotations, and a subset of the images is retrieved and ranked based on some predefined similarity measures. The retrieved images with the highest rank are illustrated.

From a detailed literature review, we see several attempts at illustrating text with pictures to help with better understanding and communication. Literature also shows efforts to translate text to a picture, text picturing, natural language visualization, etc. Hence the common features for most text-to-picture systems and approaches include the following [50], which are also illustrated in Fig. 10:

1. Input interface
2. A processing pipeline
3. Knowledge extraction tools
4. External knowledge resources to help resolve ambiguous terms
5. Image resources with semantic information
6. Matching concepts against image annotations
7. Ranking problem
8. Output rendering

4.1 Comparison of reviewed text-to-picture systems

We have surveyed several text-to-picture approaches and systems. The evolution of the text-to-picture systems can be identified as moving in two main directions [24]. First, the systems have evolved in terms of extracting text and image associations since they exploit new visual

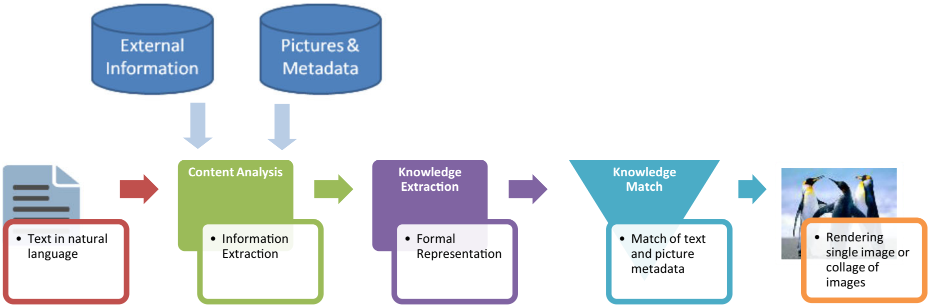


Fig. 10 Text and image processing pipeline

features fused with semantic feature associations instead of text and image annotations. Second, the systems have evolved in their produced output such that the early systems provide only one representative picture, whereas successor systems provide a set of images ordered based on the temporal flow of the input.

Table 1 below gives an overall comparison focusing on NLP, NLU, input, and output modalities of the reviewed real and functioning text-to-picture systems only. The plus/minus (±) signs indicate the support each system has for features listed in the table header.

As the table indicates, WordsEye is the only system that has a good NLU component; however, the allowed input language for describing the scenes is stilted [24]. The other systems, which have more enriched input and output interfaces, have weaker NLP, and all completely lack NLU. Many other features are shown in the following tables for a comparison between the reviewed text-to-picture approaches and systems.

We focus on text analysis models used by these systems, thus categorizing them into two groups as systems either following shallow semantic analysis or deep semantic analysis. As shown in Table 2, systems that use shallow semantic analysis models typically provide a naïve semantic parsing such as semantic role labeling. However, systems that use deep semantic analysis or linguistic approaches investigate deeper semantic parsing such as dependency parsing and semantic parsing (see Table 3).

First of all, we summarize the technical details, text processing, and rendering features of reviewed prior work with the following criteria:

- Interactivity: Indicate whether it provides the user with the means to manipulate the generated output.
- Adaptive: Refer to the system's capability of extracting information that is not given to the system a priori.
- Input: Input text can be word(s), sentence(s), paragraph(s), story, query, speech, script, regular expression, etc.
- Output: Determine the output modality (e.g., 2D picture collage or speech) of the system.
- Domain: Determine whether it is built for general or custom purposes.
- Methodology: Determine the method used by the system (data-driven, rule-based, or multi-agent methods).
- Syntactic and semantic analyses: Indicate the approaches that the system utilizes to analyze the input.
- Knowledge resource: Determine the exploited knowledge resources (e.g., lexicons or WordNet) and help to add related context to the input; image resources determine the visual resources (e.g., Google images or Flickr) and enable automated image retrieval.

