Jordanian Journal of Computers and Information Technology

December 2017 | VOLUME 03 | NUMBER 03 | ISSN 2415 - 1076 (Online) ISSN 2413 - 9351 (Print)

JJCIT

www.jjcit.org jjcit@psut.edu.jo

An International Peer-Reviewed Scientific Journal
Financed by the Scientific Research Support Fund

Jordanian Journal of Computers and Information Technology (JJCIT)

The Jordanian Journal of Computers and Information Technology (JJCIT) is an international journal that publishes original, high-quality and cutting edge research papers on all aspects and technologies in ICT fields.

JJCIT is hosted by Princess Sumaya University for Technology (PSUT) and supported by the Scientific Research Support Fund in Jordan.

AIMS AND SCOPE

The JJCIT aims to publish the most current developments in the form of original articles as well as review articles in all areas of Telecommunications, Computer Engineering and Information Technology and make them available to researchers worldwide. The JJCIT focuses on topics including, but not limited to: Computer Engineering & Communication Networks, Computer Science & Information Systems and Information Technology and Applications.

INDEXING

JJCIT is indexed in:

- ScopeMed:
  http://www.scopemed.org
- CrossRef:
  http://search.crossref.org/?q=jjcit
- OCLC WorldCat:
  http://www.worldcat.org/search?qt=worldcat_org_all&q=jjcit
- Scilit:
  http://www.scilit.net/journals/387088
- Google Scholar:
  https://scholar.google.com/citations?user=88ospLoAAAAJ&hl=en

Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 3, No. 3, December 2017.

A CASE STUDY FOR EVALUATING FACEBOOK PAGES WITH RESPECT TO ARAB MAINSTREAM NEWS MEDIA

Ala'a Al-Shaikh[1], Rizik Al-Sayyed[2] and Azzam Sleit[3]

(Received: 13-Jun.-2017, Revised: 10-Aug.-2017, Accepted: 04-Sep.-2017)

ABSTRACT

In this paper, we propose a framework to analyze and evaluate social networking pages based on usage data with respect to Arab mainstream news media. The paper introduces new metrics such as: Page Penetration and Ranking Index, as well as new evaluation methods. The framework considers the twenty-two Arab countries in addition to seven Facebook pages that belong to seven prominent Arab satellite channels. The proposed framework is used to evaluate countries for their Internet and Facebook penetration rates, as well as consumption of news through those pages. Results reveal that Arabs highly credit natively Arabic news media rather than news media that only speak Arabic. Furthermore, 65% of the Arab countries have more than 50% Facebook users who are news consumers via Facebook. Additionally, Arab countries that suffered unrest, civil war or political crises in the recent years show higher page penetration rates, such as Yemen, Syria, Egypt and Libya.

KEYWORDS

Internet, Facebook, Social networking sites, Social media, Social network analysis.

1. INTRODUCTION

Social Networking Sites (SNS) have recently become an important part in almost everybody's life. SNS are web-based services that enable their users to: (1) create accounts (profiles), (2) connect to friends, relatives, colleagues, fans, …etc. in addition to following these connections and others' connections and (3) exchange messages [1]-[2]. Facebook, Twitter, LinkedIn, Google+, Pinterest, YouTube and Instagram are examples of SNS.

Social Media (SM) are defined as a set of Internet-based applications that exploits web technologies and aims to exchange user-generated content (UGC) between participating entities [2]. SM became a ubiquitous Internet service by leveraging the widespread of SNS [3]-[4].

More than a billion users worldwide use SNS which form nearly 82% of Internet users aged 15 and older [5]. Those use SNS to: (1) establish connections, (2) exchange messages [6] and (3) share content of different types [7], such as: video, audio, UGC [5], personal or private information [8] and blogs [2]. Recently, SNS moved to the mobile computing arena introducing Mobile Social Networks [9]. The ubiquity of mobile devices helped accelerating the diffusion of social networking [10]. SNS are essential, not only for individuals, but also to businesses, educational institutions, mainstream news media, governments, …etc. that access SNS to interact with their customers (clients). However, Facebook has the greatest number of users amongst SNS in terms of number of users [6]-[11].

Nowadays, SM has become a trend in the field of news media technologies which started to penetrate newsrooms in the 1990s using websites, e-mails and mobile technologies [12]. Almost all mainstream news media today utilize SM to increase their reachability and content distribution [13], in addition to gaining a foothold in the competition [14]. Actually, mainstream media use SM to increase their audience, reach and influence [15]-[16]. Mainstream news media can reach more audience using SNS features, such as page recommendation; i.e., profiles of mainstream media may appear to users suggesting liking them [17], which in turn contributes to increasing the number of fans of those pages.

News consumers are highly affected by SM; their way of perceiving news has changed. Now, they can interact with the news that are shared using SM within a few minutes [18]-[19]. Consequently, users are

This paper is an extended version of a short paper that was presented at the international conference "New Trends in Information Technology (NTIT) 2017", 25-27 April 2017, Amman. Jordan.

1. A. Al-Shaikh is with Computer Science Department, University of Jordan, Amman, Jordan. Email: alaamsh@hotmail.com.
2. R. Al-Sayyed is with Business Information Department, University of Jordan, Amman, Jordan. Email: r.alsayyed@ju.edu.jo.
3. A. Sleit is with Computer Science Department, University of Jordan, Amman, Jordan. Email: azzam.sleit@ju.edu.jo.

not only receiving the news; they additionally can choose what to read and can comment and enter discussions with other audience or media.

In this paper, we collect usage data for 7 Facebook pages that pertain to Arab mainstream news media, these are namely: (1) Al Jazeera Channel, (2) Al Arabiyya, (3) Sky News Arabia, (4) BBC Arabic, (5) CNN Arabic, (6) France 24 Arabic and (7) Russia Today Arabic. All of those mainstream news media are satellite channels that broadcast news in Arabic. Similarly, their Facebook pages share Arabic content in different formats and are oriented to Arab countries; i.e., the 22 Arab countries of the Middle East and North Africa (MENA). The study focuses on page penetration rates and countries' news consumption via Facebook.

We use Netvizz as a tool to: (1) collect data about pages and groups on Facebook [20], (2) retrieve page posts, likes, shares and comments [21]-[22] and (3) export the collected data in standard formats [23].

The importance of this research is that we are proposing a framework that can be applied to extract usage data of Facebook pages that disseminate and share any content type which pertains to different domains, like: universities, celebrities, bloggers, SM activists, …etc.

The remainder of this paper is organized as follows: in section 2, we review some of the related work. The methodology is described in section 3. Then, the problem is formulated in section 4. In section 5, we present our experimental results. Finally, conclusion and future work are highlighted in section 6.

2. RELATED WORK

Sharing Tunisian and Egyptian revolutions' news using Twitter was discussed by G. Lotan et al. [15]. They focused on classifying users who share revolution-related content into categories. They concluded that Twitter is an important tool for spreading information. In our research, we are widening the domain of the study to include more Arab countries, some of which had revolutions during the past few years and others had not. We are different from G. Lotan et al. in that we are not interested in analyzing what users share. Rather, we are interested in how much Facebook users from the selected set of countries use Facebook to consume news and what pages they prefer.

S. Hille and P. Bakker [24] studied Facebook usage and participation of Dutch media. They discussed and studied the use of Facebook by media, users' interaction with the posts and journalists' interaction with users. They concluded that Dutch media on Facebook had very few followers compared to the popularity of traditional media and their websites. On the other hand, media were growing with a very low number of likes and comments on posts. In our paper, emphasis is on the penetration rates of Arabic mainstream news media Facebook pages. Thus, we studied how much the audience trust news shared through SNS in Arab countries and which pages have much credibility.

The impact of SM on news consumption was studied by A. Hermida et al. [13] by means of an online survey of 1600 Canadians. Their results show that social networks are important sources of news for Canadians. Differently, we followed an empirical approach. Instead of depending on user opinions that are collected by surveys, we collected real usage data by means of well-known tools. Using real usage data is more accurate than surveys, since surveys may not reflect actual attitudes due to either possible biased answers by participants or subjective options, weights and questions by the one who prepared the questionnaire. Also, the sample size used may not reflect the whole society, which was small to be able to decide in the work of A. Hermida et al. However, in our work, we used actual usage data which are obtained from credible sources. The data, also, represent all the society of users of the Internet, Facebook and the selected pages.

The association between social media and political change in Chile was investigated by S. Valenzuela et al. [25]. Authors studied the use of Facebook for news and socializing rather than using it for self-expression. In that context, authors found a strong relationship between Facebook and protest activity in the country of study. Similarly, some Arab countries witnessed revolutions, protests, unrest and sometimes regime change. It is thought that social medial played an important role in those revolutions in the context of spreading information. Our study contains more countries, some of which witnessed revolutions and others did not. We try to analyze our results on this basis to show the impact of social media on the countries that had revolutions and the countries that didn't have revolutions and to show where Facebook was more influential. S. Valenzuela et al. collected their data by means of surveys of

people aging between 18 and 29 years living in the three largest urban areas in Chile.

A. Ju et al. [26] studied how much Facebook and Twitter are effective as news platforms. They collected data about printed newspaper circulation and web traffic of each newspaper. Their study did not include all the newspapers in the US; it only included the largest ones. A good practice that A. Ju et al. did in their work was the outlier removal. Because New York Times had a very large number of Facebook and Twitter followers compared to the remaining newspapers, it was considered a statistical outlier and removed from any further analysis. There is a great difference between the methods used to collect SNS figures in our paper and those used by A. Ju et al. We used the NetVizz application to collect statistics about the selected Facebook pages. On the other hand, they used a very simple method for finding out the number of SNS users, either on Facebook or Twitter, for the selected newspapers; A. Ju et al. used to open the website of each newspaper, visit the link of the SNS account mentioned by each newspaper, and manually record the number of "likes" and "followers" for each Facebook and Twitter account, respectively. There is a criticism about the method in which they collected their SNS usage data. The authors didn't use an automated tool, such as NetVizz, which gives them detailed statistics about each page, the interactions with the pages, which countries the subscribers are from and the numbers of fans from each country. As a conclusion, the research found a positive correlation between Facebook and Twitter users, web traffic and print readership, although SM users are still the least among web and print readers.

The role of SM in newsrooms is studied by S. Lysak et al. [27] by conducting an online survey to find out which types of SM are used in newsrooms and how they are used. They concluded that SM is used as a means of raising the newsrooms' profile in the community. They also found out that news staff use SM to collect their news although those news must be verified for reliability.

3. METHODOLOGY

The problem is first formulated and the analysis and evaluation metrics are defined. We setup our study on Arab countries only; these are only 22 countries that speak the Arabic language and are in the MENA region. The targeted Facebook pages pertain to Arab mainstream media and have a high level of credibility between Arabs. NetVizz is used as a tool to collect usage data for these pages. Internet and Facebook statistics are collected from well-known worldwide sources. Datasets are created by importing the collected data into a database. Finally, calculations and analysis are made to obtain the desired results.

4. PROBLEM FORMULATION

Assuming that we have n countries and m pages, we define C as a closed set of counties, such that $C = \{c_1, c_2, \dots, c_n\}$. A country $c_i \in C$ is a tuple $c_i < Id, L, I, FB >$ such that $L, I, FB \in \mathbb{N}$, *Id* is the country code, *L* is the population, *I* is the number of Internet users and *FB* is the number of Facebook users. Similarly, Let P be a closed set of Facebook pages, such that $P = \{p_1, p_2, \dots, p_m\}$. A page $p_i \in P$ is a tuple $p_i < Id, F >$ such that $Id, F \in \mathbb{N}$, *Id* is the page identifier and *F* is the number of fans for that page.

Internet Penetration Rate (IPen) is defined as the ratio of Internet users in a specific country to its population [25]. Let I_{c_i} be the number of Internet users in country c_i and L_{c_i} the population of the same country c_i, then:

$$IPen(c_i) = \frac{I_{c_i}}{L_{c_i}}$$

Similarly, Facebook Penetration Rate (FBPen) is defined as the ratio of Facebook users in a specific country to its population [25]. Let FB_{c_i} be the number of Facebook users in country c_i and L_{c_i} the population of the same country c_i, then:

$$FBPen(c_i) = \frac{FB_{c_i}}{L_{c_i}}$$

In order to rank the pages, we propose a new metric called Page Penetration Rate (PgPen). It is the ratio of number of page fans per country to the number of Facebook users in that country. Let $f(c_i, p_j)$ be the number of page fans per country and FB_{c_i} the number of Facebook users in country c_i, then

$$PgPen(c_i, p_j) = \frac{f(c_i p_j)}{FB_{c_i}}$$

It might be feasible to group our set of countries that have similar properties into groups; for example, the grouping could be based on the geographical location. The grouping might be necessary when the number of countries is large and could be ignored when the number of countries is small. Formally, a group G_i is a subset of the country set C; that is $G_i \subset C$ such that $i > 1$.

In this paper, we propose the Ranking Matrix (R) as a means of organizing results of page penetration rates and weighting them. Mathematically, R is a $k \times k$ matrix $R(k, k)$, such that for a given group G_x, the number of rows and columns k is given by: $k = |G_x|$. Rows in R represent the countries of group *Gx* and columns are the ranks (r) obtained by page penetration rates, such that $0 < r_i \leq k$. Each rank is given a weight $w_i = k - i + 1$. A cell R_{ij} represents the number of times a country $c_i \in G_x$ achieved the rank j in terms of page penetration rate.

After the results are organized in the ranking matrix, we propose a metric to sort the countries in order of PgPen. The proposed metric is called the Ranking Index (Rx) and is calculated for each row in R separately by summing the products of each cell by its corresponding weight, then dividing the sum by k^2, such that k is the number of countries in the group.

$$Rx_{c_i} = \frac{\sum_{j=1}^{k} R_{ij} \times w_j}{k^2}$$

Finally, our objective is to sort countries in order of Internet, Facebook and page penetration rates.

5. Experimental Results

As long Population, Internet and Facebook users and page fans have a rapidly changing nature, the data in this research represent the first three quarters of the year 2016, starting from Jan. 1st to Sep. 30th. Firstly, Table 1 contains our list of the 22 Arab countries.

Table 1. List of Arab countries.

Id	Country	Population (L)	Internet Users (I)	Facebook Users (FB)
AE	United Arab Emirates	9,156,963	8,515,420	7,700,000
BH	Bahrain	1,377,237	1,278,752	800,000
DJ	Djibouti	887,861	150,000	150,000
DZ	Algeria	39,666,519	15,000,000	15,000,000
EG	Egypt	91,508,084	34,800,000	32,000,000
IQ	Iraq	36,423,395	14,000,000	14,000,000
JO	Jordan	7,594,547	5,700,000	4,800,000
KM	Comoros	788,474	60,000	60,000
KW	Kuwait	3,892,115	3,202,110	2,300,000
LB	Lebanon	5,850,743	4,545,007	3,100,000
LY	Libya	6,278,438	2,800,000	2,800,000
MA	Morocco	34,377,511	20,207,154	12,000,000
MR	Mauritania	4,067,564	714,132	370,000
OM	Oman	4,490,541	3,310,260	1,500,000
PS	Palestine	4,422,143	3,007,869	1,700,000
QA	Qatar	2,235,355	2,200,000	2,200,000
SA	Saudi Arabia	31,540,372	20,813,695	14,000,000
SD	Sudan	40,234,882	10,886,813	10,886,813
SO	Somalia	10,787,104	660,000	660,000
SY	Syria	18,502,413	5,502,250	5,502,250
TN	Tunisia	11,107,800	5,800,000	5,800,000
YE	Yemen	26,832,215	6,773,228	1,800,000

Country codes in Table 1 are represented in ISO Alpha-2 codes [28]; these are international standard codes that comprise two letters and are used as a general-purpose code [29]. Internet and Facebook users are collected from Internet World Stats website [30] and population data is collected from the World-Bank datasets [31].

NetVizz v1.41 is used to collect Facebook usage data for the specified pages from different Facebook sections and supports different formats [32].

5.1 Internet and Facebook Penetration

Internet and Facebook penetration rates are calculated and shown in Figure 1 and Figure 2, respectively.

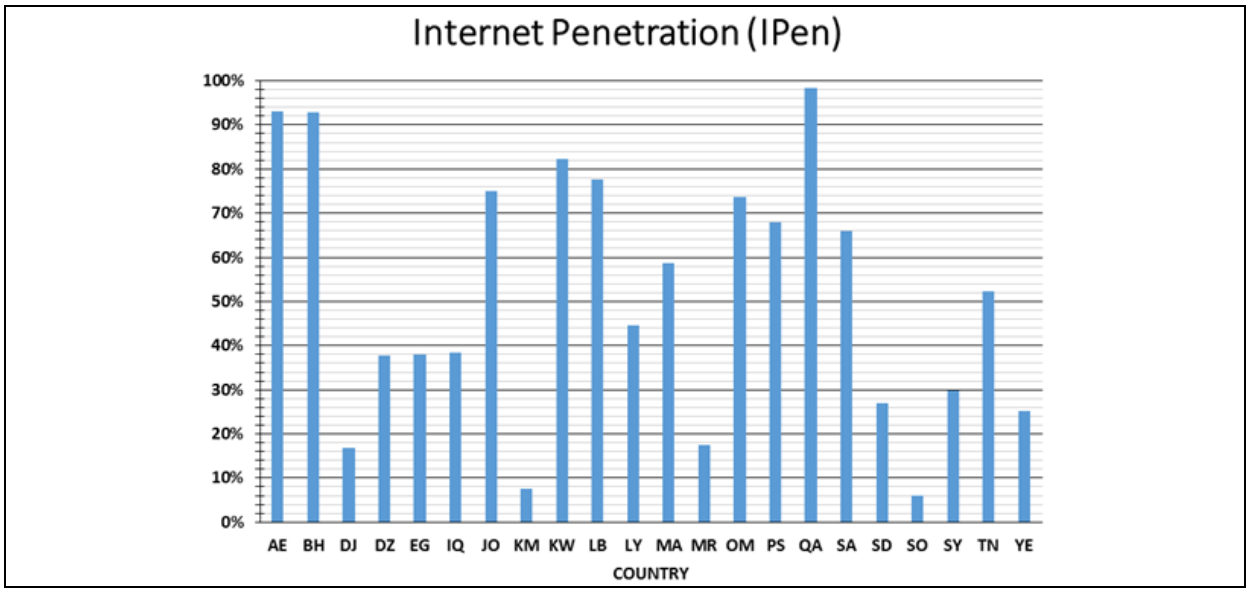


Figure 1. Internet Penetration (IPen).

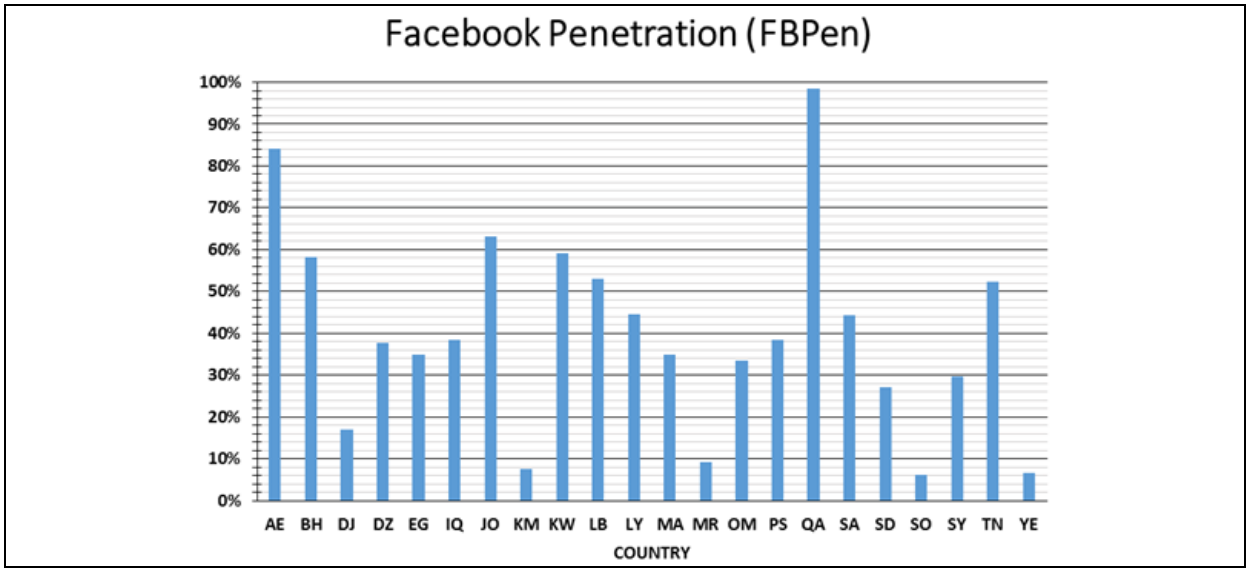


Figure 2. Facebook Penetration (FBPen).

Both results of IPen and FBPen converge; according to IPen, the first four countries were QA, AE, BH and KW, while FBPen shows QA, AE, JO and KW in the first four places. This reveals that countries with the highest Internet penetrations are almost similar to the countries with the highest Facebook penetration, which means that Facebook occupies a large amount of Internet usage, which conforms to the numbers that say that Facebook has the largest number of users amongst other SNS.

5.2 Page Penetration and Ranking

The page set contains the 7 Arab mainstream news media pages that are listed in Table 2. The same data is represented in Figure 3.

Table 2. Selected pages with number of fans for each page.

Id	Page	Fans (F)
JSC	Al Jazeera	17,360,261
ARB	Arabia	16,358,487
SKY	Sky News Arabia	8,996,886
BBC	BBC Arabic	7,248,563
F24	France 24 Arabic	5,960,249
CNN	CNN Arabic	1,798,623
RTA	RTA Arabic	9,669,311

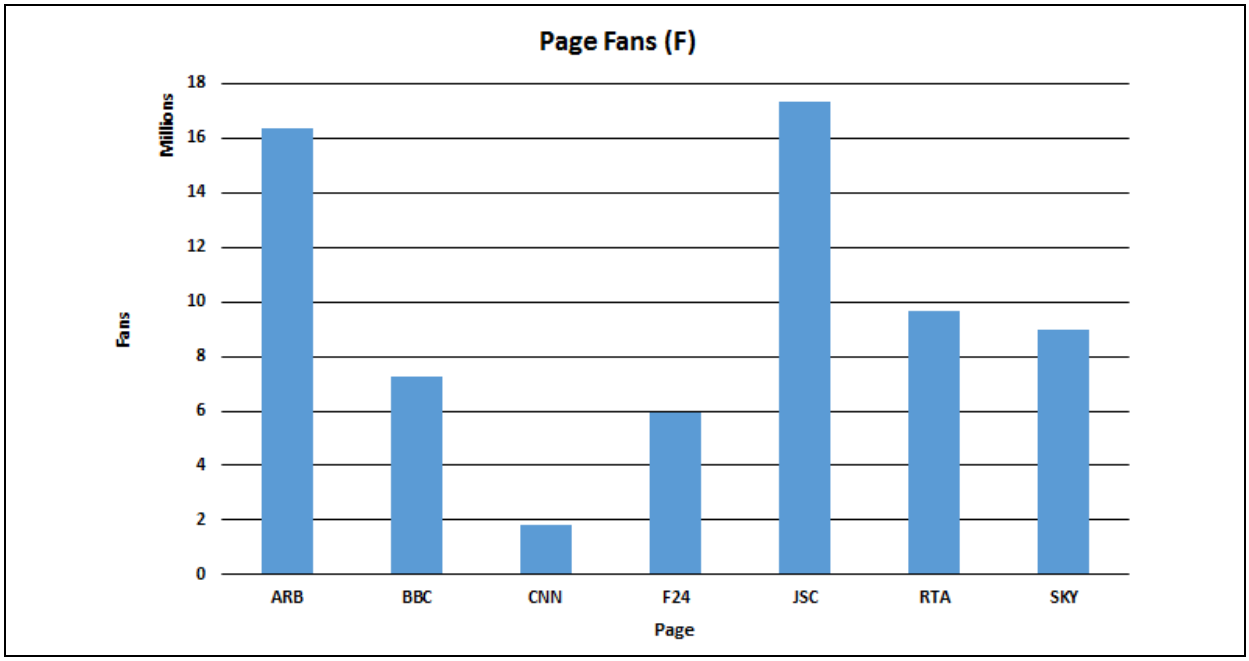


Figure 3. Page fans.

JSC comes in the first place in terms of number of fans, followed by ARB, RTA, SKY, BBC, F24 and finally CNN. Keeping in mind that both JSC and ARB were the first two Arab satellite channels, this gives an explanation to the occupation of these two pages of the first ranks in page fan percentages.

Now, we need to calculate the page penetration rate (PgPen). So, we divide the set of countries into four groups according to their geographical location.

Table 3 lists the four groups and the members of each group. It is noteworthy that both DJ and KM are not grouped, because page statistics for both countries are not available by Facebook. The figures of both countries are considered statistical outliers. Thus, they will not appear in our analysis. Consequently, we only have 20 countries.

Table 3. Countries divided into four groups.

Group	Member Countries
Arab Peninsula	AE, BH, KW, OM, QA, SA, YE
Levant	IQ, JO, LB, PS, SY
North-East Africa	EG, SD, SO
Arab Maghreb	DZ, LY, MA, MR, TN

Henceforth, we need to compute the page penetration rates (PgPen). This is computed for each page and for each country and measures the influence of a certain page in a given country. Because we divided the countries into groups as shown in Table 3, PgPen will be computed for each group separately and the results will be gathered again in one group after the Ranking Index (Rx) is calculated for each country.

Figure 4 shows the page penetration rates for the Arab Peninsula group. The group comprises 7 countries as illustrated in Table 3.

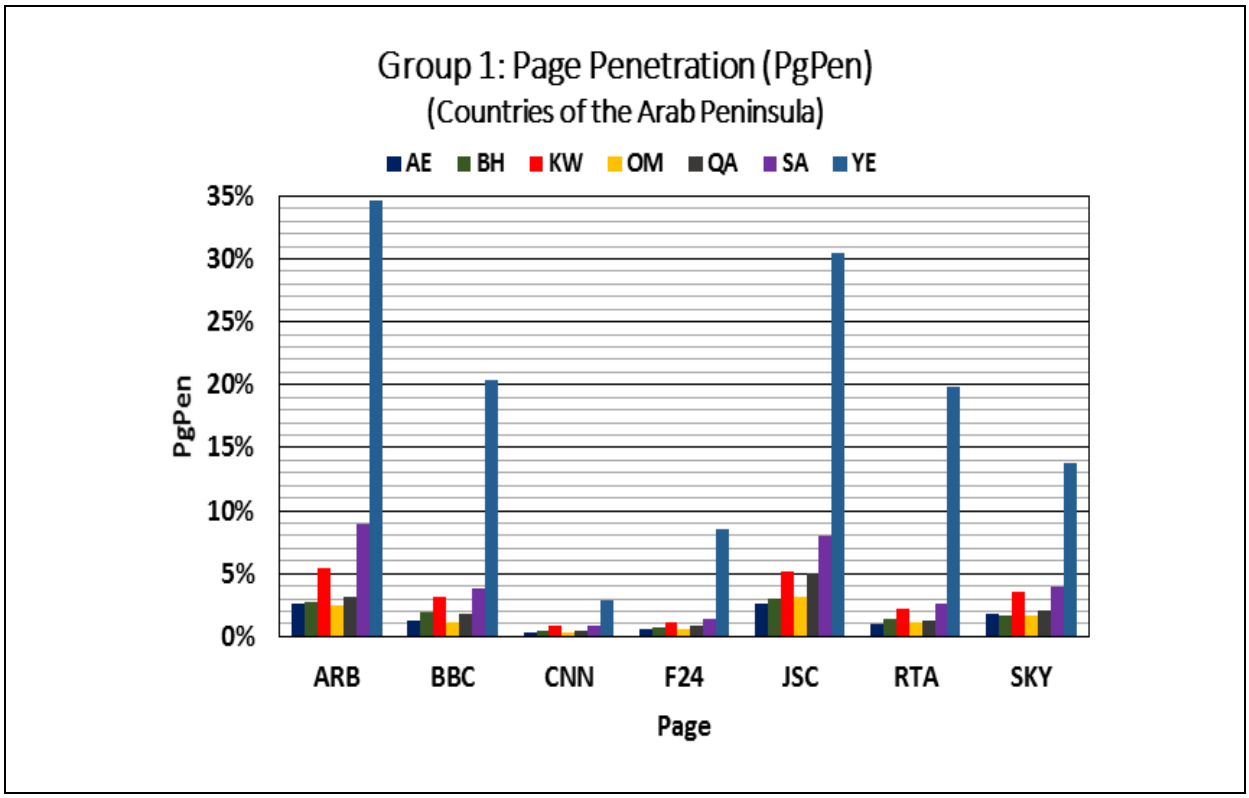


Figure 4. Page Penetration (PgPen) for the first group.

In order to rank the pages in the first group, we construct the ranking matrix as follows: we have 7 countries and thus 7 relevant ranks. Each rank is given a weight that is shown between parentheses in Table 4. We then multiply the number of times a country achieved a rank by the relevant weight of that rank and divide the total by the square of the number of countries in the group; i.e., 49 in the case of group 1. This results in the ranking index Rx for each country in the group.

Table 4. Ranking matrix of group 1.

Country	**Rank (Weight)**							**Total**	**Rx (%)**
	1 (7)	*2 (6)*	*3 (5)*	*4 (4)*	*5 (3)*	*6 (2)*	*7 (1)*		
AE	0	0	0	0	1	4	2	13	26.53
BH	0	0	0	2	3	2	0	21	42.86
KW	0	1	6	0	0	0	0	36	73.47
OM	0	0	0	0	1	1	5	10	20.41
QA	0	0	0	5	2	0	0	26	53.06
SA	0	6	1	0	0	0	0	41	83.67
YE	7	0	0	0	0	0	0	49	100

According to the ranking index (Rx) shown in Table 4, YE comes in the first place, followed by SA, KW, QA, BH, then both AE and OM are in the last two places. Results of the ranking matrix of group 1 are represented in the map shown in Figure 5.

Figure 5. Ranking of countries of group 1 represented in a map graph.

Figure 6 shows page penetration rates for the second group; Arab countries of Levant, and Table 5 shows the ranking matrix of group 2.

