Building multimedia repository for composing images perspective

Samir Elloumi¹ · Jihad Mohamad AlJa'am² · Jezia Zakraoui²

© The Author(s) 2019

Abstract

Multimedia repository is helpful for educational activities since it offers several illustrations that facilitate the learning process and the text understanding. In this paper, we propose to build a multimedia repository from collected images using object extraction techniques. Then, we associate Arabic captions to all extracted objects. These extracted objects are used to compose new scenes that could illustrate efficiently the most important events in an Arabic story. We extend, thereby the notion of image composing to our approach as the task of constructing new images based on a set of tool-boxes and a set of extracted objects as our multimedia repository of common animal behaviors. Our preliminary results show that the composed scenes using single objects provided a fair understanding of the main events of the stories as well as a coherent visual layout of all single objects. The diversity and the precision of the single objects images for the domain of animals have shown a great impact on composing new scenes either manually or dynamically.

Keywords Multimedia systems · Extracted object/single object · Text-to-picture · Images composing · Learning technologies

1 Introduction

Multimedia systems have great impact on education, learning and training systems [1, 2]. Different research works in education support the assertion that multimedia can broadly promote the performance of learners. A user can learn more efficiently from words associated with pictures, videos or audios than words only. To date, several multimedia systems have been proposed to provide illustrations for stories from an existing multimedia repository (MR). They tried to query the MR in order to retrieve the most convenient images w.r.t. search keywords. In the case of Arabic keywords, a machine translation step is required in order to be able to use existing available resources. Besides the difficulties related to the translation efficacy, it is not always obvious to have a semantically convenient matching between images' captions and the translated keywords. Many multimedia systems proposed to visually explain topics, news stream or stories by annotating

texts with pictures [3, 4], enriching textual content [5] or composing scenes [6]. There are some common limitations in the existing systems, which have not been properly addressed:

1. Existing multimedia systems can retrieve pictures automatically from the image search engine and generate illustrations [7–10]. However, manual work is required to filter out inappropriate pictures, which reveals the excessive manual efforts behind these systems as indicated in [6]. We tackle this problem with a filtering process automatically by using a deep learning captioning model to complete the missing caption if not available, for instance Google search engine (referring to JSON response object) lacks appropriate captions or at least meaningful tags about the images. If the caption for the returned images has no matching with any of the initial keywords, so they are filtered out.

✉ Jezia Zakraoui, jezia.zakraoui@qu.edu.qa; Samir Elloumi, samir.elloumi@fst.utm.tn; Jihad Mohamad AlJa'am, jaam@qu.edu.qa | ¹Faculty of Sciences of Tunis, LR11ES14, University of Tunis El Manar, Tunis, Tunisia. ²Computer Science and Engineering Department, Qatar University, 2713 Doha, Qatar.

2. Multimedia systems for illustrating Arabic text are very limited, which reflects the current technical difficulties in understanding Arabic text since for Arabic language there exists additional challenges; the first one to tackle is multimedia search using Arabic keywords which we resolve by translation the keywords into English. Unlike Arabic, for English language there are enough high-quality multimedia resources associated with well maintained captions, metadata and tags, freely usable. So, we propose to use Machine Translation (MT) to translate Arabic text to English text to overcome the problem with the lack of image resources annotated in Arabic.
3. Many pictures are available on the web, but they lack textual descriptions or captions to be included in a robust image search, so we complete the missing information by the use of a convolutional neural network (CNN) as a pre-trained model to automate the captioning process for all images as well as for extracted single object images to be included in our approach.
4. Generated images based on generative models such as Generative Adversarial Networks (GANs) [11] could be used; however, the generated images are not yet realistic, i.e. mistakable for real as argued in [12]. Besides, these models fail to generate the correct number of instances, i.e. objects with clear boundaries and spatial relationships when the input contains multiple objects, as mentioned in [13]. In some cases, the generated images have poor image quality in terms of color consistency as argued in [14] which could be improved using a regularization technique. In our case, we use rather single object images obtained through object extraction where we could resolve the problems with regard to the number of object instances and scene layout issues such as arrangement of objects in the image plane.
5. Most of existing multimedia systems can illustrate text based on retrieved images. However, none of them, to our best knowledge, considers extracting objects from retrieved images and uses them for composing new pictures from scratch, which is the key requirement in our approach to promote story understanding for Arabic text.

We believe that MR is a good option to compose a realistic and coherent illustration for a story text since, getting a well illustrative image w.r.t. a new story is not obvious even when using a well-maintained image repository. In fact, it supposes that in the MR there exists relevant illustrations w.r.t. story and the objective is to extract it. However, this hypothesis is not necessarily true for all stories. Therefore, for several stories, we could not find in MR any

relevant images describing them. The proposed approach focuses on the task of generating a coherent and realistic illustration for Arabic story text and not only on the task of image retrieval. Therefore, we propose to map text to multimedia by generating new images based on single objects that should be initially prepared in local repository. The generation process depends on the keywords, their associated images, and their drawing properties. In this paper, we are focusing more on building single objects as Multimedia Repository (MR) that contains several single object images with their associated captions.

The remainder of the paper is organized as follows. We first provide a literature review in Sect. 2 and then elaborate on our proposed method in Sect. 3. We present experimental results in Sect. 4 and evaluation in Sect. 5. Finally, we conclude this paper and discuss future directions in Sect. 6.