Table 1 Comparison of functional text-to-picture systems focusing on NLP, NLU, and modalities

System	NLP	NLU	Multimodality	
			Input	Output
Story picturing engine [31]	+	–	+	–
Text-to-picture synthesizer [60]	+	–	–	+
Utkus [12]	+	–	+	–
WordsEye [9]	+	+	–	+
CONFUCIUS [18]	+	–	–	+

Table 2 Comparison of text-to-picture systems following shallow semantic analysis

System/ Year Interactive/ Adaptive	Input/ Output	Domain/ Methodology	Syntactic/ Semantic Analysis	Knowledge Resource	Image Ranking/Layout
Story picturing engine [31]/ 2006 / No/ Yes	Paragraph/ 2D picture	General/ data-driven	Bag-of-words/ semantic association	WordNet, Terragalleria[a], AMICO[b], ARTKids[c]	Mutual reinforcement analysis/ none
Text-to-picture synthesizer [60]/ 2007, 2009/ No/ Yes	Typed text/ 2D picture collage	General/ data-driven	POS tagging/ association, semantic role labeling	Google Images, Google Web	Segment at center of largest cluster / A, B, or C region using a linear-chain CRF
Word2Image [36]/ 2008/ No/ Yes	Words being pictured/ 2D picture collage	General/ data-driven	Bag-of-words/ semantic association	Flickr	Images clustering, saliency-based rank/ images resize to cluster ranking score
Enriching textbooks with images [2]/ 2011/ No/ No	Typed paragraph/ 2D picture	School textbook/ data-driven	POS tagging, key phrase/ semantic association	NCERT[d] Corpus, Web search	Top k relevance ranking/ none
Chat with illustrator [29]/ 2014/ No/ No	Typed short text	General/ data-driven	Bag-of-words/ semantic association	Flickr	Image clustering, saliency-based rank/ size, location as optimization problem
Illustrate It! [32]/ 2017/ No/ No	Story/ 2D pictures	Children's stories/ data-driven	Bag-of-words/ hierarchical association	Arabic WordNet, OWL/Scribd[e]	Image cluster-ing, maximum conceptual graph intersection/ none

[a] http://www.terragalleria.com/

[b] http://www.amico.org/

[c] http://www.artfaces.com/artkids

[d] National Council of Educational Research and Training (NCERT), India

[e] http://www.scribd.com

Table 3 Comparison of text-to-picture systems following deep semantic analysis

System/ Year/ Interactive/ Adaptive	Input/ Output	Domain/ Methodology	Syntactic/ Semantic Analysis	Knowledge Resource	Image Ranking/ Layout
Utkus [12]/ 2012/ No/ Yes	Text/ icons collage	General/ rule-based	Chunking (verb, noun phrases)/ dependency parsing	XML-based KB, Thesaurus[a], Noun Project[b]	None/ PNG raster images of 640 × 480
Vishit [28]/ 2014/ No/ No	Text/ scene	Animals/ data-driven	POS tagging/ dependency parsing	XML-based KB, offline image repository	None/ surface extraction and spatial partitioning
WordsEye [9]/ 2011/ No/ No	Story/ 3D pictures	General/ rule-based	Statistical parsing/ dependency parsing	FrameNet[c], VigNet[d], WordNet, 3D Objects	None/ depiction rules for 3D poses
CONFUCIUS [18]/ 2006/ No/ No	1 Sentence/ 3D animation, audio	General/ rule-based	Dependency parsing/ lexical vi sual semantic	WordNet, LCS Database, FrameNet, VerbNet[e]	None/ animation engine

[a] http://speakrus.ru/dict/index.htm

[b] http://thenounproject.com

[c] English lexical database containing manually annotated sentences for semantic role labeling

[d] Extension of FrameNet

[e] English verb lexicon

- Image ranking: Determine the method of image clustering, selection, and ranking.
- Image layout: Determine the image placement and layout.