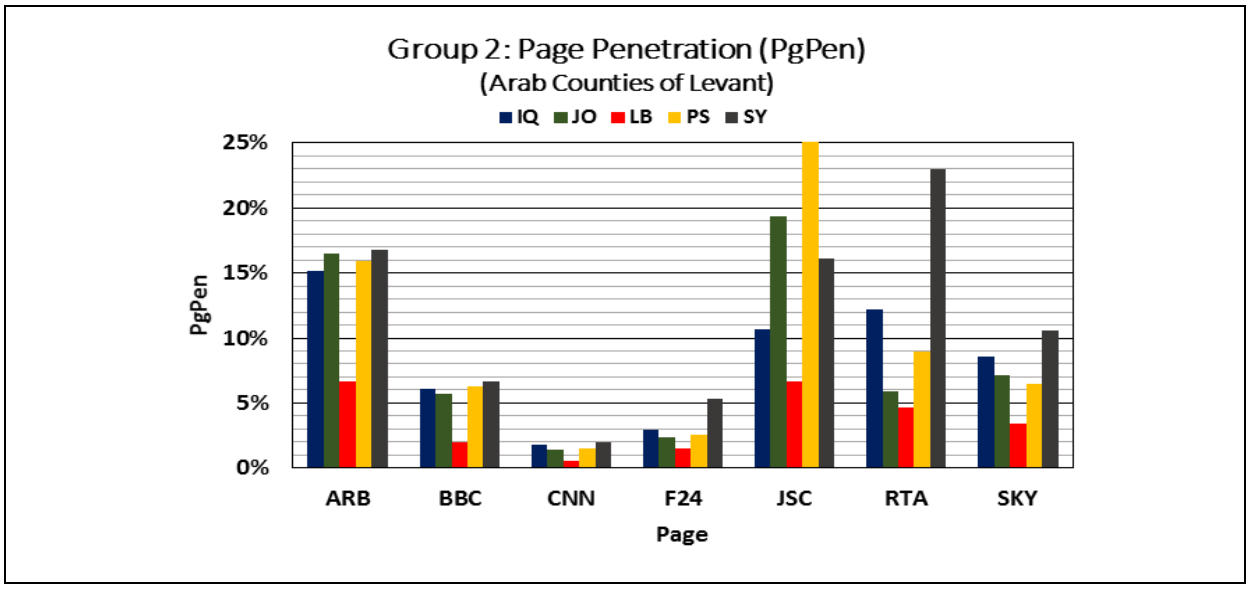


Figure 6. Page Penetration (PgPen) of group 2.

Table 5. Ranking matrix of group 2.

Country	Rank (Weight)					Total	Rx (%)
	1 (5)	*2 (4)*	*3 (3)*	*4 (2)*	*5 (1)*		
IQ	0	4	1	2	0	23	56.71
JO	0	2	1	4	0	19	54.29
LB	0	0	0	0	7	7	20
PS	1	1	4	1	0	23	65.71
SY	6	0	1	0	0	33	94.29

Based on the values of the ranking index (Rx) for the second group, SY comes in the first place, followed by both PS and IQ. JO comes in the third place and finally LB comes in the last place. These results are represented in the map shown in Figure 7.

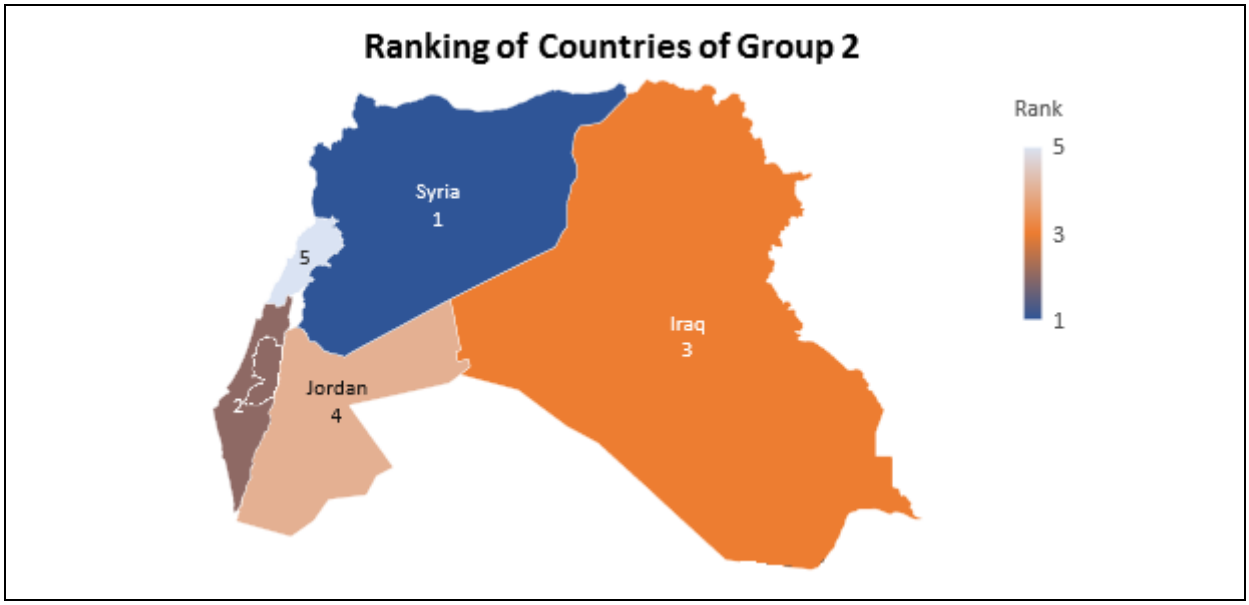


Figure 7. Ranking of countries of group 2.

In a similar way, page penetration rates of countries of the third group; North-East Africa, are shown in Figure 8.

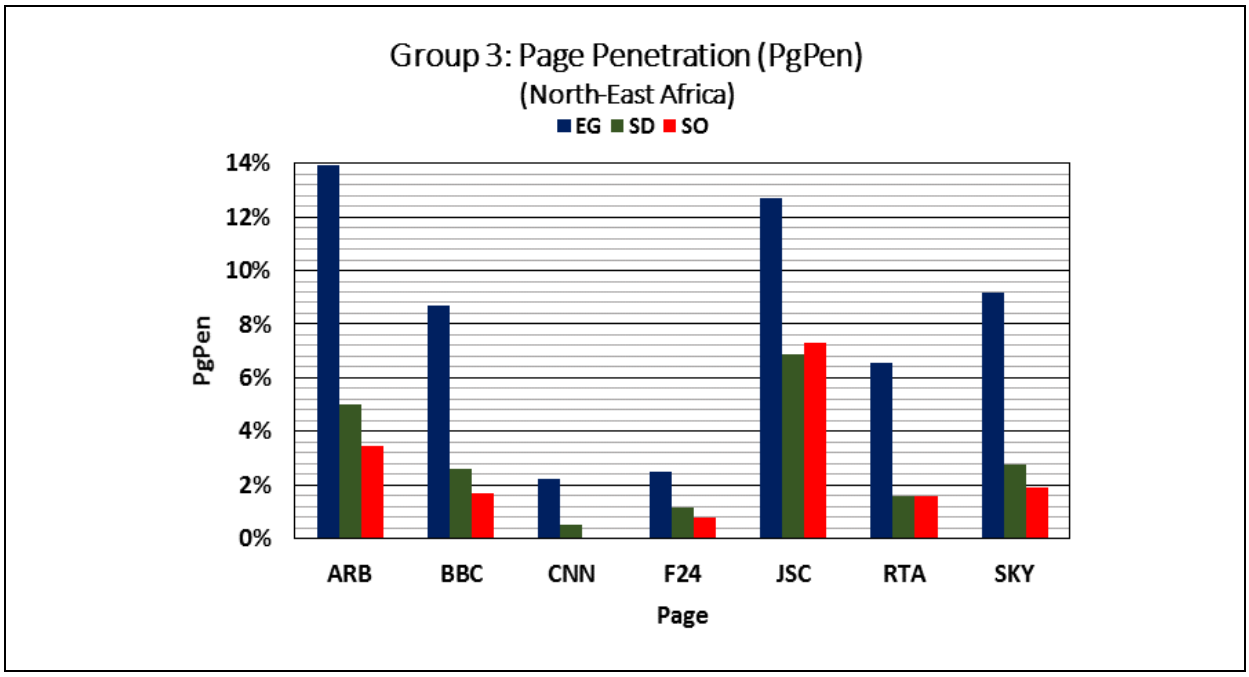


Figure 8. Page Penetration (PgPen) of group 3.

Table 6 is the ranking matrix of the countries of group 3; North-East African countries, comprising only 3 countries. According to the ranking index shown in Table 6, EG is ranked first, followed by SD and then by SO. The ranking is represented by the map graph shown in Figure 9.

Table 6. Ranking matrix of group 3.

Country	**Rank (Weight)**			**Total**	**Rx (%)**
	1 (3)	*2 (2)*	*3 (1)*		
EG	7	0	0	21	100
SD	0	5	2	12	57.14
SO	0	2	5	9	42.86

Figure 9. Ranking of countries of group 3.

Results of the last group; Arab Maghreb countries, comprising 5 countries, are shown in Figure 10.

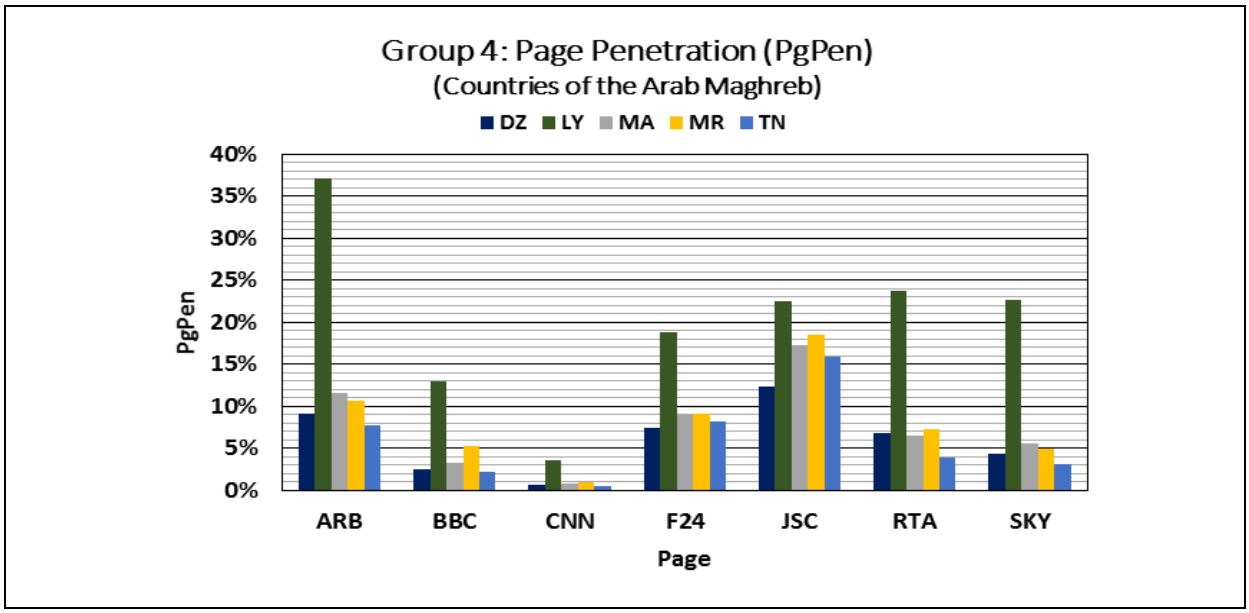


Figure 10. Page Penetration (PgPen) of group 4.

Table 7 shows the ranking matrix for this group and calculates the ranking index for each country.

Table 7. Ranking matrix of group 4.

Country	Rank (Weight)					Total	Rx (%)
	1 (5)	*2 (4)*	*3 (3)*	*4 (2)*	*5 (1)*		
DZ	0	0	1	4	2	13	37.14
LY	7	0	0	0	0	35	100
MA	0	2	4	1	0	22	62.86
MR	0	5	2	0	0	26	74.29
TN	0	0	0	1	6	8	22.86

According to Table 7, LY comes in the first place, followed by MR, then by MA and DZ. Finally comes TN. These results are represented in the map graph shown in Figure 11.

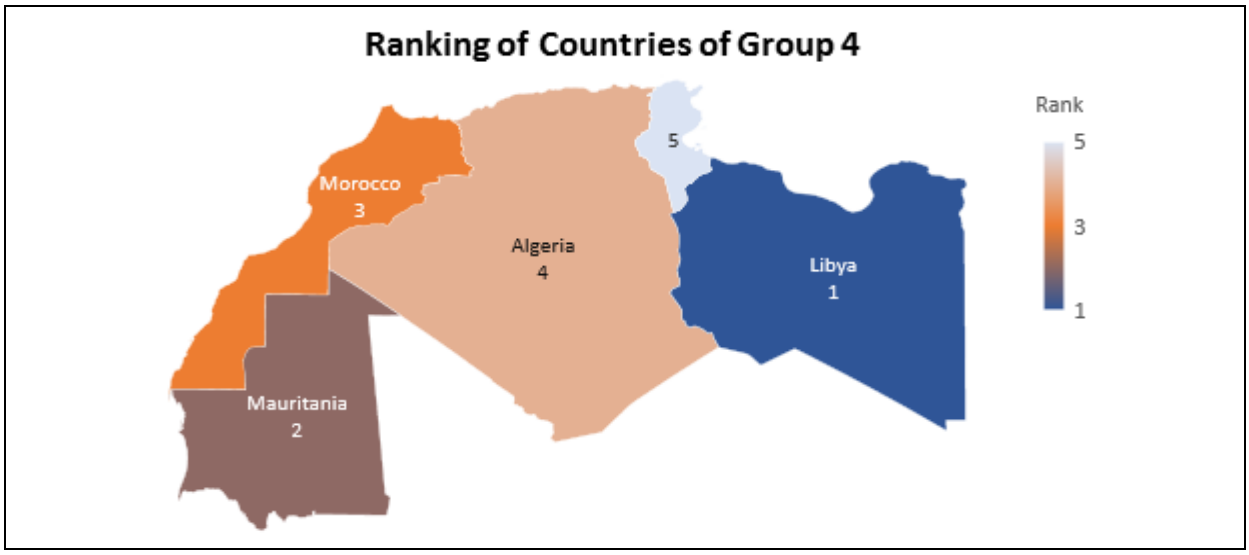


Figure 11. Ranking of countries of group 4.

Table 8 sorts all the 20 countries according to page penetration rate (PgPen) and ranking index (Rx). It contains no further calculations; it only summarizes the ranking results obtained by each ranking matrix of each group earlier.

Table 8. Sorting countries based on their Rx values.

Rank	Country	Rx (%)	Rank	Country	Rx (%)
1	YE	100	8	SD	57.14
	LY		9	JO	54.29
	EG		10	QA	53.06
2	SY	94.29	11	BH	42.86
3	SA	83.67		SO	
4	MR	74.29	12	DZ	37.14
5	KW	73.47	13	TN	22.86
6	PS	65.71	14	AE	20.41
	IQ	65.71		OM	
7	MA	62.86	15	LB	20

Finally, we represent the data listed in Table 8 in the map graph shown in Figure 12.

5.3 Discussion

Results of Internet and Facebook penetration rates shown in Figure 1 and Figure 2, respectively, show that 50% of the Arab countries achieved more than 50% Internet penetration rate, whereas nearly 32% achieved Facebook penetration rates higher than 50%. This shows a relatively high demand on Internet and its resources as suggested by A. Al-Shaikh et al. [33].

Regarding the number of fans listed in Table 2, both JSC and ARB came in the first two places. Our interpretation to these results pertains to history. JSC and ARB were the first two Arabic news satellite channels that made a debut. This gives an intuition that users believe in the maturity of these two channels, which is in turn reflected on the number of fans of their Facebook pages. Another interpretation that we can make is about the nationality of the channel; except JSC and ARB, the other channels are not originally Arab ones. They are either British, American, French or Russian, but they disseminate in Arabic. In conclusion, Arabs prefer to get their news from pages that are natively Arabic and not only Arabic-speaking.

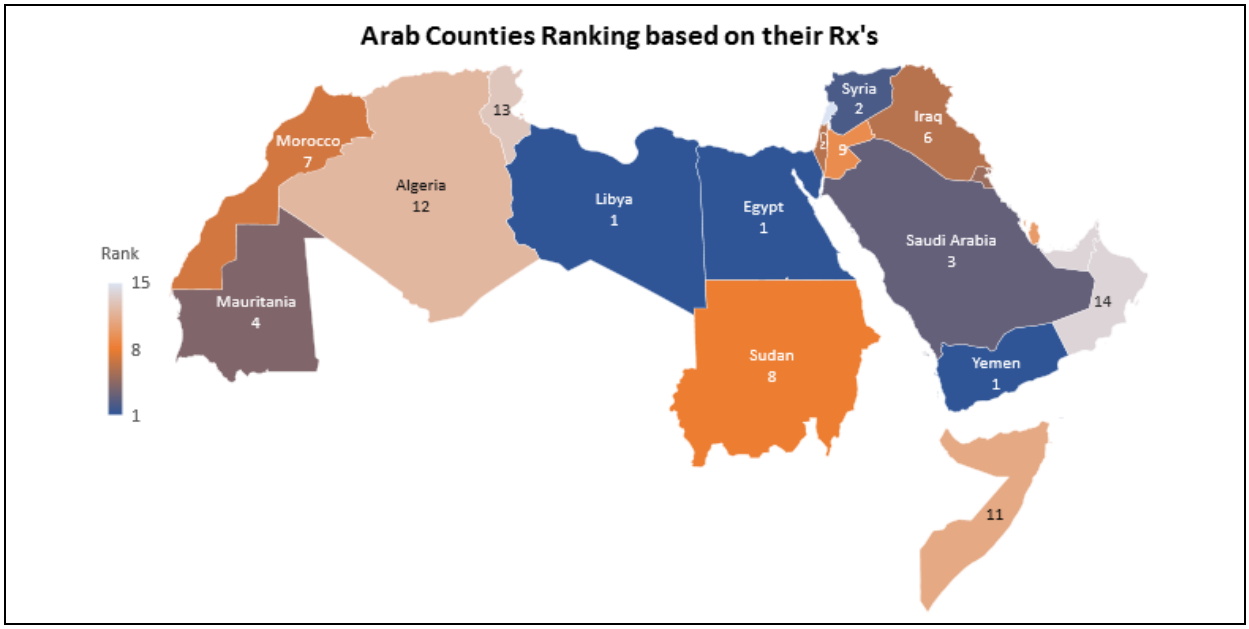


Figure 12. Ranking of countries in terms of their Ranking Indices (Rx).

According to Table 8, YE, LY and EG all share the first rank, followed by SY which came in the second rank. This leads to conclude that citizens of Arab countries that suffered political crises, coups or civil wars during the Arab uprising known by the Arab Spring are the top news consumers via Facebook.

The country that comes in the third place is SA. Despite of never witnessing any political crises, SA is leading a military coalition named the Firmness Storm against militants in YE. This is a good explanation of being in the first ranks of news consumption.

Obviously, 13 out of 20 Arab countries recorded a ranking index (Rx) higher than 50%. In other words, 65% of the Arab countries have more than half of their Facebook users follow mainstream news media on Facebook and consume their news from those pages. However, there are still other Arabic news media that have accounts on Facebook and did not take part in this study. This opens a new research area to investigate either more news media pages on Facebook or different SNS, such as Twitter, especially if we know that Tunisia was the country that ignited the spark of the Arab Spring. Nevertheless, Tunisia was of the least countries that use Facebook for news consumption with a ranking index (Rx) of 22.86%. This leads us to hypothesize that Tunisians might be using different SNS such as Twitter. This conforms to the results of G. Lotan et al. [15], who argued that Twitter was an important tool in spreading information in Egypt and Tunisia during their revolutions.

The same situation applies to Lebanon. Although it did not witness any unrest during the last years, the country is affected by the Syrian revolution which is only few kilometers away from its borders. Furthermore, Lebanon is a country known by its high levels of democracy and freedom of expression. However, being ranked in the last place in our study hypothesizes that either Lebanese are convinced by their local news agencies or they might be using different SNS, which in turn opens the door for a new research area.

It is worthwhile to mention that our findings contradict to those of S. Hille and P. Bakker [24]. We found that a relatively high percentage of Facebook users from the Arab world consume news through Facebook. This could be also due to the current situations and circumstances that some Arab countries are facing, stimulating Arabs to follow news on SM which constitutes a faster medium for spreading news, especially during those accelerating events.

6. Conclusion and Future Work

In this paper, we proposed a framework to evaluate SNS. We used Internet and Facebook penetration rates and proposed some other metrics, such as: page penetration (PgPen) and ranking index (Rx). We also introduced techniques for ranking the pages and countries. We applied the proposed framework to

Facebook pages of some Arab mainstream news media. We concluded that the credibility of natively Arabic news media is higher than that of others that are non-natively Arabic from Arabs' perspective. Also, our findings revealed that users from countries that faced civil war, unrest, political crises, …etc. are the top news consumers via Facebook. Moreover, 70% of the Arab countries have more than 50% of their Facebook users using it for news consumption. The importance of this study is that we established a framework that could be used for evaluating Facebook or any SNS pages from different domains. We can further analyze the contents of those pages or other pages to examine user trends. The same study can be applied to different SNS, like: LinkedIn, Twitter, Google+, among others.

REFERENCES

[1] D. M. Boyd and N. B. Ellison, "Social Network Sites: Definition, History and Scholarship," Journal of Computer-Mediated Communication, vol. 13, no. 1, pp. 210-230, 2007.

[2] A. M. Kaplan and M. Haenlein, "Users of the World, Unite! The Challenges and Opportunities of Social Media," Business Horizons, vol. 53, no. 1, pp. 59-68, 2010.

[3] H. G. D. Zuniga, N. Jung and S. Valenzuela, "Social Media Use for News and Individuals' Social Capital, Civic Engagement and Political Participation," Journal of Computer-Mediated Communication, vol. 17, no. 3, p. 319–336, 2012.

[4] A. Sleit and E. Al-Nsour, "Corner-based Splitting: An Improved Node Splitting Algorithm for R-tree," Journal of Information Science, vol. 40, no. 2, pp. 222-236, 2014.

[5] A. Sukhu, T. Zhang and A. Bilgihan, "Factors Influencing Information-Sharing Behaviors in Social Networking Sites," Services Marketing Quarterly, vol. 36, no. 4, pp. 317-334, 2015.

[6] E. Gilbert, S. Bakhshi, S. Chang and L. Terveen, ""I Need to Try This"?: A Statistical Overview of Pinterest," in: CHI '13: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Paris, France, 2013.

[7] Q. Xiao, W. Zhuang and M. K. Hsu, "Using Social Networking Sites: What is the Big Attraction? Exploring a Mediated Moderation Relationship," Journal of Internet Commerce, vol. 13, no. 1, pp. 45-64, 2014.

[8] N. N. Bazarova, "Contents and Contexts: Disclosure Perceptions on Facebook," in: CSCW '12: Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work, Seattle, Washington, USA, 2012.

[9] W. Dong, V. Dave, L. Qiu and Y. Zhang, "Secure Friend Discovery in Mobile Social Networks," Proceedings of IEEE INFOCOM, Shanghai, China, 10-15 April 2011.

[10] K. Sorathia and A. Joshi, "My World – Social Networking through Mobile Computing and Context Aware Application," in: Intelligent Interactive Assistance and Mobile Multimedia Computing: International Conference (IMC 2009), Rostock-Warnemunde, Germany, 9-11 November 2009. Proceedings, Springer Berlin Heidelberg, pp. 179-188, 2009.

[11] B. A. Mahafzah, A. Sleit, N. A. Hamad, E. F. Ahmad and T. M. Abu-Kabeer, "The OTIS Hyper Hexa-cell Optoelectronic Architecture," Computing, vol. 94, no. 5, pp. 411-432, 2012.

[12] A. Mare, "A Complicated but Symbiotic Affair: The Relationship between Mainstream Media and Social Media in the Coverage of Social Protests in Southern Africa," Ecquid Novi: African Journalism Studies, vol. 34, no. 1, pp. 83-98, 2013.

[13] A. Hermida, F. Fletcher, D. Korell and D. Logan, "SHARE, LIKE, RECOMMEND: Decoding the Social Media News Consumer," Journalism Studies, vol. 13, no. 5-6, pp. 815-824, 2012.

[14] W. Stassen, "Your News in 140 Characters: Exploring the Role of Social Media in Journalism," Global Media Journal - African Edition, vol. 4, no. 1, pp. 116-131, 2010.

[15] G. Lotan, E. Graeff, M. Ananny, D. Gaffney, I. Pearce and D. Boyd, "The Revolutions Were Tweeted: Information Flows during the 2011 Tunisian and Egyptian Revolutions," International Journal of Communication, vol. 5, pp. 1375–1405, 2011.

[16] A. Sleit, I. Salah and R. Jabay, "Approximating Images Using Minimum Bounding Rectangles," 1st International Conference on the Applications of Digital Information and Web Technologies (ICADIWT 2008), pp. 394-396, Ostrava, Czech Republic, 4-6 Aug. 2008.

[17] O. Murad, A. Sleit and A. Sharaiah, "Improving Friends Matching in Social Networks Using Graph

Coloring," International Journal of Computers & Technology, vol. 15, no. 8, pp. 7028-7034, 2016.

[18] C. S. Lee and L. Ma, "News Sharing in Social Media: The Effect of Gratifications and Prior Experience," Computers in Human Behavior, vol. 28, pp. 331–339, 2012.

[19] A. Sleit, A. L. A. Dalhoum, I. Al-Dhamari and A. Awwad, "Efficient Enhancement on Cellular Automata for Data Mining," Proceedings of the 13th WSEAS International Conference on Systems, pp. 616-620, 2009.

[20] Ekta, S. Dhawan and K. Singh, "Feature Extraction and Content Investigation of Facebook Users' Using Netviz and Gephi," Advances in Computer Science and Information Technology (ACSIT), vol. 3, no. 4, pp. 262-265, 2016.

[21] A. O. Larsson, "I Shared the News Today, Oh Boy," Journalism Studies, pp. 1-19, 2016.

[22] A. Sleit, "On Using B+-tree for Efficient Processing for the Boundary Neighborhood Problem," WSEAS Transactions on Systems, vol. 11, no. 11, pp. 711-720, 2008.

[23] M. Magnusson, "Facebook Usage during a Flood – A Content Analysis of Two Local Governments' Facebook Pages," The 27th Australasian Conference on Information Systems, Wollongong, pp. 1-11, 2016.

[24] S. Hille and P. Bakker, "I like News. Searching for the 'Holy Grail' of Social Media: The Use of Facebook by Dutch News Media and Their Audiences," European Journal of Communication, vol. 28, no. 6, p. 663–680, 2013.

[25] S. Valenzuela, A. Arriagada and A. Scherman, "The Social Media Basis of Youth Protest Behavior: The Case of Chile," Journal of Communication, vol. 62, no. 2, pp. 299–314, 2012.

[26] A. Ju, S. H. Jeong and H. I. Chyi, "Will Social Media Save Newspapers?" Journalism Practice, vol. 8, no. 1, pp. 1-17, 2014.

[27] S. Lysak, M. Cremedas and J. Wolf, " Facebook and Twitter in the Newsroom: How and Why Is Local Television News Getting Social with Viewers?," Electronic News, vol. 6, no. 4, pp. 187-207, 2012.

[28] "Country Codes List - ISO ALPHA-2, ISO ALPHA-3 and Numerical Country Codes - Nations Online Project," One World, [Online], Available: http://www.nationsonline.org/oneworld/country_code_list .htm. [Accessed 1-12-2016].

[29] "ISO 3166 - Country Codes - ISO," ISO, [Online], Available: http://www.iso.org/iso/home/standards/ country_codes.htm. [Accessed 1-12-2016].

[30] "Internet World Stats," [Online], Available: http://www.internetworldstats.com/. [Accessed 1-12-2016].

[31] "Population, Total Data," The World Bank, 14 10 2016, [Online], Available: http://data.worldbank.org/indicator/SP.POP.TOTL?end=2015&start=1960. [Accessed 1 12 2016].

[32] B. Rieder, "Studying Facebook via Data Extraction: The NetVizz Application," in: Proceedings of the 5th Annual ACM Web Science Conference, pp. 346-355, Paris, France, 2-4 May 2013.

[33] A. Al-Shaikh, H. Khattab, A. Sharieh and A. Sleit, "Resource Utilization in Cloud Computing As an Optimization Problem," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 7, no. 6, pp. 336-342, 2016.

ملخص البحث:

نقترح في هذه الورقة إطاراً لتحليل صفحات شبكات التواصل الاجتماعي وتقييمها بناءً على بيانات الاستخدام المتعلقة بوسائل الإعلام الإخبارية العربية السائدة. وتقدّم هذه الورقة مقاييس جديدة مثل: معدّل النفاذ الى الصفحة، ومؤشر الترتيب، ناهيك عن طرق تقييم جديدة. ويأخذ الإطار المقترح بعين الاعتبار الدول العربية، وعددها 22، الى جانب 7 صفحات فيسبوك تنتمي الى 7 قنوات فضائية عربية مشهورة. وقد تم استخدام الإطار المقترح لتقييم الدول من حيث معدّلات النفاذ الى الانترنت وفيسبوك، بالإضافة الى استهلاك الأخبار عبر تلك الصفحات.

وكشفت النتائج أن العرب يقدّرون عالياً وسائل الإعلام الإخبارية عربية الأصل على نحو يفوق تقديرهم لوسائل الإعلام الإخبارية التي تتحدث العربية فقط. علاوة على ذلك، فإن 65% من الدول العربية تمتلك نسبةً تفوق 50% من مستخدمي فيسبوك من مستهلكي الأخبار عبر فيسبوك. من جهة أخرى، تُظهر الدول العربية التي عانت من القلاقل أو الحروب الأهلية أو الأزمات السياسية في السنوات الأخيرة معدّلات نفاذ أعلى للصفحات، مثل اليمن وسورية ومصر وليبيا.

Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 3, No. 3, December 2017.

ENGLISH-ARABIC POLITICAL PARALLEL CORPUS: CONSTRUCTION, ANALYSIS AND A CASE STUDY IN TRANSLATION STRATEGIES

Alia Al-Sayed Ahmad[1], Bassam Hammo[2] and Sane Yagi[3]

(Received: 20-Jun.-2017, Revised: 19-Aug.-2017 and 04-Oct.-2017, Accepted: 07-Oct.-2017)

ABSTRACT

This study reports on the construction of a one million word English-Arabic Political Parallel Corpus (EAPPC), which will be a useful resource for research in translation studies, language learning and teaching, bilingual lexicography, contrastive studies, political science studies and cross-language information retrieval. It describes the phases of corpus compilation and explores the corpus, by way of illustration, to discover the translation strategies used in rendering the Arabic and Islamic culture-specific terms takfīr and takfīrī from Arabic into English and from English into Arabic. The Corpus consists of 351 English and Arabic original documents and their translations. A total of 189 speeches, 80 interviews and 68 letters, translated by anonymous translators in the Royal Hashemite Court, were selected and culled from King Abdullah II's official website, in addition to the textual material of the English and Arabic versions of His Majesty's book, Our Last Best Chance: The Pursuit of Peace in a Time of Peril (2011). The texts were meta-annotated, segmented, tokenized, English-Arabic aligned, stemmed and POS-tagged. Furthermore, a parallel (bilingual) concordance was built in order to facilitate exploration of the parallel corpus. The challenges encountered in corpus compilation were found to be the scarcity of freely available machine-readable Arabic-English translated texts and the deficiency of tools that process Arabic texts.

KEYWORDS

Parallel corpus, Political, English-Arabic translation, Corpus compilation, Challenges.

1. INTRODUCTION

A parallel corpus is a collection of original texts and their translations in a target language. These texts can be aligned at different levels, such as paragraph, sentence, phrase or word level. Bilingual concordances are used to display all the occurrences of a search term in the source language (SL) together with their equivalents in the target language (TL). This type of corpora plays a crucial role in research that involves two or more languages, such as machine translation, cross-language information processing, contrastive studies, language research, language learning and teaching and bilingual lexicography [1]-[6].

1.1 Parallel Corpora and Translation Studies

The benefits of parallel corpora to translation studies and translation activity are well recognized. Parallel corpora are important educational tools when they are used in translation training programs, as they provide learners with authentic examples and with a flow of language data, from which they can discover and learn the strategies employed by translators [6].