2 Literature review

The idea of creating a relationship between text and images dates back a long time. Many researches in computer science but also in other fields such as linguistics have been carried out to create such a relationship in a formal way. We review briefly well known annotated image collections, relevant approaches and multimedia systems that have been proposed to date using such kind of multimedia collections. There is a wide range of image collections; ImageNet is an annotated image dataset with WordNet noun synsets [15]. ImageCLEF is an evaluation forum for the cross-language annotation and retrieval of images [16], which also involves diverse related tools and resources. Worth mentioning image collections are MSCOCO [17], Flickr8 k [18] and its extended version Flickr30 k [19] which represent benchmark standards for image tasks, a survey of more image datasets can be found in [20].

Through the recent years, a lot of well known general multimedia datasets and other domain specific datasets have been built to support multimedia systems. In particular, image search engines such as Google image search has been employed as well such as in the work of Aramini et al. [7]. The Authors proposed different methods to extract keywords from text such as extraction of information attached to the retrieved images from the web. Other proprietary image databases have been also built such as in story picturing engine which has been proposed by Joshi et al. [21]. This system uses two large image databases, namely Terra Galleria [22] and The Art Museum Image Consortium [23], which contain a lot of high-quality photos with annotations. It consists of a pipeline of three processes: story processing and image selection, estimation

of similarity, and reinforcement-based ranking of the related images. Zhu et al. [24] proposed a Text-To-Picture system that generates pictures for unrestricted English text and ranks retrieved images using machine-learning techniques. This system was enhanced by Goldberg et al. [9] to use semantic role labeling rather than keyword extraction with picturability. Rada et al. [25] used PicNet as an illustrated dictionary for the automatic generation of pictorial representations of simple sentences that uses WordNet as a lexical resource for the automatic translation of an input text into pictures. Agrawal et al. [5] proposed techniques for retrieving images from the web to enrich textbooks with images while following predefined restrictions. Authors used a corpus of high school textbooks used in India. The evaluation showed 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 since important categories of terms are ignored. An ontology-based approach is proposed by Ustalov et al. [26]. It has been designed to operate with ontology to allow loose coupling of the system's components and to unify the interacting actors' representation and behavior while making the verification of system information resources possible.

Worth mentioning systems that created their own image datasets are [27, 28]. The former was developed by Duy et al. [27] and it creates pictures to illustrate instructions for patients. It has a pipeline of five processing steps: pre-processing, medication annotation, post-processing, image construction, and image rendering. The latter is a medical record summary system and was developed by Ruan et al. [28]. The patient's medical data is visualized spatially and temporarily based on the categorization of event categories and 6 physiological systems. An interesting approach in expressing emotions is proposed by Eva et al. [29]. Other approaches in the domain of news streaming have been proposed [3, 8]. The latter used BreakingNews, a novel dataset while the former used Flickr images. Aletras et al. [30] proposed an approach to provide images that are useful to represent general topics. The approach retrieves a collection of images for topics using online search engine. The selection of images is based on graph-based methods notably on visual information and text. Huang et al. [31] proposed a system for transforming text to visual form of fairy pictures. This system selects keywords from segmented stories, then, retrieves relevant pictures from online repositories based on their tags, and finally, the pictures are composed for illustrating the main ideas of the input. Jiang et al. [32] proposed a novel assisted instant messaging program which shows illustrations associated with textual messages. The representation of the picture is constructed from a set of most representative retrieved images. Many other, multimedia

systems have also been proposed in other works such as [33–35]. A recent multimedia system for Arabic stories is proposed in [36]. To build a multimedia repository, the system uses the Scribd [37] online book library in order to collect educational pictures which are then stored locally in binary format and marked for text extraction. The best picture is selected based on the maximum intersection between the conceptual graph of the best selected keywords and sentences from the input text and the conceptual graphs of the pictures. The system uses only cartoon images and disregards essential educational benefits from other multimedia types. And comparatively other research worked on object classification and recognition in images, in particular face recognition are investigated recently in [38–40]. These works propose new face registration algorithm and extends well known face recognition approaches based on face classification, identification and verification techniques. We are aware of these related papers. Different to them, and due to diversity and non-homogenous object layout of natural images in our MR, it is difficult to analyze our retrieved images by using these low-level visual features used in these works. In fact, various approaches in the literature tried to overcome these limitation using other techniques such as the Bag-of-Visual-Word representation combined with other techniques [41] which showed better object recognition performance, however as reported in [42], the lack of spatial information, the alignments and the overlapping of objects in natural images cannot be supported by these techniques. As a result, it is hard to use these techniques mentioned in [42] in our case as these details are the key elements in defining single objects, their relationship and their spatial arrangements in a scene. However, it is worth mentioning here that the approaches based on deep neural network (DNN) for object segmentation [43], object detection and classification are more practical to tackle the problem of any object extraction from images.