4.2 Remarks and findings obtained

Many reviews such as [24, 50, 52], and research works such as [18, 38] on text-to-picture systems highlight the following issues:

1. **Natural language understanding**: Most of the systems face many technical difficulties in understanding natural language. Therefore, they restrict the form of the input text to overcome these difficulties (e.g., one sentence in simple English is allowed as the input for a text-to-picture synthesizer) [21]. Other approaches restrict the conceptual domain to a specific domain (e.g., Vishit [28] restricts the conceptual domain to the domain of animals in different environments).
2. **Natural language processing**: Most of the systems focus on the information retrieval task and do not elaborate on the language processing aspects, including morphological, syntactic, and semantic analyses. However, in terms of language understanding and richness of the model repository, WordsEye [9] outperforms all reviewed systems.
3. **Syntax analysis**: Half of the systems use the bag-of-words representation model that treats a given text as a set of words and frequencies and disregards the syntax and word order. The rest of the reviewed systems utilize POS tagging. However, most systems do not analyze all words; some focus on just one or two parts of speech (e.g., the approach in [2] considers only nouns and adjectives).
4. **Semantic analysis**: Most of the systems lack a strong semantic analysis for capturing more general dependencies and rely on shallow parsing rather than attempting to capture the deep semantics. More specifically, most of the general-domain systems follow the shallow semantic model (e.g., Word2Image [36] matches keywords with high saliency scores to image tags and thereby introduces the notion of semantic clustering). Furthermore, only a few of the reviewed systems use KBs and ontologies for semantic analysis. Dependency analysis (or parsing), which is a linguistic-based tool, is also used in such general systems; however, it is mostly used for rule-based general domains rather than data-driven general domains.
5. **Dependency analysis:** This is a linguistic-based tool used for knowledge extraction. The Stanford Dependency Parser[5] is usually considered the gold standard for extracting linguistic (syntactic) information.
6. **Abundance of text analysis tools:** With the evolution of computing capabilities, the required linguistic tools have also evolved to the point where standard components (e.g., the Stanford Dependency Parser) can be used in the text analysis, even for Arabic text analysis. An NLP system (MADAMIRA) for Arabic text analysis has been designed by Pasha et al. [49], and includes several NLP tasks.
7. **Data-driven:** Most of the text-to-picture systems use data-driven image retrieval methods and try to solve the problem of mapping natural language to a visual representation using Web-based image search engines (e.g., Illustrate It! [32] uses Google and Scribd).

[5] http://nlp.stanford.edu/software/lex-parser.shtml.

8. **Rule-based**: Few systems use rule-based methodology; however, current data-driven systems do not outperform the rule-based systems [24]. This is probably because the data-driven systems have only been used for feasibility studies, whereas a few rule-based systems such as WordsEye are commercialized and supported by the required resources for crafting as many rules as possible.
9. **Input/Output**: In terms of inputs, only a few systems allow for a general, unrestricted natural language (e.g., WordsEye). On the other hand, systems have evolved in terms of output. The early systems provided the users with only one representative picture, as described in [31], whereas later systems have provided users with a set of images based on their relevance and have also provided an appropriate layout. More sophisticated outputs in the form of 3D animations with sound effects and displayed emotions are also available, as described in [22].
10. **External text resources**: Most of the systems used the WordNet lexicon as a typical text knowledge source in earlier works. However, a large proportion of the general-domain systems that require common-sense knowledge are not equipped with any knowledge resources. This fact highlights another fundamental problem of the current systems. They simply ignore the knowledge resources, meaning that they cannot infer in unpredictable situations and cannot be adaptive.
11. **External image resource**: Most of the systems rely on third-party image collections such as Flickr, while only a few rely on their own offline image resource with excessive preprocessing stages that include backgrounds and frames (e.g., CWI [29] relies on making excessive preparations of image resources). The visualization within this system is restricted to available images within that resource.
12. **Image annotation**: Most of the systems exploit the surrounding text of the images and the text appearing within HTML tags. Some of the systems apply an automatic annotation by collecting both images and text and then using the co-occurring text around the images to annotate them. On the other hand, there are other techniques attempting to extract the text within the image (e.g., Illustrate It! [32] transforms the image into a binary format and employs the library Tess4J[6] for optical character recognition to transform the textual content in the image into characters that are exploited for matching relevant images).
13. **Image retrieval process**: Most of the systems carry out this process by extracting concepts from the input text and then matching them against the image annotations, after which a subset of images is retrieved and ranked for a given concept based on some predefined similarity measures. In some systems, the retrieved images with the highest rank are then illustrated based on an image layout algorithm.
14. **Image layout**: Most of the systems devote significant effort to image selection, image clustering, and image layout (e.g., CWI [29] applies several image layout templates to cover grammatical relationships in a dialogue).
15. **Semantic Web**: Resources of the Semantic Web are not used, except for ontologies.
16. **Interactivity**: Most of the systems are not interactive because they lack a solid mechanism to harvest the information from user interactions and feedback.
17. **Adaptivity**: Few systems are adaptive and most of these systems also ignore a priori knowledge provided by experts or other resources.