Parallel corpora play an important role in the development of machine translation (MT) systems. There are three approaches to machine translation: linguistic knowledge, statistical and computer-assisted [7].

The first approach is rule-based and depends on such linguistic knowledge as morphological, syntactic, semantic and idiomatic knowledge of the source and target languages. In this approach, the source sentence is translated into the target sentence using a parser. This parser analyses the source sentence into its components, such as NP, VP, AdvP and PP, then replaces them with their TL equivalents with

This paper is an extended version of a short paper that was presented at the international conference "New Trends in Information Technology (NTIT) 2017", 25-27 April 2017, Amman. Jordan.

1. A. Al-Sayed Ahmad is with the English Department, University of Jordan, Amman, Jordan. Email: aliaahmadsh@gmail.com.
2. B. Hammo is with the Department of CIS, University of Jordan, Amman, Jordan. Email: b.hammo@ju.edu.jo.
3. Sane Yagi is with the Department of Linguistics, University of Jordan, Amman, Jordan. E-mail: saneyagi@yahoo.com.

the help of a dictionary. Then, the output sentence is reorganized in accordance with the linguistic rules of the target language.

The second approach is statistical machine translation. It analyses a parallel corpus, selects the SL-TL patterns that coincide most frequently and uses them in the translation. For example, in such statistically-based systems as Google Translate, parallel corpora, monolingual corpora and statistical models of their data are heavily used to automatically render texts from one language into another [8].

The third approach is computer-assisted translation, which involves an interactive process between the machine and the translator. Many types of computer-assisted translation software (e.g., electronic dictionaries and translation memories) were developed to facilitate and automate the process of translation. Bilingual and multilingual electronic dictionaries contain information about SL and TL words, such as part of speech, pronunciation and collocations. These dictionaries can be found in different forms: special devices (e.g. Atlas Modern Dictionary English-Arabic), computer software (e.g. Golden Al-Wafi), smart phone applications (e.g. Britannica Dictionary), CD-ROMs and DVD-ROMs usually sold with the printed version (e.g. Oxford Elementary Learner's Dictionary with CD-ROM) and online dictionaries (e.g. https://www.merriam-webster.com/). These dictionaries have the advantage of swiftly finding a query term.

One of the most precious computer-assisted translation resources is a translation memory that a company can create for its translators. Translation memory (TM) is a repository of translated phrases, where SL text segments are aligned with their TL equivalents. When the translator activates the translation memory and starts to translate a new document, the TM would quickly offer him/her a suggested translation for any SL segment that matches a previously translated one in its database. Building a translation memory is a "process of comparing a source text and its translation, matching the corresponding segments and binding them together as translation units" [9]. TMs save the translator's time and effort, particularly in translating documents of highly repeated texts, such as legal contracts. Some websites, such as Glosbe (https://glosbe.com/), have numerous dictionaries and translation memories that offer the user access to parallel texts in different languages including Arabic and English.

Parallel corpora have proven to be useful in developing machine translation systems. They spare time and effort and contribute to the resolution of some translation challenges. Even though machine-translation output is occasionally dull and literal, parallel corpora can capture subtle meanings of the source text (ST), idiomatic expressions and metaphors [10].

1.2 Parallel and Comparable Corpora

Parallel and comparable corpora are two types of multilingual corpora. Both parallel and comparable corpora consist of texts in two or more languages, but the first requires that there be a source language and a translation version of the same texts. The latter, on the other hand, makes no such requirement. The texts that it contains are merely of the same sampling frame (i.e., the same text size from the same genres and published in the same period of time). Parallel and comparable corpora are invaluable to translation and contrastive studies [11].

1.3 Challenges of Compiling Parallel Corpora

Developing a parallel corpus is not a straightforward task. This is due to technical and linguistic challenges encountered at most stages of construction: text selection, conversion, segmentation, stemming, alignment and annotation [12]-[15].

In the text selection process, it is challenging to find a large number of translated open-access texts that are available in the desired language combination. These texts should be machine-readable, accessible and representative samples of the use of the specific language combination [12]-[13], [16].

Moreover, some languages, such as Arabic, suffers from the scarcity and/or inefficiency of tools that are used for text conversion (e.g., OCR), segmentation, tokenization and part of speech tagging [12]-[13], [17]-[18].

Minority languages and languages of technologically underdeveloped countries have to overcome the formidable challenge of automating texts alignment [17]. It is widely acknowledged that aligning a

large amount of texts is probably prohibitively expensive [19]. That is why it is difficult to find parallel texts of any significant size that are aligned at phrase level for any language combination. Although Arabic is not a minority language, it is lacking in reliable text-alignment and text-annotation tools [12]-[13], [17], [20].

The rest of this paper is organized as follows. The next section provides a brief overview of the existing English and Arabic parallel corpora. Section 3 presents the methodology of building the corpus. Section 4 gives information about the parallel concordance. Corpus experimentation is discussed in section 5. Concluding remarks are presented in Section 6.

2. LITERATURE REVIEW

2.1 English Parallel Corpora

Parallel corpora began to appear in 1988, when Bell Communications Research and the IBM T. J. Watson Research Center compiled the Hansard corpus; the first parallel corpus of French and English [19]. It included 50 million words collected from transcriptions of the Canadian Parliament debates between 1975 and 1988 [19]. Since then, many parallel corpora projects were initiated. The corpus of European Corpus Initiative (ECI) contains about 19 million words from French, English and Spanish texts; the English-Norwegian Parallel Corpus (ENPC) consists of two million tokens that were culled from original fiction and non-fiction English and Norwegian texts and their translations. The original texts were aligned with their translations at sentence level. This corpus was aimed at carrying out comparisons between original texts and their translations, originals in both languages, translations in both languages and originals and translations in one language [21]. Later, the ENPC was joined by the German-Norwegian, French-Norwegian and Russian and Norwegian parallel corpora to form the Oslo Multilingual Corpus [21].

In the wake of ENPC, other parallel corpora that included English have been compiled (e.g., the English-Swedish Parallel Corpus, the English-French corpus, the English-German corpus and the English-Spanish corpus) [21]. These parallel corpora followed the design criteria of ENPC and shared some of its English original texts [21]. The JRC-ACQUIS Multilingual Parallel Corpus included more than one billion tokens from 22 languages [22]. In addition, the Official Journal of European Community multilingual parallel corpus involved English-German, English-Italian, English-Spanish and English-French aligned combinations [22]. The Open Parallel Corpus (OPUS) consists of nearly 352 million tokens in sixty European and Asian languages including Arabic [23]. Perhaps, these EU corpora were geared more towards research in natural language processing (NLP) than in linguistics and translation studies.

The list of parallel corpora that include English is too long to cover here. This is not the case for Arabic however.

2.2 Arabic Parallel Corpora

Arabic parallel corpora began to come into existence only in the late 1990's when the English-Arabic Parallel Egypt Corpus was developed at John Hopkins University in 1999 for the purpose of facilitating machine translation [24]. It consisted of the Qur'an in English and Arabic. Then in 2004, the English-Arabic Parallel Corpus was compiled by Al-Ajmi [17]. It contained three million words that were collected from the Kuwaiti World of Knowledge book series. This is a series of translated books about a variety of topics in history, economics, arts, science and literature. In addition to the aforementioned OPUS, a parallel Spanish-Arabic corpus was built from the annual reports of United Nations institutions for the purpose of experimenting with alignment at sentence level [12]. Samy et al. reused tools made for the Spanish language with Arabic texts. One year later, Samy et al. added English texts from the United Nations documents and developed the Arabic-Spanish-English multilingual corpus [13]. The Quranic Arabic Corpus[1] [25] is excellent for illustrating morphologically annotated classical Arabic. The English-Arabic Parallel Corpus of United Nations Texts (EAPCOUNT) was compiled by Hammouda Salhi in 2013 [26]. Finally, the Linguistic Data Consortium at the University of Pennsylvania (LDC)2 developed several English Arabic parallel corpora from broadcast conversation

[1] http://corpus.quran.com/

(e.g., talk shows), broadcast news and news wires, which amount to around 40 million words.

Although these may appear like numerous parallel corpora, most of them are either proprietary, experimental or restricted to specific text sources (e.g. News, UN documents and the Holy Quran). They are also aligned at either paragraph or sentence level but not at phrase and word levels. Most of them are not POS-tagged, as shown in Table 1. There is a clear need for properly-annotated Arabic resources that translators, learners, educators, researchers and language engineers can use free of charge.

Table 1. Examples of English-Arabic parallel corpora.

	Egypt Corpus	**The English-Arabic Parallel Corpus**	**OPUS**	**Arabic-Spanish-English Parallel Corpus**	**Quranic Arabic Corpus**	**EAPCOUNT**	**LDC**
Size (words)	77,430	3 million	352 million	3 million	77,430	5.5 million	40 million
Availability	×	×	✓	✓	✓	×	×
Medium	Written	Written	Written	Written	Written	Written	Written and spoken
Source	The Holy Quran	The World of Knowledge (a series of translated books)	OpenOffice.org documentation KDE manuals including KDE system messages PHP manuals	UN documents	The Holy Quran	UN documents	Broadcast conversation, traditional broadcast news and newswires
Alignment level	Sentence	Sentence	Sentence	Sentence	Sentence	Paragraph	Sentence and word
POS tagging	×	×	×	✓	✓	×	×
Stemming	×	✓	×	×	✓	×	×

To fill this gap, the present study developed a freely available, human-verified English-Arabic political parallel corpus (EAPPC) that contains more than one million words culled from His Majesty King Abdullah II's written and spoken texts. To the best of our knowledge, this is the first work that aligns English-Arabic parallel texts at multiple levels (i.e., sentence, clause, phrase and word levels). To make this resource even more valuable, the corpus texts have not only been stemmed and POS-tagged automatically, but also manually verified. The present corpus can be a springboard to a Jordanian English-Arabic Parallel corpus. However, the most important limitation of our parallel corpus is that it is restricted to the Arabic and English speeches, interviews, letters and book of one person and that the translation was performed by anonymous translators in the Royal Court. The size of the corpus is also a limiting factor, but the fact that this is a research in progress is a consolation.

3. BUILDING THE CORPUS

3.1 Data Selection

A preliminary survey of Arabic texts on the World Wide Web (WWW) was conducted to identify the kinds of existing Arabic texts that were translated into English and English texts that were translated into Arabic. The survey results showed that there were different types of texts such as UN documents, news (e.g., Petra News Agency and the British Broadcasting Corporation's news texts), novels (e.g., Najib Mahfouth, Ghassan Kanafani and Agatha Cristy's) and books (e.g., on history and science). However, most of these texts were not freely available. Furthermore, most available Arabic texts were found in a Portable Document Format (PDF) whose conversion into machine-readable text would result in highly corrupted content. In order to obtain high-quality textual material for the present corpus, we considered the following selection criteria:

1. Arabic data should be in Modern Standard Arabic (MSA) and must have an English translation.
2. English data should be in Standard English regardless of geographic origin and must have a translation in Modern Standard Arabic.
3. The translation must not be produced by machine or non-professional translators.
4. The data should be available in machine-readable textual format.
5. The data should be representative of MSA in general or of a particular MSA genre.
6. Texts ought not to have been used in previous parallel corpora. This avoids duplication and opens the way for our corpus to get integrated with previous English-Arabic parallel corpora in the future.
7. The copyright must permit corpus compilation.

Fortunately, the required data were found in His Majesty's official website (https://www.kingabdullah.jo/). His speeches, interviews and letters met all the selection criteria. However, we decided to add His Majesty's book, *Our Last Best Chance: The Pursuit of Peace in a Time of Peril (2011)*, despite the fact that it was not available to the researcher in electronic format. This was done for two reasons: first, to enlarge the corpus size; second, to shed light on the real problems that are often encountered by corpus compilers when they deal with non-availability of texts.

It must be acknowledged that our corpus is limited in size and in representativeness of political language. As it stands, this parallel corpus is a valuable asset; not only to political science and media specialists, but also to translation specialists.

3.2 Data Description

The present corpus consists of 351 Arabic and English original documents that were translated into the opposite direction. These documents fall into four categories: speeches, interviews, letters and one book.

Table 2. The data extracted from His Majesty's speeches, interviews and letters.

Text	**Speeches**	**Interviews**	**Letters**
Time range	1999-2015	1999-2015	1999-2015
Total number of translated texts	189	80	68
English source text	131	45	0
Arabic source text	58	35	68
Translators	Royal Court	Royal Court	Royal Court
Range of length (words)	250-2350	308-6403	200-3050
English words	200,767	210,384	45,385
Arabic words	177,444	163,140	39,730

Table 3. The King's book.

Title	**Our Last Best Chance: The Pursuit of Peace in a Time of Peril**
Publisher	Viking Press
Date of publication	2011
Place of publication	New York
No. of chapters	27
Chapter length (words)	2300-8194
English words	109491
Arabic title	فرصتنا الأخيرة: السعي نحو السلام في وقت الخطر
Publisher	Daralsaqi
Date of publication	2011
Place of publication	Beirut
Translator	Shukri Rahim
No. of chapters	27
Chapter length (words)	2300-8194
Arabic words	107634

The book was written in English and translated and published in Arabic. Tables 2 and 3 below describe the corpus texts.

3.3 Data Extraction

After obtaining permission from the Royal Hashemite Court to use the King's speeches and writings for non-commercial purposes, Arabic and English data were extracted from the official website of His Majesty King Abdullah II and from His book.

The texts on the website were initially saved in Microsoft Word document format and subsequently into utf-8 text only format. Likewise, The English version of the book was scanned and submitted to an optical character recognition (OCR) process to convert picture into text. This process is relatively fast, yet its product is not without inaccuracies. Therefore, the converted text had to be manually checked and edited. OCR was only possible for the English version of the book, but not for the Arabic one. Available OCR tools were not capable of satisfactorily converting the Arabic version of the book into any electronic text format. Thus, the Arabic version was manually retyped, checked and saved in text format to render it ready for machine processing.

3.4 Data Processing

Data processing involved six stages: metadata annotation, text segmentation, tokenization, alignment, stemming and POS tagging. These processes were completed primarily by the researchers with one checking and verifying what the other had done and ensuring consistency. On occasion, verification was sought from experts at the Arabic and English departments at the university of the researchers. As this verification was not methodical, it is the intention of the authors to conduct a systematic inter-annotator agreement study and to convert the current corpus into a gold standard that is thoroughly human-verified and validated.

3.4.1 Metadata Annotation

Metadata include text title, author, year, era, category, occasion, region or place and source language. All texts in the corpus were annotated with these eight metatags.

3.4.2 Text Segmentation

Segmentation is the process of splitting a text into smaller segments, such as paragraphs, sentences and clauses[27]. Segmenting a text allows search terms to find matches and renders texts ready for analysis [22].

Identification of sentence boundaries in the source texts (ST) and their matches in the target texts (TT) is a major challenge for the segmentation process. This is because the boundaries do not always correspond. The relations between text segments and their translations are not always in one to one correspondence. One sentence in one language might be translated into one sentence or two sentences, as illustrated in Table 4; or two sentences in ST might be translated into one in TT, as shown in Table 5. Besides, there are many examples where a clause corresponds to a sentence, as demonstrated in Table 6. Therefore, text segmentation is manually carried out at sentence, clause and phrase levels. This is done in order to obtain the best matching between ST and TT segments.

Sentence length is another issue that was taken into consideration during the segmentation process. Many lengthy sentences in ST corresponded to long ones in TT, so they were segmented into smaller meaningful chunks that would potentially recur in other texts. This was done after satisfying the best matching constraint. These chunks are similar to what Andrew Pawley calls 'conventionalized sentence stems'[28]. This makes our corpus of particular value to language learners, as they can easily recognize and learn authentic instances of sentence stems together with their translations.

3.4.3 Tokenization

Tokenization is carried out manually. Word boundary identification during the tokenization process is also a challenge. Idiomatic expressions are often treated as single dictionary entries; hence they are at the same rank as words. In many cases, two or more words are kept together as one token. This is

Table 4. One ST sentence corresponds to two TT sentences.

Source	Line	Arabic sentence (ST)	English sentence (TT)
Letter of re-designation to Ali Abul Ragheb	7	وقد تلقيت كتاب استقالتك الذي يعبر عما عرفته فيك من ولاء وانتماء وحرص على النهوض بالواجب وتحمل المسؤولية بإخلاص وتميز في الأداء وقدرة على تحقيق الإنجاز في إطار من العمل المؤسسي المستند إلى قواعد المعرفة ومواكبة روح العصر.	I received your letter of resignation, in which you articulate what is well known to us of your loyalty, of your sense of belonging and of responsibility, sincerity and excellence. Such performance has been demonstrated in the execution of your duties.

Table 5. Two ST sentences correspond to one TT sentence.

Source	Line	English sentence (ST)	Arabic sentence (TT)
The King's book Chapter 5	110	I went to study international relations at Pembroke College, Oxford. I spent a year among the grassy quads and honey-colored stone buildings of that venerable institution, studying Middle Eastern politics.	انتسبت إلى كلية "بمبروك" في جامعة أوكسفورد حيث أمضيت سنة أدرس العلاقات الدولية وسياسات الشرق الأوسط وسط تلك المربعات الخضراء الواسعة التي تحيط بها الأبنية التراثية ذات اللون العسلي، في ذلك الصرح العلمي والثقافي المهيب.

Table 6. One ST clause corresponds to one TT sentence.

Source	Line	Arabic clause (ST)	English sentence (TT)
Letter of re-designation to Ali Abul Ragheb	8	وإنني إذ أعرب عن عميق اعتزازي بما حققته هذه الحكومة من إنجازات وما تصدت له من تحديات على الصعيد الداخلي أو على الصعيد الإقليمي والدولي،	We herewith express our deep pride in what this government has accomplished and the challenges it has faced, whether locally, regionally or internationally.
	9	فإنني أتوجه بالشكر بشكل خاص لكل من عمل وأسهم في إنجاز الانتخابات البرلمانية التي أردناها غاية في النزاهة والموضوعية والشفافية.	In particular, we extend special gratitude to all those who worked and contributed to the successful completion of the parliamentary elections, which we wanted to be conducted with utmost integrity and transparency.

because a single token in Arabic might correspond to a phrase in English (e.g.,يُكَوّن corresponds to the phrase "*make up"*). On the other hand, a single token in English might correspond to two or more words in Arabic (e.g., "*cousins*" corresponds to أبناء العمومة).

3.4.4 Alignment

Accurate alignment is crucial for extracting information out of a parallel corpus. It enables users to easily and swiftly find equivalents of search terms or phrases. For example, when the user types a search term in one language, the concordance displays all occurrences of this word in that language. It also displays all the aligned equivalents in the target language. The results can be then extracted and analyzed [29].

To automatically align the corpus texts, SDL Trados WinAlign 2011 was used as an alignment tool. However, the output was unsatisfactory. Furthermore, WinAlign altered the predesignated text segmentation. Given the particular importance of alignment to the parallel corpus, it is regarded essential to manually verify its accuracy.

The text alignment was carried out at sentence, clause, phrase and word levels. However, it was not a straightforward process either. There were in the data some instances where lines in ST were left untranslated (see Table 7). Similarly, some texts in TT were inserted without any correspondence texts in ST (see Table 7). These phenomena created alignment problems.

Table 7. Examples of untranslated lines.

Line	**ST** (His Majesty's Speech at the Opening Session of the World Economic Forum, 20-May-05)	**TT**
1	No English equivalent	السلام عليكم، وأهلاً بكم في الأردن
2	Thank you Professor Schwab.	No Arabic equivalent
3	And thank you all	والشكر لكم جميعاً

To solve these problems, the lines with no equivalents were either deleted or attached to a neighboring line if they had significant contribution to the text.

The words in ST were manually matched to their equivalent words or phrases in TT to create a bilingual terminology list which would be useful for compiling bilingual dictionaries. Alignment at word level was more complex due to word order differences between Arabic and English. Words with no equivalents were left unaligned.

3.4.5 Stemming and Lemmatization

Stemming and lemmatization are beneficial for information retrieval. They reduce multiple word forms to a single word type. This multiplies the chance that the corpus users find more words that are morphologically related to a search term.

The Arabic light stems, roots and lemmas were automatically extracted using MADAMIRA. Its accuracy was reported in the literature to be respectable [18]. Then, the lemmatized texts were manually verified. English words were stemmed using the Porter stemmer [30].

3.4.6 Part of Speech Tagging

POS is of paramount concern to the linguist who wants to know how language works and how it is used. Hence, corpora are often annotated with POS. Hunston (2002) argued that POS tagging is a fundamental step in corpus exploration [5]. She stated four reasons for this. First, POS tags allow the search to be restricted to specific POS instances of a given word (e.g. searching the corpus for the noun instances of the word *play*). Second, they help identify the frequent collocates of a search term (e.g. *outdoor, creative, imaginative and pretend* usually collocate with the noun *play*, while *brilliantly, excellently, superbly, well and badly* collocate with the verb *play*). Third, POS tags allow the frequency comparison between words in different categories or genres within the corpus. Finally, POS tags enable researchers to discover the common word-class in each corpus category.

In the present corpus, Arabic words were automatically POS-tagged using MADAMIRA[18] and then manually verified.

3.4.7 Corpus Structure

EAPPC consists of Arabic and English sub-corpora, each of which is further divided into two sub-corpora that contain ST and TT in each language, as illustrated in Figure1.

4. BUILDING A PARALLEL CONCORDANCER

We built a parallel concordancer which consisted of two parts: the application through which the end-

user interacts with the corpus as shown in Figure 2 and the database which stores the parallel corpus (See Figure 3).

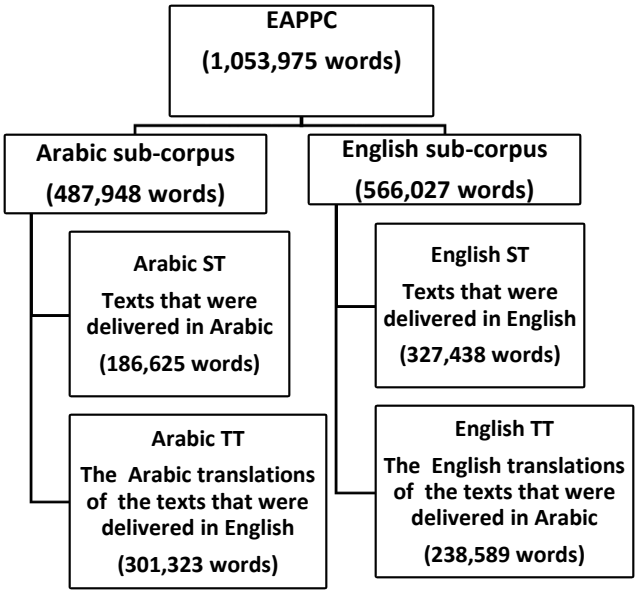


Figure 1. EAPPC structure.

Figure 2. Parallel concordancer interface.

The first step in building EAPPC was the preparation of the annotated texts. All the required data were manually arranged in three Excel sheets and then exported to a relational database as shown in Figure 3.

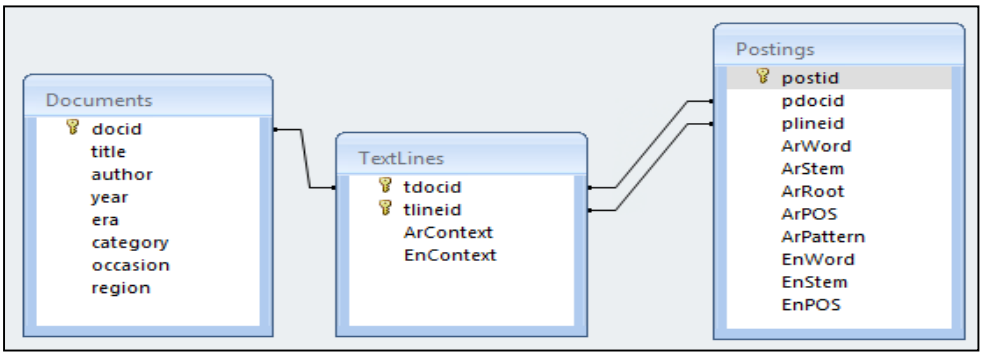


Figure 3. EAPPC relational database.

This relational database consists of three tables:

1. Documents Table, which stores information about each collected document (1 entry per document). A Document can have many text lines.
2. Text Lines Table, which stores the words of each text line in a document. Each entry (1 entry per text line) contains an ST line and its aligned equivalent TT line. A text line of each document can have many postings.
3. Postings Table, which stores information about every word in each text line (i.e., SL and TL,

English and Arabic). This includes the word itself, its stem, root, lemma and part of speech tag.

The concordance was developed using Eclipse as an integrated development environment running the Java 1.8 programming language. Both Eclipse and Java are open source software packages available for free.

5. CORPUS EXPERIMENTATION

Parallel corpora can be used to study different aspects of language, such as the features of source and target languages, the influence of SL on TL, the translation strategies used and the ideology and style of individual translators [31]-[34]. Furthermore, translators may learn strategies from parallel corpora and use them in their translation tasks [35].

Translating political texts has been labeled a complex activity [36]-[38]. Translators of such texts attempt to maintain the ideological and cultural aspects of the ST during the translation process [36], [38]. Hence, translators need to use the translation strategies and techniques that would enable them to preserve the ideology of the SL text and to cope with translation problems that surface during the translation process [38]-[39]. One of the problematic issues that often encounter translators is the translation of non-equivalence, particularly when a given concept is either unknown in the TL (a culture-specific concept) or known but is unlexicalized. Baker (1992) listed seven strategies that translators use to render non-equivalence [39]:

1. Using a hypernym.
2. Using a more neutral/ less expressive term.
3. Cultural substitution.
4. Using a loan word or a loan word plus explanation.
5. Paraphrasing.
6. Illustration and exemplification.
7. Omission of the problematic concept.

In order to demonstrate the EAPPC's utility in translation studies, the corpus was used to investigate how different Royal Court translators rendered the Arabic and Islamic culture-specific terms تكفير (takfīr)and تكفيري (takfīrī) in the speeches, interviews, letters and book of His Majesty King Abdullah II, since this is an Arabic term that illustrates non-equivalence in English.

5.1 Methodology

This research uses EAPPC, classical and modern Arabic dictionaries, *such as,* Al-ʿayn العين*(786* CE*),* Muʿjam Maqāyīs Al-luġah(1004 CE), Al- mufradāt fī Ġarīb Al-qurʾān(1109 CE), *Lisān al-ʿarab*(1311 CE) andMuʿjam Al-luġa Al-ʿarabiyah Al-muʿāsirah(2003 CE)[2].

First, the Arabic dictionaries were consulted in order to determine the meaning of the word تكفير (takfīr) and تكفيري (takfīrī). Next, the EAPPC corpus was explored using the Arabic root كفر (kfr) as a query term. The parallel corpus concordance displayed, in the keyword in context (KWIC) style, all occurrences of the term along with their translations. However, the query terms are not highlighted so that the user can easily and swiftly identify them on the screen, but we are working on a better version and it will be highlighted. Moreover, information about the text, if it is a source or a target one, is not explicitly provided. The displayed data were then exported to an Excel spreadsheet and analyzed. Finally, the study used Baker's taxonomy of translation strategies to analyze the data and to discover the adopted translation strategies.

5.2 Findings and Discussion

The term تكفير (takfīr) is an abstract noun derived from the verb of intensification, kaffara, while the term تكفيري (takfīrī), often used as a substantive adjective (i.e., a noun), is derived from the nounتكفير (takfīr). *Takfīr* has multiple senses in dictionaries. See Table 8.

[2]http://lisaan.net.

Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 3, No. 3, December 2017.

The majority of these senses have changed over time and only one sense has survived. Table 8 shows تكفير (takfīr) to have had the meanings of abasement, submissiveness, obeisance, wearing armor, expiation of sins, nodding and enthronement. Only expiation, however, has survived the ravages of time. Additionally, *Muʿjam Al-luġa Al-ʿarabiyah Al-muʿāsirah* reflects our modern conception of تكفيري (takfīrī) as the "attribute of ascribing apostasy to others".

Table 8. Senses of the term تكفير (takfīr).

Sense	Dictionary
إيْمَاءُ الذِّمي برأسِه "Nodding"	Al-ʿayn العين
تَتْوِيْجُ المَلِكِ بتاج "Enthronement"	Al-ʿayn العين
الذّلّ والخضوع "Abasement, Submissiveness"	Lisān al-ʿarab
الانحناءُ الشَّدِيد "Bowing"	Lisān al-ʿarab
ستر الشيء وتغطيته "Covering"	Al-mufradāt fī Ġarīb Al-qurʾān
أن يَتَكَفَّرَ المُحاربُ في سلاحه "Covering the body with weapons"	Lisān al-ʿarab
تُكَفِّر الخطيئة أي تمحوها "Expiation"	Lisān al-ʿarab
التَّكْفير: أن يَخْضَعَ الإنسانُ لغَيْره وينحَني ويُطَأطِئ رَأسَه قَرِيبا من الرُّكوع "Offering obeisance"	Lisān al-ʿarab
جماعة تكفيرية: جماعة متشدِّدة تنسب العصاة والمذنبين إلى الكُفر، أو عدم الإيمان بالله، أو الزّندقة "Takfīrī group: extremists who call people apostates"	Muʿjam Al-luġa Al-ʿarabiyah Al-muʿāsirah

Using EAPPC,we searched for the Arabic root (كفر) "kfr" in order to retrieve all the occurrences of the terms تكفير and تكفيري. Seventy-seven instances of derivatives of this root have been used by His Majesty. Their distribution in the corpus is shown in Figure 4 below.

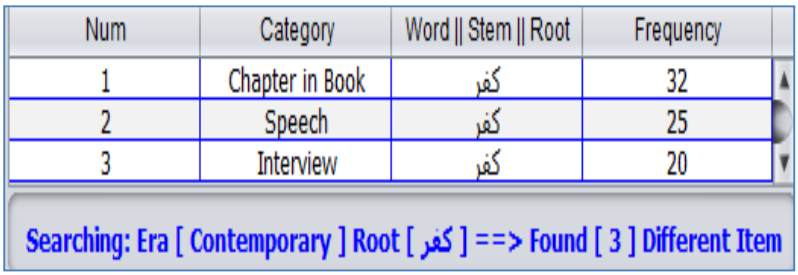


Figure 4. Search results of the root كفر (kfr) in the parallel corpus.