Recently, GANs have shown great potential in generating images and videos. Despite the success, GANs are hard to train for example the training process is usually unstable and sensitive to the choices of hyper-parameters as analyzed in this blog [44]. Given these difficulties and others in modeling details of natural images, the authors in [14] proposed Stacked Generative Adversarial Networks (StackGANs) and succeeded in generating images with photo-realistic details from text descriptions. However, when the input contains multiple objects, StackGANs fails to generate the correct number of instances with clear boundaries and spatial relationships. Moreover, for complex interactions between the main object and its surroundings, StackGANs fails to capture and express the interactions between the objects. The authors in this report [13] reported relevant examples, where the

pretrained StackGANs models fail either to generate the objects from the correct categories or to express relations between objects which could be realized by paying more attentions to the relevant words in the given text. To address this issue, Xu et al. [45] proposed an Attentional Generative Adversarial Network (AttnGAN) that enables StackGANs to generate fine-grained high-quality images via word level and sentence level conditioning.

While divers multimedia systems for English, among other resource-rich languages, has developed annotated image collections that are widely used for approaching the relationship between language and vision and especially object detection tasks, multimedia systems that can automatically provide a complete illustration for Arabic stories using single objects are missed. We target at building a MR that contains several single objects with their associated captions in the domain of animals. This repository will establish a well-annotated image repository for composing new images.

Indeed, in this paper, we present an offline multimedia repository that takes advantage of, and in fact is built from image resources on the Web, namely on ImageNet as well as on Google image search. The design of MR is flexible and extensible such that we can easily incorporate new kinds of multimedia such as video and other languages. The effectiveness of MR is demonstrated and a potential application for understanding Arabic stories by composing

images is intended in this paper. This work is an endeavor towards quantifying progress on building MR using web image retrieval.

3 The proposed system

In this section, we present the main components of the proposed system as depicted in Fig. 1. First, we detail the construction of the multimedia repository in which we collect a set of single objects. Second, we present the images composing steps.

3.1 Multimedia repository (Part 1)

Multimedia systems can help both instructors and students in pedagogical activities including, learning, imagination, and communication. Based on existing online libraries such as Google Images, ImageNet, etc., we collect a set of images and index them in a temporary database. For each image, we apply an object extraction tool in order to obtain single objects with some of their drawing properties. Subsequently, we apply an image captioning process on extracted objects/single objects to automatically assign a caption for each of them.

Fig. 1 Architecture of the proposed system. It has two main parts (1) multimedia repository, (2) images composing

3.1.1 Images retrieval

For the initial version of our MR, we suggest to select the domain of animals and to experiment several alternatives (we will compare the use of Google image search API and imageNet). Note, we consider for the current version 80 keywords to retrieve images of animals and some of their related basic behavior. Table 1 shows some sample keywords used in this work as well as their related actions and retrieved images. Note we translated the input keywords from Arabic to English manually.

3.1.2 Object extraction using Mask RCNN

Object extraction aims at identifying the foreground regions based on methods like edge detection, segmentation, and background subtraction. With the advancements of research in various other fields of science and technology over the time, Object extraction has found their application in many different areas i.e. Medical disease detection [46, 47], satellite object extraction, image retrieval, objects recognition etc.

We target at extracting foreground images which constitute key elements in defining single objects. We found the model Mask R-CNN [43] outperforming other segmentation models as a deep neural network (DNN)

for object segmentation [43], object detection and classification. The model identifies the potential regions that containing the foreground objects and extracts these objects from images without the background. Hence, the images could be reused in composing new images.

Here are the steps we followed to process the segmentation of retrieved images:

1. Collecting images from Google image search and ImageNet;
2. Executing the Mask R-CNN pre-trained model which is available on [48] to localize objects in the image with bounding boxes and extract those localized objects;
3. During testing, we extended the deep learning model to process unseen images in bulk, enabling the model to store results in the corresponding directory;
4. For each test image, we extracted all segmented objects and stored in lossy compression format (JPG);
5. To effectively utilize the segmented objects we incorporate an alpha channel to retain the transparency information on the pixel-level.
6. The transparency information is then used to adjust the focus on the single objects' part of the image by cropping the alpha channel to precisely use in the later step of image generation.

Table 1 Keywords data set

Arabic keywords	Translated keywords	Action	Retrieved images
النمر يركض	Tiger running	Run
الجمال يجلس	Camel sitting	Sit
القط يقفز	Cat jumping	Jump
الدب يأكل	Bear eating	Eat

7. Creating a transparent background for all extracted objects and saving them as.png format images

Table 2 shows a small subset of the retrieved images for given keywords including their extracted single objects from each image source. The last column results from the combination of all retained single objects after a manual evaluation. The first row of this table shows a good example of object extraction since all single objects are complete and subject for reuse in composing new images. Whereas the second row demonstrates that most of the single objects are not complete and therefore not used to compose new images, only one single object was added to MR.

Although, the effectiveness of Mask R-CNN for extracting objects, however, it fails in some cases when many objects are overlapping. Thus, the number of retained single objects related to retrieved images is low as shown in Table 4.

As a first contribution, we propose to build a local single objects repository. As a second contribution, we propose to use single objects to generate new images.

3.1.3 Captioning single objects

As the prepared single objects' image set is ready for going through the image captioning via CNN which in turn represent an image by embedding it into a fixed-length vector. Then, Recurrent Neural Network RNN especially LSTM [49] decodes the fixed-length vector to a desired output sentence by iterating a recurrence relation [50]. We used a pre-trained model as a fine-tuned checkpoint which has been trained over 3 million iterations using MSCOCO dataset [17]. All extracted single objects are stored locally as image type. All information related to them is stored

in relational database together with associated keywords and their mappings. Image captions that have been generated using a CNN pre-trained model [51] are also stored. Obtained captions are available in English only and been MT processed into Arabic.