[6] http://tess4j.sourceforge.net/

Hence, the literature shows that successful text-to-picture systems have good language understanding components, but also have fewer input channels and less intuitive visual layouts in terms of output. Contrary multimodal systems have more enriched input/output interfaces and better graphics quality, but they suffer from weaker NLP, as some of them simply ignore NLP completely, as mentioned in [18]. Overall, we have identified three main problems with the current systems. The first problem is associated with NLU, since these systems cannot capture the deep semantics embedded within the natural language descriptions. The second problem is related to visualization, which is, in turn, restricted to available images. The third problem is rooted in the fact that the current systems lack the available resources (e.g., lexicons) and the available techniques to manage and integrate open source datasets in real time.

5 Effectiveness of the survey

To the best of our knowledge, this survey is one of few reviews [24, 50] of text-to-picture systems and approaches associated with illustrating natural language. This survey has been carried out to derive the feasibility and the outcome of illustrating the Arabic language as a proof of concept. This work has presented the main problems faced by text-to-picture systems with respect to NLP, NLU, and many other requirements. For each reviewed system, we elaborated on the system's inputs and outputs, design methodology, language processes, and knowledge resources, as well as discussing the advantages and disadvantages. Many other features are shown in Table 2 and Table 3 for a clear comparison between the reviewed text-to-picture approaches and systems.

We focused on NLP and analysis models used by these systems, and thus categorized them into two groups. We concluded that systems following deep semantic analysis have higher accuracy compared to those following shallow semantic analysis. We have also shown some systems that have enriched input and output interfaces, but weaker NLP and NLU, and therefore weaker accuracy. This not only reflects the current technical difficulties in understanding natural language, but also showcases the semantic gap [33] between human perception and computer vision; i.e., semantic gaps between humans perceiving their surroundings and the computer analyzing datasets.

Furthermore, the survey showed that there is no open dataset available for the purpose of illustrating natural language, or at least for common language concepts in general. Thus, in order to overcome the semantic gap, it is important to have a deep understanding of how a language's vocabulary and its visual representations connect. Whereas some text-to-picture systems rely on many filtering algorithms and techniques in order to get appropriate materials from Web image searches, other systems create their own multimedia datasets, which has revealed the excessive manual efforts behind these systems.

In terms of input/output-modalities, early systems provided only one representative image (e.g. the story picturing engine [31]), whereas recent systems provide a set of images (e.g. Word2Image [36]). In terms of spatial and temporal relationships, all reviewed text-to-picture systems were unable to address them; this is probably because these relationships can only be visualized through dynamic systems (e.g., animation).

It should be noted that some of the reviewed systems are not available to date and are no longer enhanced (e.g., the story picturing engine [31]). Ultimately, we have concluded that text-to-picture conversion systems will not significantly improve until the machine vision and language understanding methods are improved, as argued in [24].

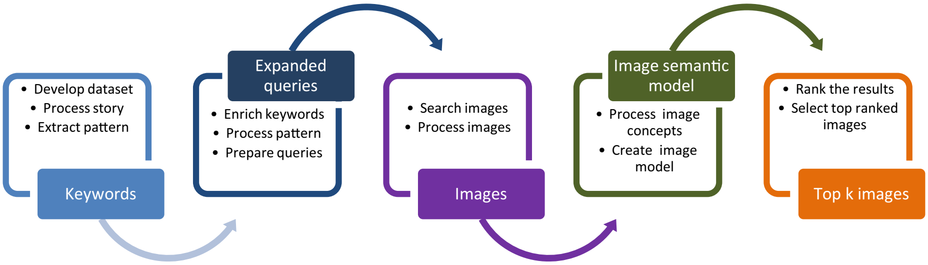


Fig. 11 Proposed approach for an automatic illustration of Arabic story text through multimedia