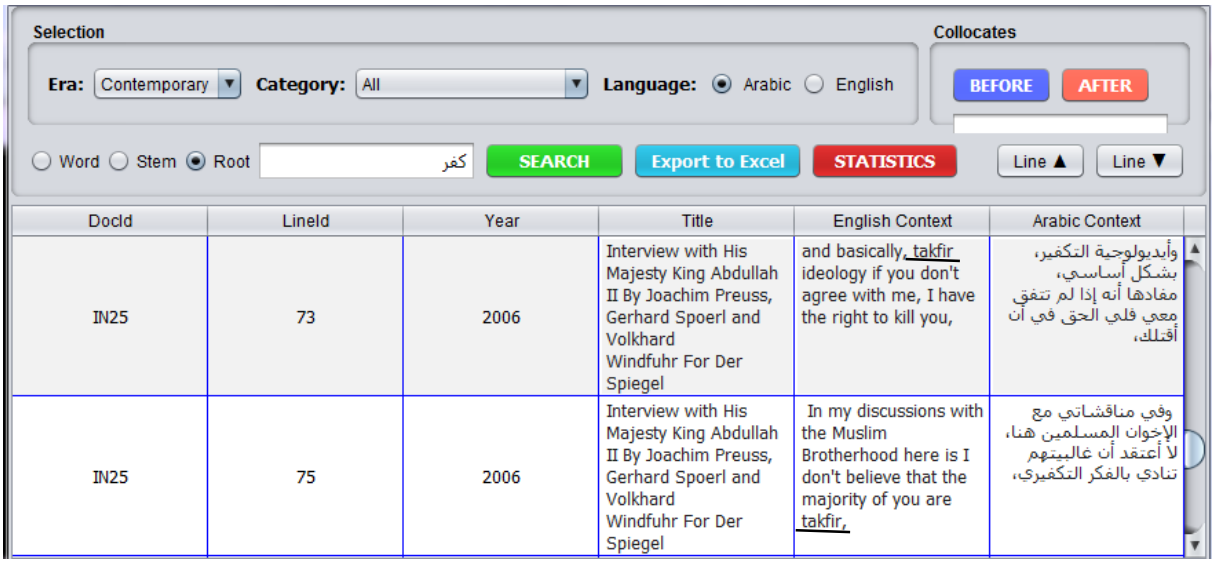
Seventy instances of these relate to the terms تكفير and تكفيري as shown in Table 9.

Table 9. Distribution of تكفير (takfīr) and تكفيري (takfīrī) instances in the parallel corpus.

Term	Speeches	Interviews	Book	Total
تكفير takfīr	21	6	3	**30**
تكفيريين takfīrīyīn/تكفيريون takfīrīyūn	-	3	24	**27**
تكفيري takfīrī	1	8	-	**9**
تكفيرية takfīrīyīah	-	-	4	**4**
Total	**22**	**17**	**31**	**70**

In order to examine how each instance of the terms تكفير (takfīr) and تكفيري (takfīrī) were rendered in Arabic and English, it is important to discover the translation strategies that were used by the King's translators. In English ST, the terms تكفير (takfīr) and تكفيري (takfīrī) were used as loanwords as illustrated in Figure 5.

The term تكفير (takfīr) in the Arabic ST component of the EAPPC occurs three times in one interview and 15 times in five speeches. In the English ST component of this corpus, the loan word *takfīr* occurs three times in one interview, six times in six speeches and three times in one chapter in the book.



DocId	LineId	Year	Title	English Context	Arabic Context
IN25	73	2006	Interview with His Majesty King Abdullah II By Joachim Preuss, Gerhard Spoerl and Volkhard Windfuhr For Der Spiegel	and basically, takfir ideology if you don't agree with me, I have the right to kill you,	وأيديولوجية التكفير، بشكل أساسي، مفادها أنه إذا لم تتفق معي فلي الحق في أن أقتلك،
IN25	75	2006	Interview with His Majesty King Abdullah II By Joachim Preuss, Gerhard Spoerl and Volkhard Windfuhr For Der Spiegel	In my discussions with the Muslim Brotherhood here is I don't believe that the majority of you are takfir,	وفي مناقشاتي مع الإخوان المسلمين هنا، لا أعتقد أن غالبيتهم تنادي بالفكر التكفيري،

Figure 5. Examples of the term تكفير (takfīr) and تكفيري (takfīrī) in His Majesty's English ST.

EAPPC evidence shows that translators adopted these strategies when rendering تكفير (takfīr) from Arabic into English: the use of loan words, loan words plus explanation, English equivalents and English equivalents with the TL terms between brackets. In many instances, translators would introduce the loan word *takfīr* with an explanation (e.g. calling others apostates) and then use it without explanation in subsequent occurrences, as shown in the following example from an interview given by His Majesty on 22 April 2006 to *Al Sabah Al Jadid Newspaper:*

ST (Arabic):

- "كما حظي بتوافق إجماعي يدين ممارسات **التكفير** التي يلجأ إليها المتطرفون لتبرير العنف."
- "ولأننا نقف ضد التطرف **والتكفير**، فقد أصبحنا مستهدفين من الجماعات الإرهابية في العراق".

TT (English)

- "This declaration condemned the practice of **takfīr (calling others apostates)** that extremists use to justify violence."
- "And because we stand against extremism and **takfīr**, we have become targets of terrorist groups in Iraq,"

Another strategy for rendering تكفير (takfīr) from Arabic into English is translation by TL equivalents, as illustrated in the following examples from His Majesty's speech at *the opening session of the International Islamic Conference* on 4 July 2005:

ST (Arabic):

"وعدم جواز **تكفير** أي مسلم من أتباعها."

TT (English):

"and that **declaring any one of them an apostate** is unacceptable.

Another strategy is using a TL equivalent with the loan word between brackets. For example, translators paraphrased تكفير (takfīr) as *apostasy* and used the loan word *takfīr* between brackets, as demonstrated in the following example from His Majesty's speech at the opening of the third extraordinary session of the *Islamic Summit* on 7 December 2005:

ST (Arabic):

"لأن عدم الاتفاق على هاتين المسألتين هو سبب الفرقة والاختلاف وتبادل تهم **التكفير**والاقتتال بين أبناء الدين الواحد"

TT (English):

"The absence of consensus on these two issues has led to divisions and differences, accusations of **apostasy (takfir)** and internecine fighting.".

Analysis shows that translators tended to use English equivalents strategy most often when rendering the term تكفير (takfīr) from Arabic into English as shown in Table 10.

Table 10. Frequencies of تكفير (takfīr) translation strategies from Arabic into English.

Translation Strategy	Frequency
The use of loan words	2
Loan words plus explanation	2
English equivalents	12
English equivalents with the TL terms between brackets	2
Total	**18**

The term *takfīrī* is the adjectival form of *takfīr* that is often used as a noun. The corpus offers 40 such instances. Twenty-eight of them occur as loan words in His Majesty's book and one in an English interview. In the subcorpus of Arabic STs, on the other hand, *takfīrī occurs* 10 times in four interviews and once in a speech. Moreover, the corpus evidence shows that thirteen occurrences of the term تكفيري (takfīrī) in the Arabic subcorpus are adjectives and twenty-seven are nouns.

The loan word *takfīrī* in English ST texts was rendered as تكفيري (takfīrī), تكفيرية (takfīrīyah), تكفيريين (takfīrīyīn) or (takfīrīyūn) تكفيريون in Arabic in accordance with the requirements of syntactic inflection.

In some instances, the word *takfīrī* is not found in the ST text but the translator understood that it was intended. In such a case, the translator made it explicit by using *takfīrī* in the TL text, as shown in the following example from His Majesty's book, *Our Last Best Chance (2011):*

ST (English):
"and we helped the Americans understand **what** to look for"
TT (Arabic):
"وقد ساعدنا الأمريكيين في التعرف على **التكفيريين**."

'and we helped the Americans recognize **takfīrī s**'

In other cases, the author referred to the word *takfīrī* by using an anaphoric pronoun. In this case, the translators explicitly used the equivalent word تكفيري (takfīrī), as illustrated in the following example from His Majesty's book, *Our Last Best Chance(2011):*

ST (English):
"Islam celebrates life; **they** seek to destroy it."
TT (Arabic):
"فإذ يحترم الإسلام الحياة الإنسانية ويصونها، لا يتردد **التكفيريون** في تدميرها والقضاء عليها.".

'Even though Islam respects and protects human life, **takfiris** do not hesitate to destroy it and quell it.'

Although the term تكفير (takfīr) has been translated into English using multiple strategies, تكفيري (takfīrī) has only been rendered using the loan word strategy, as shown in His Majesty's interview on 22 April 2006 given to *Al Sabah Al Jadid Newspaper*:

ST (Arabic):
"وجد الفكر **التكفيري** ما يغذي أهدافه البعيدة كل البعد عن قيم الإسلام الحقيقية".

TT (English):
"Takfiri thought found feeding ground for its aims that are alien to true Islamic ethics and values."

To sum up, translators tended to render the loan words *takfīr* and *takfīrī* from English into Arabic by using the same terms as they are Arabic in the first place. On the other hand, when they translated into English they employed several strategies: translation using loan words, loan words plus explanation, translation by TL equivalence and translation by equivalents with the loan word between brackets.

6. CONCLUSION

This study has described the construction of EAPPC. The ultimate aim of EAPPC is to provide translators, learners, educators, researchers and language engineers with a freely available tagged parallel corpus whose annotation has been manually verified. To illustrate its utility, we have carried out an experiment that examined the translation strategies used in rendering a culture-specific term. The

results demonstrated the ease with which knowledge about translation strategies can be gained from this parallel corpus.

REFERENCES

[1] J. Sinclair, Corpus, Concordance, Collocation, Oxford University Press, 1991.

[2] M. Baker, "Corpora in Translation Studies: An Overview and Some Suggestions for Future Research," Target, vol. 7, pp. 223-243, 1995.

[3] D. Biber, S. Conrad and R. Reppen, Corpus linguistics: Investigating Language Structure and Use, Cambridge University Press, 1998.

[4] L. Bowker, "Towards a Methodology for a Corpus-based Approach to Translation Evaluation," Meta: Journal des traducteurs Meta:/Translators' Journal, vol. 46, pp. 345-364, 2001.

[5] S. Hunston, Corpora in Applied Linguistics, Cambridge: Cambridge University Press, 2002.

[6] F. Zanettin, S. Bernardini and D. Stewart, Corpora in Translator Education, London, 2014.

[7] I. Ulitkin, "Computer-assisted Translation Tools: A Brief Review," Translation Journal, vol. 15, 2011.

[8] F. J. Och, "Statistical Machine Translation: Foundations and Recent Advances," presented at the Tutorial at MT Summit, Phuket, Thailand, 2005.

[9] L. Bowker and J. Pearson, Working with Specialized Language: A Practical Guide to Using Corpora, London, NY: Routledge, 2002.

[10] S. Goodman and K. O'Halloran, The Art of English: Literary Creativity, Basingstoke, UK: Palgrave Macmillan, 2006.

[11] K. Aijmer, B. Altenberg and M. Johansson, Languages in Contrast: Papers from a Symposium on Text-based Cross-linguistic Studies, Lund 4-5 March 1994, vol. 88, Lund Studies in English, 1996.

[12] D. Samy, A. Moreno Sandoval and J. M. Guirao, "An Alignment Experiment of a Spanish-Arabic Parallel Corpus," Proceedings of the International Conference on Arabic Language Resources and Tools (NEMLAR 2004), pp. 85-89, 2004.

[13] D. Samy, A. M. Sandoval, J. M. Guirao and E. Alfonseca, "Building a Parallel Multilingual Corpus (Arabic-Spanish-English)," Proceedings of the 5th Intl. Conf. on Language Resources and Evaluations (LREC), 2006.

[14] M. Tadić, "Building the Croatian-English Parallel Corpus," The 5th International Conference on Language Resources and Evaluation (LREC'2006), pp.523-530, Athens, Greece, 2000.

[15] S. Singh, T. McEnery and P. Baker, "Building a Parallel Corpus of English/Panjabi," Parallel Text Processing, ed: Springer, pp. 335-346, 2000.

[16] L. Rura, W. Vandeweghe and M. Montero Perez, "Designing a Parallel Corpus as a Multifunctional Translator's Aid," in XVIII FIT World Congress= XVIIIe Congrès mondial de la FIT, 2008.

[17] H. Al-Ajmi, "A New English–Arabic Parallel Text Corpus for Lexicographic Applications," Lexikos, vol. 14, pp. 326-330, 2004.

[18] A. Pasha, M. Al-Badrashiny, M. T. Diab, A. El Kholy, R. Eskander, N. Habash, et al., "MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic," LREC, pp. 1094-1101, 2014.

[19] J. Véronis, "From the Rosetta Stone to the Information Society," Parallel Text Processing, ed: Springer, pp. 1-24, 2000.

[20] L. Al-Sulaiti and E. Atwell, Designing and Developing a Corpus of Contemporary Arabic, MA Thesis, School of Computing, University of Leeds, UK, 2004.

[21] H. Hasselgård, "Contrastive Analysis / Contrastive Linguistics," The Routledge Linguistics Encyclopedia, K. Malmkjær, Ed., Third Edition, London, NY: Routledge, pp. 98-101, 2010.

[22] G. R. Yepes, "Parallel Corpora in Translator Education," Redit: Revista Electrónica de Didáctica de la Traducción y la Interpretación, pp. 65-80, 2011.

[23] J. Tiedemann and L. Nygaard, "The OPUS Corpus-Parallel & Free," Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04), Lisbon, Portugal, 2004.

[24] M. S. S. Sawalha, Open-source Resources and Standards for Arabic Word Structure Analysis: Fine Grained Morphological Analysis of Arabic Text Corpora, PhD Thesis, University of Leeds, 2011.

[25] K. Dukes and N. Habash, "Morphological Annotation of Quranic Arabic," in LREC'10, Malta, 2010.

[26] H. Salhi, "Investigating the Complementary Polysemy and the Arabic Translations of the Noun Destruction in EAPCOUNT," Meta: Journal des traducteurs Meta:/Translators' Journal, vol. 58, pp. 227-246, 2013.

[27] P. Baker, A. Hardie and T. McEnery, A Glossary of Corpus Linguistics: Edinburgh Uni. Press, 2006.

[28] A. Pawley and F. H. Syder, "Two Puzzles for Linguistic Theory: Nativelike Selection and Nativelike Fluency," Language and Communication, vol. 191, p. 225, 1983.

[29] M. Ghadessy, A. Henry and R. L. Roseberry, Small Corpus Studies and ELT: Theory and Practice, SCL 5, Philadelphia, USA: John Benjamins Publishing Co., 2001.

[30] M. F. Porter, "An Algorithm for Suffix Stripping," Program, vol. 14, pp. 130-137, 1980.

[31] M. Baker, "Corpus Linguistics and Translation Studies: Implications and Applications," Text and Technology: In honour of John Sinclair, M. Baker, G. Francis and E. Tognini Bonelli, Eds., Philadelphia, USA: John Benjamins Publishing Co., 1993.

[32] S. Laviosa, Corpus-based Translation Studies: Theory, Findings, Applications, Amsterdam – New York, NY: Rodopi B.V, 2002.

[33] M. Olohan, Introducing Corpora in Translation Studies, London, NY: Routledge, 2004.

[34] C. Fantinuoli and F. Zanettin, New Directions in Corpus-based Translation Studies, Berlin: Language Science Press, 2015.

[35] G. Shen, "Corpus-based Approaches to Translation Studies," Cross-Cultural Communication, vol. 6, pp. 181-187, 2011.

[36] A. Shunnaq, "Arabic-English Translation of Political Speeches," Perspectives: Studies in Translatology, vol. 8, pp. 207-228, 2000.

[37] G. Quentel, "Translating a Crucial Political Speech," ed: Retrieved October, 2006.

[38] K. Sárosi-Márdirosz, "Problems Related to the Translation of Political Texts," Acta Universitatis Sapientiae Philologica, pp. 159-180, 2014.

[39] M. Baker, In Other Words: A Coursebook on Translation, London, NY: Routledge, 1992.

ملخص البحث:

تتناول هذه الدراسة إنشاء مجموعة كاملة من مليون كلمة بالعربية والإنجليزية في حقل السياسة، من شأنها أن تكون مصدراً مفيداً للبحث في دراسات الترجمة، وتعلُّم اللغة وتعليمها، وعلم المعاجم ثنائية اللغة، والدراسات المقارِنة، ودراسات العلوم السياسية، واسترجاع المعلومات المتعلقة بتقاطع اللغات. وتصف الدراسة مراحل إنشاء المجموعة، وشرحها عن طريق الأمثلة، من أجل اكتشاف استراتيجيات الترجمة المستخدمة في ترجمة المصطلحين "تكفير" و"تكفيري" المتعلقين بالثقافة العربية والإسلامية من العربية الى الإنجليزية وبالعكس. تتكون المجموعة من 351 وثيقة أصلية بالعربية والإنجليزية وترجماتها. فقد تم اختيار 189 خطاباً و80 مقابلة و68 رسالة، ترجمها مترجمون في الديوان الملكي الهاشمي، من الموقع الإلكتروني الرّسمي لجلالة الملك عبدالله الثاني، إضافة الى نصّ كل من النسختين الإنجليزية والعربية لكتاب جلالته المعنون: "فرصتنا الأخيرة: السعي نحو السلام في زمن الخطر" الصادر عام 2011. وبعد استكمال مراحل إنشاء المجموعة، جرى إعداد فهرس أبجدي لتسهيل عملية تحرّي المجموعة. أما أبرز التحديات فقد تمثلت في ندرة النصوص المترجمة من العربية الى الإنجليزية وبالعكس القابلة للقراءة بواسطة الآلة، إضافة الى نقص الأدوات القادرة على معالجة النصوص العربية.

Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 3, No. 3, December 2017.

A BINARY CLASSIFIER BASED ON FIREFLY ALGORITHM

Raed Z. Al-Abdallah[1], Ameera S. Jaradat[1], Iyad Abu Doush[2] and Yazan A. Jaradat[1]

(Received: 27-Jul.-2017, Revised: 23-Sep.-2017, Accepted: 11-Oct.-2017)

ABSTRACT

This work implements the Firefly algorithm (FA) to find the best decision hyper-plane in the feature space. The proposed classifier uses a cross-validation of a 10-fold portioning for the training and the testing phases used for classification. Five pattern recognition binary benchmark problems with different feature vector dimensions are used to demonstrate the effectiveness of the proposed classifier. We compare the FA classifier results with those of other approaches through two experiments. The experimental results indicated that FA classifier is a competitive classification technique. The FA shows better results in three out of the four tested datasets used in the second experiment.

KEYWORDS

Swarm-based algorithms, Binary classification problems, Firefly algorithms.

1. INTRODUCTION

Classification means using the characteristics of an object to identify to which set of predefined classes it belongs. The classification problem has many applications, such as: medical diagnosis, news filtering, document retrieval, opinion mining, email classification and spam filtering. Binary classification is the problem of classifying a new input instance to be in one of two classes. For example, a received email is classified into either a spam or a non-spam. Another example of binary classification is when a patient is diagnosed either to be infected or not infected with a specific disease [1].

A classification technique can use a training dataset to learn how it can classify new instances. The training dataset consists of a set of training examples. Each example is a pair of an input vector of features and the desired output class value. The classification has two phases; a learning phase and a testing phase. In the learning phase, the category of an instance is identified according to its closeness of instances in the training data. A classification model is usually generated by identifying the feature values in the training data instance to one of the predefined class labels. In the testing phase, this classification model is used to explicitly select a particular class for the new instance data.

A classification algorithm can be used to identify a mathematical model to classify a new instance. A linear classification is a type of classification that uses a polynomial function of degree one to classify the new data.This linear degree function is a hyper-plane. In general, a hyper-plane of n-1 dimension is the separator in the n-dimensional space. For example, in the 3-dimensional space, the hyper-plane becomes the 2-dimensional plane. In the 2-dimensional space, the hyper-plane becomes the1-dimensional line. This hyper-plane in binary classification is used to separate the data samples in two different places in the feature space. Equation (1) below shows the general hyper-plane with n dimensions [2]:

$$w_1 x_1 + w_2 x_2 + \cdots + w_n x_n + w_{n+1} x_{n+1} = 0\,; \tag{1}$$

where $\mathrm{w_j} = (\mathrm{w_1}, \mathrm{w_2}, \ldots, \mathrm{w_n}, \mathrm{w_{n+1}})$ is the weight vector and $\mathrm{x} = (\mathrm{x_1}, \mathrm{x_2}, \ldots, \mathrm{x_n}, \mathrm{x_{n+1}})$ is the feature vector. Note that n is the dimensions of the hyper-plane and it represents the number of features. In the linear classification, we use a hyper-plane to split the data between two classes.The points in the feature space of n dimensions which are located above the hyper-plane belong to one class and the

1. R. Z. Al-Abdallah, A. S. Jaradat and Y. A. Jaradat are with the Department of Computer Science,Yarmouk University,Irbid, Jordan.
2. I. Abu Doush is with the Department of Computer Science and Information Systems, American University of Kuwait, Salmiya, Kuwait.

Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 3, No. 3, December 2017.

other points that are located below the hyper-plane belong to the second class. All points in the feature space of n dimensions that are located in half spaces above the hyper-plane are mathematically greater than zero when their values are substituted in Equation (1). In other words, they are all points that satisfy the inequality (2) to true while all other points that are located in half spaces below the hyper-plane are mathematically lesser than zero when their values are substituted in Equation 1.

$$w_1x_1 + w_2x_2 + \cdots + w_nx_n + w_{n+1}x_{n+1} > 0 \quad (2)$$

A binary linear classifier is formulated mathematically using Equation 3.

$$h(x, w, w_0) = sign(x.w + w_0)\ ; \quad (3)$$

where x is the features vector, w is the weights vector and w_0 is the base of the hyper-plane. Note that *sign* is a function which returns 1 when $x.w + w_0 > 0$ and -1 when $x.w + w_0 < 0$. The (.) is dot vectors multiplication. Algorithm (1) shows a pseudo-code to classify data with two categories using the hyper-plane equation function. Class A represents one class and class B represents the other class in the binary classifier.

Algorithm 1: Linear Classifier

1: $x = \sum_{i=1}^{n} (w_i x_i)$
2: **If** x > 0 **then**
3: **Return** Class A
4: **Else If** x < 0 **then**
5: **Return** Class B
6: **End if**

Algorithm 1. Linear classifier.

The algorithm first calculates the instance features values using Equation (1). The obtained result is used to locate the instance on the hyper-plane by comparing it with zero as a threshold.
The main classification methods can be categorized as follows:

- **Decision Trees**: a decision tree is a hierarchical decomposition of training data. A decision tree is made of decision nodes and leaf nodes. Each decision node has a condition over an attribute value. This condition is used to divide the data space into a number of branches. A leaf node represents a class that represents the decision result. The decision tree is used to classify testing data [3].
- **Rule-based Classifiers**: in a rule-based classifier, a set of IF-THEN rules are used for which the left hand side (LHS) is a condition on the feature set and the right hand side (RHS) is the predicted class label. For a given test instance, we determine the set of rules for which the test instance satisfies the condition on the LHS of the rule, then use them to determine the predicted class. Sequential Covering Algorithms (SCA) strategy is the most used strategy to induce rules from the training data. It learns a rule from a training set (conquer step), then removes from the training set the examples covered by the rule (separate step) and recursively learns another rule which covers the remaining examples [4].
- **Support Vector Machine (SVM) Classifiers**: SVM classifiers attempt to partition the data space with the use of linear or non-linear drawing between the different classes. The goal in such classifiers is to find the optimal boundaries between the different classes. In linear SVM, the optimal hyper-plane is the one that minimizes the accuracy error and maximizs the geometric margin. The geometric margin represents the minimum distance of the training samples of both classes from the separating hyper-plane [5].
- **Neural Network Classifiers**: a neural network (NN) classifier consists of units arranged in layers. Each unit takes an input, applies a liner or nonlinear function to it and then passes the output to the next layer. Each node is consisting of a set of input values (xi) and associated weights (wi). These weightings are tuned in the training phase to adapt a neural network in the learning phase [6].
- **Bayesian Classifiers:** Bayesian classifiers are probabilistic classifiers which apply Bayes' theorem. This model is then used to predict the classification of a new instance. The simplest

Bayesian classifier is the naive Bayesian classifier (NBC), which assumes that the input features are conditionally independent of each other [7].

Swarm intelligence algorithms imitate the behaviour of the swarm to obtain the optimal solution for different kinds of problems [1]. It is inspired by the collective behavior of swarms (e.g., ants, bees and a flock of birds). In the swarm, each agent interacts with other agents in a self-organizing behavior. Examples of such algorithms are: particle swarm optimization, firefly algorithm, bat algorithm and ant colony optimization [8], [22] and [32].

One of the recent swarm intelligence algorithms is firefly algorithm (FA). In this paper, FA is used as a binary linear classifier. It is used as a search algorithm to find the best weight vector of the hyper-plane classifier. The proposed algorithm is compared with five of the state of the art algorithms in terms of accuracy. The rest of this paper is organized as follows: Section 2 presents firefly algorithm. Section 3 shows a literature review of some of the classification techniques. The methodology used and the proposed firefly classifier are described Section 4. Section 5 considers the experimental results using five binary datasets. The discussion and a comparison with the state of the art methods are presented in section 6. Finally, we sketch the conclusion and the future work in Section 7.

2. FIREFLY ALGORITHM

FA is inspired by the flashing behaviour in the matting phase of fireflies' life cycle in nature. It is developed by Xin-She Yang at Cambridge University in late 2008 [9]. The fundamental function of flashing light in fireflies is to attract a mate. A male or female firefly light glows brighter in order to make itself more attractive for a mate. The FA algorithm is presented in algorithm (2). FA uses the following three rules [11]:

- A firefly is attracted to other fireflies regardless of their sex, because all fireflies are unisex.
- Attractiveness is proportional to their brightness, thus for any two flashing fireflies, the less bright one will move towards the brighter one. Both attractiveness and brightness are decreasing as the distance between the two fireflies increases. If no one is brighter than a particular firefly, then it moves randomly.
- The brightness or light intensity of a firefly is determined by the objective function of the optimization problem.

Algorithm 2: Firefly Algorithm	
1:	Initialize parameters α, β, γ, t=0, Bs=0
2:	$P^{(0)} = InitializeFA()$ // Initialize Randomly Firefly population
3:	**While** t < Max-Iteration **do**
4:	FitnessFA($P^{(t)}$) // calculate fitness value for each solution
5:	Bs = BestFA($P^{(t)}$) // order population then find best solution
6:	$P^{(t+1)} = MoveFA(P^{(t)})$ // Firefly movement
7:	t= t+1
8:	$\alpha^{(t)} = NewAlpha$ //calculate new alpha value
9:	**End While**
10:	**Output** Bs

Algorithm 2. Firefly algorithm [10].

In Algorithm 2, alpha (α) is the random movement parameter that controls the step length of the random movement, γ is the fixed light absorption coefficient, β is the brightness, t is the iteration number and Bs is the best solution.
InitializeFA() function in line (2) is used for initializing the fireflies' population randomly, where each individual contains two attributes; a position and a fitness. The while-loop (lines 3-9) starts with the *FitnessFA* function in line (3) which is used to calculate the quality of all population solutions. Then, *BestFA* function in line (5) is used to sort the population of fireflies according to their fitness values. After that, the *MoveFA* function in line (6) is used to perform a move of the firefly position (the details are presented in algorithm (3)) [11]. Finally, the *NewAlpha* function in line (8) is used to decrease the initial value of parameter alpha (α) as the iteration increases. The firefly search process is repeated until we reach Max-Iteration steps. After the loop is terminated, the best solution is obtained [10].

Algorithm 3: $MoveFA(P^{(t)})$

1: **For** i = 1 **To** n **do** // n is population size
2: **For** j = 1 **To** n **do**
3: $x_i = P_i^{(t)}$.position // array of positions for firefly I at t iteration
4: $x_j = P_j^{(t)}$. Position
5: $P_i^{(t+1)} = P_i^{(t)}$
6: **If**(f(x_i)<f(x_j)) **then** //f is attractiveness function
7: $$r_{ij} = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2}$$
8: // move firefly i towards j
9: $x_i^{t+1} = x_i^t + \beta_0 e^{-\gamma r_{ij}^2}(x_j - x_i) + \alpha_t(rand - 0.5)$ // rand is a random number
10: $P_i^{(t+1)}.position = x_i^{t+1}$
11: **End if**
12: **End For**
13: **End For**
14: **Return**($P^{(t+1)}$)

Algorithm 3. Firefly movement.

Algorithm 3 shows the steps for the function *MoveFA*, where line (6) tests the attractiveness (brightness) between two fireflies using the fitness function to determine which firefly is moving and to which one. The details about the fitness function are shown in algorithm (4). The firefly with less brightness will move towards the brighter firefly. The r_{ij} in line (7) is the distance between any two fireflies i and j at $x_i$and x_j positions which is calculated using the Cartesian distance. x_{ik} is the k^{th} component of the spatial coordinate x_i vector of i^{th} firefly and x_{jk} is the k^{th} component of the spatial coordinate x_j vector of j^{th} firefly.

The new position x_i^{t+1} of the moving firefly i at t+1 iteration is calculated by line (9) where the step size of the moving firefly i depends on the last two terms which are added to the current position for firefly i at t iteration. The second term is used to control the step size due to the attraction of a firefly towards the intensity of the light (brightness) by neighboring fireflies. Brightness here is inversely proportional to the distance between the two fireflies due to exponential function characteristics.

The brightness is decreasing as the two firefly distance increases. The third term is a randomization vector of random variables, where α is the random movement parameter that controls the step length of the movement. Note that β_0 is the attraction factor at $r_{ij} = 0$ and γ is the light absorption coefficient. For most cases β_0 = 1, α ∈ [0, 1] and γ = 1 [11]. Finally, line 14 returns the new population after the movement phase is completed.

In this paper, we propose a novel FA to be used as a binary linear classifier. The hyper-plane that separates the data into two classes by searching for the best values of a weight vector using Equation (1) is used for the classification decision. The proposed classifier is evaluated on five datasets. Then, it is compared with other classification techniques [12,13]. We use FA for many reasons. Firstly, FA is efficient, because it does not need complex computations and has a limited number of parameters. Moreover, FA is a stochastic meta-heuristic algorithm that can be applied for solving the optimization problems. Thirdly, since firefly algorithm is population-based meta-heuristic, it improves multiple candidate solutions to guide the search [10].

3. LITERATURE REVIEW

There are many classification techniques to generate classifiers: particle swarm classifiers [12], decision tree classifiers [14] and artificial neural network (ANN) classifiers [15]. In this literature review, we present a summary of selected work on using swarm intelligence and ANN for classification.

Sousa et al. [12] proposed the use of Particle Swarm Optimizer (PSO) as a new tool for classification. Three different particle swarm algorithms were implemented and tested against genetic algorithm and tree induction algorithm (J48). The results proved that PSO is competitive when compared with the other techniques.

Zahiri and Seyedin [16] proposed an Intelligent Particle Swarm classifier (IPS classifier) for finding the decision hyper-plane to classify patterns of different classes in the feature space using PSO algorithm. The IPS classifier used an intelligent fuzzy controller which was designed to improve the performance and efficiency of the proposed classifier by adapting three important parameters of PSO (swarm size, neighbourhood size and constriction coefficient). Three pattern recognition problems with different feature vector dimensions were used to demonstrate the effectiveness of the proposed classifier. The experimental results showed that the performance of the IPS-classifier is comparable to or better than the k-nearest neighbour (k-NN) and multi-layer perception (MLP) classifiers. In another work, Martens et al. [17] proposed a new ant-based classification technique named AntMiner+. The key difference between the proposed AntMiner+ and the previous AntMiner versions is the use of a better performing MAX-MIN ant system. Furthermore, AntMiner+ controlled the commonly encountered problem in Ant Colony Optimization (ACO), which is setting the parameters as the new method automatically sets the algorithm parameters. The experiments showed that AntMiner+ accuracy is superior when compared to the other AntMiner versions and that its results are competitive or better than the results achieved by other classification techniques.