Table 3 shows some generated captions for retrieved images as well as those respectively extracted objects. It is worth mentioning that in some cases the accuracy of captioning process of the single objects outperforms the accuracy of the captioning of the original retrieved images as shown in the first row of Table 3. Thus, a minimal enhancement in captioning step is observed, which in turn underlines the importance of such objects extraction tools. However, for other images the accuracy did not change as shown by the second row of Table 3. The third row of the same table shows that the accuracy decreases since the generated caption after the object extraction step is not accurate. In summary, we cannot admit from this preliminary work that extracting objects would lead to better accuracy for automatic image captioning. However, inadequate image captions are results from the weakness of the used image captioning pretrained model which showed a moderate accuracy for English and a lower accuracy for machine translated Arabic as analyzed in [52]. To resolve it, we plan to train the model using a well-annotated dataset in Arabic.

3.2 Images composing (part 2)

In our Image Story Generator tool, image composing is designed in two ways, manual image composing and automatic image composing. In our current tool version, we compose new images or pictures manually, allowing thereby flexible working with the tool. We briefly describe how a user composes new images using our

Table 2 A subset of retrieved images from two image sources, and the resulting multimedia repository MR

Sample Keyphrases	Translated keyphrases	Google image search API		ImageNet		Multimedia repository MR
		Original retrieved images	Single objects	Original retrieved images	Single objects	
يركض النمر	Tiger running
الجمال ينحني	Camel crouching

Table 3 Captioning retrieved images and single objects

Translated keyphrases	Original retrieved images	Generated captions for original retrieved images	Extracted single objects	Generated captions for extracted single objects	Caption accuracy Evaluation
Cat jumping		A dog jumping in the air to catch a frisbee		A cat that is looking up at a bird	Better accuracy
Tiger running		A close up of a zebra in a field with trees in the background		A cat laying on top of a pile of rocks in the background	Same accuracy
Elephant walking		An elephant is standing in a grassy field		A large elephant standing next to a baby elephant	Low accuracy

tool. First, a user input keywords in an input field on the top of the tool main interface and hits enter. Note we use single keyword only at this current version. The retrieved single objects from MR are displayed in a panel on the Graphical Toolbox on the left side of the main interface as indicated in Fig. 2. The user or the teacher can drag and drop the main interface, locate images and resize images thereby composing a new scene describing the input sentence. The teacher can successively search for other single objects images doing same steps as described. Thus, the teacher can show the final

illustration to the students. The newly created picture can be stored locally. After the image has been stored the system extracts images sizes, images positions, etc. and save this information as drawing properties. This information will be used further for generating new pictures/images dynamically.

Note the option to change the pictures at any stage is also provided to guarantee a flexible learning environment. Figure 2 shows a screen shot of our tool where two single objects are added attempting to describe the given input keywords.

Fig. 2 Manual composing images output for the sentence "Tiger is running behind a deer"

4 Experimental settings

We conducted experiments on two image sources: Google image search and ImageNet. On both image sources, we gave as input same keywords. Note in ImageNet only single words are allowed and each single keyword is associated with many images. Table 4 summarizes the statistics of the two datasets and the combination of both into a unified MR.

In the experimental set up we start by following settings.

- *Input text data set* we prepared 80 keywords i.e. key-phrases extracted from short simple stories in the domain of animals, a subset is shown in Table 5.
- *Database* has been set to store retrieved images and keywords and their mappings. Image captions have been generated using a CNN pre-trained model [51]. Obtained captions are available in English only and being MT processed in Arabic. These data are also stored in the database.
- *Image search* we used Google image search [53] and ImageNet [15] to search images related to input text dataset. It is worth mentioning that we use these images in our prototype for illustrative purpose only. In the future, it is planned to set up a dataset of appropriate educational images for better learning and understanding.
- *Object extraction model*, we use *Mask R-CNN* which is a pre-trained model for tensorflow published on github [48].
- *Deep learning model* we use *im2txt*, which is a pre-trained model for tensorflow published on github [51]. It is a model developed by Google that takes an image as input and creates a caption for it.

After the setting step, we have performed following main tasks to build MR:

1. Extracting keywords and key-phrases using Arabic text segmentation, tokenization and part-of-speech tagging;
2. Translating extracted keywords manually;

3. Retrieving images and creating database to store the mapping of keywords to images to serve as preliminary image pool;
4. Extracting objects from all retrieved images using Mask R-CNN pre-trained model;
5. Retaining correct single objects manually;
6. Captioning single objects using a Convolutional Neural Network (CNN) pre-trained model;
7. Saving single objects locally and all associated information in our relational database.
8. Integrating and loading single objects in the tool for manual composing images

5 Evaluation

The objective of this work is to build an MR for Arabic text. We target at enhancing our current proposed system version that is still under development and improvement. We are currently processing sentences with simple structure where keywords are extracted, translated and then searched in two different image sources. We show the statistics of our MR and then we conduct a comparative evaluation as well as a user evaluation.