6 Conclusion and future work

We reviewed important prior works about text-to-picture systems and approaches. We also compared them according to input knowledge resolution, knowledge resources, knowledge extraction, image selection, and matching and output rendering. For the most part, existing systems favor linguistic-focused text processing. The objective of this review is to investigate the feasibility of "automatic visualization of Arabic story text through multimedia," which, in turn, involves text processing and analysis with the available tools. Moreover, this review allows us to refine our problem statement and to identify relevant topics for further research.

So far, we propose the following approach: First, processing the story to get a semantic representation of the main characters and events in each paragraph. Second, constructing expanded queries for each paragraph using the output of the previous step. Third, through an image search, finding a list of the top picture candidates. Exploring the results, a user or instructor can eventually refine the results of the automatic illustration step by selecting a few suitable pictures to compose the final visualization for each paragraph. Figure 11 illustrates these steps.

According to the survey, there were answers to some initial research questions and findings. However, the following questions have been newly identified for investigation:

1. What are the optimal method and tools of knowledge extraction specifically for Arabic text?
2. What are the challenges for processing Arabic text and how should unresolved issues be worked around? Do the input language and the conceptual domain both need restriction?
3. What are the standalone components from MADAMIRA[7] /Stanford[8] /Farasa[9] that improve performance at processing Arabic text?
4. Are there alternative approaches for knowledge extraction or combinations of existing approaches?
5. Can consideration be given to statistical-based (or corpus-based) techniques, linguistic-based (e.g., parsing) techniques, or deep learning techniques?
6. What are the options for sourcing 2D or 3D objects?
7. What are the algorithms for matching these objects?
8. How should limited semantic processing of object tags be dealt with?
9. What are the available resources for resolving semantic ambiguities?

[7] https://camel.abudhabi.nyu.edu/madamira/
[8] https://nlp.stanford.edu/projects/arabic.shtml
[9] http://qatsdemo.cloudapp.net/farasa/

10. What are the alternative (re-usable) corpora, ontologies, lexicons, or data from the Semantic Web, etc. for Arabic text?

The literature shows that text-to-picture systems perform relevant tasks that can improve language understanding and universal communication, which makes them an interesting and challenging interdisciplinary research problem. Indeed, according to [24], text-to-picture systems can be improved by exploiting better semantic processing, image processing, and associative learning techniques.

Acknowledgements This work was made possible by NPRP grant #10-0205-170346 from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors.

Funding Information Open Access funding provided by the Qatar National Library.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

1. Adorni G, Di Manzo M, Giunchiglia F (1984) Natural Language Driven Image Generation, in *Proceedings of the 10th International Conference on Computational Linguistics and 22Nd Annual Meeting on Association for Computational Linguistics*
2. Agrawal R, Gollapudi S, Kannan A, Kenthapadi K (2011) Enriching Textbooks with Images, in *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*
3. Bernardi R, Cakici R, Elliott D, Erdem A, Erdem E, Ikizler N, Keller F, Muscat A, Plank B (2016) Automatic Description Generation from Images: A Survey of Models, Datasets, and Evaluation Measures
4. Bobick A, Intille SS, Davis JW, Baird F, Pinhanez C, Campbell LW, Ivanov Y, Schtte A, Wilson AD (1999) "The KidsRoom: A Perceptually-Based Interactive and Immersive Story Environment," *Presence,* pp. 369–393
5. Boonpa SRS, Charoenporn T (2017) Relationship extraction from Thai children's tales for generating illustration, *2nd International Conference on Information Technology (INCIT)*
6. Carney RN, Levin JR (2002) "Pictorial Illustrations Still Improve Students' Learning from Text," in *J.R. Educational Psychology Review*
7. Chong W, Blei D, Li F (2009) "Simultaneous image classification and annotation," in *IEEE Conference on Computer Vision and Pattern Recognition*
8. Coelho F, Ribeiro C (2011) "Automatic Illustration with Cross-media Retrieval in Large-scale Collections," in *Content-Based Multimedia Indexing (CBMI)*
9. Coyne B, Sproat R (2001) Wordseye: an automatic text-to-scene conversion system," in *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*
10. Csomai A, Mihalcea R (2007) Linking educational materials to encyclopedic knowledge, in *Proceedings of the 2007 Conference on Artificial Intelligence in Education: Building Technology Rich Learning Contexts That Work*
11. Delgado D, Magalhães J, Correia N (2010) Automated illustration of news stories, in *Proceedings of the 2010 IEEE Fourth International Conference on Semantic Computing*
12. Dmitry U (2012) A Text-to-Picture System for Russian Language," in *Proceedings 6th Russian Young Scientist Conference for Information Retrieval*
13. Dmitry U, Alexandr K (2012) An Ontology-Based Approach to Text-to-Picture Synthesis Systems, in *Proceedings of the Second International Workshop on Concept Discovery in Unstructured Data*
14. S. Dupuy, A. Egges, V. Legendre and P. Nugues (2001) Generating a 3d simulation of a car accident from a written description in natural language: The carsim system, in *Workshop on Temporal and Spatial Information Processing*