Assarzadeh et al. [18] introduced Harmony Search Algorithm-based classifier. The experimental results showed that the performance of the HS-classifier is better than the k-nearest neighbour classifier, particle swarm, genetic algorithm and imperialist competitive algorithm-based classifier. Mantas and Abellán [13] presented a modified version of C4.5, called Credal-C4.5. The modified version of C4.5 used an imprecise probability based on mathematical theory and uncertainty measures. Credal-C4.5 estimated the features probabilities and the class variable using imprecise probabilities. It used imprecise information gain ratio which is a new split criterion. Credal-C4.5 built smaller and better performance trees than the classic C4.5 classifier.

Gandomi et al. [19] introduced FA for solving mixed continuous/discrete structural optimization problems taken from the literature regarding welded beam design, pressure vessel design, helical compression spring design, reinforced concrete beam design, stepped cantilever beam design and car side impact design. The optimization results indicated that FA is more efficient than other meta-heuristic algorithms, such as particle swarm optimization, genetic algorithms, simulated annealing and differential evolution.

Durkota [20] used a modified version of FA to solve the class of discrete problems named Quadratic Assignment Problems (QAP), where the solutions are represented as permutations of integers. In this algorithm, the continuous functions like attractiveness, distance and movement are mapped into newly developed discrete functions. The experimental results were obtained on 11 different QAP problems.

Sayadi et al. [21] proposed a new discrete firefly meta-heuristic for minimizing the make span for the permutation shop scheduling problem. They compared the results of the proposed algorithm with those of other existing ant colony optimization techniques. The results indicated that firefly algorithm outperforms the ant colony for some well-known benchmark problems.

Jati [23] applied FA on the symmetric traveling salesman problem. In this algorithm, a permutation representation is used, where an element of the array represents a city and the index represents the order of a tour. The firefly move is generated using inversion mutation. The simulation results indicated that the proposed algorithm performed very well for some traveling salesman problem instances when compared with other memetic algorithms.

Alweshah [24] proposed a hybrid firefly algorithm with artificial neural network (FA-ANN) for time series problems. The hybrid approach is tested on 6 benchmark UCR time series data sets. The experimental results revealed that the proposed FA-ANN can effectively solve time series classification problems. In another work, Alweshah and Abdullah [25] proposed a method that hybridizes the firefly algorithm with simulated annealing (SFA). They also investigated the effectiveness of using Lévy flight within the firefly algorithm (LFA) to better explore the search space. Moreover, they integrated SFA with Lévy flight (LSFA) to improve the algorithm performance. The

algorithm was tested on 11 standard benchmark datasets. The experimental results indicated that the LSFA shows better performance than the SFA and LFA. Moreover, the LSFA is able to obtain better results in terms of classification accuracy when compared with other algorithms from the literature.

Alweshah et al. [26] proposed an improved probabilistic neural network model that employs biogeography-based optimization to enhance the accuracy of classification. Their proposed approach was tested on 11 standard benchmark medical datasets. The results showed that the classification accuracy of the proposed model outperforms that of the traditional probabilistic neural network model.

Faris et al. [27] investigated the efficiency of the Lightning Search Algorithm (LSA) in training Neural Network. The investigated LSA-based trainer was evaluated on 16 popular medical diagnosis problems. The algorithm was compared to BP, LM and 6 other evolutionary trainers. The statistical test conducted proved that the LSA-based trainer is significantly superior in comparison with current algorithms on the majority of datasets.

Aljarah et al. [28] proposed a new training algorithm based on whale optimization algorithm (WOA). A set of 20 datasets with different levels of difficulty have been chosen to test the proposed WOA-based trainer. The obtained results were compared with those of a back-propagation algorithm and six evolutionary algorithms. The results proved that the proposed trainer is able to outperform current algorithms on the majority of datasets.

4. METHODOLOGY

In this section, we give a brief description of the dataset used in evaluating the proposed classifier. After that, we provide details on how the dataset is partitioned to generate the training and testing subsets. Finally, we list the steps to develop the firefly-based classifier. Figure 1 shows the overall research design.

4.1 Datasets

We use the following binary class datasets to test the proposed firefly classifier. The datasets are obtained from the University of California at Irvine (UCI) Machine Learning Repository as follows:

1. Wisconsin Breast Cancer (Original) (WBC) dataset: this dataset is collected from the University of Wisconsin Hospitals between 1989 and 1991. It is commonly used among researchers who use machine learning methods in order to classify patients with breast cancer using a set of attributes. WBC contains 699 instances, 241 instances are malignant class and 458 instances are benign class. Each instance has 9 attributes and each attribute is represented as an integer between 1 and 10.
2. Haberman's Survival dataset: this dataset contains cases from a study that was conducted between 1958 and 1970 at the University of Chicago's Billings Hospital on the survival of patients who had undergone surgery for breast cancer. It contains 306 instances and each instance has 3 attributes. The class of an instance is either survived or died.
3. Statlog (Heart) is a multivariate dataset: it has 13 categorical attributes. It contains 270 instances. Each instance has a class which is either the absence or the presence of heart disease.
4. Liver Disorders dataset: it has 6 attributes. The first 5 variables are all blood tests which are considered to be sensitive to liver disorders. The last attribute is the drinks number of half-pint equivalents of alcoholic beverages drunk per day. The dataset contains 345 instances.
5. Connectionist Bench (Sonar, Mines vs. Rocks) dataset: this dataset contains 208 instances and 111 patterns obtained by bouncing sonar signals off a metal cylinder at various angles and under various conditions. It also contains 97 patterns obtained from rocks under similar conditions.

4.2 Dataset Partitioning Criteria

A ten-fold cross-validation procedure is used to supply the testing and training datasets. The original dataset is partitioned into ten data subsets. Each partition T_i is used as a testing set and the remaining 9 partitions are grouped together to build a training set. Then, we run the FA 30 times. In each time, we

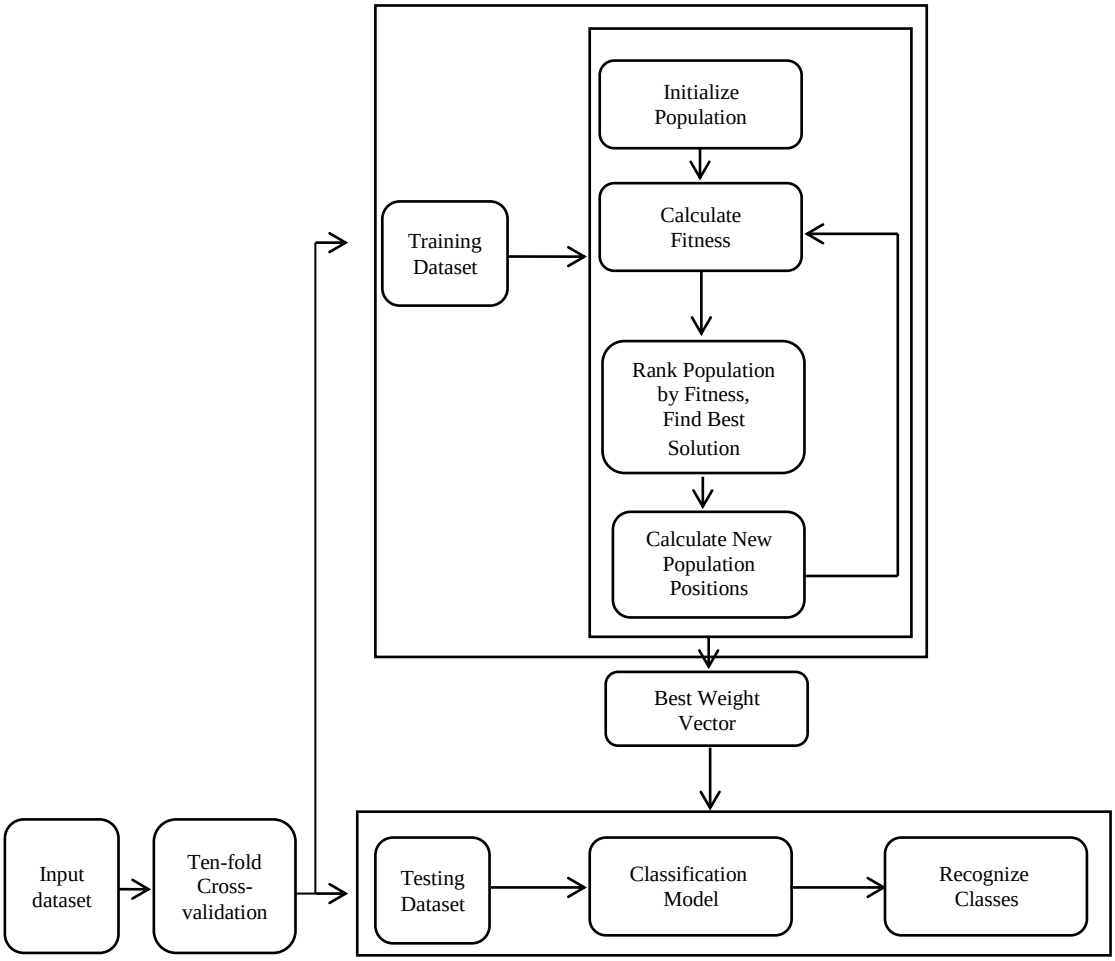


Figure 1. Framework of firefly algorithm classifier.

use T_i as a test set and the union of the others T_{ij} where $i \neq j$ as training set [28]. We use the number of partitions to be 10 to avoid costly computation for higher portioning values. The total number of models calculated in our work is 100 x 10-fold for each run.

4.3 Fitness Function

The objective function is used to score the quality of a solution. Our Fitness Function is the classifier accuracy, which is calculated using Equation (4).

$$Accuracy = \frac{\#\ of\ correctly\ classified\ objects\ in\ test\ set}{\#\ of\ objects\ in\ test\ dataset} * 100 \quad (4)$$

4.4 Learning Phase

In this phase, we apply firefly algorithm on the training dataset to infer a hyper-plane classifier model by searching for the weights vector that constructs the hyper-plane. This weights vector is the learned knowledge from learning phase and it is reused in the testing phase. The following are the steps:

A. **Initialize Population:** each firefly is a candidate solution, each candidate solution $w = (w_1, w_2, \dots, w_n, w_{n+1})$ is the weights vector. These weights vectors are initialized randomly.

B. **Calculate Fitness:** in this step, we determine the fitness (attractiveness) of each of the fireflies in the population using fitness function. The fitness calculation pseudo-code is presented in algorithm 4, where the position of each firefly is the weight vector for a candidate solution and the fitness is really the accuracy of the hyper-plane that the firefly holds its weights vector.

   In algorithm 4, the $P_L^{(t)}$ position in line (2) is the weights vector for the firefly L at the t iteration. These positions or weights are initialized randomly. In line (4), we substitute both

Algorithm 4: FitnessFA($P^{(t)}$)

1: **For** L = 1 **To** n **do** // n is population size
2: $w = P_L^{(t)}$.position // position is array of weights initialized randomly of firefly L at t
3: **For** k = 1 To n do // n is population size
4: $$y = \sum_{k=1}^{n} (w_k x_k)$$
5: **If** y > 0 **Then**
6: Predict = Class A
7: **Else**
8: Predict = Class B
9: **End if**
10: **If** Predict == Actual_Class **Then**
11: Correct = Correct + 1
12: **End if**
13: **End For**
14: $P_L^{(t)}$. Fitness= Correct/n * 100
15 **End For**

Algorithm 4. Calculate fitness.

weight vector of the current firefly and the x features vectors of the training dataset to obtain a value for each instance in the training dataset. Then, we use the obtained value to predict where this instance is located in the space of the hyper-plane by comparing it with zero as a threshold (lines 5-9). We use class A to represent one class and class B to represent the other class. We compare the obtained prediction with the actual one in order to increase correct counter of the prediction (lines 10-12). Then, we use this counter to calculate the fitness for each firefly. The fitness is the accuracy of the hyper-plane.

C. **Rank Population:** in this step, we rank the fireflies according to their fitness value.

D. **Find Current Best Solution**: after ranking the fireflies according to fitness, we return the current best solution which is the weights vector that has the maximum fitness.

E. **Calculate New Population Positions:** in this step, we calculate the new positions of the moving fireflies, where fireflies with the less fitness move towards fireflies with greater fitness (brightness). Equation 5 is used to calculate the new position of the firefly when it performs this move.

$$x_i^{t+1} = x_i^t + \beta_0 e^{-\gamma r_{ij}^2}(x_j - x_i) + \alpha_t(rand - 0.5)\,; \quad (5)$$

where α is the random movement parameter that controls the step length of the random movement, γ is a fixed light absorption coefficient and β is the brightness that depends on the distance between the two fireflies i and j. The r_{ij} is the distance between the two fireflies i and j at x_i and x_j positions, which is calculated using the Cartesian distance. The new position x_i^{t+1} of the moving firefly i at t+1 iteration depends on the step size of the moving firefly i and the current position for firefly i at t iteration. The step size of the moving firefly i relies on the attraction of firefly i towards firefly j. The brightness is inversely proportional to the distance between the two fireflies due to exponential function characteristics. The brightness is decreasing as the distance between the two fireflies increases. α is the random movement parameter that controls the step length of the random movement to make firefly i jump not too far away from firefly j. In the Equation, rand is a random number generator which is uniformly distributed between [0, 1]. In order to make the numbers within the range [-0.5, 0.5], 0.5 value is subtracted from rand. The pseudo-code which shows the steps of the new move is presented in algorithm (3).

F. **Checking for Stopping Criteria**: if the stopping criteria are satisfied, return the best weight Vector; otherwise repeat from step B.

4.5 Testing Phase

In this phase, we use the best weight vector learned from the learning phase to construct the binary

linear classification model that satisfies Equation (1) to predict the class of instaces in the dataset to be tested using algorithm 1.

5. PARAMETER SETUP AND RESULTS

5.1 Parameter Settings

Firefly classifier is implemented using MATLAB 2014a. The experiments are executed on an Intel core i5 processor running with 8 GB of RAM under Microsoft Windows 7. Table 1 presents the parameter setting that we used for the experiments.

Table 1. Firefly parameter setting.

Parameter	value
Termination condition	100
Number of fireflies	100
Attractiveness B_0	2
Light absorption coefficient (Gamma)	1
Randomization parameter (alpha)	0.2
Alpha constant	0.98
Number of times we run the algorithm on each dataset	30

5.2 Performance Metrics

The results obtained by the firefly-based classifier are evaluated using two performance metrics. These are: classification accuracy and k-fold cross-validation (KCV) accuracy. Classification accuracy is calculated using the previous formula (4) [29].

For the k-fold cross-validation (KCV) accuracy evaluation, the original sample is randomly partitioned into k sub-samples. A single sub-sample of the k sub-samples is used as validation data for testing and the remaining k-1 sub-samples are used as training data. The cross-validation process is then repeated k times; each one of the k sub-samples is used exactly once as the validation data. The average of the k results from the k-folds gives the KCV test accuracy of the algorithm [30]-[31] . Our k-fold cross-validation is a 10-fold cross-validation. The proposed algorithm runs 30 times on each dataset using a 10-fold cross-validation. The result from each run is reported and the average of the 30 runs is calculated.

5.3 Experimental Results

We perform two different experiments to evaluate our proposed classifier. The first experiment is applied on Wisconsin Breast Cancer (Original) dataset, while the second experiment is applied on Haberman's Survival, Statlog (Heart), Liver Disorders and Sonar datasets.

The two experiments are applied using a 10-fold and the accuracy of the 30 runs is averaged. Then, these results are compared with those of different algorithms from the state of the art, which are: DPSO, CPSO, LDW PSO, GA, J48, MID3 and CCDT [12]-[13].

In the first experiment, we applied the FA classifier using different population sizes: 25, 50, 100, 200 and 300 on Wisconsin Breast Cancer (WBC) dataset. After that, we calculated the average of the accuracy of the obtained results. Table 2 presents the experimental results of the 10-fold average accuracy. The accuracy of the classifier increases as the population size increases. A population size larger than 100 makes the algorithm exploration not well focused and the algorithm performance is then degraded.

The results from the proposed algorithm when applied on WBC dataset are compared with those of five algorithms presented in [12] which are: DPSO, CPSO, LDW PSO, GA and J48. Table 4 and Figure 2 show the comparison between the proposed classifier accuracy when compared with these techniques.

Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 3, No. 3, December 2017.

Table 2. Effect of different population sizes on the performance of FA when applied on WBC dataset.

Population	Accuracy
25	92 ± 1
50	92 ± 2
100	95 ± 3
200	93 ± 2
300	91 ± 2
Average	93

Table 3. FA result for a population size of 100 on WBC dataset.

	Accuracy	Sensitivity	Specificity	Positive Predictive	Negative Predictive
Minimum	94.14	95.01	88.47	93.18	93.3
Maximum	97	97.92	94.25	96.28	97
Average	95.41333	96.489	91.20533	94.55733	95.44033
Standard deviation	0.728282	0.777927	1.413973	0.816693	1.088744

Table 4. Classification accuracy of FA compared with other techniques on WBC dataset.

Population Size	DPSO	CPSO	LDW PSO	GA	J48	FA
25	92 ± 3	92 ± 3	92 ± 5	92 ± 4	93	92 ± 1
50	92 ± 3	92 ± 4	92 ± 6	92 ± 3		92 ± 2
100	92 ± 3	92 ± 4	92 ± 2	92 ± 4		92 ± 3
200	92 ± 5	92 ± 4	92 ± 4	92 ± 3		92 ± 2
300	92 ± 3	92 ± 3	92 ± 4	92 ± 3		92 ± 2
AVG	94	93	93	93		93

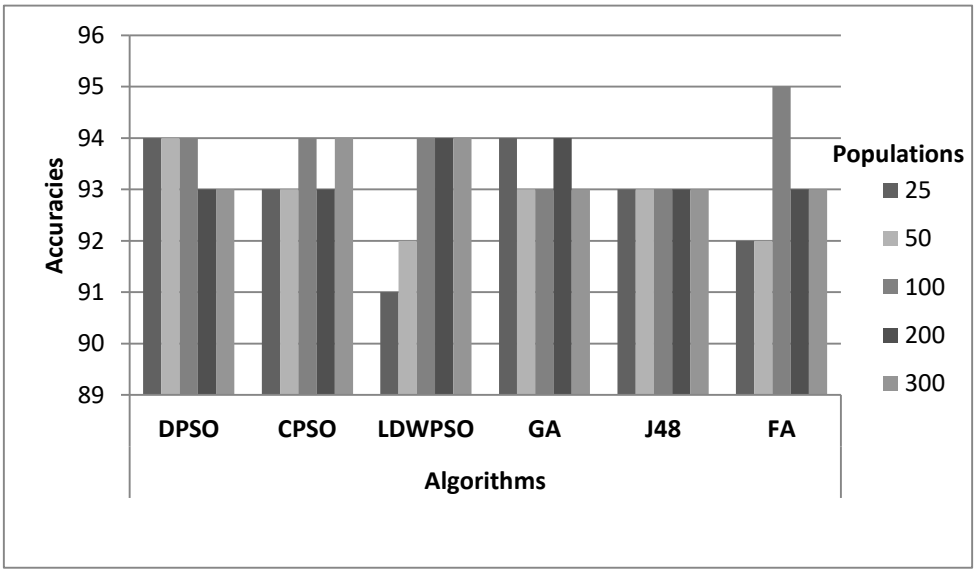


Figure 2. Accuracies of Firefly compared with other techniques on WBC dataset.

According to this experiment, we set the population size of the proposed algorithm to be 100 and the algorithm is repeated 30 times. We calculate the accuracy, sensitivity, specificity, positive predictive and negative predictive obtained from the classifier for each run. The sensitivity of each class is calculated as TP/(TP+FN) and the specificity of each class is calculated as TN/(TN+FP). Note that TP is true positive, FN is false negative, TN is true negative and FP is false positive. The minimum, maximum, average and standard deviation values are reported. Table 3 shows these results.

The second experiment was applied on Haberman's Survival dataset, Statlog (Heart) dataset, Liver

Disorders dataset and connectionist Bench dataset. We compared our results with the results from [13]. Table 4 and figure 3 show the accuracy of the proposed algorithm when compared with [13]. The parameter settings are kept as in the first experiment. We repeat the second experiment using a population size of 100 for 30 runs. We calculate the accuracy, sensitivity, specificity, positive predictive and negative predictive obtained from the classifier for each run. The minimum, maximum, average and standard deviation values are reported. Table 5, Table 6, Table 7 and Table 8 show the results for the Haberman, Heart-tatlog, Liver Disordes and Sonar datasets respectively.

Table 5. FA result for a population size of 100 on Haberman dataset.

	Accuracy	Sensitivity	Specificity	Positive Predictive	Negative Predictive
Minimum	71.55	80.09	26.79	86.7	24.09
Maximum	77.08	83.85	50.12	92.29	43.48
Average	74.9807	82.3093	39.698	89.3153	33.9067
Standard deviation	1.3321	0.9891	5.9915	1.6653	5.6787

Table 6. FA result for a population size of 100 on Heart-tatlog dataset.

	Accuracy	Sensitivity	Specificity	Positive Predictive	Negative Predictive
Minimum	76.3	80.47	66.07	75.7	64.71
Maximum	82.22	86.69	75.74	85.85	81.67
Average	79.6547	83.7443	70.2407	79.465	74.608
Standard deviation	1.5897	1.6157	2.3346	2.646	4.3328

Table 7. FA result for a population size of 100 on Liver Disordes dataset.

	Accuracy	Sensitivity	Specificity	Positive Predictive	Negative Predictive
Minimum	62.95	68.71	60.14	38.34	81.75
Maximum	71.71	77.05	65.23	56.7	89.74
Average	68.2203	72.8753	62.809	48.256	85.7943
Standard deviation	2.530924	2.5029	1.3808	4.7153	2.2573

Table 8. FA result for a population size of 100 on Sonar dataset.

	Accuracy	Sensitivity	Specificity	Positive Predictive	Negative Predictive
Minimum	95.8	95.79	95.95	95.87	96.04
Maximum	98.26	97.87	97.66	98.09	98.15
Average	97.01	96.9933	97.065	97	96.945
Standard deviation	0.639	0.4865	0.4635	0.5010	0.5898

Table 9. Classification accuracies of FA compared to those of different classification algorithms.

Dataset	FA	C4.5	Credal-C4.5	MID3	CCDT
Haberman	**74.98**	70.52	73.89	70.62	73.59
Heart-tatlog	79.65	76.78	80.04	77.93	**82.11**
Liver Disordes	**68.22**	65.37	64.18	65.75	56.85
Sonar	**97.01**	73.42	71.47	73.53	73.92

6. DISCUSSION

We tested the FA algorithm with different population sizes 25,50,100,200 and 300. But, the one that gives the best result is 100, because the algorithm exploration is focused as the population has enough diversity.

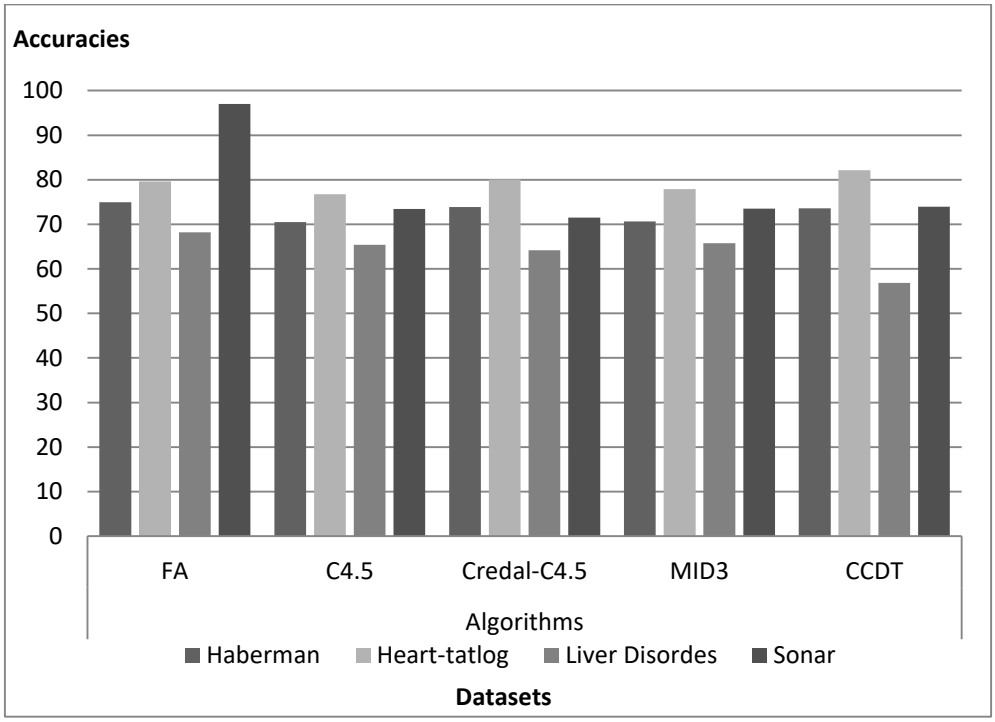


Figure 3. Accuracies of Firefly with a population of 100 compared with those of other techniques on four different datasets.

The experimental results and the accuracy values show that the classification accuracy of the proposed FA when applied on Wisconsin Breast Cancer data using different population sizes reached an average accuracy of 93%, which is similar to the results obtained from CPSO, LDW PSO, GA and J48 algorithms presented in [12]. On the other hand, the DPSO algorithm has a slightly better average accuracy value of 94%. So, our FA results are competitive when compared to those of other algorithms from the state of the art.

Table 9 shows that the classification accuracy of FA is competitive when compared with those of the algorithms presented in [13] when applied to different datasets. The FA showed better results when compared with [13] in three out of the four tested datasets, which are: Haberman, Liver Disordes and Sonar.

The results of applying FA classifier are promising. FA has advantages over other mentioned algorithms, as it speeds up the convergence rate in a small number of iterations. Applying FA leads to achieve good accurate results in an efficient manner. This is because it attracts low-quality solutions to be merged with good solutions, which makes the whole population to be automatically sub-divided into subgroups. Each group is swarmed around its local best and that speeds up the convergence rate. Then, among all these local best solutions, the best solution from the population is quickly found. Another reason that makes FA more efficient is that the parameters in the algorithm can be tuned to control the randomness as the iterations proceed. This makes convergence speed up [10].

7. CONCLUSION AND FUTURE WORK

In this work, we presented the Firefly algorithm as a binary linear classifier. The FA optimizes the parameter values for hyper-planes in the feature space. Experimental results show that the Firefly classifier achieves good accuracy when compared with other algorithms. The experimental results and the accuracy values are compared with those of the classifiers CPSO, LDW PSO, GA, J48, C4.5, Credal-C4.5, MID3 and CCDT.

The results prove that the firefly classifier is a competitive classifier. Therefore, the FA approach can be used for other more complex classification problems. Through conducting experiments, we have come to some conclusions regarding the application of the FA. First, increasing the population size improves the accuracy of the firefly classifier, but the algorithm performance is degraded when we reach a population size of more than 100. Second, increasing the number of iterations over 100 in the

two experiments has not increased the algorithm accuracy. This proves that firefly classifier achieves very good results with fast convergence. Lastly, the results show that the Firefly classifier is a reasonably good classifier when it is compared with other state of the art classifiers [12]-[13].

As a future work, we suggest analyzing the algorithm when it is applied to a multiclass dataset. Another future work can be studying the effect of tunning different parameters of the FA.

REFERENCES

[1] C. Aggarwal and C. Zhai, "A Survey of Text Classification Algorithms in Mining Text Data," Springer, pp. 163-222, 2012.

[2] W. Clancey, "Heuristic Classification," Artificial Intell. Journal, vol. 27, no. 3, pp. 289-350, 1985.

[3] J. Quinlan, "Induction of Decision Trees," Machine Learning, vol. 1, no. 1, pp. 81-106, 1986.

[4] G. Pappa and A. Freitas, Automating the Design of Data Mining Algorithms, Springer Verlag, Berlin Heidelberg, 2010.

[5] M. Mavroforakis and S. Theodoridis, "A Geometric Approach to Support Vector Machine (SVM) Classification," IEEE Transactions on Neural Networks, vol. 17, no. 3, pp. 671-682, 2006.

[6] A. Ng and M. Jordan, "On Discriminative vs. Generative Classifiers: A Comparison of Logistic Regression and Naive Bayes," Advances in Neural Information Processing Systems, vol. 2, pp. 841-848, 2002.

[7] P. Langley and K. Thompson, "An Analysis of Bayesian Classifiers, " Aaai, 1992.

[8] Jr . Fister, X. Yang, I. Fister, J. Brest and D. Fister, "A Brief Review of Nature-inspired Algorithms for Optimization," arXiv preprint arXiv:1307.4186 , 2013.

[9] X. Yang, "Firefly Algorithms for Multimodal Optimization," International Symposium on Stochastic Algorithms, Stochastic Algorithms: Foundations and Applications (SAGA 2009), Springer, pp. 169-178, 2009.

[10] I. Fister, X. Yang and J. Brest, "A Comprehensive Review of Firefly Algorithms," Swarm and Evolutionary Computation, vol. 13, pp. 34-46, 2013.

[11] X. Yang, "Firefly Algorithm, Stochastic Test Functions and Design Optimization," International Journal of Bio-Inspired Computation,vol. 2, no. 2, pp. 78-84, 2010.

[12] T. Sousa, A. Silva and A. Neves, "Particle Swarm-based Data Mining Algorithms for Classification Task," Parallel Computing, vol. 30, no. 5, pp. 767-783, 2004.

[13] C. Mantas and J. Abellán, "Credal-C4. 5: Decision Tree Based on Imprecise Probabilities to Classify Noisy Data," Expert Systems with Applications, vol. 41, no. 10, pp. 4625-4637, 2014.

[14] N. Bhargava, G. Sharma, R. Bhargava and M. Mathuria,"Decision Tree Analysis on j48 Algorithm for Data Mining," Proceedings of International Journal of Advanced Research in Computer Science and Software Engineering, vol. 3, no. 6, pp. 1114-1119, 2013.

[15] F. Partovi and M. Anandarajan, "Classifying Inventory Using an Artificial Neural Network Approach," Computers & Industrial Engineering, vol. 41, no. 4, pp. 389-404, 2002.

[16] S. Zahiri and S. Seyedin, "Intelligent Particle Swarm Classifiers," Iranian Journal of Electrical and Computer Engineering, vol. 4, no. 1, pp. 63-70, 2005.

[17] D. Martens, M. De Backer, R. Haesen, J. Vanthienen, M. Snoeck and B. Baesens, "Classification with Ant Colony Optimization," IEEE Transactions on Evolutionary Computation, vol. 11, no. 5, pp. 651-665, 2007.

[18] Z. Assarzadeh, S. Monadjemi, P. Moallem and S. Hashemi, "Harmony Search-based Classifier," Int. J. Acad. Res. Appl. Sci., vol. 5, no. 1, pp. 108-119, 2015.

[19] A. Gandomi, X. Yang and A. Alavi, "Mixed Variable Structural Optimization Using Firefly Algorithm," Computers & Structures, vol. 89, no. 23, pp. 2325-2336, 2011.

[20] K. Durkota, Implementation of a Discrete Firefly Algorithm for the QAP Problem within the Seage Framework, BSc Tthesis, Faculty of Electrical Engineering, Czech Technical University, 2011.