As Table 4 shows, MR contains 770 single objects which are saved physically in folders indexed by their keywords as well as in relational database for further processing steps. Many of the extracted objects were not reusable due to sliced object parts during the extraction task and thus eliminated. Subsequently, in the evaluation section, we will make a comparison between the outputs from Google image search API and imageNet.

Table 5 lists all keywords and their counts in MR. As we see for each keyword we retained only those single object images that are most representative and having a coherent visual layout.

5.1 Comparative evaluation

We show some of the image results in Table 6 where the retrieved images from Google image search API and we compared them to the composed ones created using our tool and MR for the corresponding sample sentences. We compare the output images from Google image search API

Table 4 Statistics of our experimented data sets

Image source	#Keywords	#Retrieved images	#Extracted objects	#Retained extracted objects
Google image search API	80	320	1490	535
ImageNet	80	96	400	235
MR	80	416	1890	770

Table 5 Statistics of our single objects and keywords

Keyword Id	Keywords	#single object images
1	Tiger, running	12
2	Tiger, jumping	10
3	Tiger, walking	11
4	Tiger, eating	6
5	Tiger, drinking	12
6	Deer, running	19
7	Deer, walking	13
8	Deer, sleeping	5
9	Deer, drinking	6
10	Deer, eating	8
11	Zebra, drinking	12
12	Zebra, standing	12
13	Zebra, running	10
14	Zebra, walking	8
15	Zebra, sleeping	10
16	Bear, walking	1
17	Bear, eating	12
18	Bear, standing	14
19	Bear, drinking	3
20	Bear, swimming	16
22	Camel, walking	15
23	Camel, sitting	15
24	Camel, drinking	8
25	Camel, running	12
26	Camel, crouching	4
27	Rabbit, eating	14
28	Rabbit, running	15
29	Rabbit, sleeping	9
30	Rabbit, standing	11
31	Rabbit, graving	5
32	Turtle, eating	4
33	Turtle, walking	5
34	Turtle, drinking	4
35	Turtle, graving	4
36	Turtle, diving	5
37	Monkey, jumping	20
38	Monkey, eating	16
39	Monkey, sleeping	16
40	Monkey, drinking	6
41	Monkey, running	10
42	Bull, drinking	12
43	Bull, running	14
44	Bull, eating	9
45	Bull, walking	14
46	Bull, struggling	7
47	Crocodile, hunting	8
48	Crocodile, swimming	8
49	Crocodile, catching	6

Table 5 (continued)

Keyword Id	Keywords	#single object images
50	Crocodile, diving	8
51	Crocodile, devouring	7
52	Elephants, Shattering	8
53	Elephants, Walking	16
54	Elephants, eating	20
55	Elephants, running	16
56	Elephants, drinking	17
57	Fox, falling	4
58	Fox, catching	14
59	Fox, eating	7
60	Fox, running	12
61	Fox, walking	14
62	Horse, running	23
63	Horse, eating	7
64	Horse, jumping	18
65	Horse, walking	14
66	Crow, standing	24
67	Crow, eating	6
68	Crow, flying	14
69	Crow, drinking	4
70	Crow, walking	9
71	Cat, jumping	14
72	Cat, drinking	23
73	Cat, hunting	18
74	Cat, running	16
75	Cat, walking	15
76	Butterfly, rambling	9
77	Butterfly, flying	12
78	Butterfly, eating	8
79	Butterfly, drinking	5
80	Butterfly, sleeping	6

and our newly built MR, both in respect of human evaluation. We evaluate our approach with regard to (1) using the combination of different web image sources and to (2) using single object images for composing new scenes or pictures to promote understanding of Arabic text through images.

First, it is worth mentioning that image results from Google image search API are more significant in number and relevance, while the image results from ImageNet are limited, however better clustered. Nevertheless, the combination of both image sources enriched our repository in many aspects. Second, regarding the image composing, we argue that using single objects to compose new scene to represent textual input has shown an enhancement compared to retrieved images from Google image search API, since in some cases we could represent with

Table 6 A comparison between retrieved images from Google image search and composed images thorough MR

Story ID	Sentences	Retrieved images from API Google	Composed images
1	يركض النمر وراء قطع من الغزلان The tiger is running behind a herd of deer
2	كان الدبان الجائعان يجولان بالغابة The two hungry bears were wandering through the forest
3	صارع التمساح الحمير الوحشية The crocodile grapples with zebras
4	الأحمر و الثيران و النمر يشربون الماء من البحيرة Zebras, bulls and tigers drink water from the lake
5	سارت الجمال ثم جثت على الرمل The camels walked and crouched on the sand
6	حطمت القيلة بيت الأرانب Elephants smashed the rabbits house

Table 6 (continued)

7	تركض الحصين و الحمير الوحشية وراء الفيلة في الملعب The horses and the zebras run behind the elephants in the stadium
8	مجموعة من القطط تقفز وراء فراشات صغيرة A group of cats jump behind small butterflies
9	تأكل الأرانب و السلحفاة الخس The rabbits and the turtles eat lettuce
10	يجري الثعلب وراء أرنب صغير The fox is running behind the little rabbit
11	تركض الحصين و الثيران في الملعب The horses and the bulls are running in the playground
12	الثيران تأكل أوراق الأشجار The bulls eat leaves of trees