15. Duy B, Carlos N, Bruce EB, Qing ZT (2012) Automated illustration of patients instructions, *Journal of the American Medical Informatics Association,* pp. 1158–1167
16. Elhoseiny M, Elgammal A (2015) Text to Multi-level MindMaps: A Novel Method for Hierarchical Visual Abstraction of Natural Text, in *Multimedia Tools and Applications*
17. Erhan OV, Alexander T, Samy B, Dumitru E (2016) Show and tell: lessons learned from the 2015 {MSCOCO} image captioning challenge. IEEE Trans Pattern Anal Mach Intell 39(4):652–663
18. Eunice MM (2006) *automatic conversion of natural language to 3D animation,* University of Ulster
19. Ganguly D, Calixto I, Jones G (2015) "Overview of the Automated Story Illustration Task at FIRE 2015," in *FIRE*
20. Goldberg A, Dyer CR, Eldawy M, Heng L (2008) "Easy As ABC?: Facilitating Pictorial Communication via Semantically Enhanced Layout," in *Proceedings of the Twelfth Conference on Computational Natural Language Learning*
21. Goldberg AB, Rosin J, Zhu X, Dyer CR (2009) "Toward text-to-picture synthesis," in *Proc. NIPS 2009 Symposium on Assistive Machine Learning for People with Disabilities*
22. Hanser E, Kevitt PM, Lunney T, Condell J (2009) Scenemaker: automatic visualisation of screenplays, in *Annual Conference on Artificial Intelligence*
23. Hanser E, Kevitt PM, Lunney T, Condell J (2010) NewsViz: Emotional Visualization of News Stories, in *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*
24. Hassani WLK (2016) Visualizing natural language descriptions: a survey. ACM Comput Surv 17:1–34
25. Huang CJ, Li CT, Shan MK (2013) "VizStory: Visualization of Digital Narrative for Fairy Tales," in *Proceedings - 2013 Conference on Technologies and Applications of Artificial Intelligence*
26. Huimei H, Ying L, Xingquan Z (2018) Convolutional neural network learning for generic data classification. Inf Sci 477:448–465
27. Ibrahim M, Waseem A, Rania E (2018) Bi-gram term collocations-based query expansion approach for improving Arabic information retrieval. Arab J Sci Eng 43:7705–7718
28. Jain P, Darbari H, Bhavsar VC (2014) "Vishit: A Visualizer for Hindi Text," in *Proceedings - 2014 4th International Conference on Communication Systems and Network Technologies*
29. Jiang Y, Liu J, Lu H (2014) Chat with illustration. Multimedia Systems 22:5–16
30. Joshi D, Wang JZ, Li J (2004) "The Story Picturing Engine: Finding Elite Images to Illustrate a Story Using Mutual Reinforcement," in *Proceedings of the 6th ACM SIGMM International Workshop on Multimedia Information Retrieval*
31. Joshi D, Wang JZ, Li J (2006) The Story Picturing Engine—a system for automatic text illustration," *ACM Trans. Multimedia Comput. Commun. Appl.,* pp. 68–89
32. Karkar AG, Alja'am JM and Mahmood A (2017) "Illustrate It! An Arabic Multimedia Text-to-Picture m-Learning System," in *IEEE Access*
33. Kastner M, Ide I, Kawanishi Y (2018) "Estimating the visual variety of concepts by referring to Web popularity," in *Multimed Tools Appl*
34. Larson C, Peterson B (1999) *Interactive Storytelling in a Multimodal Environment,* Aalborg University, Institute of Electronic Systems
35. Li J, Wang JZ, Wiederhold G (2000) "IRM: integrated region matching for image retrieval," in *Proceedings of the Eighth ACM International Conference on Multimedia*
36. Li H, Tang J, Li G, Chua TS (2008) "Word2Image: Towards Visual Interpretation of Words," in *MM'08 - Proceedings of the 2008 ACM International Conference on Multimedia, with co-located Symposium and Workshops*
37. Lin P, Huang Y, Chen C (2018) Exploring imaginative capability and learning motivation difference through picture E-book. IEEE Access 6:63416–63425
38. Ma ME (2002) CONFUCIUS: An intelligent multimedia storytelling interpretation, Technical report, School of Computing and Intelligent Systems, University of Ulster
39. Ma M, Kevitt PM (2005) "Visual Semantics and Ontology of Eventive Verbs," in *Natural Language Processing – IJCNLP 2004: First International Joint Conference*
40. Ma Y, Liu Y, Xie Q, Li L (2018) CNN-feature based automatic image annotation method. Multimed Tools Appl. https://doi.org/10.1007/s11042-018-6038-x
41. MacCartney B (2014) *Understanding Natural Language Understanding,* ACM SIGAI Bay Area Chapter Inaugural Meeting
42. Mihalcea R, Chee WL (2008) Toward communicating simple sentences using pictorial representations. Mach Transl 22:153–173
43. Mihalcea R, Csomai A (2007) "Wikify!: Linking documents to encyclopedic knowledge," in *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*