[21] M. Sayadi, R. Ramezanian and N. Ghaffari-Nasab, "A Discrete Firefly Meta-heuristic with Local

Search for Makespan Minimization in Permutation Flow Shop Scheduling Problems," International Journal of Industrial Engineering Computations, vol. 1, no. 1, pp. 1-10, 2010.

[22] R. Sawalha and I. Abu Doush, "Face Recognition Using Harmony Search-based Selected Features," International Journal of Hybrid Information Technology, vol. 5, no. 2, pp. 1-16, 2012.

[23] G. Jati, "Evolutionary Discrete Firefly Algorithm for Travelling Salesman Problem," Springer, pp. 393-403, 2011.

[24] M. Alweshah, "Firefly Algorithm with Artificial Neural Network for Time Series Problems," Research Journal of Applied Sciences, Engineering and Technology, vol. 7, no. 19, pp. 3978-3982, 2014.

[25] M. Alweshah and S. Abdullah, "Hybridizing Firefly Algorithms with a Probabilistic Neural Network for Solving Classification Problems," Applied Soft Computing, vol. 35, pp. 513-524, 2015.

[26] M. Alweshah, A. Hammouri and S. Tedmori, "Biogeography-based Optimization for Data Classification Problems," International Journal of Data Mining, Modelling and Management, vol. 9, no. 2, pp.142-162, 2017.

[27] H. Faris, I. Aljarah, N. Al-Madi and S. Mirjalili, "Optimizing the Learning Process of Feed Forward Neural Networks Using Lightning Search Algorithm," International Journal on Artificial Intelligence Tools, vol. 25, no. 6, 2016.

[28] I. Aljarah, H. Faris and S. Mirjalili, "Optimizing Connection Weights in Neural Networks Using the Whale Optimization Algorithm," Soft Computing, pp.1-15, 2016.

[29] P. Cheewaprakobkit, "Study of Factors Analysis Affecting Academic Achievement of Undergraduate Students in International Program," Proceedings of the International Multi-conference of Engineers and Computer Scientists, 2013.

[30] N. Diamantidis, D. Karlis and E. Giakoumakis, "Unsupervised Stratification of Cross-validation for Accuracy Estimation," Artificial Intelligence Journal, vol. 116, no. 2, pp. 1-16, 2000.

[31] O. Inan, M. Uzer and N. Yılmaz, "A New Hybrid Feature Selection Method Based on Association Rules and PCA for Detection of Breast Cancer," International Journal of Innovative Computing, Information and Control, vol. 9, no. 2, pp. 727-729, 2013.

[32] S. ElMustafa, A. Jaradat, I. Abu Doush and N. Mansour, "Community Detection Using Intelligent Water Drops Optimization Algorithm," International Journal of Reasoning-based Intelligent Systems, vol. 9, no. 1, pp. 52-65, 2017.

ملخص البحث:

توظف هذه الدراسة خوارزمية "يَراعاتٍ" من أجل إيجاد المستوى الفائق للقرار الأفضل في فضاء الخصائص. يستخدم المصنف المقترح تقنية تثبُّتٍ متقاطعٍ مبنية على تقسيم ذي عشرة أضعاف لكل من طوري التدريب و الاختبار المستخدمَين في عملية التصنيف.

وقد جرى استخدام خمس مسائل مرجعية ثنائية لتمييز الأنماط بأبعاد مختلفة لمتجهات الخصائص، وذلك من أجل عرض فاعلية المصنف المقترح. كذلك، تمت مقارنة نتائج مصنِّف خوارزمية "اليراعات" المقترح مع نتائج طرق أخرى، وذلك عبر التجربة. فقد تم إجراء تجربتين لهذا الغرض، وأثبتت نتائجهما أن مصنف خوارزمية "اليراعات" منافس جيد لتقنيات التصنيف الاخرى. وتجدر الإشارة أن المصنف المقترح حقق نتائج أفضل في ثلاثٍ من مجموعات البيانات الأربع المختبرة في التجربة الثانية.

Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 3, No. 3, December 2017.

ARABIC HANDWRITTEN CHARACTER RECOGNITION BASED ON DEEP CONVOLUTIONAL NEURAL NETWORKS

Khaled S. Younis[1]

(Received: 24-Jun.-2017, Revised: 06-Sep.-2017 and 21-Oct.-2017, Accepted: 31-Oct.-2017)

ABSTRACT

The automatic analysis and recognition of offline Arabic handwritten characters from images is an important problem in many applications. Even with the great progress of recent research in optical character recognition, a few problems still wait to be solved, especially for Arabic characters. The emergence of Deep Neural Networks promises a strong solution to some of these problems. We present a deep neural network for the handwritten Arabic character recognition problem that uses convolutional neural network (CNN) models with regularization parameters such as batch normalization to prevent overfitting. We applied the Deep CNN for the AIA9k and the AHCD databases and the classification accuracies for the two datasets were 94.8% and 97.6%, respectively. A study of the network performance on the EMNIST and a form-based AHCD dataset were performed to aid in the analysis.

KEYWORDS

Convolutional neural network, Deep learning, Optical character recognition, Arabic handwritten character recognition, EMNIST.

1. INTRODUCTION

The field of optical character recognition (OCR) is very important, especially for offline handwritten recognition systems. Offline handwritten recognition systems are different from online handwritten recognition systems [1]. The ability to deal with large amounts of script data in certain contexts will be invaluable. One example of these applications is the automation of the text transcription process applied on ancient documents considering the complex and irregular nature of writing [2]. Arabic optical text recognition is experiencing slow development compared to other languages [3].

One problem with recognizing the Arabic alphabet is that many characters have similar shapes but with varying locations of dots relative to the main part of the character. Figure 1 shows the isolated alphabet of the Arabic language. As can be seen at the top row, the three characters on the left have a similar main part but the dot on the “Kha” is above while, for the “Jiim”, the dot is below the main part and the “Haa” has no dots at all. It is noteworthy that handwritten characters are more challenging, since human writers tend to combine dots and use dashes instead or change the shape of characters as can be seen in Figure 2, which shows 48 handwritten samples of the same letter "Ayn" that were used in previous work [4].

Moreover, the Arabic alphabet is widely used by many people from different countries including all Arab countries in addition to being used in the Persian, Urdu and Pashto languages. It would be great to use Arabic handwritten character recognition (AHCR) to convert many documents into digital format that can be accessed electronically. Applications include: reading postal addresses off envelops and automatically sorting mail, helping the blind to read, reading customer-filled forms (government forms, insurance claims, application forms), automating offices, archiving and retrieving text and improving human-computer interfaces.

Deep Learning (DL) is a new application of machine learning for learning representation of data. DL algorithms have taken the top place in the object recognition field due to the great performance improvement they have provided [5], [30].

This paper is an extended version of a short paper that was presented at the international conference "New Trends in Information Technology (NTIT) 2017", 25-27 April 2017, Amman. Jordan.

1. Khaled S. Younis is with the Department of Computer Engineering, The University of Jordan, Amman, Jordan. Email: Younis@ju.edu.jo.

خ	ح	ج	ث	ت	ب	ا
kha	haa	jiim	thaa	taa	baa	alif
ص	ش	س	ز	ر	ذ	د
saad	shiin	siin	zaay	raa	thaal	daal
ق	ف	غ	ع	ظ	ط	ض
qaaf	faa	ghayn	ayn	thaa	taa	daad
ي	و	ه	ن	م	ل	ك
yaa	waaw	ha	nuun	miim	laam	kaaf

Figure 1. The Arabic alphabet.

Figure 2. Sample of 48 handwritten "Ayn" letters.

Deep Learning (DL) is a new application of machine learning for learning representation of data. DL algorithms have taken the top place in the object recognition field due to the great performance improvement they have provided [5], [30].

Convolutional Neural Networks (CNNs) are a type of neural networks that are applied in many fields and provide efficient solutions in many problems, where there is some translation invariance like some applications of object recognition and speech recognition. However, CNN DL solutions require a lot of training samples, which places computational requirements on the system. Nevertheless, the accelerating progress and availability of low-cost computer hardware, high-speed networks and software for high-performance distributed computing encouraged the use of computationally expensive techniques. For example, Cecotti [6] used Graphical Processing Units (GPUs) and High-Performance Clusters (HPCs) to classify isolated characters of 9 databases using computationally expensive techniques.

There are several frameworks for Deep Learning. One of the most popular libraries is TensorFlow, that was released by Google in 2015 [7]. It is an open source code written in C++ programming language and capable of using GPUs very well. Another simpler framework is Keras [8], which is a higher-level API built on top of TensorFlow. Keras uses Python for programming, which makes writing programs easier than native TensorFlow codes.

Therefore, in this paper, we will discuss the building of a robust CNN DL model for solving the problem of AHCR using TensorFlow/Keras. This model is expected to outperform traditional AHCR algorithms that depend on feature extraction and classification and can be applied on huge and

different databases efficiently without the need for feature engineering and extremely long training time.

The contributions of this research are: (i) Reviewing state-of-the-art research in AHCR (ii) Proposing robust architecture for CNN DL for solving the AHCR problem (iii) Studying the effect of different regularization techniques and network parameters on the performance on very large size AHCR databases (iv) Utilizing the functionalities offered by TensorFlow and Keras libraries for AHCR and (v) Achieving the highest accuracy on the AHCR problem.

The rest of the paper is organized as follows; Section 2 discusses related work in the field of AHCR. Section 3 describes the motivation for the proposed solution as well as the different components of its architecture, Section 4 introduces the experiments performed in a scientific way, including the results obtained and Section 5 discusses conclusions derived from the results and presents plans for future work.

2. RELATED WORK

Algorithms designed to recognize handwritten characters are still less successful than those for printed characters, mainly due to the diversity in handwritten character shapes and forms. Arabic character recognition is an important problem, since it is a step that may be needed in the more challenging Arabic word or sentence recognition problem [9]. Character segmentation to separate the word into characters is another challenging problem. The character recognition problem is related to the simpler problem of Arabic numeral recognition which has recently attained great results [10].

Various methods have been proposed and high recognition rates are reported for the handwritten English and Chinese characters. However, in this section, we are going to present only the most competitive related work solving the AHCR problem.

Many algorithms in the past concentrated on finding structural features (such as the presence of loops, the orientation of curves, …etc.) or statistical features (such as moments, histogram of gray level distribution, …etc.) [11]. These features try to maximize the interclass variability while minimizing the intra-class variability and were fed to a classifier.

Some algorithms are considered segmentation-based recognition systems; this means that their experimental results depend on segmenting the words before recognizing the characters. The IFN/ENIT database was used in Al-abodi and Li [12], who had proposed a recognition system based on geometrical features of Arabic characters. The average recognition accuracy is 93.3%. Other works that used IFN/ENIT for segmentation-based character recognition such as [33] also achieved similar performance using three main modules: preprocessing, feature extraction and recognition. However, we will not discuss such system in this paper. Even though the IFN/ENIT database [28] is available, it is designed for classifying words and letter segmentation is required before character recognition is performed. In addition, it is considered small and does not contain enough representative samples. Therefore, it was deemed unsuitable for CNN DL architecture evaluation.

Arabic characters have different forms depending on the location of the characters in the word. Hidden Markov Models (HMM) assume each letter is a state and using the context leads to better classification of Arabic handwritten word as in [37]. Nevertheless, recent work [39] discusses the limitations of HMM in terms of the need for manual extraction of features, which requires prior knowledge of the language and is robust to handwriting diversity and complexity. CNN applications to direct word recognition have been discussed in [39] and [41]. Use of Bidirectional Long Short Term Memory (BLSTM) networks is proved useful in other languages, but the application to Arabic language is of great interest. However, the lack of very large datasets and the layout of Arabic text are causing problems in implementation. One can also use character segmentation followed by recognition. For the latter, they used LSTM [38] with convolution to construct bounding boxes for each character. We then pass the segmented characters to a CNN for classification and then reconstruct each word according to the results of classification and segmentation. It was shown that character segmentation had given better performance confirming the intuition that the much smaller scope of the model's initial feature representation problem for characters as opposed to words and final labeling problem helped boost the performance. There is great diversity in handwriting for

particular words/characters among writers, thus making the task of recognizing all of the different ways in which a character or word is written very challenging. An important aspect is the availability of huge dataset to train the deep network. Since there is no such database for Arabic words, the analysis of isolated character recognition is important and may be included in the system for segmentation-based word or sentence recognition.

In 2014, Torki et al. [13] built their own database of about nine thousand characters. They called it the AlexU Isolated Alphabet (AIA9k) database. Then, they extracted three window-based gradient-based descriptors: Histogram of Oriented Gradients (HOG) [14], Speeded-Up Robust Features (SURF) [15] and Scale Invariant Feature Transform (SIFT) [16]. In addition, they extracted two texture-based descriptors and tried 4 classifiers (Logistic regression, ANN, SVM-Linear and SVM-RBF) on their database. The best achieved accuracy was 94.28% using SVM-RBF on SIFT features. The 75 characters that were misclassified are shown in Figure 3. While there are some characters that are not that difficult to classify, some characters are indeed confusing.

Figure 3. Misclassified characters from the AIA9K dataset using the method of [13].

In 2015, Lawgali [11] published a survey about Arabic Character Recognition and none of the algorithms mentioned used deep learning. However, also in 2015, Elleuch [9] introduced an Arabic handwritten character recognition using Deep Belief Neural Networks. It does not require any feature engineering. The input is simply the raw data or grayscale pixel values of the images. The approach was tested on the HACDB database [17] that contains 6600 shapes of handwritten characters written by 50 persons. The dataset is divided into a training set of 5280 images and a test set of 1320 images. The result was promising on the character recognition task with 97.9% accuracy but discouraging on the word recognition database with an accuracy of less than 60%.

In 2017, Elleuch [18] continued working on the DBN and stack of feature extractors, such as Restricted Boltzmann Machine (RBM) and Auto-Encoder and reported the results on character recognition that was in fact similar (97.8%) to the previous work. These are very promising results and demonstrate the superiority of DL methods in AHCR.

Nevertheless, it must be mentioned that the HACDB database is considered an easy and clean database and there are well defined main parts of the different letter forms among 66 different classes. On the other hand, in character recognition, it is harder to classify similar letters which are only different by a dot. HACDB are much easier to classify and have three times classes as the AIA9k database.

In 2017, El-Sawi et al. [19] collected the Arabic Handwritten Character Dataset (AHCD) of 16800 images of isolated characters. They built a CNN Deep learning architecture to train and test the dataset. They used optimization methods to increase the performance of CNN. Their proposed CNN gave an average 94.9% classification accuracy on testing data.

We can see that a few techniques of deep learning have proven their usefulness for the AHCR problem and hence we will explain in the Motivation section next why we decided to propose the following architecture.

3. PROPOSED ARCHITECTURE

In this section, we describe the criterion behind the decisions taken during the design phase and the architecture parameters for the models used for solving the AHCR problem.

3.1 Motivation

It is foreseen that, due to the success of modern neural network architectures, state-of-the-art handwritten recognition systems will either go towards hybrid systems (Deep Networks with some segmentation and feature extraction) or pure neural recognizers featuring deep architectures [20]. The discussed related work has demonstrated the inefficiency of selecting the right feature and going through the preprocessing stages. The goal of this paper is to study the CNN DL approach that particularly makes use of the Convolution layer to leverage three ideas that help improve the classification network: sparse interaction, parameter sharing and equivalent representation [21].

We will apply a robust CNN architecture to the Arabic characters AIA9k and AHCD databases as a case study with enough samples to validate the assumptions and give meaningful feedback. We will use CNN capability to extract features and train for recognition instead of extracting a large set of gradient or textural features as it was done in Torki et al. [13]. Moreover, we will use recent techniques of optimization and regularization, such as Batch Normalization [22] and Varying Learning Rate, to deal with training neural network issues. These were not used in the work of El-Sawy et al. [19], for example, as they used fixed learning rate and didn't normalize the mini-batches during training. Moreover, by making a large CNN network with many layers, it becomes more capable to detect more features automatically. Hence, using different numbers of convolutional layers with different numbers of filters should help us achieve better accuracy.

To solve the problem of latency in processing the data, GPUs are used as suggested by Ciresan et al. [23], who trained and tested the CNN network using a committee of classifiers and reduced the error rate of MNIST dataset [24] to 2.7%. For this reason, we decided to choose Keras and TensorFlow as the developing environment, since there is a GPU-enabled TensorFlow with support for CUDA acceleration.

3.2 CNN Architecture

Convolutional neural networks can convert the input structure through each layer of the network seamlessly to extract automatically the features of the images.

CNNs are based on a mathematical operation called convolution. A convolution is a multiplication operation of each pixel in the image with each value in the kernel, which is in turn another matrix and then summing the products. The key advantage of using the convolution operation is generating many images from the original image that enhance different features extracted from the original image, which leads to making the classification process more powerful [29].

In CNN, we use different types of layers as shall be explained shortly. First, the convolution layer, also called a feature extractor, extracts features from the input image. Initially, CNN does not know where exactly the features (shapes) in the image will be located; so, it tries to find them everywhere in the image by using a matrix called filter. Each filter represents a specific feature. CNN applies the convolution operation by a sliding filter in the image and multiplies each pixel in the image with each value in the filter. Then, this operation is repeated for other features (filters) and the output of this layer will be a set of filtered images [29].

In modern deep learning libraries, some consider a second layer called "Nonlinearity layer". In this layer, the Rectified Linear Unit (ReLU) activation function of the neurons is implemented to produce an output after each convolution [34]. ReLU is an element-wise operation (applied per pixel) to introduce non-linearity in our network. Since convolution is a linear operation, element-wise matrix multiplication and addition, so we add nonlinearity using ReLU. This operation converts each negative pixel in a feature map into zero and keeps each positive pixel.

Batch Normalization is a normalization and regularization technique proposed by Ioffe and Szegedy [22] to address the following issues that appear during the training process of deep neural networks:

1. Internal Covariate Shift: which refers to the change in the distribution of input of each layer

(features) that is affected by parameters in all input layers in which a small change in the network can significantly affect the entire network; and

2. Vanishing Gradient in saturating nonlinear functions: such as tanh and sigmoid, which are prone to get stuck in the saturation region as the network grows deeper despite the proposed solutions to carefully initialize the network, using small learning rate or replacing these functions by ReLU function.

Our system suggests the use of Batch Normalization as a part of the network architecture and it was experimentally proven to cause an improvement in terms of speed and accuracy. The Batch Normalization layer is added just before the nonlinearity and especially after the convolutional layers to limit its output away from the region of saturation using the mean and variance.

The Pooling or subsampling layer reduces the dimensionality of each filtered image, but preserves the most important features in the previous layer. Pooling can be of different types: Maximum, Average Sum, …etc. The output will have the same number of images, but they will each have fewer pixels. This is also helpful in managing the computational load [36]. The pooling operation is demonstrated in Figure 4. However, it is argued that max-pooling can be redundant and could be replaced by purely using convolutional layer with increased stride without loss in accuracy [35].

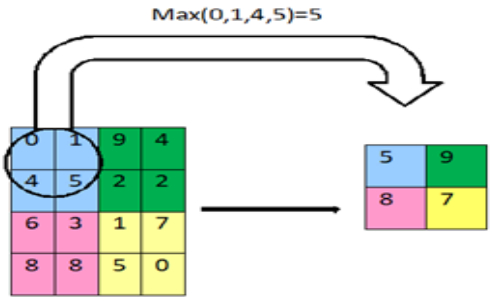


Figure 4. Pooling operation.

Dropout layers are also used in convolutional neural networks with the aim of reducing overfitting. This layer "drops out" a random set of neurons in that layer by setting their activation to zero. It makes sure that the network can generalize to test data by getting weights that are insensitive to training samples. Dropout is used during training with different percentages of total number of neurons in each layer [26].

Finally, the fully connected layers are the basic building blocks of traditional neural networks. They treat the input as one vector instead of two dimensional arrays. Full connection implies that every neuron in the previous layer is connected to every neuron in the next layer. The output from convolutional and pooling layers represents high level features and fully connected layers used to classify material (input images) into the appropriate class based on the training of the dataset [36].

Figure 5 shows the base architecture of the AHCR proposed network. Other modifications will be explained in the next sections.

As can be seen, the general network we designed has three convolutional layers followed by a fully connected layer as hidden layers. Max pooling can be ignored. The first layers are the input layers that take input of shape 28x28 or 32x32 pixels of grayscale characters depending on the size of input samples (database), then a convolutional layer of 24 filter map of size 6x6 and stride 1, followed by batch normalization, ReLU activation function and dropout of 1.0.

Dropout technique at the end of some layers is used with a "keep probability" parameter of 0.5. This means that at each training iteration, half of the neurons of the last layer get activated while the other half activation is set to zero. This tends to prevent our network from overfitting by not building a model so tightly tied to the training samples.

In the next layers, the same order of layers is used with increased number of convolutional filter map of 48 filters of size 5x5 and stride 2 and 64 filters of size 4x4 and stride 2, respectively.

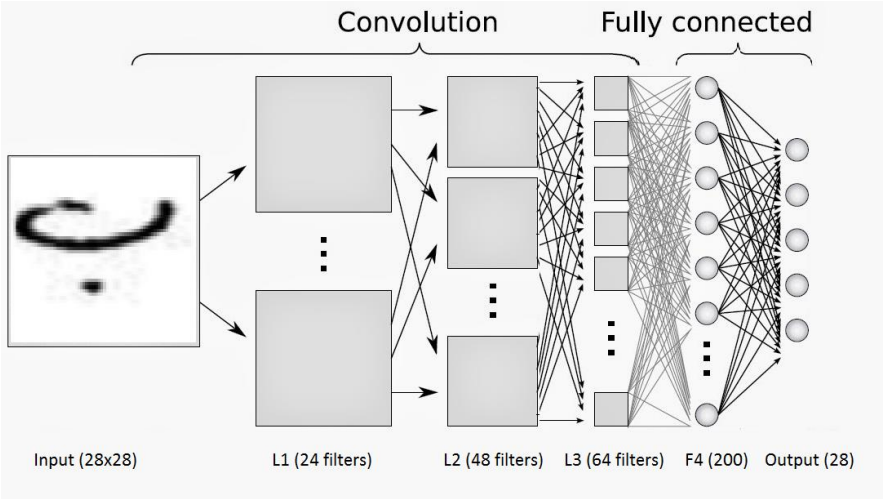


Figure 5. Proposed CNN architecture for the AHCR problem.

Finally, a fully connected layer of 200 neurons is used before the output layer of 28 neurons to match the number of classes (Arabic alphabet). We used the Softmax activation function to output probabilities between 0 and 1 for each class representing the confidence that a certain character belongs to a specific class.

For updating the weights during training, we used the Categorical Cross-Entropy [25] as a cost function which is the appropriate cost function for multi-class classification problems. We used Adam Optimizer to find the minima of the cost function with a varying learning rate [27] that recalculates its value after each batch.

4. EXPERIMENTS

4.1 Databases

In this section, we describe the different publicly available datasets that were used to evaluate the proposed network.

4.1.1 Arabic Handwritten Character Dataset (AHCD)

The dataset is composed of 16,800 characters written by 60 participants; the age range is from 19 to 40 years and 90% of participants are right-handed. Each participant wrote each character (from "Alef" to "Yeh") ten times. The forms were scanned at a resolution of 300 dpi. The database is partitioned into two sets: a training set (13,440 characters to 480 images per class) and a test set (3,360 characters to 120 images per class) [19].

4.1.2 AlexU Isolated Alphabet (AIA9K) Dataset

This dataset introduces a compact 9K novel dataset of 28 classes that represent isolated Arabic handwritten alphabet of 32x32 pixels [13]. AIA9K dataset was collected from 107 volunteer writers, between 18 and 25 years old, who are B.Sc. or M.Sc. students. The writers were 62 females and 45 males. Each writer wrote all of the Arabic letters 3 times. The total valid number of collected characters is 8,737 letters; this novel dataset can be requested from the authors of the paper mentioned in [13]. A sample of the dataset is shown in Figure 3. These are 75 characters that were misclassified in one of the experiments in [13].

4.2 Results

This section introduces the results obtained by the proposed AHCR network. Two subsections will describe the results of applying the network to classify AIA9K dataset and the AHCD datasets, subseq-uently. A subsection will compare the results obtained using the proposed approach with those of other approaches. Next, we will describe the results of applying the proposed network on Latin (English) characters. Finally, we will generate a derived database from the AHCD database, where

only samples of each group of characters had the same shape (major stroke), then we will discuss the application of the proposed methodology on this dataset.

4.2.1 Results of the AHCR System Using The AHCD Dataset

The results we obtained after testing the proposed network described in sub-section 3.2 on the AHCD Arabic isolated alphabet dataset are described here. We divided the data into three parts; training, validation and testing, with ratios of 70%, 15% and 15% for each set, respectively. Then, we ran training for 10 epochs and 100 batch sizes. We obtained an accuracy of 92% on the test set at the end of the training.

We increased the number of epochs to 20 and 28, respectively. Notable improvements have been obtained and accuracy increased to 93% and 94.5%, respectively. In the next step, we increased the number of filters of the first convolutional layer from 24 to 72, the second convolutional layer from 48 to 144, the third convolutional layer from 64 to 192 and increased the number of the fully connected layer neurons from 200 to 400. Test accuracy improved to 94.7%.

Analysis of the difference in accuracies of training (100%), validation (97.5%) and testing (94.7%) revealed a gap that is an indication of overfitting. One way of improving generalization is to increase the size of training data. A simple shift of the input image to the left by 1 pixel will result in a total different input for the network, while it does not affect the actual class. Data augmentation techniques are a way to artificially expand the dataset. Some popular augmentation examples are horizontal flips, vertical flips, random crops, translations and rotations. Data augmentation was deemed necessary to improve the network performance. We increased the number of training images from ~13k to ~80k using translation in both horizontal and vertical directions by 3 pixels, rotation of +10 and -10 degrees and by adding Gaussian noise with zero mean and a standard deviation of 5. Horizontal or vertical flipping was deemed unsuitable for our application. However, we have seen significant performance gain merely by using translation which was reflected in obtaining a testing data accuracy of 96.7%. On the other hand, when we used all the 80k images and after training for 18 epochs, accuracy jumped to 97.6%.

Figure 6 shows that the training and validation accuracies changed during training for 18 epochs on 80k augmented dataset. It is obvious that the gap between the two metrics was insignificant after the initial epochs. The small jump on epoch 18 was an indication of the early stopping function that avoids overfitting and stop training.

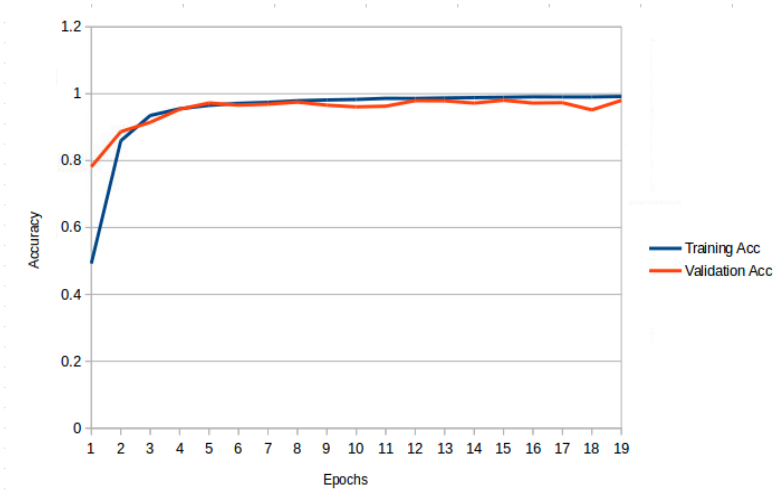


Figure 6. Training accuracy and validation accuracy during training.

Figure 7 shows the training and validation loss curves as functions of training epochs. It is worth mentioning that validation loss decreased significantly at epoch 4 and stayed very small until the end of epoch 19.

Using varying learning rate also helped reach lower minimum of the loss function. This technique is very essential in many optimization algorithms.

Before we used data augmentation to train the network, many experimental setups have been tested to improve the network performance [21] as summarized below:

- We tried to increase the network capacity by increasing the size of the network via adding another fully connected layer of 200 neurons. However, testing accuracy decreased.
- Test accuracy decreased to 93.5% when we reduced the number of neurons of the fully connected layer to 150 neurons.
- Test accuracy increased to 94.7% when we tried to double the number of filters to the three convolutional layers to 48, 96 and 128, respectively and increased the number of neurons of the fully connected layer from 200 to 300.
- No change in performance occurred when we tripled the number of convolutional layer filters and made the fully connected layer neurons 400, which indicates that the network size was adequate.
- Thresholding the grayscale images to convert them into binary images decreased the accuracy to 92.1%, which highlights the benefits of grayscale information.

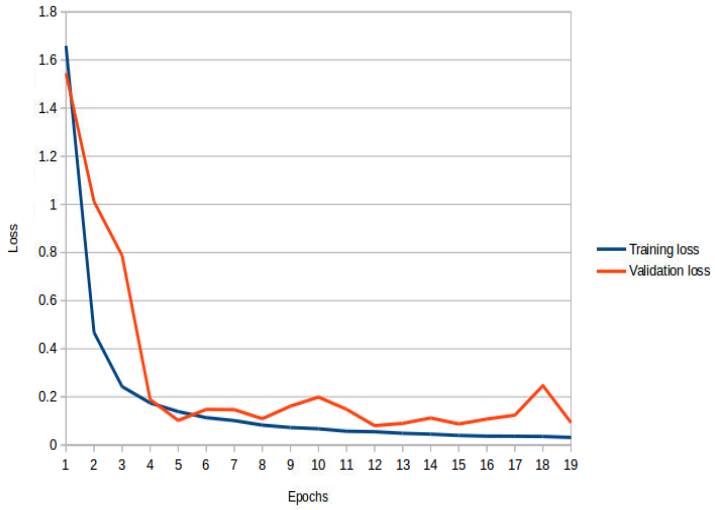


Figure 7. Training loss and validation loss.

In terms of the effect of changing the regularization parameters:

- Test accuracy decreased a lot when we trained the same network without using Batch Normalization and learning rate update on layer 4.
- Test accuracy decreased to 92% when we removed dropout from layer 4.
- Test accuracy decreased to 94.3% and 91.9%, when we changed dropout values to 0.5 and 0.8, respectively, in convolutional layers.

When we repeated the tests using the same architecture, different random parameter initialization resulted in slightly different results. For example, we obtained test accuracies like 96.3% and 95,6% on the same architecture and same number of training epochs.

To study the misclassified samples, we generated confusion matrices and saved all misclassified samples with an indication of original class and the assigned class, see Table 1. It was clear that most of the confusion comes from characters with similar morphology like “Daal” vs. “Raa” and “Zaay” ⟨ز⟩ vs. “Thaal” ⟨ذ⟩, or characters with diacritics (like dots) such as “Raa” ⟨ر⟩ vs. “Zaay” ⟨ز⟩ or “Jiim” ⟨ج⟩ vs. “Haa” ⟨ح⟩ and “Kha” ⟨خ⟩. This suggests the usefulness of another level of classifier system that is trained on classifying these shapes. Figure 8 illustrates some of the testing images that were misclassified due to the dot position (e.g. “Ghayn” ⟨غ⟩ to “Ayn” ⟨ع⟩), or to the number of dots (e.g “Qaaf” ⟨ق⟩ to “Faa” ⟨ف⟩), or to the curvature of the character (e.g. “Daal” ⟨د⟩ to “Raa” ⟨ر⟩).