Table 6 (continued)

13	أكل النمران و الأسد اللحم Two tigers and one lion eat meat
14	تجري الكلاب والنعالب وراء الغزلان The dogs and the foxes run behind the deer
15	السلحفاة تغطس في البحيرة The turtles dive in the lake
16	وقف الدبان بشجرات بها خلايا نحل The two bears stand on two trees with beehives
17	وقفت الغربان على شجرة عالية The crows are standing on a high tree
18	قفزت القردة فوق أرانب نائمة The monkey jumped over sleeping rabbits

Table 6 (continued)

19	أكلت الجمال العشب The camels eat grass
20	طارت الغربان مع الفراشات فوق الشجرة The crows flew with the butterflies over the tree

MR better illustrations than we retrieved by Google as shown in Table 6. Moreover, we are also able to customize the actors of a story, to add actors, to locate them and to resize them, etc., to compose a coherent story through pictures, for instance, row#1 to row#9 show that adding more actors promote the understanding of the input sentence rather than using the images retrieved image from Google where the number of actors is not represented at all. Row#5 shows that different positions and orientations for actors is illustrated all in one scene. All in all, we can represent spatial information through localizing object images in different positions and in different sizes which promotes understanding of the story. Row#8 shows an example of object overlapping where different body parts of cats are overlapped, but the boundaries still clear. Besides, we can draw different actions of actors and their relationships to each other in one scene only. Thus, this tool saved us time and images resources.

In summary, and with regard to the mentioned limitations in exiting systems in Sect. 1, our approach saved us (1) time and manual effort in filtering retrieved images and in completing missing captions. Besides, our approach included (2) enough high-quality multimedia resources associated with well maintained captions in English and Arabic. (3) A robust image search is set, so that automated captioning process is performed for all images and for all extracted single object images. (4) The use of single object images obtained through object extraction resolved many issues such as the problems with regard to number of instances and scene layout issues such as arrangement

of objects in the image plane. Many remarks and findings have been concluded; most of them contributed to extracting single objects from retrieved images and using them in our Image Story Generator tool for composing new scene, which is the key requirement in our approach to promote story understanding for Arabic text. It is worth mentioning, that our Image Story Generator tool using our MR outperforms Google image retrieval in representing spatial information, plural words and different categories of objects without any excessive preprocessing compared to the case of StackGANs where 80% of the birds' images which have a small object-image ratio have been processed. In our case, we believe that object-image ratio is a necessary detail in natural images which should be kept in order to support understanding of realistic scenes and thereby realistic stories.

To support the details of natural images, we target at enriching our MR to cover a wide range of single objects and diverse actions. However, we still face some limitations and problems in composing new images such as the quality of extracted single objects, choosing background, etc. Nevertheless, our work strongly argues that illustrations for text through image retrieval only will not produce a meaningful illustration to help understanding with pictures unless we propose a better solution that helps conveniently namely through manual composing new images and in future work through automatic composing and automatic generating images.

We list below major limitations in our current version that should be improved:

Table 7 User satisfaction values for composed scenes/images and image retrieval

Story Id	User satisfaction with image retrieval				User satisfaction with our image composting tool			
	Composed main actors	Composed main events	Composed spatial and temporal information	Overall	Composed main actors	Composed main events	Composed spatial and temporal information	Overall
1	4.5	3.2	2.2	3.3	5	3.7	3.2	4
2	4	4	4	4	5	5	5	5
3	4.5	4.5	4.5	4	4	4	3	3.6
4	2.2	3.7	4	3.3	5	5	5	5
5	3.7	2.5	4.2	3.5	5	5	5	5
6	5	1	0	2	4.5	2.2	2.2	3
7	4	4	2	3.3	5	5	5	5
8	3	4	4	3.6	5	4	5	4.7
9	3.5	2.2	2.7	3	5	3	4	4
10	5	5	5	5	5	5	5	5
11	3.2	3.7	4	3.7	5	5	5	4.8
12	3	3.7	3.7	3.5	5	5	4	4.8
13	5	5	5	5	5	5	5	5
14	3	2	2	2.3	5	5	5	5
15	5	5	5	5	5	5	5	5
16	5	5	5	5	5	5	5	4.8
17	4.2	4.7	3.2	4.1	5	5	4.5	5
18	3	5	2.5	3.5	4	4	4	4
19	5	4.5	5	4.8	5	5	5	5
20	5	5	4.5	4.8	4	4	4	4
Avg	4.0	3.9	3.6	3.8	4.8	4.5	4.4	4.5

- Search result, to improve search results from Google search engine or ImageNet, the keyword query should be expanded with synonyms, since some results may be retrieved using synonym;
- Keyword extraction, we should revise the techniques for keywords extraction by considering the advanced aspects of NLP and text understanding;
- Improving the Arabic text parsing in purpose to handle texts of more than one sentence;
- Enhancing the linguistic Arabic text analysis to handle more parts of speech such as adjectives;
- A filtering algorithm must be developed to remove inappropriate images;
- Improving the quality of extracted single objects such as sharpen the edges, removal of object collision, add different backgrounds.