44. Mihalcea R, Tarau P (2004) "TextRank: Bringing Order into Texts," in *Proceedings of the 2004 conference on empirical methods in natural language processing*
45. Milne D (2010) *Applying Wikipedia to Interactive Information Retrieval,* University of Waikato
46. Mwiny M (2013) *Understanding Simple Stories through Domain-based Terms Extraction and Multimedia Elements,* Qatar University
47. Na L, Xia Y (2018) Affective image classification via semi-supervised learning from web images," *Springer, Multimedia Tools and Applications,* pp. 1–18
48. Nabil A, Noureddine E-n, Said AO, Mohammed M (2018) Using unsupervised deep learning for automatic summarization of Arabic documents. Arab J Sci Eng 43:7803–7815
49. Pasha A, Mohamed AB, Mona D, El KA, Ramy E, Nizar H, Manoj P, Owen R and Ryan R (2014) "MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation*
50. Poots JK, Bagheri E (2017) Overview of text annotation with pictures. IT Professional 20(01):36–44
51. Poots JK, Cercone N (2017) First steps in the investigation of automated text annotation with pictures. Big Data and Information Analytics 2:97–106
52. Poots JK, Cercone N (2017) First steps in the investigation of automated text annotation with pictures
53. Ramisa A, Yan F, Moreno-Noguer F, Mikolajczyk K (2016) BreakingNews: Article Annotation by Image and Text Processing, *ArXiv e-prints*
54. Ruan W, Appasani N, Kim K, Vincelli J, Kim H, Lee W (2018) Pictorial Visualization of EMR Summary Interface and Medical Information Extraction of Clinical Notes, in *IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA)*
55. Sampath H, Sivaswamy J, Indurkhya B (2010) "Assistive systems for children with dyslexia and autism," *SIGACCESS Access. Comput.,* pp. 32–36
56. Sumi K, Nagata M (2006) Animated Storytelling System via Text, in *Proceedings of the 2006 ACM SIGCHI International Conference on Advances in Computer Entertainment Technology*
57. Talsania A, Modha S, Joshi H, Ganatra A (2015) Automated Story Illustrator, in *Post Proceedings of the Workshops at the 7th Forum for Information Retrieval Evaluation*,
58. Yamada A, Yamamoto T, Ikeda H, Nishida T, Doshita S (1992) Reconstructing Spatial Image from Natural Language Texts," in *Proceedings of the 14th Conference on Computational Linguistics*
59. Zhiyu W, Peng C, Lexing X, Wenwu Z, Yong R, Shiqiang Y (2014) Bilateral correspondence model for words-and-pictures Association in Multimedia-Rich Microblogs. *ACM Trans Multimedia Comput Commun Appl* 34:1–21
60. Zhu X, Goldberg AB, Eldawy M, Dyer CR, Strock B (2007) "A text-to-picture synthesis system for augmenting communication," in *Proceedings of the 22nd national conference on Artificial intelligence - Volume 2*