4.2.2 AHCR Using AIA9K Dataset

We repeated the same set of experiments using our CNN architecture but on the AIA9K database. We divided the data into three parts; training, validation and testing, with ratios of 70%, 15% and 15% for

each set, respectively. Then, we ran training for 17 epochs and obtained a classification accuracy of 93.4%. Changing dropout value from 0.5 to 0.75 caused test accuracy to reach 94.2%. Increasing the number of filters in a similar way to the approach described in sub-section 4.2.1 improved accuracy to 94.65% after 29 epochs. Finally, changing the fully connected layer dropout keep probability from 0.75 to 0.8 and training for 32 epochs improved test accuracy once more to reach 94.8%.

Table 1. Classification accuracy for each AHCD character followed by the count of each wrongly assigned character based on the confusion matrix.

Alif:714, (Acc=0.992), Miim:3, Kaaf:2, Daal:1
Baa:712, (Acc=0.989), Taa:3, Haa:2, Yaa:2, Kaaf:1
Taa:699, (Acc=0.971), Thaa:17, Nuun:3, Kaaf:1
Thaa:696, (Acc=0.967), Taa:13, Kaaf:6, Daad:2, Qaaf:2, Faa:1
Jiim:716, (Acc=0.994), Haa:4
Haa:707, (Acc=0.982), Kha:7, Jiim:6
Kha:708, (Acc=0.983), Haa:12
Daal:700, (Acc=0.972), Raa:10, Laam:3, Waaw:3, Thaal:2, Zaay:1, Kaaf:1
Thaal:688, (Acc=0.956), Daal:15, Zaay:7, Waaw:6, Thaa:1, Zhaa:1, Kaaf:1, Nuun:1
Raa:704, (Acc=0.978), Daal:7, Zaay:3, Waaw:3, Thaa:1, Kha:1, Kaaf:1
Zaay:661, (Acc=0.918), Thaal:31, Raa:24, Daal:2, Kha:1, Zhaa:1
Siin:717, (Acc=0.996), Saad:2, Daal:1
Shiin:716, (Acc=0.994), Qaaf:3, Siin:1
Saad:710, (Acc=0.986), Daad:5, Siin:2, Ha:2, Kaaf:1
Daad:693, (Acc=0.963), Saad:18, Shiin:4, Thaa:2, Faa:2, Kha:1
Taa:718, (Acc=0.997), Siin:1, Zhaa:1
Zhaa:690, (Acc=0.958), Taa:29, Ghayn:1
Ayn:691, (Acc=0.960), Ghayn:15, Miim:6, Haa:5, Kha:2, Taa:1
Ghayn:720, (Acc=1.000)
Faa:713, (Acc=0.990), Qaaf:4, Jiim:2, Kaaf:1
Qaaf:686, (Acc=0.953), Faa:25, Kaaf:4, Yaa:2, Taa:1, Jiim:1, Zaay:1
Kaaf:718, (Acc=0.997), Saad:1, Ha:1
Laam:715, (Acc=0.993), Zhaa:4, Daal:1
Miim:705, (Acc=0.979), Ha:9, Alif:5, Daad:1
Nuun:674, (Acc=0.936), Kaaf:18, Taa:15, Faa:7, Thaa:1, Raa:1, Daad:1, Zhaa:1, Qaaf:1, Yaa:1
Ha:706, (Acc=0.981), Waaw:13, Faa:1
Waaw:702, (Acc=0.975), Daal:7, Raa:6, Ha:5
Yaa:697, (Acc=0.968), Ayn:8, Haa:4, Faa:3, Jiim:2, Qaaf:2, Baa:1, Siin:1, Saad:1, Kaaf:1

Figure 8. Misclassified test samples due to (i) dot position – "Ghayn" to "Ayn" (ii) Dot shape – "Qaaf" to "Faa" (iii) Character curvature – "Daal" to "Raa".

Similar to the tests performed in sub-section 4.2.1, adding an additional fully connected layer of the same size of the final fully connected layer decreased test accuracy to 94.3%. We tried different experiential adjustments to the network parameters to change the network characteristics, which didn't provide improvements.

Analysis of misclassified testing samples revealed similar observations to those in sub-section 4.2.1, but it also showed that some writers had actually written the "Haa" not in its isolated format ⟨ ه ⟩ , but rather in its shape at the start of the word; initial form ⟨ هـ ⟩ . In addition, some mislabeled characters were observed.

4.3 Comparison to Other Approaches

In this sub-section, the performance of our system is compared with those of other Arabic handwriting

recognition systems that use AIA9K and AHCD datasets. Relevant results are presented in Table 2. The AIA9K and AHCD databases are available for the general public for research purposes and this makes system comparisons meaningful.

We concentrated our efforts on the largest databases; namely, AIA9K and AHCD. For AIA9K, our proposed system outperformed the system in [23] and was easier to implement and there was no need to evaluate many sets of engineered features and classifier combinations.

As for the AHCD database, our system was able to obtain a high percentage of accuracy of 97.6% as implemented using Keras/TensorFlow. The reported accuracy of the system in [19] is lower by 2.7%.

One of the features of working with Keras is that selection of hyper-parameters based on experts' opinions or thumb rules comes preprogrammed, so that programmers experiment with default parameters.

This demonstrates that our proposed algorithm is favorable to the state-of-the-art when we are dealing with the largest database available for AHCR. Our results are better than those of all of the experiments performed by Torki et al. [13]. Our design obtained better accuracy, since we started with a good architecture and added ritualization, tweaked the network parameters, increased the number of layers and adopted a data augmentation mechanism.

Table 2. Comparison between proposed approach and other approaches on the same datasets.

Authors	Database	Training Data & Testing Data	Classification Accuracy
Torki et al. [13]	AIA9k	8738 images 85% training 15% testing (by gender)	94.28%
El-Sawi et al. [19]	AHCD	16800 images 13440 Training images 3360 Testing images	94.9%
Proposed Approach	AIA9k	8738 images 85% training 15% testing	94.8%
Proposed Approach	AHCD	16800 images 13440 Training images 3360 Testing images	97.6%

We tried to build a famous network architecture called Residual Networks (ResNet) [31]-[32]. This network contains 18 layers and achieved an accuracy of 92% in the first few experiments. It is a promising architecture that we intend to study further.

4.4 Comparison to Latin Alphabet Classification

To study the misclassified characters, a comparison of the proposed network performance on Latin Alphabet (modern English characters) was deemed suitable to know if the network has the same failures as compared to the misclassification in Arabic characters. Does the error come from the proximity of the morphology of certain characters?

In this sub-section, we demonstrate the performance of our system on a database called EMNIST dataset, that is a set of handwritten character digits derived from the National Institute of Standards and Technology (NIST) Special Database 19 [40]. It has a dataset of Latin (English) alphabet that contains 145600 images of English alphabet (A-Z and a-z) sorted into 26 classes, where each class has both uppercase and lowercase characters converted into a 28x28 pixel image format. The dataset is divided into a training set of 124800 and a testing set of 20800 images.

We applied the same network architecture on that dataset and training was completed in 15 epochs with a test accuracy of 95%. This is far better than the best accuracy of 85% using 10,000 hidden layer neurons of the ELM network trained using the Online Pseudo-Inverse Update Method (OPIUM)

reported in [40].

However, the goal of this experiment was merely to study whether the trained network will have problems with characters that have the same morphology. Table 3 lists the classification accuracy of each of the characters along with the number of times that each character was misclassified as another. It is obvious that the confusion is mostly seen between characters that have similar morphology, such as “l” and “I”, “G” and “Q” and “U’ and “V”. This confirmed our hypothesis that failures occurred in classifying Arabic characters like (“Alif” and “Miim”) or (“Baa”, “Taa” and “Thaa”) - See Table 1 - due to similar shape. Hence, data augmentation and increasing the network capacity were deemed a necessary to improve performance.

Table 3. Classification accuracy for each EMNIST character followed by the count of each wrongly assigned character based on the confusion matrix.

A:762, (Acc.=0.953), Q:6, D:4, F:4, H:4, N:4, C:3, O:3, U:3, Z:3, G:1, R:1, X:1, Y:1
B:787, (Acc.=0.984), E:2, H:2, Z:2, C:1, D:1, G:1, N:1, O:1, R:1, S:1
C:793, (Acc.=0.991), E:5, G:1, U:1
D:770, (Acc.=0.963), O:17, A:4, Q:3, B:1, C:1, J:1, N:1, P:1, T:1
E:785, (Acc.=0.981), C:9, L:2, A:1, I:1, P:1, W:1
F:784, (Acc.=0.980), T:8, E:2, G:2, P:2, I:1, R:1
G:664, (Acc.=0.830), Q:104, A:10, S:5, C:3, B:2, E:2, F:2, J:2, Y:2, D:1, N:1, O:1, P:1
H:765, (Acc.=0.956), N:17, L:5, B:4, K:3, X:2, R:1, T:1, U:1, W:1
I:558, (Acc.=0.698), L:223, J:10, C:2, E:2, R:2, B:1, O:1, V:1
J:756, (Acc.=0.945), I:22, D:5, F:3, S:3, T:3, X:2, Y:2, G:1, L:1, U:1, V:1
K:786, (Acc.=0.983), B:3, R:3, X:3, H:2, E:1, L:1, T:1
L:652, (Acc.=0.815), I:129, C:7, H:6, J:2, B:1, D:1, R:1, Y:1
M:794, (Acc.=0.993), N:4, K:1, W:1
N:780, (Acc.=0.975), M:5, R:5, H:3, W:2, X:2, I:1, U:1, V:1
O:784, (Acc.=0.980), D:10, A:2, U:2, C:1, Q:1
P:793, (Acc.=0.991), D:4, E:1, L:1, Q:1
Q:730, (Acc.=0.912), G:43, A:12, O:5, E:2, F:2, I:2, D:1, U:1, Y:1, Z:1
R:774, (Acc.=0.968), V:5, Y:5, K:4, T:3, X:3, C:1, E:1, I:1, N:1, O:1, Z:1
S:790, (Acc.=0.988), G:4, A:2, J:2, D:1, N:1
T:785, (Acc.=0.981), F:2, K:2, X:2, A:1, B:1, C:1, E:1, J:1, O:1, R:1, Y:1, Z:1
U:769, (Acc.=0.961), V:16, Y:3, C:2, N:2, W:2, A:1, H:1, J:1, L:1, O:1, S:1
V:736, (Acc.=0.920), U:36, Y:16, R:7, J:1, L:1, Q:1, X:1, Z:1
W:788, (Acc.=0.985), N:5, H:2, M:2, U:2, V:1
X:791, (Acc.=0.989), K:3, Y:2, H:1, N:1, P:1, V:1
Y:780, (Acc.=0.975), X:9, R:3, J:2, T:2, D:1, G:1, K:1, V:1
Z:795, (Acc.=0.994), C:1, F:1, I:1, J:1, L:1

4.5 Classification Performance on Form-based Arabic Characters

Since it was shown that most misclassification of the system was due to similarity in the shape of characters, it was deemed suitable to carry out some tests where we keep only one character of every group of letters that have the same form. For example, various consonants that have the same form (or major stroke), such as “Baa” ⟨ ب⟩ , “Taa” ⟨ ت⟩ and “Thaa” ⟨ ث⟩ and only distinguished by pointing diacritics (the number and location of dots). Other examples are "Jiim” ⟨ ج⟩ , “Haa” ⟨ ح⟩ and “Kha” ⟨ خ⟩ . In this sub-section, we keep only one sample of each of these groups of similar letters and study the classification accuracy. We kept only the letter “Baa” from the first group and the Letter “Haa” from the second group, …etc. The reduced alphabet of Arabic Letters by form is 16 characters listed in Table 4. The classification accuracy has increased to 98.22% as expected. Looking at Table 4 reveals that misclassification occurred again because writers tend to write “Daal” in a relaxed smooth curved shape that looks like “Raa”, but the opposite is not true. The other main source of misclassification is the loop in “Waa” that is misclassified as “Faa”. This suggests that a classifier system that classifies in two stages; the first classifies digits by form and the second tries to look at diacritics and other distinct features to carefully distinguish between characters of similar morphology, would prove useful.

5. CONCLUSIONS

Automated software systems for the recognition of Arabic characters could have huge applications in many industry and government sectors. In this paper, we presented a Deep Learning system based on convolutional neural network that is capable of classifying Arabic handwritten characters with a state-of-the-art classification accuracy of 94.8% and 97.6% on the AIA9k and AHDC datasets, respectively.

Table 4. Classification accuracy for each of the form-based Arabic character dataset followed by the count of each wrongly assigned character based on the confusion matrix.

Alif:719, (Acc.=0.999), Yaa:1
Baa:718, (Acc.=0.997), Ayn:1, Laam:1
Haa:704, (Acc.=0.978), Ayn:13, Alif:1, Baa:1, Siin:1
Daal:682, (Acc.=0.947), Raa:29, Waaw:6, Laam:3
Raa:715, (Acc.=0.993), Daal:3, Faa:1, Yaa:1
Siin:714, (Acc.=0.992), Baa:2, Saad:2, Ayn:1, Faa:1
Saad:712, (Acc.=0.989), Siin:8
Taa:719, (Acc.=0.999), Yaa:1
Ayn:713, (Acc.=0.990), Taa:2, Miim:2, Alif:1, Haa:1, Faa:1
Faa:717, (Acc.=0.996), Kaaf:2, Yaa:1
Kaaf:713, (Acc.=0.990), Yaa:3, Alif:2, Faa:1, Ha:1
Laam:715, (Acc.=0.993), Kaaf:4, Faa:1
Miim:695, (Acc.=0.965), Alif:12, Ha:7, Raa:2, Saad:2, Baa:1, Faa:1
Ha:692, (Acc.=0.961), Waaw:11, Faa:7, Taa:6, Saad:3, Siin:1
Waaw:668, (Acc.=0.928), Faa:27, Daal:9, Ha:9, Raa:4, Alif:1, Taa:1, Ayn:1
Yaa:719, (Acc.=0.999), Alif:1

The system employs some regularization techniques (Dropout and Batch Normalization), which provide performance and efficiency improvements. The system was implemented using TensorFlow and Keras framework.

As a future work, in addition to working more with ResNets, we plan to apply the algorithm on a larger and more diverse database. This could be done by merging more than one source of database. We plan to implement deeper networks with state-of-the-art techniques for regularization and optimization. In the next stages, we are planning to use segmentation-free techniques to be able to train our network with larger datasets that contain Arabic handwritten connected characters (words) to make use of the context using LSTM networks. Ultimately, this research will open the door to great opportunities expanding the applications of deep learning using these powerful libraries to problems of ancient Arabic handwritten character recognition.

ACKNOWLEDGEMENTS

This project is partially supported by grants of Hamdi Mango Center for Scientific Research (HMCSR) and the Deanship of Scientific Research at the University of Jordan, Amman, Jordan.

REFERENCES

[1] R. Plamondon and S. N. Srihari, "On-line and Off-line Handwriting Recognition: A Comprehensive Survey," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 2, no. 1, pp. 63-84, 2000.

[2] A. Belaïd and N. Ouwayed, Segmentation of Ancient Arabic Documents, in "Guide to OCR for Arabic Scripts," Eds. Volker Märgner and Haikal El Abed, Springer-Verlag, London, pp. 103-122, 2011.

[3] G. Abandah, M. Khedher and K. Younis, "Handwritten Arabic Character Recognition Using Multiple Classifiers based on Letter Form," Proceedings of the 5th IASTED International Conference on Signal Processing, Pattern Recognition and Applications (SPPRA 2008), pp. 128-133, Innsbruck, Austria, 13-15 Feb. 2008.

[4] G. Abandah, M. Khedher and K. Younis, "Evaluating and Selecting Features for Recognizing Handwritten Arabic Characters," Tech. Report, Computer Eng. Dept., The Univ. of Jordan, Sep. 2007.

[5] G. Hu, Y. Yang, D. Yi, J. Kittler, W. Christmas, S. Li and T. Hospedales, "When Face Recognition Meets with Deep Learning: An Evaluation of Convolutional Neural Networks for Face Recognition," Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 142–150, Santiago, Chile, 13–16 December 2015.

[6] H. Cecotti, "Hierarchical K-nearest Neighbor with GPUs and a High-performance Cluster: Application to Handwritten Character Recognition," International Journal of Pattern Recognition and Artificial Intelligence, vol. 31, no. 2, pp. 1–24, 2017.

[7] M. Abadi et al. "Tensorflow: Large-scale Machine Learning on Heterogeneous Distributed Systems," arXiv preprint arXiv:1603.04467, 2016.

[8] F. Chollet and Keras, GitHub Repository, [Online], Available: https://github.com/fchollet/keras, GitHub, 2015.

[9] M. Elleuch, N. Tagougui and M. Kherallah, "Arabic Handwritten Characters Recognition Using Deep Belief Neural Networks," Proc. of the 12th International Multi-Conference on Systems, Signals & Devices (SSD15), pp. 1-5, Mahdia, Tunisia, 16-19 March 2015.

[10] H. Alwzwazy et al., "Handwritten Digit Recognition Using Convolutional Neural Networks," International Journal of Innovative Research in Computer and Communication, vol. 4, no. 2, 2016.

[11] A. Lawgali, "A Survey on Arabic Character Recognition, " International Journal of Signal Processing, Image Processing and Pattern Recognition, vol. 8, no. 2, pp. 401-426, 2015.

[12] J. Al Abodi and X. Li, "An Effective Approach to Offline Arabic Handwriting Recognition," Pattern Analysis and Applications, vol. 40, no. 6, pp. 1883-1901, 2014.

[13] M. Torki et al., "Window-based Descriptors for Arabic Handwritten Alphabet Recognition: A Comparative Study on a Novel Dataset," arXiv preprint arXiv:1411.3519, 2014.

[14] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), vol. 1, pp. 886-893, 2005.

[15] H. Bay, A. Ess, T. Tuytelaars and L. van Gool, "Surf: Speeded-up Robust Features," Journal of Computer Vision and Image Understanding, vol. 110, no. 3, pp. 346-359, 2008.

[16] D. Lowe, "Distinctive Image Features F-ROM Scale-invariant Key Points," International Journal of Computer Vision, vol. 60, pp. 91–110, 2004.

[17] A. Lawgali et al., "HACDB: Handwritten Arabic Characters' Database for Automatic Character Recognition," European Workshop on Visual Information Processing (EUVIP), pp. 255-259, 2013.

[18] M. Elleuch et al., "Optimization of DBN using Regularization Methods Applied for Recognizing Arabic Handwritten Script," International Conference on Computational Science (ICCS 2017), vol. 108, pp. 2292-2297, Zurich, 12-14 June 2017.

[19] A. El-Sawy, M. Loey and H. El-Bakry, "Arabic Handwritten Characters Recognition Using Convolutional Neural Network," WSEAS Transactions on Computer Research, vol. 5, pp. 11-19, 2017.

[20] G. Fink, "1st International Workshop on Arabic Script Analysis and Recognition (ASAR 2017)," [Online], Available: http://asar.ieee.tn/speakers/, Nancy, France, April 3-5, 2017.

[21] I. Goodfellow, Y. Bengio and A. Courville, Deep Learning, MIT Press, pp. 335-339, [Online], Available: http://www.deeplearningbook.org.

[22] S. Loffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," [Online], Available: https://arxiv.org/abs/1502.03167v3, 2015.

[23] Dan C. Cireșan et al., "Handwritten Digit Recognition with a Committee of Deep Neural Nets on GPUs," arXiv preprint arXiv:1103.4487, 2011.

[24] Y. Lecun, L. Bottou, Y. Bengio and P. Haffner, "Gradient-based Learning Applied to Document Recognition," Proceedings of IEEE, vol. 86, no. 11, pp. 2278–2324, 1998.

[25] P. Golik, P. Doetsch and H. Ney, "Cross-entropy vs. Squared Error Training: A Theoretical and Experimental Comparison," Interspeech, vol. 13, 2013.

[26] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," Journal of Machine Learning Research, vol. 15, no. 1, pp. 1929-1958, 2014.

[27] D. P. Kingma and B. Adam, "A Method for Stochastic Optimization," [Online], Available: https://arxiv.org/abs/1412.6980.

[28] M. Pechwitz, S. S. Maddouri, V. Mrgner, N. Ellouze and H. Amiri, "Ifn/enit - database of Handwritten Arabic Words," Colloque Inter. Francophone sur lEcrit et le Document (CIFED), pp. 129–136, 2002.

[29] Y. Le Cun, K. Kavukvuoglu and C. Farabet, "Convolutional Networks and Applications in Vision," Proc. of the IEEE International Symposium on Circuits and Systems (ISCAS'10), Paris, France, 2010.

[30] K. Younis and A. Alkhateeb, "A New Implementation of Deep Neural Networks for Optical Character Recognition and Face Recognition," Proc. of the New Trends in Information Technology (NTIT-2017), The University of Jordan, 25-27 April 2017.

[31] Ke Zhang, "Residual Networks of Residual Networks: Multilevel Residual Networks," IEEE Transactions on Circuits and Systems for Video Technology, IEEE Transactions on Latex Class Files, vol. 14, no. 8, Aug. 2016.

[32] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770-778, 2016.

[33] C. Boufenar, M. Batouche and M. Schoenauer, "An Artificial Immune System for Offline Isolated Handwritten Arabic Character Recognition. Evolving Systems," Springer-Verlag, pp.1-17, 2016.

[34] V. Nair and G. Hinton, "Rectified Linear Units Improve Restricted Boltzmann Machines," Proceedings of the 27th International Conference on Machine Learning (ICML-10), pp. 807–814, Haifa, June 2010.

[35] J. Springenberg, A. Dosovitskiy, T. Brox and M. Riedmiller, "Striving for Simplicity: The All Convolutional Net," arXiv preprint arXiv:1412.6806, 2014.

[36] A. Krizhevsky, I. Sutskever and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," Advances in Neural Information Processing Systems (NIPS), pp. 1097-1105, 2012.

[37] I. Ahmad and G. Fink, "Class-based Contextual Modeling for Handwritten Arabic Text Recognition," International Conference on Frontiers in Handwritten Recognition (ICFHR), pp. 554-559, 2016.

[38] S. Hochreiter and J. Schmidhuber, "Long Short-term Memory," Journal of Neural Computation, vol. 9, no. 8, pp. 1735-1780, 1997.

[39] B. Balci, D. Saadati and D. Shiferaw, "Handwritten Text Recognition Using Deep Learning," CS231n: Convolutional Neural Networks for Visual Recognition, Stanford University, Course Project Report, Spring 2017.

[40] G. Cohen, S. Afshar, J. Tapson and A. Van Schaik, "EMNIST: An Extension of MNIST to Handwritten Letters," arXiv:1702.05373v2, 1 March 2017.

[41] A. Ray, S. Rajeswar and S. Chaudhury, "Text Recognition Using Deep BLSTM Networks," Proc. of the 8th International Conference on Advances in Pattern Recognition (ICAPR), pp. 1-6, Kolkata, India, 4-7 Jan. 2015.

ملخص البحث:

يمثل التحليل الآلي وتمييز الأحرف العربية المخطوطة باليد من الصّور مشكلة مهمة في كثير من التطبيقات. وعلى الرغم من التقدم الهائل في البحوث المرتبطة بتمييز الأحرف باستخدام الضوء في الآونة الأخيرة، فما زالت هناك بعض المشكلات التي تنتظر الحلّ، وبخاصة الأحرف العربية. حيث ان ظهور الشبكات العصبية العميقة يُعدّ أمراً واعداً جداً لإيجاد حلول لبعض تلك المشكلات. في هذه الورقة، نقدم شبكة عصبية عميقة لحلّ مشكلة تمييز الأحرف العربية المخطوطة باليد. تَستخدم الشبكة المقترحة نماذج التفافية من الشبكات العصبية مع متغيراتٍ خاصة بالتنظيم، مثل التسوية على دفعات، للحيلولة دون فَرْط الملاءَمة. وقد طبقنا الشبكة العصبية العميقة المقترحة على قواعد البيانات AIA9K وAHCD؛ إذْ كانت دقة التصنيف 94.8% و97.6% على الترتيب. وقد تمت دراسة أداء الشبكة على EMNIST ومجموعة البيانات AHCD المبنية على الشكل للمساعدة في التحليل.

Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 3, No. 3, December 2017.

IMPROVING THE PERFORMANCE OF NO-REFERENCE IMAGE QUALITY ASSESSMENT ALGORITHM FOR CONTRAST-DISTORTED IMAGES USING NATURAL SCENE STATISTICS

Yusra Al-Najjar[1] and Chen Soong Der[2]

(Received: 17-Oct.-2017, Revised: 30-Nov.-2017 and 15-Dec.-2017, Accepted: 17-Dec.-2017)

ABSTRACT

This study was conducted to explore the role of two color features in improving the performance of the existing No-Reference Image Quality Assessment Algorithms for Contrast-Distorted Images (NR-IQA-CDI). The used color features were Colorfulness and Naturalness of color expressed in CIELab and CIELuv color spaces. Test images used were the public benchmark databases that contain contrast-distorted images - TID2013, CID2013 and CSIQ. Experiments for the exploration were conducted in two stages: the preliminary stage and the comprehensive stage. The results of preliminary study showed that the features of colorfulness and naturalness of color can improve the prediction of human opinion score which relies mainly on the feature of brightness-only contrast. The results inspired to more comprehensive study where the Natural Scene Statistics (NSS) of these two features were estimated by modelling the probability distribution function (pdf) of 16,873 test images from a public database called SUN2012. The results based on k-fold cross validation with k ranging from 2 to 10 showed that the performance of NR-IQA-CDI can be improved by adding the NSS of these features.

KEYWORDS

NR-IQA-CDI, Colorfulness, Naturalness, Contrast Distortion, NSS.

1. INTRODUCTION

Image quality assessment (IQA) is an important study area in image processing and computer vision [1], since it requires assessing the performance of various image processing algorithms. Human Visual System (HVS) is the ultimate receiver and interpreter of image content; therefore, subjective assessment is considered the most reliable assessment method. However, subjective assessment is time-consuming, expensive and requires a lot of effort. To overcome this limitation, many Image Quality Assessment algorithms (IQAs) have been proposed over the past decade. The aim of IQA is to predict image quality in a manner that is consistent with the results of subjective assessment [2]-[3].

The IQAs available are classified into three categories according to the level of access to the reference image - the distortion - free or perfect-quality image - which are: Full-Reference IQAs (FR-IQAs) in which there is full access to the reference image, Reduced-Reference IQAs (RR-IQAs) which have access to only some of the information about the reference image [4]-[5] and the No-Reference IQAs (NR-IQAs) that require no information about any reference image [6]. In many applications, where there is no information about the reference image, the NR-IQA is highly desired [1]. One of such applications is the assessment of contrast-distorted image (CDI). Contrast distortion happens during image acquisition [1], where acquisition devices are not perfect or lighting is poor. This might cause loss of contrast and small details. In such case, this acquired image will be the original image, but it cannot be used as a reference image with perfect quality.

In general, contrast change – distortion – is an important aspect in the field of image evaluation [1]. Nevertheless, despite all the work that has been proposed, there is still a lack for an algorithm that evaluates a contrast-distorted image that has no reference. This study aims to explore some elements that affect the evaluation of contrast distorted image by using the NSS of the color features.

1. Y. Al-Najjar is with the Faculty of Information and Communication Technology, Universiti Tenaga Nasional, Malaysia. Email: yusra.najjar.2011@gmail.com
2. Chen S. Der is with the Faculty of Information and Communication Technology, Universiti Tenaga Nasional, Malaysia. Email: sc2339@gmail.com

Recently image quality assessment algorithms have been proposed dealing with different types of distortion, such as compression, blur or noise, but little with contrast distortion. No one can claim that there is an IQA algorithm that is perfectly consistent with human opinion. Some of them were close to subjective assessment, but for one database only. So, studies are still going on to find a better IQA algorithm. To help with this, in our research, we focused on studying elements of the algorithm – basic bricks - instead of building the assessment algorithm and the main target database was TID2013 database, which recent IQA algorithms failed to evaluate [1]. Depending on what others have reached in this field, we decided to start working on improving the performance of these NR-IQA-CDI algorithms. And to do that, we needed to specify new features as will be explained in this research.

Following sections include a comprehensive review of the work carried out on NR-IQA algorithms, research methodology that included brief explanation of colorfulness and naturalness of color, performance metrics used in the experiments, databases used, statistical tests and procedure of the experiments, followed by an evaluation of results, and finally a conclusion.

2. LITERATURE REVIEW

When NR-IQA algorithms started, they were not based on specific types of distortion, which means that they were general image quality assessment algorithms detecting distortion in general; but, what if the distortion was specific: compression, blur, noise or contrast alone. Some of the algorithms were specialized to assess images with one distortion type such as blur [7], whereas others were general algorithms. Our main concern in this research is contrast distortion. Currently, it is noticed that general NR metrics usually follow two main thrusts; being natural scene statistics (NSS)-based or learning-based. NSS-based metrics extract properties from the image using statistical properties in the natural image, while learning-based metrics are involved in learning and testing based on neural networks or support vector regression (SVR). These two types are not quite different, where NSS-based algorithms assume a certain statistical system in the spatial domain of natural scene [8].

In [9], Gu et al. proposed a new no-reference (NR)/ blind sharpness metric in the autoregressive (AR) parameter space. This metric was established by analyzing AR model parameters. They calculated the energy and contrast differences in the locally estimated AR coefficients in a pointwise way, then they quantified image sharpness with percentile pooling to predict the overall score. They validated the effect of the technique on subsets of blurring artefacts from four large-scale image databases (LIVE, TID2008, CSIQ and TID2013). Despite the claim that the metric was good, but it was not tested using contrast-distorted images.

In [10], the authors proposed a metric based on Curvelet No-Reference Transform (CNR), which outperformed other full reference metrics, such as SSIM and PSNR, in predicting the level of distortion noise, JPEG compression or blur in natural images. But, this metric did not take into consideration contrast change (distortion); neither global nor local. Later in 2010, [11], Lua et al. proposed a no-reference metric that used contourlet transform based on an improved model of NSS - CNSS – established with contourlets. They claimed that this algorithm was superior to the conventional NSS model and could be applied to any distortion. This algorithm was general and was applied only to LIVE database. It was not tested on global or local contrast-distorted images. Li et al. in [12] used phase congruency, entropy and gradient as image features to assess image quality depending on the general regression neural network (GRNN). Li et al. used LIVE database and divided it into five datasets. Five-fold cross-validation was used. This metric was a general metric, concentrated on one database and did not consider contrast change, whether global or local.

Distortions change the statistical properties found in natural images. Moorthy et al. in [13] proposed a blind IQA framework and integrated algorithm based on natural scene statistics called DIIVINE index. This metric did not compute specific distortion features, but it extracted statistical features. This metric dealt with compression, noise, blur and fading. The metric was trained over LIVE database and evaluated using only TID2008. DIIVINE metric did not deal with contrast change and was limited to one global enhanced database for testing. In 2011, researchers started to work on contrast-distorted images. In [14], the authors claimed that the contrast quality is determined by two metrics; the histogram flatness (HF) and the Histogram Spread (HS). They claimed that low-contrast images have a low HS value, whereas high-contrast images have a high HS value. This means that HS can

differentiate between low- and high-contrast images. The images used were natural and medical images. The HS metric can specify whether the image requires more enhancement or not, but not specifying its quality. This metric was not tested on databases; besides, this metric relies on partial access to the reference image, which makes it a reduced reference metric and not a no-reference metric.