5.2 User evaluation

We evaluate our tool based user satisfaction of the manually composed images and if they accurately represent the main events of the given stories/keywords. The relevance and the suitability of each created image are obtained through two users. They are provided with 20 stories together with the created images/scenes. They were asked

to judge if the created images were a representation of the main actors and events, and to provide a rating on a scale of 0 (completely unsuitable) to 5 (very suitable). The results are shown in Table 7; the overall user satisfaction with the output is equal to 4.5 for our tool compared to 3.8 for image retrieval. This clearly demonstrates that, users were satisfied at a certain level with the composed images than with the retrieved images.

6 Conclusion

A single objects multimedia repository is presented in this paper. The proposed MR solves the limitations of existing multimedia systems by providing single objects illustrations, considering common concepts including actions and relations in the domain of animals. The current version covers a narrow range of single objects depicting the common behavior of animals like food activities and motion. There are still many single objects missing. However, we are continually collecting more multimedia content to enrich our multimedia repository. Through our experimentations and evaluations, we found out that our work strongly argues that illustrations through image retrieval only, still suffer from many problems and that our tool in

which, we identify the key components for abstracting the information of multiple object instances to generate plausible scene composition, can be used to overcome the related problems.

Future work in this direction may be to employ more image resources in different domains and to introduce diverse multimedia content, such as 3D virtual reality scenes, videos, and audio. In particular, we will investigate in next steps the potential of automatically generating new pictures i.e. automate images composing. Motivated by the success of StackGANs, we would like to spend more time training a stackGAN model using an annotated dataset in Arabic, as we believe our approach has not reached its full capacity due to the time constraint of this project. We plan to explore more options such as more flexible attention mechanisms and more expressive convolutional neural network models to capture and to express the interactions between objects and more details of the text input.

Acknowledgement Open Access funding provided by the Qatar National Library.

Funding This work was made possible by NPRP grant #10-0205-170346 from the Qatar National Research Fund (a member of Qatar Foundation).

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Human and animal rights This research uses publicly available data and it doesn't involve human participants or animals.

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

- Carney RN, Levin JR (2002) Pictorial illustrations still improve students' learning from text. *J R Educ Psychol Rev* 14:5–26
- Lin P, Huang Y, Chen C (2018) Exploring imaginative capability and learning motivation difference through picture e-book. *IEEE Access* 6:63416–63425
- Ramisa A, Yan F, Moreno-Noguer F, Mikolajczyk K (2016) Breaking news: article annotation by image and text processing. *ArXiv e-prints*
- Zakraoui J, Saleh M, Ja'am JA (2019) Text-to-picture tools, systems, and approaches: a survey. *Multimed Tools Appl* 1–27
- 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*, Glasgow, pp 1847–1856
- 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*, Bhopal, pp 886–890
- Aramini S, Ardizzone E, Mazzola G (2015) Automatic illustration of short texts via web images. In: *Proceedings of the 6th international conference on information visualization theory and applications (IVAPP-2015)*
- Delgado D, Magalhães J, Correia N (2010) Automated illustration of news stories. In: *Proceedings of the 2010 IEEE fourth international conference on semantic computing*, Pittsburgh, pp 73–78
- Goldberg AB, Rosin J, Zhu X, Dyer CR (2009) Toward text-to-picture synthesis. In: *Proceedings of the NIPS 2009 symposium on assistive machine learning for people with disabilities*
- Li H, Tang J, Li G, Chua T-S (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*, Vancouver, pp 813–816
- Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. In: *Advances in neural information processing systems 27*, Curran Associates, Inc., pp 2672–2680
- Reed S, Akata Z, Yan X, Logeswaran L, Schiele B, Lee H (2016) Generative adversarial text to image synthesis. In: *ICML*, New York, USA
- Fu A, Yiju H (2017) Text-to-image generation using multi-instance stackgan
- Zhang H, Xu T, Li H, Zhang S, Wang X, Huang X, Metaxas D (2017) StackGAN++: realistic image synthesis with stacked generative adversarial networks. *arXiv preprint arXiv:1710.10916*, 2017
- Jia D, Wei D, Richard S, Li-Jia L, Kai L, Li-F-F (2009) ImageNet: a large-scale hierarchical image database. In: *2009 IEEE conference on computer vision and pattern recognition*, Miami, FL
- ImageCLEF/LifeCLEF Cross—multimedia retrieval in CLEF [Online]. <https://www.imageclef.org/>. Accessed 8 Feb 2018
- MSCOCO [Online]. <http://cocodataset.org/#home>. Accessed 1 Mar 2018
- Micah H, Young Peter Y, Julia H (2013) Framing image description as a ranking task: data, models and evaluation metrics. *J Artif Intell Res* 47:853–899
- Peter Y, Alice L, Micah H, Julia H (2014) From image descriptions to visual denotations: new similarity metrics for semantic inference over event descriptions. *Trans Assoc Comput Linguist* 2:67–78
- Bernardi R, Cakici R, Elliott D, Erdem A, Erdem E, Ikizler-Cinbis N, Keller F, Muscat A, Plank B (2016) Automatic description generation from images: a survey of models, datasets, and evaluation measures. *J Artif Intell Res* 55:409–442
- 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*
- Terragalleria [Online]. <https://www.terrageria.com/>. Accessed 1 Dec 2018
- Art Museum Image Consortium [Online]. <http://www.amico.org/home.html>. Accessed 1 Feb 2019
- 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*, Vancouver, vol 2, pp. 1590–1595
- Mihalcea R, Chee WL (2008) Toward communicating simple sentences using pictorial representations. *Mach Transl* 22:153–173