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Jezia Zakraoui received the master’s and Ph.D. degree in computer science from Vienna University of Technology, Austria, in 2012. She is currently working with the Department of Computer Science and Engineering, Qatar University. Her research interests include semantic Web, text mining, artificial intelligence, deep learning and machine learning.

Moutaz Saleh received his Ph.D. in computer Science from University Kebangsaan Malaysia, M.Sc. in Distributed Computing from University Putra Malaysia, and B.Sc. in Computer Engineering from Yarmouk University in Jordan. He worked as a Cisco certified network engineer in Jeraisy Computer and Communications Services (JCCS) in Riyadh/KSA and MobileCom Telecommunications in Amman/Jordan from 2001 to 2003. Dr. Moutaz is currently a Lecturer in the department of computer science and engineering, college of engineering at Qatar University and a Cisco Regional Academy Instructor. He is also a Cisco Certified Internetwork Expert (CCIE) and a Cisco Certified Instructor Academy (CCAI) serving at Cisco Regional Academy in Qatar. His research interests include assistive technology, learning systems, networking, distributed and real-time systems. Dr. Moutaz has authored or co-authored over 50 refereed journal and conference papers in reputed international journals and conferences. He has served as a technical program committee for Elsevier Computer Communications, EURASIP Wireless Communications and Networking, and IEEE (LCN, ICT, AICCSA, APCC, GCC, ICC, EDUCON, and TALE).

Jihad Mohamed Alja'am received the Ph.D. degree, MS. degree and BSc degree in computing from Southern University (The National Council for Scientific Research, CNRS), France. He worked on the connection machine CM5 with 65,000 microprocessors in the USA. He was with IBM-Paris as Project Manager and with RTS-France as IT Consultant for several years. He is currently with the Department of Computer Science and Engineering at Qatar University as full professor. He organized many workshops and conferences in France, USA and the GCC countries. His current research interests include multimedia, assistive technology, learning systems, human–computer interaction, stochastic algorithms, artificial intelligence, information retrieval, and natural language processing. Dr. Alja'am is a member of the editorial boards of the *Journal of Soft Computing*, *American Journal of Applied Sciences*, *Journal of Computing and Information Sciences*, *Journal of Computing and Information Technology*, and *Journal of Emerging Technologies in Web Intelligence*. He acted as a scientific committee member of different international conferences (ACIT, SETIT, ICTTA, ACTEA, ICLAN, ICCCE, MESM, ICENCO, GMAG, CGIV, ICICS, and ICOST). He is a regular reviewer for the ACM computing review and the journal of supercomputing, IEEE ACCESS (Associate Editor). He has collaborated with different researchers in Canada, France, Malaysia, GCC and USA. He published so far 159 papers, 8 books chapters in computing and information technology which are published in conference proceedings, scientific books, and international journals. Prof. ALJA'AM is the main organizer and general chair of the international conference on computer and applications. He is leading a research team in multimedia and assistive technology and collaborating in the Financial Watch and Intelligent Document Management System for Automatic Writer Identification and MOALEM projects. Prof. ALJA'AM received the 2015 ACM Transactions on Multimedia Computing, Communications and Applications (TOMM) Nicolas D. Georganas Best Paper Award. And the best research paper of the 10th annual international conference on computer games multimedia & allied technologies (Singapore, 2016).