There have been NR metrics that were designed for general purpose usage. The authors of [15] and [16] built statistical models of mean subtracted contrast normalized (MSCN) coefficients and spatial relationship between neighbouring pixels. The model was trained on features obtained from both natural and distorted images and on human judgments of the quality of these images. Therefore, BRISQUE metric in [15] was limited to the types of distortion it has been tuned to. It worked fine for noise, blur and compression distortion, but its performance degraded significantly for contrast-distorted images, whereas for [16] there was a limited number of databases, and by comparison, the NIQE Index was not tied to any specific distortion type.

In 2013 [17], Gu et al. proposed a reduced reference metric for contrast-change images called RIQMC metric depending on information residual between the input and the distorted images as well as the first four order statistics of the distorted image histogram. Gu et al. used CID2013, TID2013 and CSIQ public databases. This RIQMC metric was devised based on phase congruency and information statistics of histogram, acquiring superior performance beyond existing models and managing to enhance original natural images. Despite that the metric achieved an impressive performance, the major drawback was that there was partial access to the reference image and that access was not based on natural image statistics (NSS) model. It inevitably needs a single number (the entropy) from original natural image. Entropy is considered an important feature, while it cannot represent the local information of the image. Xue et al. used statistics associated with gradient magnitude and Laplacian features to measure image quality [18]. Xue et al. proposed a blind image quality metric BIQA that predicts image quality by analyzing the image statistics in some transformed domains, such as discrete cosine transform domain and wavelet domain. Used databases were LIVE, TID2008 and CSIQ. Despite that the results for BIQA were good, it dealt with distortion in general such as blur and compression, but not with contrast-distortion.

In 2015 [1], Fang et al. proposed a no-reference image quality metric for contrast-distorted images depending on natural scene statistics (NSS). They employed many images to build an NSS model based on moment features. The authors used the three public databases CID2013, TID2013 and CSIQ. The results of their experiments were good for one of the databases, but not for TID2013 database, which called for more research.

Li et al. in [7] used discrete orthogonal moments to evaluate blur effect. They proposed a blind blur assessment metric concentrating on blur distortion only. They used four public databases, which are LIVE, CSIQ, TID2008 and TID2013, in their experiments. The authors claimed that blur affects the magnitudes of moments of an image based on discrete Tchebichef moments. The gradient of a blurred image is first computed to account for the shape, which is more effective for blur representation. Then, the gradient image is divided into equal-size blocks and the Tchebichef moments are calculated to characterize the image shape. The energy of a block was computed as the sum of squared non-DC moment values. Finally, the proposed image blur score is defined as the variance-normalized moment energy, which is computed with the guidance of a visual saliency model to adapt to the characteristics of human visual system. This metric attempts to model indicators of quality for the distortion in question – blur - and hence was unsuitable for use for contrast change.

In [19], Liu et al. proposed a no-reference image quality assessment method depending on mutual information of wavelet domain named (MIQA-II), by computing neighbouring pixels using LIVE [20] database. This method showed good results in identifying distortion. But, this method was based only on LIVE database without taking contrast change into consideration. In [21], Gu et al. reported a new large dedicated contrast-changed image database (CCID2014) which includes 655 images and associated subjective ratings recorded from 22 inexperienced observers and proposed a reduced reference image quality metric for contrast change (RIQMC) using phase congruency and statistics information of the image histogram. Validation of the proposed model was conducted on contrast related to CCID2014, TID2008, CSIQ and TID2013 databases. Despite that the metric showed good results, it required information from the original image – which was entropy – in addition to that it was

not tested on locally enhanced images. Depending on this metric, Wu et al. in [22] proposed a no-reference metric for contrast-distorted image assessment. They extracted five statistical features from the distortion image and two features were extracted from the phase congruence (PC) map of distortion image. These features and human mean opinion scores (MOS) of training images were jointly utilized to train a model of Support Vector Regression (SVR). Despite that the results were close to RIQMC results, these results depended on one database only which is CCIID2014. In [23], Gu et al. presented a no-reference image quality assessment metric for contrast-distorted images. In the metric, they searched for local details. They first removed predicted regions in an image, claiming that the unpredicted ones are of much information. Then, they computed entropy of unpredicted areas of maximum information by visual saliency. Gu et al. used CID2013, CCID2014, CSIQ, TID2008 and TID2013 databases. According to the results achieved, there was still weakness in predicting quality, especially for TID2013, where the results was 0.64.

Lately, Shokrollahi et al. proposed a contrast-changed image quality (CCIQ) metric including a local index, named edge-based contrast criterion (ECC) and three global measures [24]. They did not consider the metric as a full-reference metric, since the original image is not regarded to have the ideal quality. They claimed that it follows a new paradigm in image quality assessment. Experimental results on the three benchmark databases CID2013, TID2013 and TID2008 demonstrate that the proposed metric outperforms the state-of-the-art methods. This metric showed good evaluation for CID2013 and TID2013 when compared to full-reference metric (PSNR), reduced reference metric (RIQMC, [17]) and no-reference metric (NR-CDIQA, [1]). Regardless of the results, this metric is still referring to the original – reference - image.

In [25], Gu et al. developed a blind/no-reference (NR) model for assessing the perceptual quality of screen content pictures with big data learning. In this model, they extracted four types of features descriptive of picture complexity, screen content statistics, global brightness quality and sharpness of details. The efficacy of the new model was compared with existing blind picture quality assessment algorithms applied on screen content image databases. The proposed model gave a promising performance.

In [26], Gu et al. investigated the problem of image quality assessment (IQA) and enhancement using machine learning. In their work, they developed a new NR-IQA model by extracting 17 features from the given image by analyzing contrast, sharpness, brightness, among other features. They validated the efficiency of their metric using nine datasets. Another contribution of the authors was image enhancement to be based on quality optimization. The authors conducted histogram modifications to modify image brightness and contrast to a proper level. They claimed that their framework can enhance image contrast and lightness.

3. RESEARCH METHODOLOGY

Experiments for exploring colorfulness and naturalness features were conducted in two stages: the preliminary stage, that explores the features of contrast and color using the raw data; this stage was fast and easy. Depending on the results from the preliminary stage, the comprehensive stage - using the NSS of the features - was conducted. The comprehensive stage required more time and work. The study implementation of this study started with:

1- Getting the distribution of local contrast, colorfulness and naturalness for SUN2012 database using *dfittool( )* in MATLAB software.
2- Getting the probability distribution function (pdf) of these features.
3- The implementation was conducted over three public databases CID2013 [27], [17], TID2013 [28] and CSIQ [29].
4- The regression method used was Support Vector Regression (SVR) with cross-validation for k = 2 to 10.
5- The performance metrics and tests used are mentioned later in sub-sections 3.3 and 3.5.

3.1 Colorfulness Feature

Colorfulness is a low-level feature of color; it is a visual sensation by which the perceived colorfulness of any part of an object appears to be less or more chromatic [30]. Images were converted CIELab

color space which contains all perceivable colors. CIELab was used, because it has an infinite range of colors that exceeds those in RGB color model to simulate human vision. In addition, it is device-independent. See Figure 1.

Figure 1. To the right is the image in RGB and to the left is the image in CIELab color space.

3.2 Naturalness Feature

Naturalness is a high-level feature of color; it is composed of hue, saturation, chroma and luminance, which makes it a good complement to colorfulness. For naturalness computation, images were converted into CIELuv color space that uses previously mentioned color components.

3.3 Performance Metrics Used in the Experiments

For measuring the correlation with MOS, Pearson Linear Correlation Coefficient (PLCC) and Root Mean Squared Error (RMSE) were used to predict accuracy. For monotonicity, Spearman Rank Order Correlation Coefficient (SROCC) was used.

3.4 Databases

Databases used in the experiments were: first the CID2013 database which was established for contrast-distorted images. It contains 15 references and 400 contrast-enhanced images. Twenty-two people have shared in providing quality score for MOS. Second, the TID2013 database that contains 25 reference images with different distortion types including contrast distortion. The last was the CSIQ database with 30 reference images and different types of distortion. CSIQ database contains 116 contrast-distorted images. Subjective tests on the three databases followed the recommendations of ITU-R BT 500-12 [31]. A sample of images used in the research is displayed in Figure 2.

Figure 2. Sample of original (to the left) and contrast-distorted images. Top row from CSIQ database, middle row from CID2013 database and bottom row from TID2013 database.

3.5 Statistical Test

Paired *t*-test was conducted for each of the four performance metrics based on the metric's value obtained before and after modification in either feature or regression. The output of t-test was the p-value; the probability that the differences between two groups of data in a pair were not statistically significant; the higher the probability, the more likely the differences are not statistically significant. In

this study, the *p*-value less than 0.05 was interpreted as the differences being statistically significant. The results were reported in the form of percentage of difference for each of the databases and average differences over all databases.

3.6 Procedure of the Experiments

Conducted experiments in the preliminary stage included:

1. Computing local contrast, colorfulness and naturalness for each image in the three public databases.
2. Preparing the five moment features for each database (mean, standard deviation, entropy, skewness and kurtosis) for color features.
3. Using Support Vector Regression (SVR) for predicting the output data to be compared with target data of MOS.
4. Computing PLCC, SROCC and RMSE correlations performance metrics.
5. Computing t-test for each performance metric.

3.6.1 Preliminary Stage

Contrast vs. (Contrast + Colorfulness) and Contrast vs. (Contrast + Naturalness)

This experiment measures the correlation between MOS and local contrast feature before and after adding each of color features - colorfulness and naturalness - to the contrast.

1. Image data were partitioned using cross-validation *cvpartition*( ) function for k-fold from 2 to 10 to minimize bias.
2. In each round – from the k sets – each set acts as a test set only once, whereas the remaining are training sets and same for all sets.
3. PLCC, SROCC and RMSE performance metrics were used to find the correlation between MOS and local contrast.
4. Each k group prediction was averaged.
5. Percentage of difference before and after adding the (raw or NSS) of the feature was computed.
6. The statistical parametric t-test was performed to test the null hypothesis.

3.6.2 Comprehensive Stage

(Five moment features) vs. (five moment features + pdf of Colorfulness feature) and (five moment features) vs. (five features + pdf of Naturalness feature)

To compute colorfulness of color, the image was converted from RGB color space into CIELab color space, then Chroma of the color was computed according to the equation:

$$C^* = (a^{*2} + b^{*2})^{0.5} \quad \text{Equation (1)}$$

where C* is the computed Chroma. Colorfulness of color was computed according to the following equation:

$$Ck = mean(C^*) + std(C^*) \quad \text{Equation (2)}$$

The steps to compute the naturalness of color were to convert the image from RGB color space into CIELuv color space. Images were first converted from RGB color space into XYZ color space, then the parameters for Luv were computed according to the equations:

$$L^* = \begin{cases} 116(\frac{Y}{Y_n})^{\frac{1}{3}} - 16 & if\ \frac{Y}{Y_n} > 0.008856 \\ 903.3\left(\frac{Y}{Y_n}\right) & if\ \frac{Y}{Y_n} \leq 0.008856 \end{cases} \quad \text{Equation (3)}$$

$$u^* = 13\,(L^*)(\acute{u} - \acute{u_n}) \quad \text{Equation (4)}$$

$$v^* = 13\,(L^*)(\acute{v} - \acute{v_n}) \quad \text{Equation (5)}$$

$$\text{where } \acute{u} = \frac{4X}{X+15Y+3Z}\,, and\ \acute{v} = \frac{9Y}{X+15Y+3Z} \quad \text{Equation (6)}$$

L^* scales from 0 to 100.

Then the computed variables of luminance, hue and saturation were used in computing naturalness of color for each image as follows:

$$C^* = (u^{*2} + v^{*2})^{0.5} \quad \text{Equation (7)}$$

$$h_{uv} = \arctan(\frac{v^*}{u^*}) \quad \text{Equation (8)}$$

$$S_{uv} = \frac{C^*}{L^*} \quad \text{Equation (9)}$$

Finally, the naturalness of color was computed according to the equation:

$$\mathrm{N_{ij}} = \exp(-0.5(\frac{\mathrm{S_{ij}} - \mu_\mathrm{j}}{\sigma_\mathrm{j}})^2) \quad \text{Equation (10)}$$

This experiment measures the correlation between MOS and the probability of the five features (mean, standard deviation, entropy, skewness and kurtosis) before and after adding the probability of each feature of colorfulness and naturalness. Steps followed were the same as in the preliminary experiment.

4. EVALUATION OF RESULTS

The results of the preliminary stage were encouraging for moving to the comprehensive stage. So, the results displayed here concern the comprehensive stage only.

Table 1 and Table 2 summarize the results. Table 1, row 1 shows the percentage of difference in each of the performance metrics for adding the colorfulness feature. Table 2, row 1 shows the *p*-values of the paired *t*-tests. As seen in Table 1, there was a good improvement in the results of the experiment when using TID2013 database – this database was a target for improvement. PLCC and SROCC increased by 3.04% and 4.92% and RMSE decreased slightly by 0.94%. As seen in Table 2, all the three *p*-values for TID2013 database were less than 0.05, indicating that the differences in all three performance metrics were statistically significant.

As for CID2013, there were marginal decrements in PLCC and SROCC by 0.36% and 0.59%, respectively, with a marginal increment in RMSE by 1.08%. As seen in Table 2, all the three *p*-values for CID2013 were less than 0.05, indicating that the differences in all the three performance metrics were statistically significant. Nevertheless, it is worth noting that the magnitudes of change in the three metrics were very marginal.

For CSIQ database, there were good increments in PLCC and SROCC by 13.69% and 13.42%, respectively, while there was a good decrement in RMSE by 11.44%. The statistical test results in Table 2 indicated that the differences in the three performance metrics were statistically significant.

For the average results over the three databases, there were good increments in PLCC and SROCC by 5.46% and 5.92%, respectively, with a decrement in RMSE by 3.77%. The results of statistical tests showed that there were statistically significant improvements in PLCC, SROCC and RMSE, since the *p*-values were less than 0.05. Overall analysis indicated that adding colorfulness could improve the performance of NR-IQA-CDI.

This was for adding colorfulness feature. As for adding naturalness feature, Table 1, row 2 shows the percentage of difference in each of the performance metrics, whereas Table 2, row 2 shows the *p*-values of the paired t-tests. As seen in Table 1, there was an improvement in the results of experiment using TID2013 database. PLCC increased by 3.53%, while SROCC slightly increased by 0.17%. RMSE decreased by 1.42%. As seen in Table 2, the *p*-values for PLCC and RMSE were less than 0.05, indicating that the differences in these two performance metrics were statistically significant.

However, there was no statistically significant difference in SROCC.

As for CID2013 database, there were slight increments in PLCC and SROCC by 0.43% and 0.73%, respectively, with a marginal decrement in RMSE by 0.8%. As seen in Table 2, all the three *p*-values for CID2013 were less than 0.05, indicating that the differences in all three performance metrics were statistically significant.

For CSIQ database, there were increments in PLCC and SROCC by 17.05% and 18.67%, respectively and there was a decrement in RMSE by 12.88%. The statistical test results indicated that the differences in all three performance metrics were statistically significant.

For the average results over the three databases, there were increments in PLCC and SROCC by 7.00% and 6.53%, respectively and a decrement in RMSE by 5.03%. The results of statistical tests showed that there were statistically significant differences in PLCC, SROCC and RMSE, because the *p*-values were less than 0.05. Overall analysis indicated that adding naturalness could improve the performance of NR-IQA-CDI.

Coming to adding both colorfulness and naturalness together, results were displayed as follows:

Table 1, row 3, shows the percentage of difference in each of the performance metrics and Table 2, row 3, shows the *p*-values of the paired t-tests. As seen in Table 1, there was an improvement in the results of experiment using TID2013 database, which was a target for improvement. PLCC and SROCC increased by 5.19% and 5.28%, respectively. RMSE decreased by 1.79%. As seen in Table 2, for all three performance metrics - PLCC, SROCC and RMSE - *p*-values for TID2013 were less than 0.05, indicating that the differences in all three performance metrics were statistically significant.

As for CID2013 database, there was a slight decrement in PLCC by 0.07%, whereas SROCC slightly increased by 0.19%, with a marginal decrement in RMSE by 0.67%. As seen in Table 2, SROCC and RMSE *p*-values for CID2013 were less than 0.05, indicating that the differences in these performance metrics were statistically significant.

For CSIQ database, there were good increments in PLCC and SROCC by 19.18% and 20.50%, respectively. RMSE showed a good decrement by 15.47%. The *p*-values for statistical test results were less than 0.05, indicating that the differences in the three performance metrics were statistically significant.

Averaging the results over the three databases, there were increments in PLCC and SROCC by 8.10% and 8.65%, respectively, with a decrement in RMSE by 5.53%. The results of statistical tests showed that there was statistically significant improvement in PLCC, SROCC and RMSE, because all the *p*-values were less than 0.05. Overall, the results indicated that adding both colorfulness and naturalness of color could help improve the performance of NR-IQA-CDI.

Comparing adding each color feature alone to adding both together gave the results listed in Table3 and Table 4. Table 3 shows the percentage of difference in each of the performance metrics and Table 4 shows the *p*-values of the paired t-tests for adding colorfulness or naturalness versus adding both.

Table 3 shows that PLCC increased by 2.42% and 1.04% when adding both features as compared to adding only colorfulness and adding only naturalness, respectively. The *p*-values of statistical tests were less than 0.05, indicating that the differences in the performance metrics for adding both features were statistically significant.

SROCC increased by 2.48% and 2.03% when adding both features as compared to adding only colorfulness and adding only naturalness, respectively. The *p*-values of statistical test results were less than 0.05, indicating that the differences in the performance metrics for adding both features were statistically significant.

RMSE decreased by 1.95% and 0.59% when adding both features as compared to adding only colorfulness and adding only naturalness, respectively. The *p*-value of statistical test for both features versus colorfulness only was less than 0.05, indicating that the differences in the performance metrics were statistically significant. There was no statistically significant difference for adding both features versus adding only naturalness, but this has only a small effect.

Table 1. Percentage of difference in the performance after adding natural scene statistics of both colorfulness and naturalness to the five features each one separated and both together.

Feature/Database / Performance Metric	TID2013			CID2013			CSIQ			AVERAGE		
	PLCC	SROCC	RMSE	PLCC	SROCC	RMSE	PLCC	SROCC	RMSE	PLCC	SROCC	RMSE
Colorful-ness	3.04%	4.92%	-0.94%	-0.36%	-0.59%	1.08%	13.69%	13.42%	-11.44%	5.46%	5.92%	-3.77%
Natural-ness	3.53%	0.17%	-1.42%	0.43%	0.73%	-0.80%	17.05%	18.67%	-12.88%	7.00%	6.53%	-5.03%
Both Features (Colorfulness & Naturalness)	5.19%	5.28%	-1.79%	-0.07%	0.19%	0.67%	19.18%	20.50%	-15.47%	8.10%	8.65%	-5.53%

Table 2. *P*-values for the performance for adding natural scene statistics of both colorfulness and naturalness to the five features each one separated and both together.

Feature /Database / Performance Metric	TID2013			CID2013			CSIQ			AVERAGE		
	PLCC	SROCC	RMSE	PLCC	SROCC	RMSE	PLCC	SROCC	RMSE	PLCC	SROCC	RMSE
Colorful-ness	0.0019	0.0002	0.0030	0.0002	0.0001	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0003
Natural-ness	0.0000	0.3060	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0003	0.0000
Both Features (Colorfulness & Naturalness)	0.0001	0.0001	0.0001	0.1480	0.0482	0.0016	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Table 3. Percentage of difference in the performance after adding both features vs. colorfulness and both features vs. naturalness over the three public databases.

Features	PLCC	SROCC	RMSE
Both vs. Colorfulness	2.42%	2.48%	-1.95%
Both vs. Naturalness	1.04%	2.03%	-0.59%

Table 4. *P*-values for paired t-test on the difference in the performance of NR-IQA-CDI after adding both features vs. adding only one of them.

	PLCC	**SROCC**	**RMSE**
Both vs. Colorfulness	$1.6454x10^{-06}$	$7.5066x10^{-05}$	$1.1542x10^{-09}$
Both vs. Naturalness	$2.3174x10^{-03}$	$6.5035x10^{-04}$	$2.1236x10^{-01}$

5. CONCLUSION

This study tested the hypotheses that: “adding colorfulness of color could improve the performance of NR-IQA-CDI in predicting MOS” and “adding naturalness of color could improve the performance of NR-IQA-CDI in predicting MOS”.

Results denoted that there was a significant positive difference – improvement in the performance of prediction – for adding only colorfulness of color feature, for adding only naturalness of color feature and even for adding both of color features together in evaluating the no-reference image quality. Adding both color features versus adding either one of them showed – in general – that adding both features gives better results.

The variety of images in each database and the number of images played a part in the results. But, in spite of this, the results came promising for the proposed features in predicting the quality of images that have no reference.

REFERENCES

[1] Y. Fang, K. Ma, Z. Wang and W. Lin, "No-Reference Quality Assessment of Contrast-Distorted Images Based on Natural Scene Statistics," IEEE Signal Processing Letters, vol. 22, no. 7, pp. 838-842, 2015.

[2] W. Lin and C. -C. J. Kuo, "Perceptual Visual Quality Metrics: A survey," Visual Communication and Image Representation, vol. 22, no. 4, pp. 297-312, 2011.

[3] A. C. B. Zhou Wang, Modern Image Quality Assessment (Synthesis Lectures on Image, Video, & Multimedia Processing), San Mateo, CA, USA: Morgan & Claypool Publishers, 2006.

[4] Z. Wang, A. Bovik, H. Sheikh and E. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity," IEEE Trans. Image Processing, CA, 2004.

[5] H. R. Sheikh and A. C. Bovik, "Image Information and Visual Quality," Transactions on Image Processing, Montreal, Que., Canada, 2006.

[6] R. Hassen, Z. Wang and M. Salama, "No-Reference Image Sharpness Assessment Based on Local Phase Coherence Measurement," The IEEE International Conference on Acoustics, Speech and Signal Processing, Dallas, TX, USA, 2010.

[7] L. Li, W. Lin, X. Wang, G. Yang, K. Bahrami and A. C. Kot, "No-Reference Image Blur Assessment Based on Discrete Orthogonal Moments," in IEEE Transactions on Cybernetics, Cybern, 2016.

[8] D. L. Ruderman and W. Bialek, "Statistics of Natural Images: Scaling in the Woods," Physical Review Letters, vol. 73, no. 6, pp. 551-558, 1995.

[9] K. Gu, G. Zhai, W. Lin, X. Yang and W. Zhang, "No-Reference Image Sharpness Assessment in Autoregressive Parameter Space," Shanghai Municipal Commission of Economy and Informatization, China, Shanghai, 2015.

[10] J. Shen, Q. Li and G. Erlebacher, "Curvelet-Based No-Reference Objective Image Quality Assessment," Proceedings of the 27th Picture Coding Symposium (PCS'09), NJ, USA, 2009.

[11] W. Lua, K. Zing, D. Tao, Y. Yuan and X. Gao, "No-Reference Image QualityAassessment in Contourlet Domain," Neurocomputing, vol. 73, no. 4-6, pp. 784-794, January 2010.

[12] C. Li, A. C. Bovik and X. Wu, "Blind Image Quality Assessment Using a General Regression Neural Network," IEEE Transactions on Neural Networks, 2011.

[13] A. K. Moorthy and Alan Conrad Bovik, "Blind Image Quality Assessment: From Natural Scene Statistics to Perceptual Quality," IEEE Transactions on Image Processing, 2011.

[14] A. K. Tripathi, S. Mukhopadhyay and A. K. Dhara, "Performance Metrics for Image Contrast," International Conference on Image Information Processing (ICIIP), Shimla, India, Nov. 2011.

[15] A. Mittal, A. K. Moorthy and A. C. Bovik, "No-Reference Image Quality Assessment in the Spatial Domain," IEEE Transactions on Image Processing, August, 2012.

[16] A. Mittal, R. Soundarajan and A. C. Bovik, "Making a “Completely Blind” Image Quality Analyzer," IEEE Signal Processing Letters, pp. 209-212, 2013.

[17] K. Gu, G. Zhai, X. Yang, W. Zhang and M. Liu, "Subjective and Objective Quality Assessment for Images with Contrast Change," Image Processing (ICIP), Melbourne, VIC, Australia, Sept. 2013.

[18] W. Xue, X. Mou, L. Zhang, A. C. Bovik and X. Feng, "Blind Image Quality Assessment Using Joint Statistics of Gradient Magnitude and Laplacian Features," IEEE Transactions on Image Processing, 2014.

[19] B. Liu, L.-X. Liu, H.-P. Dong and Y.-G. Lin, "Blind Image Quality Assessment Based on Mutual Information," Electronics, Communications and Networks, vol. 382, pp. 127-136, 29 June 2016.

[20] H. Sheikh, Z. Wang, L. Cormack and A. Bovik, "LIVE Image Quality Assessment Database Release 2," [Online]. Available: http://live.ece.utexas.edu/research/quality. [Accessed 2014].

[21] K. Gu, G. Zhai, W. Lin and M. Liu, "The Analysis of Image Contrast: From Quality Assessment to Automatic Enhancement," Transactions on Cybernetics (T-)CYP, vol. 46, no. 1, pp. 284 - 297, Jan 2016.

[22] J. Wu, Z. Xia, Y. Ren and H. Li, "No-Reference Quality Assessment for Contrast-Distorted Image," International Conference on Image Processing Theory Tools and Applications (IPTA), Oulu, Finland, Dec. 2016.

[23] K. Gu, W. Lin, G. Zhai, X. Yang, W. Zhang and C. W. Chen, "No-Reference Quality Metric of Contrast-Distorted Images Based on Information Maximization," IEEE Transactions on CYBERNETICS, 2016.

[24] A. Shokrollahi, A. Mahmoudi-Aznaveh and B. M.-N. Maybodi, "Image Quality Assessment for Contrast Enhancement Evaluation," International Journal of Electronics and Communications (AEÜ), pp. 61-66, 2017.

[25] K. Gu, J. Zhou, J.-F. Qiao, G. Zhai, W. Lin and A. C. Bovik, "No-Reference Quality Assessment of Screen Content Pictures," IEEE Transactions on Image Processing, China, Singapore, 2017.

[26] K. Gu, D. Tao, J.-F. Qiao and W. Lin, "Learning a No-Reference Quality Assessment Model of Enhanced Images with Big Data," IEEE Transactions on Neural Networks and Learning Systems, Singapore, 2017.

[27] T. Virtanen, M. Nuutinen, M. Vaahteranoksa, P. Oittinen and J. Häkkinen, "CID2013: A Database for Evaluating No-Reference Image Quality Assessment Algorithms," IEEE Transactions on Image Processing, vol. 24, no. 1, pp. 390-402, 2015.

[28] N. Ponomarenko, O. Ieremeiev, V. Lukin, K. Egiazarian, L. Jin, J. Astola and B. Vozel, "Color Image Database TID2013: Perculiarities and Preliminary Results," The 4th European Workshop on Visual Information, 2013.

[29] E. C. Larson and D. M. Chandler, "Categorical Image Quality (CSIQ) Database," 2010. [Online], Available: http://vision.okstate.edu/csiq.

[30] M. Faichild, Color Appearance Models, John Wiley & Sons, p. 87, 2013.

[31] I. R. Assembly, "Methodology for the Subjective Assessment of the Quality of Television Pictures ITU - Radiocom Sector," International Telecommunication Union, 2013.

ملخص البحث:

أجريت هذه الدراسة بهدف استكشاف دور اثنتين من خصائص اللون في تحسين أداء خوارزميات تقييم جودة الصور عديمة المرجع القائمة بالنسبة للصور مشوهة التباين. هاتان الخاصيتان هما: حيوية اللون وطبيعته. أما الصور التي تم اختبارها فكانت من قواعد بيانات مرجعية تشتمل على صور مشوهة التباين. وتم إجراء التجارب في مرحلتين هما: المرحلة الأولية، والمرحلة الشاملة. وقد بينت نتائج المرحلة الأولية من الدراسة أن خاصيتي حيوية اللون وطبيعيته يمكنهما تحسين التوقع الخاص بالرأي البشري الذي يرتكز على خاصية التباين المرتبط بشدة الإضاءة فقط.

وقد شجعت نتائج المرحلة الأولية على القيام بدراسة شاملة من أجل تقدير إحصائيات المشهد الطبيعي للخاصيتين المذكورتين، وذلك عبر نمذجة دالة توزيع الاحتمال لِـ 16873 صورة اختبار من قاعدة البيانات SUN2012 . وأوضحت النتائج المبنية على التأييد التقاطعي المضاعف، حيث يتراوح معامل المضاعفة من 2 إلى 10، أن أداء خوارزميات تقييم جودة الصور عديمة المرجع يمكن تحسينه بإضافة إحصائيات المشهد الطبيعي لخاصيتي حيوية اللون وطبيعته إلى الصور المختبرة.

EDITORIAL BOARD SUPPORT TEAM

LANGUAGE EDITOR

Haydar Al-Momani

EDITORIAL BOARD SECRETARY

Eyad Al-Kouz

JJCIT ADDRESS

WEBSITE: www.jjcit.org

EMAIL: jjcit@psut.edu.jo

ADDRESS: Princess Sumaya University for Technology, Khalil Saket Street, Al-Jubaiha.

B.O. BOX: 1438 Amman 11941 Jordan.

TELEPHONE: +962-6-5359949.

FAX: +962-6-7295534.

المجلة الأردنية للحاسوب و تكنولوجيا المعلومات

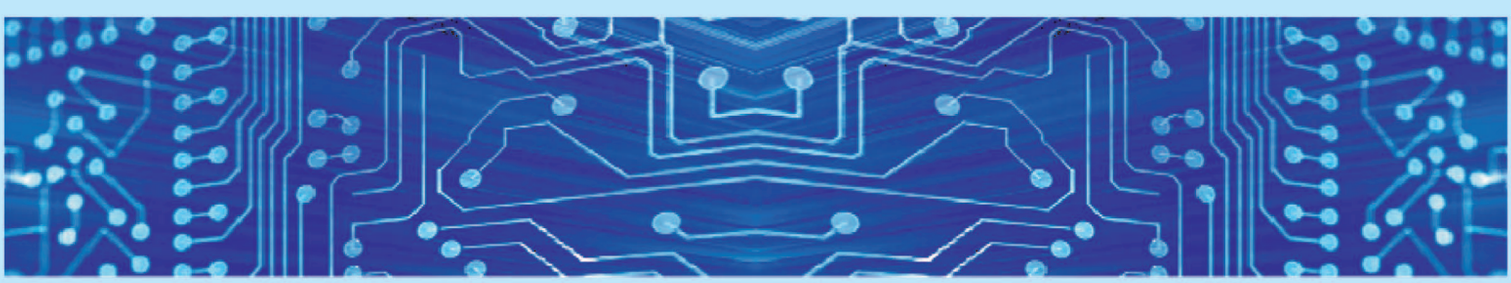
ISSN 2415 - 1076 (Online)
ISSN 2413 - 9351 (Print)

كانون الأول ٢٠١٧ المجلد ٣ العدد ٣

JJCIT

www.jjcit.org

jjcit@psut.edu.jo



مجلة علمية عالمية متخصصة محكمة

تصدر بدعم من صندوق دعم البحث العلمي