26. Dmitry U (2012) A Text-to-picture system for Russian language. In: Proceedings 6th Russian young scientist conference for information retrieval, Yaroslavl, pp 35–44
27. Duy B, Carlos N, Bruce EB, Qing Z-T (2012) Automated illustration of patients instructions. *J Am Med Inform Assoc* 2012:1158–1167
28. 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), Ottawa, pp 1–6
29. Eva H, Mc KP, Tom L, Joan C (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, Stroudsburg, pp 125–130
30. Aletras N, Stevenson M (2013) Representing topics using images. In: HLT-NAACL
31. Huang C, Li C, Shan M (2013) Automatic generation of visual story for fairy tales with digital narrative. In: Conference on technologies and applications of artificial intelligence
32. Jiang Y, Liu J, Lu H (2014) Chat with illustration. *Multimed Syst* 22:5–16
33. Hong R, Zha Z-J, Gao Y, Chua T-S, Wu X (2012) Multimedia encyclopedia construction by mining web knowledge. *Signal Process* 93:2361–2368
34. Ganguly D, Calixto I, Jones G (2015) Overview of the automated story illustration task at FIRE 2015. In: Post proceedings of the workshops at the 7th forum for information retrieval evaluation, Gandhinagar, pp 63–66
35. Boonpa SRS, Charoenporn T (2017) Relationship extraction from Thai children's tales for generating illustration. In: 2nd international conference on information technology (INCIT), Nakhonpathom, pp 1–5
36. Karkar AG, Alja'am JM, Mahmood A (2017) Illustrate it! An Arabic multimedia text-to-picture m-learning system. *IEEE Access* 5:12777–12787
37. Scribd [Online]. <http://www.scribd.com>. Accessed 5 Jan 2018
38. R. Naeem, A. T. Imtiaz, S. Muhammad, A. Nouman, M. Anzar and R. Sohail (2019) Three-dimensional face recognition using variance-based registration and subject-specific descriptors. *Int J Adv Robot Syst* 16(3):1729881419851716
39. Naeem R, Imtiaz AT, Muhammad S, Anzar M, Sohail R, Saadat HD, Nouman A, Muhammad U, Mirza JAB, Usman M (2019) Deeply learned pose invariant image analysis with applications in 3D face recognition. *Math Prob Eng* pp. Article ID 3547416, 21 pages
40. Zafar B, Ashraf R, Ali N, Ahmed M, Jabbar S, Naseer K, Ahmad A, Jeon G (2018) Intelligent image classification-based on spatial weighted histograms of concentric circles. *Comput Sci Inf Syst* 15(3):615–633
41. Nouman A, Khalid BB, Robert S, Savvas AC, Zeshan I, Muhammad R, Hafiz AH (2016) A novel image retrieval based on visual words integration of SIFT and SURF. *PLoS ONE* 11(6):e0157428
42. Ali N, Zafar B, Iqbal M, Sajid M, Younis M, Dar S et al (2019) Modeling global geometric spatial information for rotation invariant classification of satellite images. *PLoS ONE* 14(7):e0219833
43. Kaiming H, Georgia G, Piotr D, Ross G (2017) Mask R-CNN. In: IEEE international conference on computer vision (ICCV), Venice, pp 2980–2988
44. Medium Data Science [Online]. https://medium.com/@jonathan_hui/gan-why-it-is-so-hard-to-train-generative-advisory-networks-819a86b3750b. Accessed 01 June 2019
45. Xu T, Zhang P, Huang Q, Zhang H, Gan Z, Huang X, He X (2018) AttnGAN: fine-grained text to image generation with attentional generative adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1316–1324
46. Qaiser T et al (2018) HER2 challenge contest: a detailed assessment of automated HER2 scoring algorithms in whole slide images of breast cancer tissues. *Histopathology* 72(2):227–238
47. Qaiser T, Tsang Y, Taniyama D, Sakamoto N, Nakane K, Epstein D, Rajpoot N (2018) Fast and accurate tumor segmentation of histology images using persistent homology and deep convolutional features. *CoRR*, vol. abs/1805.03699
48. Matterport: Mask RCNN [Online]. https://github.com/matterport/Mask_RCNN. Accessed 4 Jan 2019
49. Hochreiter S, Schmidhuber J (1997) Long short-term memory. *Neural Comput* 9(8):1735–1780
50. Erhan OV, Alexander T, Samy B, Dumitru (2016) Show and tell: lessons learned from the 2015 {MSCOCO} image captioning challenge. *CoRR*
51. Shallue C (2018) Show and tell: a neural image caption generator [Online]. <https://github.com/tensorflow/models/tree/master/research/im2txt>. Accessed 01 May 2018
52. Zakraoui J, Elloumi S, Alja'am J, Yahia S (2019) Improving Arabic text to image mapping using a robust machine learning technique. *IEEE Access* 7:18772–18782
53. Developers G: Google custom search [Online]. <https://developers.google.com/apis-explorer/#search/customsearch/>. Accessed 01 Apr 2018

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